



ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΑΘΗΝΩΝ

ΤΜΗΜΑ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΠΜΣ ΔΙΟΙΚΗΣΗ, ΑΝΑΛΥΤΙΚΗ ΚΑΙ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ

ΕΠΙΧΕΙΡΗΣΕΩΝ

(M.Sc. in Business Administration, Analytics and Information Systems)

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

**ΕΞΕΡΕΥΝΩΝΤΑΣ ΤΗ ΣΧΕΣΗ ΜΕΤΑΞΥ ΚΕΡΔΟΦΟΡΙΑΣ ΚΑΙ
ΠΟΙΟΤΙΚΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΤΩΝ ΕΡΓΩΝ ΛΟΓΙΣΜΙΚΟΥ
ΣΤΟΝ ΙΣΤΟ: ΑΝΑΛΥΣΗ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ**

ΧΡΙΣΤΙΝΑ ΡΕΓΚΛΗ

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ

ΒΑΣΙΛΕΙΟΣ ΛΑΖΑΡΟΥ

ΑΘΗΝΑ

2023

Copyright © Ρέγκλη Χριστίνα, 2023

Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Η έγκριση της διπλωματικής εργασίας από το Τμήμα Οικονομικών Επιστημών (ΠΜΣ Διοίκηση, Αναλυτική και Πληροφοριακά Συστήματα Επιχειρήσεων) του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών δεν υποδηλώνει απαραίτητως και αποδοχή των απόψεων του συγγραφέα εκ μέρους του Τμήματος.

Υπεύθυνη Δήλωση

«Ο υποψήφιος βεβαιώνει ότι η υποβληθείσα εργασία είναι προσωπική εκτός από τα σημεία όπου γίνεται αναφορά στις εργασίες άλλων»

Ρέγκλη Χριστίνα

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να εκφράσω τη βαθύτατη εκτίμησή μου στον καθηγητή Βασίλειο Λαζάρου για την καθοδήγηση, την υποστήριξη, την ενθάρρυνση και τις συμβουλές του κατά τη διάρκεια της έρευνας και τη συγγραφή της διπλωματικής εργασίας.

Επίσης θα ήθελα να ευχαριστήσω θερμά όλους τους ανθρώπους που εργάζονται στην εταιρία από την οποία άντλησα τα δεδομένα, για το χρόνο που αφιέρωσαν για να απαντήσουν σε πολλές ερωτήσεις μου και για τη συμβολή τους στην ίδια τη συλλογή των δεδομένων.

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΙΚΟΝΕΣ.....	4
ΠΙΝΑΚΕΣ	4
ΔΙΑΓΡΑΜΜΑΤΑ	4
ΠΕΡΙΛΗΨΗ	6
1. ΕΙΣΑΓΩΓΗ	8
2. ΘΕΩΡΗΤΙΚΟ ΜΕΡΟΣ & ΑΝΑΣΚΟΠΗΣΗ	11
2.1 Ιδιομορφίες των Έργων Μηχανικής Ιστού	11
2.2 Μεθοδολογίες Διαχείρισης Έργων	11
2.3 Παράγοντες Κινδύνου Έργων Λογισμικού.....	13
2.4 Διαχείριση Κόστους Έργου	15
2.4.1 Έλεγχος του Κόστους	16
2.5 Εκτίμηση Προσπάθειας.....	18
2.5.1 Εκτίμηση Κόστους.....	18
2.5.2 Μετρικές Μεγέθους Έργων	19
2.5.3 Παράγοντες που Επηρεάζουν την Προσπάθεια Ανάπτυξης Λογισμικού	20
2.5.4 Μέθοδοι Εκτίμησης Πόρων	21
2.5.5 Τύπος Δεδομένων Εκτίμησης Προσπάθειας.....	24
2.5.6 Προκλήσεις στις Εκτιμήσεις των Έργων Ανάπτυξης Λογισμικού	25
2.5.7 Κριτική Αξιολόγηση Μεθόδων.....	26
2.6 Διαδεδομένα Μοντέλα Εκτίμησης στον Κλάδο της Τεχνολογίας Λογισμικού	26
2.7 Σχετικές Έρευνες – Ανασκόπηση	28
2.8 Συμπεράσματα	29
3. ΜΕΘΟΔΟΛΟΓΙΑ ΕΡΕΥΝΑΣ	32
3.1 Συλλογή Δεδομένων	32
3.2 Επιλογή και Περιγραφή Μεταβλητών	33
3.3 Περιγραφική Στατιστική των Δεδομένων.....	41
3.3.1 Περιγραφή Συχνοτήτων	41
3.3.2 Περιγραφή Σχέσεων.....	52
3.4 Μέθοδος Ανάλυσης των Δεδομένων	58
3.4.1 Ανάλυση Συστάδων	58
3.4.3 Επιλογή & Καταλληλότητα Αλγορίθμων Συσταδοποίησης.....	58
3.4.4 Καταλληλότητα Αλγορίθμων Συσταδοποίησης	62
3.5 Προετοιμασία και Προ-επεξεργασία των Δεδομένων	65

3.5.1 Κλιμάκωση των Δεδομένων πριν τη Συσταδοποίηση	67
4. ΑΠΟΤΕΛΕΣΜΑΤΑ.....	70
4.1 K-means	70
4.1.1 Προσδιορισμός & Αξιολόγηση Αριθμού Συστάδων	70
4.1.2 Οπτικοποίηση των Clusters	73
4.1.3 Αξιολόγηση Λύσεων.....	75
4.2 Hierarchical Clustering	78
4.2.1 Οπτικοποίηση των Λύσεων	80
4.2.2 Αξιολόγηση των Λύσεων.....	84
4.3 PAM Clustering	84
4.3.1 Ο Βέλτιστος Αριθμός Συστάδων	84
4.3.2 Οπτικοποίηση των Συστάδων	86
4.3.3 Αξιολόγηση των Συστάδων	88
4.4 Διερεύνηση & Προσδιορισμός των Χαρακτηριστικών των Clusters	88
4.4.1 Περιγραφή των Χαρακτηριστικών των Clusters.....	88
4.4.2 Οπτικοποίηση των Χαρακτηριστικών των Clusters	92
4.5 Συγκριτική Αξιολόγηση των Clusters.....	97
5. ΣΥΜΠΕΡΑΣΜΑΤΑ & ΠΡΟΤΑΣΕΙΣ.....	98
5.1 Στρατηγική Α: Στοχεύοντας σε μεγάλα & πολύπλοκα έργα	100
5.2 Στρατηγική Β: Στοχεύοντας σε μεγάλο αριθμό μικρών & χαμηλής πολυπλοκότητας έργων.....	102
5.3 Η Επιλογή Στρατηγικής	104
5.4 Επίλογος.....	105
5.5 Προτάσεις για έρευνα	106
6. ΒΙΒΛΙΟΓΡΑΦΙΑ	108

ΕΙΚΟΝΕΣ

Εικ. 1. Αποτελέσματα των Έργων ανάλογα με τη Μεθοδολογία που χρησιμοποιείται.	13
Εικ. 2. Τεχνικές Εκτίμησης της Προσπάθειας βάσει Δεδομένων	23

ΠΙΝΑΚΕΣ

Πίν. 1. Οι Μεταβλητές & η Περιγραφή τους	34
Πίν. 2. Δείγμα από τα Δεδομένα.....	40
Πίν. 3. Συγκεντρωτικά Στατιστικά για τις Ποσοτικές Μεταβλητές	47
Πίν. 4. Οι Επιλεγθείσες Μεταβλητές.....	66
Πίν. 5. Συγκεντρωτικά Στατιστικά των Ποσοτικών Μεταβλητών του Cluster 1.....	90
Πίν. 6. Συγκεντρωτικά Στατιστικά των Ποσοτικών Μεταβλητών του Cluster 2.....	91
Πίν. 7. Συγκεντρωτικά Στατιστικά των Ποσοτικών Μεταβλητών του Cluster 3.....	92

ΔΙΑΓΡΑΜΜΑΤΑ

Διάγρ. 1. Συχνότητες Μεταβλητής "Company Size"	41
Διάγρ. 2. Συχνότητες Μεταβλητής "Project Type Category"	42
Διάγρ. 3. Συχνότητες Μεταβλητής "Website Technology Platform"	43
Διάγρ. 4. Συχνότητες Μεταβλητής "Complexity Status"	44
Διάγρ. 5. Συχνότητες Μεταβλητής "Seniority Dev1"	45
Διάγρ. 6. Συχνότητες Μεταβλητής "Change Requests"	46
Διάγρ. 7. Ιστόγραμμα Μεταβλητής "Duration Months"	47
Διάγρ. 8. Ιστόγραμμα Μεταβλητής "Spent Time" & "Estimated Effort Hours"	48
Διάγρ. 9. Ιστόγραμμα Μεταβλητής "Baseline" & "Cost"	49
Διάγρ. 10. Ιστόγραμμα Μεταβλητής "Hourly Cost"	50
Διάγρ. 11. Ιστόγραμμα Μεταβλητής "Profit"	50
Διάγρ. 12. Ιστόγραμμα Μεταβλητής "Profit Margin" & "Profit Ratio"	51
Διάγρ. 13. Ιστόγραμμα Μεταβλητής "Revenue"	51
Διάγρ. 14. Ιστόγραμμα Μεταβλητής "HTML_Templates"	52
Διάγρ. 15. Boxplots Ποσοτικών Μεταβλητών ανά Company Size.....	53
Διάγρ. 16. Boxplots Ποσοτικών Μεταβλητών ανά Project Type	54
Διάγρ. 17. Boxplots Ποσοτικών Μεταβλητών ανά Web Technology Platform	56
Διάγρ. 18. Boxplots Ποσοτικών Μεταβλητών ανά Complexity Status	57
Διάγρ. 19. Gap Statistic, Silhouette Width & Elbow Method for K-means.....	72
Διάγρ. 20. K-means, k =2, με το Outlier	73
Διάγρ. 21. K-means, k =2, χωρίς το Outlier	74
Διάγρ. 22. K-means, k =2, k=3, k=4.....	75
Διάγρ. 23. K-means Silhouette Width για k =2	76

Διάγρ. 24. K-means Silhouette Width για $k=3$	77
Διάγρ. 25. K-means Silhouette Width για $k=4$	77
Διάγρ. 26. Hierarchical Clustering with Outlier	78
Διάγρ. 27. Hierarchical Clustering Complete Method	79
Διάγρ. 28. Hierarchical Clustering Single Method.....	79
Διάγρ. 29. Hierarchical Clustering Average Method	80
Διάγρ. 30. Hierarchical Clustering, $k=2$	81
Διάγρ. 31. Silhouette Width Hier. Cl., $k=2$	81
Διάγρ. 32. Hierarchical Clustering, $k=3$	82
Διάγρ. 33. Silhouette Width Hier. Cl., $k=3$	82
Διάγρ. 34. Hierarchical Clustering, $k=4$	83
Διάγρ. 35. Silhouette Width Hier. Cl., $k=4$	83
Διάγρ. 36. Gap Statistic, Silhouette Width & Elbow Method for PAM.....	86
Διάγρ. 37. PAM clustering for $k=2$, $k=3$, $k=4$	87
Διάγρ. 38. Μέσοι Όροι Ποσοτικών Μεταβλητών ανά Cluster	93
Διάγρ. 39. Κατανομή Ποσοτικών Μεταβλητών ανά Cluster	94
Διάγρ. 40. Συχνότητες Ποιοτικών Μεταβλητών ανά Cluster.....	95
Διάγρ. 41. Σχετικές Συχνότητες Ποιοτικών Μεταβλητών ανά Cluster	96

ΠΕΡΙΛΗΨΗ

ΕΛΛΗΝΙΚΗ

ΕΞΕΡΕΥΝΩΝΤΑΣ ΤΗ ΣΧΕΣΗ ΜΕΤΑΞΥ ΚΕΡΔΟΦΟΡΙΑΣ ΚΑΙ ΠΟΙΟΤΙΚΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΤΩΝ ΕΡΓΩΝ ΛΟΓΙΣΜΙΚΟΥ ΣΤΟΝ ΙΣΤΟ: ΑΝΑΛΥΣΗ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ

Είναι τεκμηριωμένο σε όλη τη σχετική βιβλιογραφία ότι η εκτίμηση της προσπάθειας έχει αναδειχθεί ως ο πιο κρίσιμος και ζωτικός παράγοντας για τα έργα ανάπτυξης λογισμικού. Δεδομένης της εξατομικευμένης φύσης αυτών των έργων και ειδικά των έργων λογισμικού ιστού, και της έλλειψης επαρκών δεδομένων για τη χρήση μοντέλων πρόβλεψης (κέρδους, εσόδων ή εργατοωρών), σκοπός της παρούσας μελέτης είναι να διερευνήσει τη σχέση μεταξύ της υψηλής κερδοφορίας των έργων και άλλων διαθέσιμων ποιοτικών χαρακτηριστικών των έργων ώστε μέσω της τμηματοποίησης των έργων να μπορεί να παρέχει στρατηγικές κατευθύνσεις σχετικά με την αποτελεσματική διαχείριση των έργων και την κατάλληλη στόχευση των έργων - πελατών. Για τη διεξαγωγή αυτής της έρευνας, συλλέχθηκαν δεδομένα από πρόσφατα ολοκληρωμένα διαδικτυακά έργα μιας εταιρείας που δραστηριοποιείται αποκλειστικά στον κλάδο της ανάπτυξης έργων λογισμικού στον ιστό. Η μέθοδος που εφαρμόστηκε για την έρευνα και την προσέγγιση του προβλήματος είναι η μέθοδος συσταδοποίησης. Η ανάλυση των συστάδων παρείχε πολύτιμες πληροφορίες σχετικά με την ποιότητα των έργων και με βάση αυτές τις πληροφορίες μπορούν να δοθούν κατευθύνσεις στη διοικητική ομάδα και τους επαγγελματίες που δεν αφορούν μόνο τη συγκεκριμένη εταιρεία από την οποία ελήφθησαν τα δεδομένα, αλλά κάθε εταιρεία που δραστηριοποιείται στον τομέα της ανάπτυξης τέτοιων έργων, καθώς αποτελούν περισσότερες προσεγγίσεις στρατηγικής με στόχο την αύξηση κερδοφορίας.

Λέξεις-κλειδιά: εκτίμηση κόστους έργων λογισμικού, διαχείριση έργων, διαχείριση κόστους έργων, πρόβλεψη προσπάθειας έργων, συσταδοποίηση έργων λογισμικού, ανάλυση συστάδων

ΑΓΓΑΙΚΗ

EXPLORING THE RELATIONSHIP BETWEEN PROFITABILITY AND QUALITATIVE CHARACTERISTICS OF WEB DEVELOPMENT PROJECTS: CLUSTERING ANALYSIS

Effort estimation is a critical factor in software development, particularly in web software projects, where customization is prevalent, and data may not be sufficient for prediction models. This study aims to investigate the relationship between project profitability and other qualitative project characteristics. By clustering recently completed web projects of a web development company, the research provides strategic guidance for effective project management and client targeting. The cluster analysis provides valuable information about project quality, relevant not only to the specific company but also to any company in the web software project development field. The findings offer more strategic approaches to increase profitability and can be applied to guide management teams and professionals in project development.

Keywords: web-based Project Management, software effort prediction, software project management, effort estimation, cluster analysis

1. ΕΙΣΑΓΩΓΗ

Η ευρεία αποδοχή της τεχνολογίας και η πανταχού παρουσία του διαδικτύου έχει οδηγήσει με τη σειρά της στη μεγάλη άνθιση και επέκταση του κλάδου ανάπτυξης λογισμικού (web development industry). Και καθώς ο επιχειρηματικός κόσμος μετακινείται online και αποκτά ψηφιακή παρουσία, αναμένεται ότι αυτός ο κλάδος θα συνεχίσει να αναπτύσσεται με ραγδαίους ρυθμούς. Η τεράστια και αυξανόμενη ζήτηση από τις επιχειρήσεις, μικρομεσαίες ή μεγάλες, να μετακινήσουν εξ' ολοκλήρου ή εν μέρει, τις επιχειρηματικές τους δραστηριότητες στο διαδίκτυο έχει δημιουργήσει προκλήσεις για τον κλάδο του web development γενικότερα και για τη διαχείριση έργων (project management) ειδικότερα. Αυτές οι προκλήσεις είναι κρίσιμης σημασίας για τις επιχειρήσεις που δραστηριοποιούνται στο κλάδο του web development και αναδύονται συχνά κατά τη διάρκεια ενός project σε οποιαδήποτε φάση. Η μελέτη και η έρευνα των προκλήσεων αυτών, από διάφορους ενδιαφερομένους, ακαδημαϊκούς, ερευνητές ή και από τους ίδιους τους επαγγελματίες, είναι πολύ σημαντική στο να βρεθούν τρόποι προσέγγισης που θα ελαχιστοποιήσουν το επιχειρηματικό ρίσκο που προκύπτει από τις προκλήσεις στο κλάδο του web development.

Με την έναρξη ενός επαγγελματικού έργου ανάπτυξης λογισμικού εγείρονται σύντομα ορισμένα κρίσιμα ερωτήματα, όπως: Πόσο θα κοστίσει το έργο; Πότε θα ολοκληρωθεί; Πόση προσπάθεια πρέπει να επενδυθεί σε αυτό; Θα μπορέσει η επένδυση να αποσβεστεί σύντομα; Ποια είναι τα στοιχεία που χρειάζονται πραγματικά οι πελάτες μας; Η δυνατότητα απάντησης σε αυτά τα ερωτήματα και σε ορισμένα άλλα συναφή είναι ζωτικής σημασίας για το σχεδιασμό επιχειρηματικών στρατηγικών (π.χ. οικονομικών ή εμπορικών, μεταξύ άλλων) από τα αποτελέσματα του έργου. Οι απαντήσεις στα προαναφερθέντα ερωτήματα πρέπει να καθορίζουν όλες τις αποφάσεις, όπως η έναρξη ενός έργου ή κάποιου άλλου, ο τύπος του προϊόντος που πρέπει να αναπτυχθεί και τα χρήματα που πρέπει να επενδυθούν σε αυτό.

Είναι γνωστό ότι η εκτίμηση και ο σχεδιασμός ενός έργου ανάπτυξης λογισμικού είναι μια πολύπλοκη διαδικασία. Για την αντιμετώπιση αυτής της πρόκλησης, οι παραδοσιακές τεχνικές εκτίμησης επικεντρώνονται σε μια προγνωστική προσέγγιση, η οποία απαιτεί ένα σταθερό και οικείο περιβάλλον που μάλλον δεν ανταποκρίνεται στην πραγματικότητα.

Παρόλο που έχει σημειωθεί αρκετή έρευνα στο τομέα της εκτίμησης πόρων των έργων ανάπτυξης λογισμικού η ακριβέστερη εκτίμηση εξακολουθεί να αποτελεί προβληματισμός για τον κλάδο. Η ακριβής πρόβλεψη των πόρων είναι πολύ σημαντική επειδή οι ανακριβείς εκτιμήσεις μπορούν να οδηγήσουν σε σπατάλη χρόνου και οικονομικών πόρων. Το κύριο ζήτημα της εκτίμησης κόστους του έργου είναι να ληφθεί πρώτα μια ακριβής εκτίμηση μεγέθους του έργου. Μια συνολική εκτίμηση περιλαμβάνει την πρόβλεψη των πόρων σε εργατοώρες, το κόστος του έργου σε αξία νομίσματος και την ημερολογιακή διάρκεια που χρειάζεται για να ολοκληρωθεί το έργο. Σύμφωνα με τον Corazza et al, η αποτελεσματική εκτίμηση κόστους ενός έργου ανάπτυξης λογισμικού είναι αυτή που παρακολουθείται από την ομάδα προγραμματιστών (development team), τον διαχειριστή του έργου (Project Manager), γίνεται αποδεκτή από όλα τα ενδιαφερόμενα μέρη και βασίζεται σε εμπειρία προηγούμενων έργων.

Λόγω των αυξημένων απαιτήσεων για πιο ακριβείς και αποτελεσματικές τεχνικές που μπορούν να λειτουργήσουν και σε πιο περίπλοκες συνθήκες και περίπλοκα έργα λογισμικού, υπάρχει ένα σημαντικό πεδίο βελτίωσης τέτοιων τεχνικών. Ακόμη η ποικιλομορφία των μέτρων μεγέθους που προτείνονται στις διάφορες μελέτες και η απουσία αυτοματοποιημένων εργαλείων για την απλούστευση της διαδικασίας συλλογής δεδομένων, δημιουργεί δυσμενείς συνθήκες για έρευνα αλλά ταυτόχρονα αναδεικνύει νέες κατευθύνσεις στην έρευνα.

Η μέθοδος που θα εφαρμοστεί για την έρευνα και τη προσέγγιση του προβλήματος είναι η μέθοδος της συσταδοποίησης (clustering). Δεδομένης της εξατομικευμένης φύσης αυτών των έργων, της έλλειψης ικανών δεδομένων για χρήση μοντέλων πρόβλεψης (κέρδους, εσόδων ή εργατοωρών), θα προσεγγίσουμε το πρόβλημα εξερευνώντας τη σχέση μεταξύ της επιτυχίας των έργων σε όρους κερδοφορίας και των άλλων διαθέσιμων ποιοτικών χαρακτηριστικών τους, με στόχο την τελική τμηματοποίηση των έργων, ώστε να δώσει κατευθύνσεις στην επιχείρηση για το ποια έργα φαίνεται να δίνουν αξία στην εταιρία και στη συνέχεια με βάση αυτή τη γνώση να στοχεύσει τέτοιου είδους έργα. Η ιδέα είναι ότι αντί να εστιάζουν στη διαδικασία πρόβλεψης των ανθρωποωρών δηλαδή του κόστους σε micro επίπεδο για να μπορέσουν να διαχειριστούν καλύτερα τους πόρους και τα κόστη σε σχέση με τα έσοδα στα πλαίσια του project management, να εστιάσουν την «ποιότητα» των έργων που έχει αξία για την επιχείρηση. Μία τέτοια macro προσέγγιση σε σχέση με το είδος των έργων που έχει νόημα να αναλαμβάνουν, δίνει στρατηγική κατεύθυνση στην εταιρία που η αξία της μακροπρόθεσμα υπερκαλύπτει τις

προκλήσεις και τις δυσκολίες στην επιμέρους διαχείριση τους βραχυπρόθεσμα. Από την έρευνα προκύπτουν πολύτιμες πληροφορίες για τους ανθρώπους του “management” των εταιριών που αφορούν την καλύτερη διαχείριση των έργων και τη στρατηγική στόχευση πελατών- έργων.

Το επόμενο μέρος της εργασίας οργανώνεται ως εξής: στην ενότητα 2, παρουσιάζεται το θεωρητικό υπόβαθρο της διαχείρισης έργων, που εστιάζει διαχείριση κόστους και στις εκτιμήσεις των ανθρωποωρών εργασίας. Παρουσιάζονται επίσης διαδεδομένα μοντέλα πρόβλεψης και μια σύντομη ανασκόπηση σχετικών ερευνών. Στην ενότητα 3, παρουσιάζεται η μεθοδολογία της έρευνας και οι αλγόριθμοι που επιλέχθηκαν για την ανάλυση. Στην ενότητα 4 παρουσιάζονται τα αποτελέσματα της έρευνας για κάθε αλγόριθμο που εφαρμόστηκε και στην τελευταία ενότητα 5 συγκεντρώνονται τα συμπεράσματα και προτάσεις για περεταίρω βελτίωση της έρευνας.

2. ΘΕΩΡΗΤΙΚΟ ΜΕΡΟΣ & ΑΝΑΣΚΟΠΗΣΗ

2.1 Ιδιομορφίες των Έργων Μηχανικής Ιστού

Web Engineering VS Software Engineering Projects

Σήμερα, η άνοδος του Διαδικτύου και η παγκόσμια και διασυνδεδεμένη οικονομία έχουν αυξήσει τις ανάγκες για ταχεία προσαρμογή στις μεταβαλλόμενες ανάγκες των πελατών. Αυτές οι εξελίξεις προέκυψαν παράλληλα με την αποδοχή της Μηχανικής Ιστού (Web Engineering) ως κλάδου της Μηχανικής Λογισμικού (Software Engineering). Η Μηχανική Ιστού μπορεί να οριστεί ως ένα σύνολο μεθόδων, τεχνικών και εργαλείων στη Μηχανική Λογισμικού που βοηθούν μια ομάδα προγραμματιστών να δημιουργήσει συστήματα στον Παγκόσμιο Ιστό. Υπάρχουν διάφορα χαρακτηριστικά που διαφοροποιούν τα έργα Ιστού από τα υπόλοιπα έργα ανάπτυξης λογισμικού:

- Πολύπλοκη δομή πλοήγησης.
- Κρίσιμες απαιτήσεις διασύνδεσης - διεπαφής (όπως άγνωστοι χρήστες ή διαθεσιμότητα, μεταξύ άλλων).
- Πτυχές ασφάλειας.
- Αύξηση της αποδοτικότητας της συντήρησης, αποφυγή διακοπών λειτουργίας.
- Παράδοση το συντομότερο δυνατό.
- Μείωση του "χρόνου διάθεσης στην αγορά" (time-to-market).
- Προσαρμογή στις ταχέως μεταβαλλόμενες απαιτήσεις.

Είναι σημαντικό να επισημανθεί ότι ορισμένα από τα προαναφερθέντα χαρακτηριστικά δεν είναι αποκλειστικά των έργων Διαδικτύου (Web Development Projects) και μπορούν να εμφανιστούν και σε έργα που δεν ανήκουν στον Παγκόσμιο Ιστό. Παρ' όλα αυτά, η συνύπαρξη όλων αυτών μαζί ταυτόχρονα μπορεί να προσδιοριστεί ως μια ιδιομορφία των έργων Διαδικτύου.

2.2 Μεθοδολογίες Διαχείρισης Έργων

Τα έργα ανάπτυξης λογισμικού στο διαδίκτυο χαρακτηρίζονται, μεταξύ άλλων, από τη χρήση σύντομων κύκλων ανατροφοδότησης, οι οποίοι μπορεί να προκαλούν συχνές

αλλαγές στις απαιτήσεις. Εκτός αυτού, τα ιδιαίτερα χαρακτηριστικά των διαδικτυακών διεπαφών απαιτούν στενή συνεργασία και επαναλαμβανόμενο σχεδιασμό, γεγονός που θα προκαλέσει δυσκολίες, εάν το έργο εκτιμάται με όρους απόλυτων μεγεθών.

Σε τέτοια περιβάλλοντα, οι ευέλικτες μεθοδολογίες ανάπτυξης λογισμικού (Agile) με συνεχή παρακολούθηση, μέτρηση και συχνές παρεμβάσεις που βασίζονται κυρίως στη χρήση εμπειρικών διαδικασιών, μετατρέπονται σε μια σταθερή εναλλακτική λύση για τους οργανισμούς του κλάδου προκειμένου να σχεδιάσουν και να εκτιμήσουν τα έργα ανάπτυξης λογισμικού στον Παγκόσμιο Ιστό (Web projects). Αυτές οι μεθοδολογίες (Agile) προσφέρουν ένα κατάλληλο πλαίσιο για τα ιδιαίτερα χαρακτηριστικά των έργων ανάπτυξης λογισμικού στο Διαδίκτυο, όπως η γρήγορη ανταπόκριση στις αλλαγές, η προσαρμοστικότητα και η μείωση του χρόνου ανάπτυξης (web effort). Επιπλέον, η κλασική προσέγγιση όσον αφορά την εκ των προτέρων συλλογή απαιτήσεων (Waterfall) απαιτεί ένα σταθερό περιβάλλον, κάτι που δεν ισχύει στην περίπτωση των έργων αυτών, όπου οι απαιτήσεις αλλάζουν ταχύτερα. Ο σταδιακός και επαναληπτικός τρόπος επεξεργασίας των απαιτήσεων με την ευέλικτη μέθοδο (Agile Methodology) μπορεί να ταιριάζει καλύτερα στη συγκεκριμένη περίπτωση.

Η κλασική προσέγγιση της διαχείρισης έργων αναφέρει ότι ένα έργο επιτυγχάνει όταν συνδυάζει την επίτευξη των στόχων που έχουν τεθεί σε μεταβλητές όπως το κόστος, το χρονοδιάγραμμα και το πεδίο εφαρμογής (scope). Ακολουθώντας αυτά τα κριτήρια, η Standish Group διεξάγει τις γνωστές έρευνες CHAOS για τον έλεγχο της επιτυχίας των έργων. Ορίζουν ως επιτυχημένο έργο αυτό που εκτελείται εγκαίρως, εντός του προϋπολογισμού και περιλαμβάνει τα αρχικά καθορισμένα χαρακτηριστικά. Όταν πρόκειται για τεχνολογικά έργα, τα ευέλικτα έργα πετυχαίνουν εκεί που τα παραδοσιακά έργα αποτυγχάνουν. Στην πραγματικότητα, η πιο πρόσφατη μελέτη Standish Group Chaos Study από το 2021 δείχνει ότι τα ευέλικτα έργα έχουν 3 φορές περισσότερες πιθανότητες επιτυχίας από τα έργα Waterfall. Και τα έργα Waterfall έχουν 2x περισσότερες πιθανότητες να αποτύχουν. Στην εικόνα 1 παρακάτω, παρουσιάζεται το ποσοστό επιτυχίας των έργων ανάλογα με τον τύπο της μεθοδολογίας που χρησιμοποιείται:

PROJECT SUCCESS RATES



Εικ. 1. Αποτελέσματα των Έργων ανάλογα με τη Μεθοδολογία που χρησιμοποιείται

Τα στοιχεία ενός σχεδίου διαχείρισης έργου περιλαμβάνουν, μεταξύ άλλων, τα εξής:

Σχέδιο διαχείρισης χρονοδιαγράμματος: Το σχέδιο διαχείρισης χρονοδιαγράμματος καθορίζει τα κριτήρια και τις δραστηριότητες για την ανάπτυξη, την παρακολούθηση και τον έλεγχο του χρονοδιαγράμματος. Το σχέδιο διαχείρισης χρονοδιαγράμματος παρέχει διαδικασίες και ελέγχους που θα επηρεάσουν την εκτίμηση και τη διαχείριση του κόστους.

Σχέδιο διαχείρισης κινδύνων: Το σχέδιο διαχείρισης κινδύνων παρέχει την προσέγγιση για τον εντοπισμό, την ανάλυση και την παρακολούθηση των κινδύνων. Το σχέδιο διαχείρισης κινδύνων παρέχει διαδικασίες και ελέγχους που θα επηρεάσουν την εκτίμηση και τη διαχείριση του κόστους.

2.3 Παράγοντες Κινδύνου Έργων Λογισμικού

Αν και τα έργα θεωρούνται απαραίτητα στοιχεία σε πολλούς κλάδους, οι επαγγελματίες και οι ερευνητές έχουν συνειδητοποιήσει ότι τα υψηλά ποσοστά αποτυχίας των έργων κοστίζουν σε πολλές επιχειρήσεις αρκετά εκατομμύρια δολάρια ετησίως. Οι παράγοντες κινδύνου είναι αστάθμητα γεγονότα που μπορούν να αποτελέσουν σοβαρή απειλή για την επιτυχή ολοκλήρωση ενός έργου λογισμικού. Αρκετές μελέτες επισημαίνουν ότι η

αναποτελεσματική διαχείριση των κινδύνων είναι ένας από τους κύριους λόγους καθυστέρησης του χρονοδιαγράμματος σε έργα λογισμικού.

Παραδείγματα των πιο συνηθισμένων παραγόντων κινδύνου είναι: ασαφείς ή μεταβαλλόμενες απαιτήσεις, υποτιμημένη πολυπλοκότητα του έργου, ασταθές οργανωτικό περιβάλλον, έλλειψη διοικητικής υποστήριξης, κακή δέσμευση του χρήστη και ελλείψεις προσωπικού. Άλλες μελέτες επικεντρώθηκαν στη μέτρηση (του επιπέδου) του κινδύνου, στην κατανόηση της φύσης και των τύπων των κινδύνων και στον μετριασμό των συνιστωσών του κινδύνου. Οι Wallace κ.ά. (2004) εξέτασαν τη βιβλιογραφία στον τομέα και προσδιόρισαν έξι διαστάσεις των κινδύνων έργων λογισμικού: οργανωτικό περιβάλλον, χρήστης, απαιτήσεις, πολυπλοκότητα έργου, σχεδιασμός και έλεγχος και κίνδυνος ομάδας. Ανέλυσαν τις σχέσεις μεταξύ των διαστάσεων του κινδύνου και της συνολικής απόδοσης του έργου. Πιο πρόσφατα, οι Menezes et al. (2019) εντόπισαν 148 παράγοντες κινδύνου μέσω συστηματικής βιβλιογραφικής ανασκόπησης και τους ταξινόμησαν σύμφωνα με την ταξινομία SEI (Software Engineering Institute). Σχετικές έρευνες έχουν επίσης γίνει με τη χρήση στατιστικής ανάλυσης για την αξιολόγηση των παραγόντων κινδύνου σε έργα λογισμικού.

Στον τομέα της διαχείρισης έργων, οι αποτυχίες παρατηρούνται συχνότερα από τις επιτυχίες, ειδικά στον τομέα της πληροφορικής (James, 1997, Korac-Boisvert and Kouzmin, 1995).

Οι ερευνητές έχουν εντοπίσει διάφορους παράγοντες για τις αποτυχίες έργων, και οι σημαντικότεροι είναι οι εξής οι ακόλουθοι:

- Δέσμευση και συμμετοχή/υποστήριξη της ανώτατης διοίκησης
- Κατανομή των περιορισμένων πόρων
- Επικοινωνία μεταξύ των διαφόρων ενδιαφερομένων μερών
- Διαμόρφωση και δομή της ομάδας
- Κοινωνική συνοχή στην ομάδα και πολυπλοκότητα του έργου και οργανωτική κουλτούρα

Οι μελέτες σχετικά με τις αποτυχίες έργων είναι απαραίτητες, καθώς κάθε μια, σε γενικές γραμμές, συνδέεται με διάφορες οικονομικές και μη οικονομικές απώλειες που μπορούν να ανακόψουν την ανάπτυξη άλλων πιθανών έργων. Επιπλέον, η μάθηση από την αποτυχία ενός έργου μπορεί να διαδραματίσει βασικό ρόλο στο σχεδιασμό και την

παρακολούθηση μελλοντικών έργων για να εξασφαλιστεί η μακροπρόθεσμη επιτυχία κάθε οργανισμού που στοχεύει σε συνεχή βελτίωση (Kannan et al., 2017, Zheng et al., 2018). Γενικά, οι αποτυχίες έργων οδηγούν σε απώλειες πόρων διαφόρων ειδών, οι οποίες θα μπορούσαν να είναι λόγος για τα κίνητρα των ερευνητών στον τομέα αυτό (Espiner, 2007- McManus and WoodHarper, 2008- Bianco et al., 2016). Υπάρχουν πολλές μελέτες που διερευνούν τους λόγους/παράγοντες για τις αποτυχίες των έργων και προτείνουν μέτρα αποφυγής και άλλα προληπτικά μέτρα.

Όλα τα έργα φέρουν κινδύνους, δεδομένου ότι αποτελούν μοναδικά εγχειρήματα με διαφορετικό βαθμό πολυπλοκότητας που αποσκοπούν στην επίτευξη αποτελεσμάτων. Αυτό γίνεται σε ένα πλαίσιο περιορισμών και παραδοχών, ενώ παράλληλα ανταποκρίνονται στις προσδοκίες των ενδιαφερομένων μερών που μπορεί να είναι αντικρουόμενες και μεταβαλλόμενες. Οι οργανισμοί θα πρέπει να επιλέγουν να αναλαμβάνουν τον κίνδυνο ενός έργου με ελεγχόμενο και σκόπιμο τρόπο, προκειμένου να δημιουργούν αξία εξισορροπώντας τον ρίσκο και την ανταμοιβή.

Η διαχείριση έργων, όπως αναφέρθηκε και παραπάνω αποσκοπεί μεταξύ άλλων στον εντοπισμό και τη διαχείριση των κινδύνων. Το σχέδιο διαχείρισης κινδύνων παρέχει την προσέγγιση για τον εντοπισμό, την ανάλυση και την παρακολούθηση των κινδύνων. Επίσης παρέχει διαδικασίες και ελέγχους που θα επηρεάσουν την εκτίμηση και τη διαχείριση του κόστους. Όταν δεν γίνεται κατάλληλη διαχείριση, οι κίνδυνοι αυτοί έχουν τη δυνατότητα να προκαλέσουν απόκλιση του έργου από το σχέδιο και αποτυχία στην επίτευξη των καθορισμένων στόχων του έργου. Κατά συνέπεια, η αποτελεσματικότητα της διαχείρισης κινδύνων έργου συνδέεται άμεσα με την επιτυχία του έργου.

2.4 Διαχείριση Κόστους Έργου

Η διαχείριση του κόστους του έργου περιλαμβάνει τις διαδικασίες που εμπλέκονται στον προγραμματισμό, την εκτίμηση, τον προϋπολογισμό, τη χρηματοδότηση, τη διαχείριση και τον έλεγχο του κόστους, έτσι ώστε το έργο να μπορεί να ολοκληρωθεί εντός του εγκεκριμένου προϋπολογισμού. Οι διαδικασίες διαχείρισης κόστους έργου είναι οι εξής:

Διαχείριση κόστους σχεδίου: Η διαδικασία καθορισμού του τρόπου με τον οποίο θα εκτιμηθεί, θα προϋπολογιστεί, θα διαχειριστεί, θα παρακολουθηθεί και θα ελεγχθεί το κόστος του έργου.

Εκτίμηση του κόστους: Η διαδικασία ανάπτυξης μιας προσέγγισης των χρηματικών πόρων που απαιτούνται για την ολοκλήρωση των εργασιών του έργου.

Καθορισμός του προϋπολογισμού: Η διαδικασία συγκέντρωσης του εκτιμώμενου κόστους των επιμέρους δραστηριοτήτων ή πακέτων εργασίας για τον καθορισμό μιας εγκεκριμένης βάσης κόστους.

Έλεγχος του κόστους: Η διαδικασία παρακολούθησης της κατάστασης του έργου για την επικαιροποίηση του κόστους του έργου και τη διαχείριση των αλλαγών στη γραμμή βάσης κόστους.

Η διαχείριση του κόστους του έργου αφορά κυρίως το κόστος των πόρων που απαιτούνται για την ολοκλήρωση των δραστηριοτήτων του έργου και θα πρέπει να εξετάζει την επίδραση των αποφάσεων του έργου στο επακόλουθο επαναλαμβανόμενο κόστος χρήσης, συντήρησης και υποστήριξης του προϊόντος, της υπηρεσίας ή του αποτελέσματος του έργου. Για παράδειγμα, ο περιορισμός του αριθμού των αναθεωρήσεων σχεδιασμού μπορεί να μειώσει το κόστος του έργου, αλλά μπορεί να αυξήσει το κόστος λειτουργίας του προϊόντος που προκύπτει.

2.4.1 Έλεγχος του Κόστους

Οι τεχνικές ανάλυσης δεδομένων που μπορούν να χρησιμοποιηθούν για τον έλεγχο του κόστους περιλαμβάνουν μεταξύ άλλων:

Ανάλυση κερδισμένης αξίας (Earned Value Analysis). Η ανάλυση κερδισμένης αξίας συγκρίνει τη βασική γραμμή μέτρησης της απόδοσης με την πραγματική απόδοση του χρονοδιαγράμματος και του κόστους. Η μέθοδος αυτή ενσωματώνει τη βασική γραμμή εύρους με τη βασική γραμμή κόστους και τη βασική γραμμή χρονοδιαγράμματος για να σχηματίσει τη βασική γραμμή μέτρησης απόδοσης. Πιο αναλυτικά αναπτύσσει και παρακολουθεί τρεις βασικές διαστάσεις:

Προγραμματισμένη αξία (Planned Value). Η προγραμματισμένη αξία (PV) είναι ο εγκεκριμένος προϋπολογισμός που αποδίδεται σε προγραμματισμένες εργασίες, δηλαδή για τις εργασίες που πρέπει να πραγματοποιηθούν για να ολοκληρωθεί μια δραστηριότητα ή ένα στοιχείο της δομής ανάλυσης εργασιών. Αυτός ο προϋπολογισμός κατανέμεται ανά φάση κατά τη διάρκεια της ζωής του έργου, αλλά σε ένα δεδομένο

χρόνο, η προγραμματισμένη αξία ορίζει το πραγματικό έργο που θα έπρεπε να έχει επιτευχθεί.

Κερδισμένη Αξία (Earned Value): Η κερδισμένη αξία (EV) είναι ένα μέτρο της πραγματοποιηθείσας εργασίας που εκφράζεται σε όρους του προϋπολογισμού που έχει εγκριθεί για την εν λόγω εργασία. Είναι ο προϋπολογισμός που σχετίζεται με το εγκεκριμένο έργο που έχει ολοκληρωθεί. Η κερδισμένη αξία που μετράται πρέπει να σχετίζεται με τη βασική γραμμή μέτρησης της απόδοσης και δεν μπορεί να είναι μεγαλύτερη από τον εγκεκριμένο προϋπολογισμό προγραμματισμένης αξίας για ένα στοιχείο. Η κερδισμένη αξία χρησιμοποιείται συχνά για τον υπολογισμό του ποσοστού ολοκλήρωσης ενός έργου.

Πραγματικό κόστος (Actual Cost): Το πραγματικό κόστος (AC) είναι το πραγματοποιηθέν κόστος για τις εργασίες που εκτελούνται σε μια δραστηριότητα κατά τη διάρκεια μιας συγκεκριμένης χρονικής περιόδου. Είναι το συνολικό κόστος που πραγματοποιήθηκε για την εκτέλεση των εργασιών που μετρά η κερδισμένη αξία. Το πραγματικό κόστος πρέπει να αντιστοιχεί ως προς τον ορισμό σε αυτό που είχε προϋπολογιστεί στην προγραμματισμένη αξία και μετρήθηκε στην κερδισμένη αξία (π.χ. μόνο άμεσες ώρες, μόνο άμεσες δαπάνες ή όλες οι δαπάνες, συμπεριλαμβανομένων των έμμεσων δαπανών). Το πραγματικό κόστος δεν έχει ανώτατο όριο, και θα πρέπει να μετράται ό,τι πρέπει να δαπανηθεί για την επίτευξη της κερδισμένης αξίας.

Ανάλυση Απόκλισης (Variance analysis) : Η ανάλυση αποκλίσεων είναι η εξήγηση (αιτία, αντίκτυπος και διορθωτικές ενέργειες) για τις αποκλίσεις κόστους (Cost Variance = $\text{Earned Value} - \text{Actual Cost}$), χρονοδιαγράμματος (Schedule Variance = $\text{Earned Value} - \text{Planned Value}$) και απόκλισης κατά την ολοκλήρωση. Οι αποκλίσεις κόστους και χρονοδιαγράμματος είναι οι πιο συχνά αναλυόμενες μετρήσεις. Για έργα που δεν χρησιμοποιούν επίσημη ανάλυση κερδισμένης αξίας, παρόμοιες αναλύσεις αποκλίσεων μπορούν να πραγματοποιηθούν συγκρίνοντας το προγραμματισμένο κόστος με το πραγματικό κόστος για τον προσδιορισμό των αποκλίσεων μεταξύ της γραμμής κόστους και της πραγματικής απόδοσης του έργου. Περαιτέρω ανάλυση μπορεί να πραγματοποιηθεί για τον προσδιορισμό της αιτίας και του βαθμού της απόκλισης σε σχέση με τη γραμμή βάσης του χρονοδιαγράμματος και τυχόν διορθωτικές ή προληπτικές ενέργειες που απαιτούνται. Οι μετρήσεις απόδοσης κόστους χρησιμοποιούνται για την αξιολόγηση του μεγέθους της απόκλισης σε σχέση με την αρχική γραμμή βάσης κόστους. Το ποσοστιαίο εύρος των αποδεκτών αποκλίσεων τείνει να μειώνεται όσο περισσότερες

εργασίες πραγματοποιούνται. Τα πιο καθιερωμένα παραδείγματα ανάλυσης αποκλίσεων είναι:

Απόκλιση χρονοδιαγράμματος (Schedule Variance): Η απόκλιση του χρονοδιαγράμματος (SV) είναι ένα μέτρο της απόδοσης του χρονοδιαγράμματος, εκφρασμένη ως η διαφορά μεταξύ της κερδισμένης αξίας (Earned Value) και της προγραμματισμένης αξίας (Planned Value). Είναι το ποσοστό κατά το οποίο το έργο είναι μπροστά ή πίσω από την προγραμματισμένη ημερομηνία παράδοσης, σε μια δεδομένη χρονική στιγμή. Πρόκειται για ένα δείκτη της απόδοσης του χρονοδιαγράμματος σε ένα έργο. Ισούται με την κερδισμένη αξία (EV) μείον την προγραμματισμένη αξία (PV). Η απόκλιση του χρονοδιαγράμματος είναι ένα χρήσιμο μέτρο ενδεικτικό του πότε ένα έργο καθυστερεί ή προηγείται του βασικού του προγράμματος. Η απόκλιση του προγράμματος θα είναι τελικά μηδενική όταν ολοκληρωθεί το έργο, επειδή θα έχουν κερδηθεί όλες οι προγραμματισμένες αξίες (Planned Value). Εξίσωση : $SV = EV - PV$

Απόκλιση κόστους (Cost Variance): Η απόκλιση κόστους (CV) είναι το ποσό του ελλείμματος ή πλεονάσματος του προϋπολογισμού σε μια δεδομένη χρονική στιγμή, εκφρασμένη ως η διαφορά μεταξύ της κερδισμένης αξίας (Earned Value) και του πραγματικού κόστους (Actual Cost) . Είναι ένα μέτρο της απόδοσης του κόστους σε ένα έργο. Ισούται με την κερδισμένη αξία (EV) μείον το πραγματικό κόστος (AC). Η απόκλιση κόστους στο τέλος του έργου θα είναι η διαφορά μεταξύ του προϋπολογισμού κατά την ολοκλήρωση και του πραγματικού ποσού που δαπανήθηκε. Η απόκλιση κόστους (CV) είναι ιδιαίτερα κρίσιμη επειδή δείχνει τη σχέση της πραγματικής απόδοσης με το κόστος που δαπανήθηκε. Η αρνητική απόκλιση κάνει συχνά δύσκολη την ανάκαμψη του έργου. Εξίσωση: $CV = EV - AC$.

2.5 Εκτίμηση Προσπάθειας

2.5.1 Εκτίμηση Κόστους

Η εκτίμηση κόστους είναι η διαδικασία ανάπτυξης μιας προσέγγισης του κόστους των πόρων που απαιτούνται για την ολοκλήρωση των εργασιών του έργου. Το βασικό πλεονέκτημα αυτής της διαδικασίας είναι ότι καθορίζει τους χρηματικούς πόρους που

απαιτούνται για το έργο. Η διαδικασία αυτή εκτελείται περιοδικά καθ' όλη τη διάρκεια του έργου, ανάλογα με τις ανάγκες.

Είναι σαφές και εμπεριστατωμένο από διάφορους ερευνητές σε όλη τη σχετική βιβλιογραφία ότι η εκτίμηση του έργου σε πόρους (effort estimation) έχει αναδειχθεί ως ο πιο κρίσιμος και ζωτικής σημασίας παράγοντας για έργα ανάπτυξης λογισμικού. Η σημαντικότητα αυτού του παράγοντα έγκειται ακριβώς στην ακρίβεια να προβλέπει τους πόρους που χρειάζονται. Οι εκτιμήσεις αυτές όμως στην πραγματικότητα επηρεάζονται κυρίως από τις εκτιμήσεις του μεγέθους του έργου (size). Επομένως η ακρίβεια στο μέγεθος δημιουργεί την ακρίβεια και στην εκτίμηση των πόρων.

2.5.2 Μετρικές Μεγέθους Έργων

Η εκτίμηση του μεγέθους ενός έργου είναι μία ξεχωριστή οντότητα στο project management που εξαρτάται από διάφορες ανεξάρτητες ιδιότητες. Αυτές οι μεταβλητές ονομάζονται συντελεστές μέτρησης μεγέθους και χρησιμοποιούνται συγκεκριμένα για να αντιπροσωπεύσουν λειτουργικά χαρακτηριστικά, χαρακτηριστικά διαστάσεων ή πολυπλοκότητας ενός έργου. Η τυποποιημένη μονάδα ή κλίμακα που χρησιμοποιείται από τους επαγγελματίες για να εκφράσουν το μέγεθος της εφαρμογής ονομάζεται μετρήσεις μεγέθους. Οι μετρήσεις για το μέγεθος του έργου (web size metrics), όπως αναφέρεται και στη βιβλιογραφία, καθορίζεται από πληθώρα προδιαγραφών λειτουργικών απαιτήσεων που αναμένεται η εφαρμογή να ικανοποιεί και δεν υπάρχει κάποιο πρότυπο που μπορεί να καθοδηγήσει τους project managers στη σωστή εξαγωγή, επεξεργασία και επιλογή μέτρων μεγέθους (size measures) για την ανάπτυξη έργων λογισμικού. Ωστόσο, συναντάμε στη βιβλιογραφία διάφορες μελέτες και νέες προσεγγίσεις ως προς το sizing των έργων στο web development και την περιπλοκότητα (complexity) με στόχο την ανάπτυξη μοντέλων εκτίμησης του web effort, και να δώσουν τη δυνατότητα στους επαγγελματίες να εκτιμήσουν το μέγεθος ενός έργου με μεγαλύτερη ακρίβεια. Παρόλα αυτά, δεν υπάρχει καθολικότητα στην επιλογή τους. Η δυναμική φύση και η μεταβαλλόμενη αρχιτεκτονική των έργων ανάπτυξης λογισμικού, καθιστούν την επίτευξη ακριβών αποτελεσμάτων εκτίμησης, μία μεγάλη πρόκληση για το Project Management, που διακυβεύει και την οικονομική ευημερία της εταιρείας.

Το μέγεθος ενός έργου (size) είναι ένα από τα κύρια συστατικά της εξίσωσης κόστους για το web development, με το κέρδος και το ρίσκο του απρόοπτου ως άλλες δυο

μεταβλητές κόστους. Η εκτίμηση των πόρων (effort estimation) είναι μια συστηματική διαδικασία που χρησιμοποιείται για την πρόβλεψη της ποσότητας του ανθρώπινου δυναμικού που απαιτείται για την ανάπτυξη του έργου. Αυτό συνήθως εκφράζεται σε ανθρωποώρες, ή ανθρωποημέρες ή ακόμα και σε ανθρωπομήνες. Επομένως η ακριβής εκτίμηση οριοθετεί άμεσα και τους περιορισμούς του προϋπολογισμού που σχετίζονται με την ανάπτυξη του έργου, δηλαδή τους περιορισμούς που σχετίζονται με εκτιμήσεις των εσόδων και εξόδων του έργου. Οι κακές εκτιμήσεις προκαλούν είτε καθυστέρηση στην παράδοση του έργου είτε υπερβάσεις στον προϋπολογισμό. Και οι δύο περιστάσεις έχουν αρνητικές επιπτώσεις στην επιχείρηση. Όπως αναφέρθηκε και παραπάνω η ακρίβεια στο sizing του έργου οδηγεί σε ακριβή εκτίμηση των απαιτούμενων πόρων και ομοίως σε ακριβή εκτίμηση της διάρκειας του έργου. Συνεπώς προκύπτει και ακριβής εκτίμηση του κόστους του έργου, που είναι το ζητούμενο για την κερδοφορία σε επίπεδο έργου αλλά και σε επίπεδο επιχείρησης.

2.5.3 Παράγοντες που Επηρεάζουν την Προσπάθεια Ανάπτυξης Λογισμικού

Προηγούμενες μελέτες ανέλυσαν πληθώρα παραγόντων που επηρεάζουν την προσπάθεια ανάπτυξης λογισμικού (οι λεγόμενοι συντελεστές προσπάθειας – effort drivers). Οι Trendowicz κ.ά. (2014) εξέτασαν και χώρισαν τους πιο συχνά χρησιμοποιούμενους οδηγούς προσπάθειας σε τέσσερις κατηγορίες: *παράγοντες προσωπικού* (δηλ. ικανότητες και εμπειρία της ομάδας), *παράγοντες διαδικασίας* (δηλ. ποιότητα των εφαρμοζόμενων μεθόδων, εργαλείων και τεχνολογιών), *παράγοντες έργου* (δηλ. πόροι έργου και δραστηριότητες διαχείρισης) και *παράγοντες προϊόντος* (δηλ. προσπάθεια για ανάλυση απαιτήσεων, σχεδίαση και υλοποίηση κώδικα). Οι παράγοντες του προσωπικού και οι παράγοντες του έργου είναι οι πιο συχνά αναφερόμενοι παράγοντες προσπάθειας σε έργα με ευέλικτη μεθοδολογία (agile projects). Όπως παρατηρήθηκε σε προηγούμενες μελέτες, η ακρίβεια των μεθόδων εκτίμησης της προσπάθειας εξαρτάται από την επιλογή των σχετικών παραγόντων και την εξάλειψη των ασχέτων και παραπλανητικών παραγόντων.

Η ακριβής εκτίμηση των πόρων είναι δραστηριότητα κλειδί στην ανάπτυξη έργων λογισμικού. Επηρεάζει άμεσα το χρόνο και το κόστος των έργων. Οι υποεκτιμήσεις ή οι υπερεκτιμήσεις έχουν σημαντικές πιθανότητες να αποδειχθούν επικίνδυνες για τις επιχειρήσεις του κλάδου. Από τη μία μεριά οι υποεκτιμήσεις οδηγούν σε επιπτώσεις όπως υποστελέχωση και έλλειψη πόρων, αδυναμία εκπλήρωσης προθεσμιών, ή και αποτυχία

του ίδιου του έργου να εκπληρώσει το σκοπό του. Από την άλλη, οι υπερεκτιμήσεις από την άποψη του ανθρώπινου δυναμικού και άρα του κόστους του έργου, επιβαρύνουν τον προϋπολογισμό της επιχείρησης. Μελέτες αναφέρουν ότι υπάρχει μέχρι και 40% υπερεκτιμήσεις σε έργα ανάπτυξης λογισμικού. Επομένως οι ακριβείς και αποτελεσματικές εκτιμήσεις είναι σημαντική δραστηριότητα και παίζει βασικό ρόλο στη διαδικασία ανάπτυξης ενός έργου λογισμικού καθώς μειώνει το ρίσκο, συμβάλλει στην έγκαιρη παράδοση του έργου και εντός προϋπολογισμού. Επιπλέον θέτουν τη βάση για όλες τις δραστηριότητες κατά τη διάρκεια του έργου, συμπεριλαμβανομένης της διαχείρισης έργου (project management), του προγραμματισμού εργασιών και του προϋπολογισμού.

Υπάρχουν πολλές τεχνικές εκτίμησης, αλγοριθμικές ή μη αλγοριθμικές προσεγγίσεις, μοντέλα κτλ., οι οποίες έχουν όλες στόχο να βελτιστοποιήσουν την ακρίβεια στις εκτιμήσεις των πόρων για να βοηθήσουν τους επαγγελματίες σε αυτή τη διαδικασία της αποδοτικότερης διαχείρισης των έργων.

2.5.4 Μέθοδοι Εκτίμησης Πόρων

Η εκτίμηση των πόρων ενός έργου είναι ένας βασικός παράγοντας αποδοτικής διαχείρισης του έργου και η επίτευξη μιας πολύ ακριβούς εκτίμησης που αντιπροσωπεύει τον πραγματικό φόρτο εργασίας είναι ζητούμενο για τους επαγγελματίες δεδομένου ότι μία ακατάλληλη ή μη ρεαλιστική εκτίμηση διακυβεύει την επιτυχία ενός έργου. Σε μία μελέτη όπου ρωτήθηκαν περισσότεροι από 1000 προγραμματιστές και επαγγελματίες πληροφορικής, βρέθηκε ότι οι τρεις κύριοι λόγοι αποτυχίας ενός έργου ανάπτυξης λογισμικού σχετίζονται με την εκτίμηση των πόρων. Μια ανακριβής εκτίμηση έχει φυσικά αρνητικές συνέπειες που οδηγεί σε λανθασμένη κατανομή των πόρων.

Οι εκτιμήσεις των πόρων τέτοιων έργων εμπίπτουν σε δύο βασικές κατηγορίες: αυτές οι οποίες είναι βασισμένες σε δεδομένα και αυτές που βασίζονται στην κρίση και εμπειρία των ειδικών. Στην πρώτη, χρησιμοποιούνται ιστορικά δεδομένα από παρόμοια ολοκληρωμένα έργα, που περιέχουν ένα σύνολο χαρακτηριστικών του έργου πέρα από τις μεταβλητές μεγέθους του έργου, για να εκτιμήσουν τους πόρους που απαιτούνται σε ένα τρέχον έργο. Στη δεύτερη περίπτωση ζητείται η συμβουλή των εμπειρογνομόνων στη διαδικασία εκτίμησης, για να εκτιμηθούν οι πόροι που απαιτούνται για την ολοκλήρωση ενός μέρους ή και ολόκληρου του έργου.

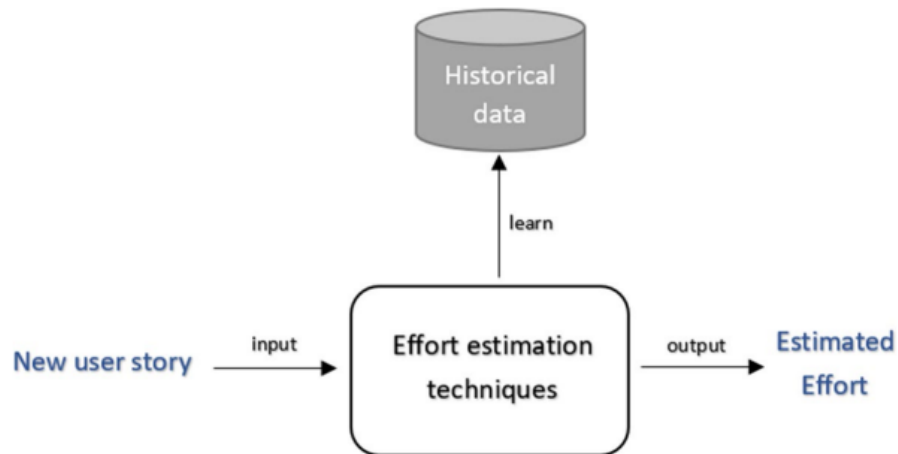
Εκτιμήσεις με βάση την εμπειρία των ειδικών

Όπως αναφέρθηκε και παραπάνω οι εκτιμήσεις αυτές βασίζονται στην εμπειρία των ειδικών, όπως επίσης και στην ανάλυση των διαθέσιμων πληροφοριών για ένα έργο, συμπεριλαμβανομένων των απαιτήσεων, του ίδιου του μεγέθους του έργου, των διαθέσιμων πόρων και εργαλείων. Η εκτίμηση μπορεί να γίνει από έναν μόνο εμπειρογνώμονα, ο οποίος μπορεί να χρησιμοποιήσει μία στρατηγική από κάτω προς τα πάνω (bottom up) για την εκτίμηση του έργου. Σε αυτή τη στρατηγική, ο εκτιμητής (συνήθως ο Project Manager), αναλύει και επιμερίζει ολόκληρο το έργο ή τις κύριες εργασίες σε επιμέρους μικρότερες εργασίες. Αυτές στη συνέχεια υπολογίζονται και εκτιμώνται ξεχωριστά. Τέλος, το άθροισμα αυτών των εκτιμήσεων, αντιπροσωπεύει τη συνολική εκτίμηση των πόρων του έργου. Επίσης η εκτίμηση του έργου μπορεί να γίνει και από μία ομάδα ειδικών. Έχει βρεθεί ότι οι ομαδικές εκτιμήσεις, είναι λιγότερο αισιόδοξες από αυτές που γίνονται από έναν μόνο ειδικό, πιο ρεαλιστικές και αντανakλούν καλύτερα την πραγματικότητα, δηλαδή τις δεδουλευμένες ώρες του έργου. Αυτό εξηγείται από το γεγονός ότι οι εις βάθος συζητήσεις δίνουν την δυνατότητα να εξερευνηθούν οι δραστηριότητες του έργου με μεγαλύτερη λεπτομέρεια και επομένως οι διαχειριστές του έργου καταλήγουν σε πιο ακριβείς εκτιμήσεις σχετικά με το μέγεθος της προσπάθειας που χρειάζεται για την υλοποίηση του έργου. Όσον αφορά στις ομαδικές εκτιμήσεις, αυτές μπορούν είναι είτε αδόμητες είτε δομημένες. Στις αδόμητες ομαδικές συζητήσεις, οι ειδικοί απλώς συζητούν χωρίς να ακολουθούν κάποια συγκεκριμένη προσέγγιση, μέχρι να καταλήξουν σε μια κοινή απόφαση για την εκτίμηση του έργου. Στις δομημένες εκτιμήσεις, ειδικοί βασίζονται σε συστηματικές προσεγγίσεις και τεχνικές.

Εκτιμήσεις βασισμένες σε δεδομένα

Στις μεθόδους που βασίζονται σε δεδομένα, οι επαγγελματίες χρησιμοποιούν ιστορικά δεδομένα από παρόμοια ολοκληρωμένα έργα για να κάνουν εκτιμήσεις. Αυτά τα δεδομένα, πέρα από δείκτες μεγέθους του έργου, περιλαμβάνουν και μεταβλητές σχετικά με τα χαρακτηριστικά του έργου. Το σύνολο αυτό των δεδομένων είναι το βασικό συστατικό στοιχείο για κάθε τέτοιου είδους τεχνική εκτίμησης. Ακόμη και αν ο εφαρμοσμένος αλγόριθμος ή μέθοδος έχουν εξαιρετικές αποδόσεις στις αναλύσεις, μία ακριβής εκτίμηση δεν μπορεί να επιτευχθεί αν έχει εξαχθεί από μη ρεαλιστικές πληροφορίες και μεταβλητές. Επομένως η κατάλληλη επιλογή των ανεξάρτητων μεταβλητών (effort drivers) μπορεί να οδηγήσει σε επιτυχείς εκτιμήσεις και το

αντίστροφο. Αυτές οι εκτιμήσεις που βασίζονται σε δεδομένα, γίνονται χρησιμοποιώντας μεθόδους είτε κατασκευής μοντέλων (model-based) ή αναλογιών (analogy-based methods). Οι μέθοδοι που βασίζονται σε μοντέλα χρησιμοποιούν τα ιστορικά δεδομένα σαν είσοδο (input) στο μοντέλο, το οποίο έχει σαν έξοδο (output) την ανθρώπινη προσπάθεια (web effort value), π.χ. ανθρωποώρες.



Εικ. 2. Τεχνικές Εκτίμησης της Προσπάθειας βάσει Δεδομένων

Σε αντίθεση, οι μέθοδοι που βασίζονται στην αναλογία χρησιμοποιούν ιστορικά δεδομένα ως βάση περίπτωσης για να αναζητήσουν ένα σχεδόν πανομοιότυπο προηγούμενο έργο, που ονομάζεται αναλογικό, και στη συνέχεια χρησιμοποιήσουν την πραγματική του προσπάθεια (actual effort) ως εκτίμηση στο τρέχον έργο.

Ωστόσο όσον αφορά σε άλλες «έξυπνες» προσεγγίσεις η μηχανική μάθηση (Machine Learning) χρησιμοποιείται ευρέως σε μελέτες εκτίμησης των πόρων των έργων. Η μάθηση μηχανής είναι μία μέθοδος βασισμένη σε μοντέλα που εκπαιδεύει μηχανές να σκέφτονται παρόμοια με τους ανθρώπους μαθαίνοντας από προηγούμενη εμπειρία και γνώση (Dam 2019). Η μηχανική εκμάθηση έχει δύο κύριες τεχνικές: εποπτευόμενη μάθηση (supervised learning) και χωρίς επίβλεψη μάθηση (unsupervised learning) (Alloghani et al. 2020). Στην πρώτη οι μεταβλητές εισόδου και εξόδου είναι γνωστές και η μηχανή θα εκπαιδευτεί σε μία συνάρτηση μεταβλητών για να προβλέψει την έξοδο μία νέας μεταβλητής εισόδου. Η ταξινόμηση (classification) και η παλινδρόμηση (regression) είναι οι δύο τεχνικές εποπτευόμενης μάθησης που χρησιμοποιούνται στη μηχανική μάθηση. Στη δεύτερη, ο αλγόριθμος μαθαίνει και ανακαλύπτει τα κρυμμένα μοτίβα (pattern detection) από δεδομένα χωρίς ετικέτα (unlabeled data). Η ευρέως

χρησιμοποιούμενη μέθοδος για την εκτίμηση των έργων ανάπτυξης λογισμικού είναι η εποπτευόμενη μάθηση (supervised learning).

2.5.5 Τύπος Δεδομένων Εκτίμησης Προσπάθειας

Όσον αφορά στα δεδομένα που χρησιμοποιούνται στις μεθόδους μηχανικής μάθησης για την εκτίμηση της ανθρώπινης προσπάθειας των έργων, υπάρχουν δύο τύποι: τα δομημένα και τα αδόμητα δεδομένα (structured or unstructured data).

Δομημένα

Τα δομημένα δεδομένα έρχονται συνήθως με τη μορφή μιας σχεσιακής βάσης δεδομένων, δηλαδή περιέχουν τόσο κατηγορικές όσο και ποσοτικές μεταβλητές. Αυτά τα χαρακτηριστικά πρέπει να έχουν επιλεγεί προσεκτικά, καθώς η επιτυχία ενός μοντέλου εκτίμησης των έργων εξαρτάται από τέτοια χαρακτηριστικά. Ένας βασικός περιορισμός των προσεγγίσεων μηχανικής μάθησης που ασχολούνται με αυτόν τον τύπο δεδομένων είναι ότι τα χαρακτηριστικά – μεταβλητές καθορίζονται χειροκίνητα. Επομένως, είναι προϋπόθεση η σωστή επιλογή των μεταβλητών για την ανάπτυξη ενός ακριβούς μοντέλου. Αυτή η χειροκίνητη επιλογή των χαρακτηριστικών καταναλώνει περισσότερο χρόνο και επιπλέον απαιτείται γνώση του κλάδου – τομέα (Choetkiertikul et al. 2019). Επιπρόσθετα, τα χαρακτηριστικά αυτά θα οριστούν λαμβάνοντας υπόψη το πλαίσιο των δεδομένων. Επομένως ενδέχεται να μην είναι κατάλληλα και να μην ταιριάζουν σε άλλα πλαίσια, για την εφαρμογή του μοντέλου. Ανεξάρτητα όμως των παραπάνω περιορισμών, η χρήση αυτών των τύπων δεδομένων, δίνει την ευελιξία στους επαγγελματίες να ορίσουν χαρακτηριστικά με ποιοτικά κριτήρια, όπως η εμπειρία των προγραμματιστών, η πολυπλοκότητα του έργου ή αλλά ενδεικτικά του μεγέθους του έργου.

Αδόμητα

Κατά την εκτίμηση των έργων, τα αδόμητα δεδομένα περιέχουν μόνο κειμενικές πληροφορίες (textual information) σχετικά με τα έργα. Μερικές φορές περιέχουν και εντολές προγραμματισμού. Το μοντέλα μηχανικής μάθησης που ασχολούνται με αδόμητα δεδομένα προσπαθούν να προβλέψουν την ανθρώπινη προσπάθεια από περιγραφές κειμένου.

2.5.6 Προκλήσεις στις Εκτιμήσεις των Έργων Ανάπτυξης Λογισμικού

Οι προκλήσεις που αντιμετωπίζουν οι επαγγελματίες του κλάδου στην επίτευξη ακριβών εκτιμήσεων είναι οι εξής:

Έλλειψη εμπειρίας στα μέλη της ομάδας. Ένας από τους λόγους των ανακριβών εκτιμήσεων είναι το γεγονός ότι η εκτιμήσεις αναλαμβάνονται από μια άπειρη ομάδα (Chongpakdee και Vatanawood 2017), που δεν είναι εξοικειωμένη με τις εργασίες του έργου ή δεν έχει επαρκή γνώση σχετικά με τον τρόπο μέτρησης της προσπάθειας για την υλοποίηση του έργου (López-Martínez et al. 2018).

Εξωτερική Πίεση. Αυτή η πίεση προέρχεται από τους διευθυντές ή τους πελάτες και μπορεί να οδηγήσει σε υπερεκτίμηση ή υποεκτίμηση της απαιτούμενης προσπάθειας για τις εργασίες (Altaleb et al. 2020). Για παράδειγμα προκειμένου να ικανοποιηθεί ο πελάτης ή ο διευθυντής του έργου, μπορεί να απαιτηθούν εργατοώρες για την παράδοση του έργου στον επιθυμητό χρόνο, προκαλώντας υποεκτίμηση της προσπάθειας που με τη σειρά της μπορεί να οδηγήσει σε κακής ποιότητας λογισμικό (Popli and Chauhan 2014)(Altaleb et al. 2020).

Αλληλουχία εργασιών. Ορισμένοι ερευνητές πιστεύουν ότι η ακολουθία με την οποία εκτιμώνται οι εργασίες ενός έργου έχει αντίκτυπο στους εκτιμητές και συνεπώς, επηρεάζει την ακρίβεια της εκτίμησης. Οι Grimstad και Jorgensen (2009) υποστηρίζουν ότι, αν υπάρχει πραγματική επίδραση από τη σειρά που ακολουθείται, μπορεί να υπάρχει μια βέλτιστη ακολουθία με την οποία να παράγεται η πιο ακριβής εκτίμηση. Αυτοί διεξήγαγαν ένα πείραμα σε δύο ομάδες: (1) εκτίμηση απαιτήσεων μικρού μεγέθους και (2) εκτίμηση απαιτήσεων μεγάλου μεγέθους. Στη συνέχεια και οι δύο ομάδες κλήθηκαν να εκτιμήσουν απαιτήσεις μεσαίου μεγέθους. Τα αποτελέσματα δείχνουν ότι η ομάδα που εκτιμούσε μικρές απαιτήσεις είχε την τάση να υποεκτιμά τις μεσαίου μεγέθους απαιτήσεις, ενώ η άλλη ομάδα είχε την τάση να υπερεκτιμά. Ένα παρόμοιο πρόσφατο πείραμα στο οποίο συμμετείχαν 104 επαγγελματίες λογισμικού (Jørgensen και Halkjelsvik 2020) ανέφερε ότι οι μεγάλες εργασίες υποτιμώνται όταν αντιμετωπίζονται μετά από μικρές εργασίες, ενώ μια μικρή εργασία υπερεκτιμάται όταν μια εκτίμηση για μια μεγάλη εργασία έχει προηγηθεί.

2.5.7 Κριτική Αξιολόγηση Μεθόδων

Μια πιθανή λύση στις προκλήσεις που αναφέρθηκαν είναι μια αυτοματοποιημένη μέθοδος εκτίμησης της προσπάθειας με βάση τα δεδομένα (data-driven), για τους ακόλουθους λόγους. Πρώτον, οι μέθοδοι με βάση τα δεδομένα μπορούν να εξαλείψουν ή να διορθώσουν τη μεροληψία της εκτίμησης με βάση τον άνθρωπο (human-based estimation). Ένας άνθρωπος μπορεί να επηρεάζεται από διάφορες πτυχές από τις οποίες μια μέθοδος βασισμένη σε δεδομένα δεν επηρεάζεται, για παράδειγμα, η αλληλουχία εργασιών και η εξωτερική πίεση, όπως αναφέρθηκε παραπάνω. Άλλες πτυχές που μπορεί να αυξήσουν την ανθρώπινη μεροληψία περιλαμβάνουν τον τρόπο σκέψης, τις δεξιότητες, την εκπαίδευση και τον οργανωτικό ρόλο (Jorgensen and Grimstad, 2012). Δεύτερον, η συμμετοχή εμπειρογνομόνων δεν απαιτείται κάθε φορά που μια ομάδα πρέπει να εκτιμήσει την προσπάθεια ενός έργου (Trendowicz and Jefery, 2014). Η τεχνική με βάση τα δεδομένα μαθαίνει και αποκτά γνώσεις από προηγούμενα έργα (ιστορικές δεδομένα), επομένως, θα ξεπεράσει τους περιορισμούς μιας άπειρης ομάδας χωρίς την ανάγκη για τη συμμετοχή των εμπειρογνομόνων. Τρίτον, μια αυτοματοποιημένη μέθοδος με βάση τα δεδομένα κοστίζει λιγότερο από την εκτίμηση βάσει εμπειρογνομόνων (Trendowicz and Jefery, 2014). Τέλος, σε σύγκριση με την εκτίμηση από ειδικούς, οι μέθοδοι που βασίζονται στα δεδομένα είναι περισσότερο επαναχρησιμοποιήσιμες και προσαρμόσιμες σε νέα έργα με παρόμοιο πλαίσιο (Trendowicz and Jefery, 2014).

2.6 Διαδεδομένα Μοντέλα Εκτίμησης στον Κλάδο της Τεχνολογίας Λογισμικού

Όπως παρουσιάστηκε προηγουμένως, η εκτίμηση της προσπάθειας και του κόστους ενός έργου ανάπτυξης λογισμικού είναι μια σύνθετη και βασική διαδικασία που πρέπει να διεξάγεται για να εκτιμηθεί κατά πόσον ένα έργο έχει καλή σχέση ποιότητας-απόδοσης. Πολλά μοντέλα έχουν προταθεί τα τελευταία 40 χρόνια για την αντιμετώπιση αυτού του προβλήματος. Αυτά μπορούν να ταξινομηθούν σε δύο κύριες κατηγορίες: Μοντέλα που δεν βασίζονται σε αλγόριθμους και μοντέλα που βασίζονται σε αλγόριθμους. Στην παρούσα ενότητα παρουσιάζεται μία επισκόπηση ορισμένων από τα πιο δημοφιλή μοντέλα που βασίζονται σε αυτές τις προσεγγίσεις.

Τα μοντέλα που βασίζονται σε αλγόριθμους (Algorithm-based models) χρησιμοποιούν μαθηματικές προσεγγίσεις για τον υπολογισμό της προσπάθειας του έργου ως συνάρτηση της σημαντικών παραγόντων κόστους. Ορισμένα από τα πιο σημαντικά μοντέλα που βασίζονται σε αλγόριθμους παρατίθενται και περιγράφονται παρακάτω:

COCOMO (Constructive Cost Model): Προτάθηκε αρχικά από τον Boehm το 1981 και η πραγματική του έκδοση με την ονομασία COCOMO II, δημοσιεύθηκε το 2000. Αυτό το μοντέλο χρησιμοποιεί έναν βασικό τύπο παλινδρόμησης και τα αποτελέσματά του βασίζονται κυρίως στο μέγεθος του κώδικα (που δίνεται σε χιλιάδες γραμμές κώδικα). Το μοντέλο COCOMO υπολογίζει τη διάρκεια και το κόστος ενός έργου λογισμικού βασιζόμενο στο μέγεθος του προϊόντος και την ποιότητα της ομάδας ανάπτυξης.

Function Points (Σημεία λειτουργίας): Αυτό το μοντέλο, που προτάθηκε από τους Albrecht και Gaffney (1979), παρέχει ένα μέτρο του προγράμματος με βάση τη λειτουργικότητα. Προσδιορίζονται οι λειτουργικές απαιτήσεις χρηστών του λογισμικού και ο συνολικός αριθμός των σημείων λειτουργίας που εξαρτώνται από την κάθε μία από αυτές, κατηγοριοποιείται σε έναν από αυτούς τους πέντε τύπους: έξοδοι, ερωτήματα αναζήτησης (inquiries), είσοδοι, εσωτερικά αρχεία και εξωτερικές διεπαφές. Μόλις η λειτουργία προσδιοριστεί και κατηγοριοποιηθεί σε έναν τύπο, στη συνέχεια αξιολογείται ως προς την πολυπλοκότητά της και της αποδίδεται ένας αριθμός σημείων λειτουργίας.

Use Case Point Technique (Περίπτωση Χρήσης): Προτάθηκε από τον Karner το 1993, το μοντέλο αυτό σχετίζεται με την Περίπτωση Χρήσης. Καθώς η Περίπτωση Χρήσης προσφέρει το λειτουργικό πεδίο εφαρμογής, η ανάλυση μπορεί να προσφέρει πολύτιμη γνώση για το μέγεθος ενός έργου ανάπτυξης.

Παρακάτω παρατίθενται και περιγράφονται ορισμένα από τα κυριότερα μοντέλα που δεν βασίζονται σε αλγόριθμους:

Estimation by analogy - Εκτίμηση με αναλογία: Η μέθοδος αυτή συνίσταται στη δημιουργία αναλογιών μεταξύ του πραγματικού έργου και του κόστους προηγούμενων και παρόμοιων έργων.

Delphi technique - Τεχνική Delphi: Η μέθοδος αυτή επικεντρώνεται στην άποψη μιας ομάδας εμπειρογνομόνων. Εκτελούνται διάφοροι γύροι διαβουλεύσεων με τη χρήση ερωτηματολογίων παρέχοντας στους συμμετέχοντες μια ενοποιημένη περίληψη των αποτελεσμάτων του προηγούμενου γύρου. Στη συνέχεια, οι εμπειρογνώμονες επανεξετάζουν τη γνώμη τους σύμφωνα με τα προηγούμενα αποτελέσματα. Υπάρχει

υπάρχει μια παραλλαγή αυτής της μεθόδου, που ονομάζεται Wideband Delphi, η οποία διαδόθηκε από τον Boehm, η οποία περιλαμβάνει μεγαλύτερη συμμετοχή από την πλευρά των εμπειρογνομόνων, καθώς οι εκτιμήσεις συζητούνται από κοινού σε μια κοινή συνεδρίαση.

Expert judgment method - Μέθοδος κρίσης εμπειρογνομόνων: Η μέθοδος αυτή αξιοποιεί την εμπειρία μιας ομάδας εμπειρογνομόνων για τον υπολογισμό ενός εκτιμώμενου κόστους. Η τεχνική αυτή χρησιμοποιείται μαζί με την τεχνική Delphi προκειμένου να βελτιωθεί και να συστηματοποιηθεί η γνώμη των εμπειρογνομόνων που ερωτώνται.

2.7 Σχετικές Έρευνες – Ανασκόπηση

Στην ενότητα αυτή παρουσιάζονται προηγούμενες ανασκοπήσεις των τεχνικών εκτίμησης της προσπάθειας ανάπτυξης λογισμικού. Μια προηγούμενη μελέτη στο Wen et al. (2012) εξετάζει τη χρήση της μηχανικής μάθησης στην εκτίμηση της προσπάθειας, συμπεριλαμβανομένης της ακρίβειας και του πλαισίου της. Αναλύοντας 84 πρωτογενείς μελέτες, οι συγγραφείς διαπίστωσαν ότι τα μοντέλα μηχανικής μάθησης τείνουν να είναι ακριβέστερα από τις μεθόδους που δεν χρησιμοποιούν μοντέλα. Ορισμένες ανασκοπήσεις αφορούν μελέτες εκτίμησης της προσπάθειας που επικεντρώνονται σε έναν συγκεκριμένο τύπο εφαρμοσμένης τεχνικής. Για παράδειγμα, οι Idria et al. (2016) αναλύουν εκείνες που κατασκευάζουν τεχνικές εκτίμησης της προσπάθειας σε σύνολο. Μετά την εξέταση 24 πρωτογενών μελετών, οι συγγραφείς διαπίστωσαν ότι οι τεχνικές αυτές ξεπερνούν τους περιορισμούς των μεμονωμένων τεχνικών εκτίμησης, εντοπίζοντας δύο κύριους τύπους (ομοιογενείς και ετερογενείς). Διαπίστωσαν επίσης ότι, συνήθως, οι τεχνικές συνόλου υπερτερούν έναντι των μεμονωμένων τεχνικών. Μια ανασκόπηση των Dave και Dutta (2014) εξέτασε τη χρήση του νευρωνικού δικτύου στην εκτίμηση της προσπάθειας λογισμικού. Ανέλυσαν 21 πρωτογενείς μελέτες όσον αφορά στο μοντέλο που χρησιμοποιήθηκε, στο σύνολο δεδομένων και στη μέθοδο αξιολόγησης. Το συμπέρασμα ήταν ότι δίνεται αυξημένη προσοχή στα μοντέλα νευρωνικών δικτύων για την εκτίμηση της προσπάθειας. Σημειώνουν επίσης ότι οι περισσότερες πρωτογενείς μελέτες χρησιμοποίησαν ένα μικρό σύνολο δεδομένων, όπως μόνο 63 έργα.

Σημαντική ποσότητα έρευνας έχει πραγματοποιηθεί για την εκτίμηση πόρων των έργων ανάπτυξης λογισμικού, χρησιμοποιώντας προσεγγίσεις μηχανικής μάθησης. Οι κυριότερες προσεγγίσεις μηχανικής μάθησης που χρησιμοποιούνται είναι τα Τεχνητά

Νευρωνικά Δίκτυα (Artificial Neural Networks), η Ασαφής Λογική (Fuzzy Logics), οι Γενετικοί Αλγόριθμοι (Genetic Algorithms) και τα Δέντρα Παλινδρόμησης (Regression Trees) για την εκτίμηση της προσπάθειας έργων λογισμικού. Ωστόσο μελετώντας την βιβλιογραφία μέχρι και σήμερα, αποκαλύπτεται η έλλειψη σε δεδομένα πραγματικής ζωής από τον κλάδο ανάπτυξης λογισμικού που να είναι σύμφωνα με τις τρέχουσες μεθόδους ανάπτυξης λογισμικού, καθώς επίσης και η ανάγκη για άλλες αξιόπιστες μετρήσεις που μπορούν να χρησιμοποιηθούν για την εκτίμηση των πόρων.

Υπάρχουν διάφορες μέθοδοι επικύρωσης που θα μπορούσαν να ληφθούν υπόψη για την ενίσχυση των μελετών για την επικύρωση των αποτελεσμάτων των εκτιμήσεων της προσπάθειας ανάπτυξης έργων λογισμικού. Οι κυριότερες μέθοδοι επικύρωσης είναι οι διασταυρούμενες επικυρώσεις (Cross Validations), η μέθοδος Jackknife και η επαναληπτική μέθοδος (Iterative method). Οι ερευνητικές τάσεις έχουν υποδείξει την ανάγκη διερεύνησης και επαύξησης των τεχνικών εκτίμησης. Περαιτέρω, οι μετρήσεις μεγέθους και άλλες προσεγγίσεις μηχανικής μάθησης, όπως τα δέντρα παλινδρόμησης, θα μπορούσαν επίσης να διερευνηθούν χρησιμοποιώντας σύνολα δεδομένων πραγματικής ζωής.

Εν κατακλείδι, οι τεχνικές μηχανικής μάθησης αλγορίθμου (Machine Learning) έχουν διερευνηθεί ευρέως για τη δημιουργία μοντέλων πρόβλεψης, ικανά να εκτιμήσουν τους απαιτούμενους πόρους ενός έργου, καθώς και να βελτιώσουν την ακρίβεια άλλων τεχνικών εκτίμησης. Τα Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Network), ως μοντέλο μηχανικής μάθησης, και το Μέσο Μέγεθος Σχετικού Σφάλματος (Mean Magnitude of Relative Error), ως μέτρο ακριβείας του μοντέλου έχουν χρησιμοποιηθεί ευρέως στις σχετικές μελέτες. Τα Νευρωνικά Δίκτυα και η Μηχανή Διανυσματικής Υποστήριξης (Support Vector Machine) είναι οι δύο τεχνικές που ξεχωρίζουν σε απόδοση συγκριτικά με άλλα μοντέλα μηχανικής μάθησης στις περισσότερες μελέτες.

2.8 Συμπεράσματα

Έχουμε ήδη αναφέρει ότι δύο από τα κύρια χαρακτηριστικά των έργων διαδικτύου (web projects) είναι η γρήγορη προσαρμογή στις αλλαγές και οι πολύπλοκες διεπαφές. Από τη μία πλευρά, η γρήγορη προσαρμογή στις αλλαγές συνεπάγεται αρκετές τροποποιήσεις στις απαιτήσεις. Η εκτίμηση έργου των μοντέλων που βασίζονται σε αλγόριθμους επικεντρώνεται στο εκτιμώμενο απόλυτο μέγεθος (αριθμός γραμμών κώδικα,

λειτουργικών απαιτήσεων ή περιπτώσεων χρήσης) του έργου, το οποίο αποκομίζεται στην αρχή μετά από μια ισχυρή φάση συλλογής απαιτήσεων. Αυτό μπορεί να οδηγήσει στην απόπειρα "παγώματος" των απαιτήσεων στο τέλος αυτής της φάσης. Αυτός ο περιορισμός μπορεί να μην ταιριάζει απόλυτα σε ορισμένα έργα στον Διαδίκτυο, όπου, όπως παρουσιάστηκε προηγουμένως, οι απαιτήσεις δεν είναι πλήρως γνωστές σε αυτές τις πρώιμες φάσεις και οι ανάγκες των χρηστών θα μπορούσαν να αλλάξουν γρήγορα κατά τη διάρκεια του έργου. Από την άλλη πλευρά, δεδομένου ότι η ύπαρξη πολύπλοκων διεπαφών είναι ένα άλλο χαρακτηριστικό των συστημάτων Ιστού, μια συνεργατική και επαναληπτική διαδικασία μεταξύ του χρήστη και της ομάδας προγραμματιστών που καθοδηγείται από μακέτες, διαγράμματα και δοκιμαστικές εφαρμογές θα είναι απαραίτητη για το σχεδιασμό αυτών των διεπαφών. Λόγω αυτού του γεγονότος, η εκ των προτέρων εκτίμηση με βάση τεχνικές όπως τα σημεία λειτουργίας ή τα σημεία περιπτώσεων χρήσης μπορεί να είναι δύσκολο να εφαρμοστεί. Τέλος, τα μη αλγοριθμικά μοντέλα που βασίζονται στην αναλογία, μπορούν να ταιριάζουν καλύτερα στα έργα του Παγκόσμιου Ιστού, καθώς δεν λαμβάνουν υπόψη το απόλυτο μέγεθος του έργου. Ωστόσο, αυτό το είδος εκτίμησης συνήθως εξετάζει την προσπάθεια και όχι την παραδοτέα αξία, η οποία είναι ζωτικής σημασίας για την επιλογή των λειτουργικών χαρακτηριστικών που πρέπει να υλοποιηθούν.

Οι υπάρχουσες μέθοδοι εκτίμησης μπορούν γενικά να ταξινομηθούν σε προσεγγίσεις που βασίζονται σε εμπειρογνώμονες και σε προσεγγίσεις που βασίζονται σε μοντέλα. Οι μέθοδοι που βασίζονται σε εμπειρογνώμονες βασίζονται στην ανθρώπινη εμπειρογνωμοσύνη για την επιλογή των σχετικών παραγόντων και είναι η πιο δημοφιλής τεχνική τόσο σε ευέλικτα (agile) όσο και σε παραδοσιακά έργα (waterfall). Οι μέθοδοι που βασίζονται σε μοντέλα αξιοποιούν δεδομένα από προγενέστερα έργα που καλύπτουν ένα ορισμένο αρχικό σύνολο παραγόντων προκειμένου να προσδιορίσουν ένα υποσύνολο παραγόντων που είναι σχετικοί. Και οι δύο μέθοδοι έχουν σημαντικούς πρακτικούς περιορισμούς όταν εφαρμόζονται μεμονωμένα. Οι αποφάσεις των εμπειρογνομόνων βασίζονται σε μεγάλο βαθμό στην υποκειμενική εκτίμηση, η οποία μπορεί να οδηγήσει σε ανακρίβειες και ασυνέπειες μεταξύ των εκτιμήσεων. Η αποτελεσματικότητα των μεθόδων που βασίζονται σε δεδομένα, από την άλλη πλευρά, εξαρτάται σε μεγάλο βαθμό από την ποσότητα και την ποιότητα των διαθέσιμων δεδομένων. Επομένως, είναι πιθανό ότι καμία στρατηγική δεν θα είναι η βέλτιστη για

όλες τις συνθήκες. Ως εκ τούτου, σε πρόσφατες μελέτες προτείνεται ο συνδυασμός μεθόδων που βασίζονται σε εμπειρογνώμονες και σε δεδομένα.

Τα μοντέλα μηχανικής μάθησης επίσης έχουν κερδίσει έδαφος ως εναλλακτική λύση στις μεθόδους εκτίμησης που βασίζονται σε μοντέλα. Έχουν επιτύχει πολλά υποσχόμενα αποτελέσματα στην εκτίμηση της προσπάθειας για έργα λογισμικού.

Εν κατακλείδι, είναι πιθανόν ότι η ιδιαίτερη φύση των έργων ανάπτυξης λογισμικού στο Διαδίκτυο δεν έχει ληφθεί υπόψη κατά τη διαδικασία εκτίμησης της προσπάθειας και του κόστους. Αντιθέτως, αντιμετωπίστηκαν όπως τα κλασικά έργα ανάπτυξης λογισμικού και εφαρμόστηκαν οι ίδιες προσεγγίσεις. Ωστόσο, όπως εξηγήθηκε, τα έργα Ιστού διαφέρουν από τα κλασικά έργα ανάπτυξης και έχουν να αντιμετωπίσουν μοναδικά ζητήματα, τόσο από τεχνική άποψη (π.χ. πολύπλοκες διεπαφές, πτυχές πλοήγησης ή πολύπλοκη συντήρηση, μεταξύ άλλων) όσο και από διαχειριστική άποψη (π.χ. ευμετάβλητες απαιτήσεις, σύντομα χρονοδιαγράμματα, μειωμένος χρόνος διάθεσης στην αγορά ή σύντομοι κύκλοι ανατροφοδότησης), τα οποία δεν μπορούν να αντιμετωπιστούν αποκλειστικά μέσω των κλασικών προσεγγίσεων εκτίμησης.

3. ΜΕΘΟΔΟΛΟΓΙΑ ΕΡΕΥΝΑΣ

3.1 Συλλογή Δεδομένων

Για τη διεξαγωγή της παρούσας έρευνας, συλλέχθηκαν δεδομένα από πρόσφατα ολοκληρωμένα έργα (web-based projects) εταιρίας που δραστηριοποιείται αποκλειστικά στο κλάδο ανάπτυξης έργων στο διαδίκτυο. Η συλλογή των δεδομένων πραγματοποιήθηκε σε δύο φάσεις:

1. Αδόμητες Συνεντεύξεις

Προκειμένου να αποκτηθεί ολοκληρωμένη γνώση των έργων και των χαρακτηριστικών τους, πραγματοποιήθηκαν συνεντεύξεις με τους Project Managers των έργων, τον Οικονομικό Διευθυντή και τον Chief Executive Officer της εταιρείας. Σε αυτές τις συνεντεύξεις, προσδιορίστηκαν ακριβώς οι προκλήσεις και τα προβλήματα που αντιμετωπίζει η εταιρεία σχετικά με τη διαχείριση των έργων, την κερδοφορία τους, και τη διαχείριση του κόστους τους. Επιπλέον, συζητήθηκαν ποιοτικά και ποσοτικά χαρακτηριστικά των έργων και άλλες μεταβλητές που επηρεάζουν σημαντικά την αποδοτική διαχείριση. Επίσης καταγράφηκαν και οι παράγοντες που εκτιμούν από την εμπειρία τους ότι επηρεάζουν σημαντικά την πιθανότητα επιτυχίας των έργων και τον χρόνο υλοποίησής τους. Τέλος αποτυπώθηκε η ανάγκη για μελέτη και αξιολόγηση των έργων ώστε να προσδιοριστούν αυτά που φαίνεται ότι φέρουν το μεγαλύτερο κέρδος για την επιχείρηση.

2. Προσδιορισμός των μεταβλητών και συλλογή των δεδομένων των έργων.

Με βάση τις πληροφορίες που συλλέχθηκαν από τις αδόμητες συνεντεύξεις, προσδιορίστηκαν συγκεκριμένες μεταβλητές προς ανάλυση και εφόσον ήταν διαθέσιμες και έπειτα άρχισε η καταγραφή και συγκέντρωση των δεδομένων.

Πιο συγκεκριμένα συλλέχθηκαν μεταβλητές που διακρίνονται σε 5 ομάδες:

- **ποιοτικά χαρακτηριστικά** των έργων, όπως το μέγεθος εταιρείας του πελάτη (Company_Size), ο τύπος κατηγορίας του project (Project_Type_Category), η τεχνολογία πλατφόρμας (Website_Technology_Platform).
- Χαρακτηριστικά που αφορούν σε **οικονομικά στοιχεία των έργων**, όπως το Κόστος με βάση την εκτίμηση (Baseline), το Πραγματικό Κόστος (Cost), Κέρδος € (Profit), Περιθώριο Κέρδους (Profit_Margin), η αναλογία του κέρδους σε σχέση

με το κόστος (Profit Ratio), τα έσοδα (Revenue) και το μέσο ωριαίο κόστος της εργατοώρας (Hourly_Cost).

- Χαρακτηριστικά που αφορούν στην **χρονική διάρκεια του έργου**, όπως οι πραγματικές καταγεγραμμένες ανθρωποώρες που καταναλώθηκαν (Spent_Time), η χρονική διάρκεια ενός έργου με βάση το τιμολόγιο έναρξης και λήξης (Duration_Months), και η εκτίμηση του έργου σε ανθρωποώρες (Estimated_Effort_Hours).
- Ένα χαρακτηριστικό που αφορά στην **ομάδα των προγραμματιστών**, δηλαδή το επίπεδο του κύριου προγραμματιστή (Seniority_Dev1) που δούλεψε το έργο.
- Και τέλος χαρακτηριστικά των έργων δηλωτικά του **μεγέθους του έργου**, όπως η πολυπλοκότητα (Complexity_Status), ο αριθμός των html templates που χρησιμοποιήθηκαν (HTML_Templates), και ο μέσος αριθμός των αλλαγών των απαιτήσεων κατά τη διάρκεια του έργου (Change_Requests).

3.2 Επιλογή και Περιγραφή Μεταβλητών

Συνολικά συγκεντρώθηκαν δεδομένα από 103 web projects με 17 μεταβλητές, όπως φαίνονται στον Πίνακα 1 παρακάτω:

Όνομα Μεταβλητής	Τύπος Μεταβλητής	Περιγραφή
Company_Size	factor	Το μέγεθος της εταιρίας από την οποία προέρχεται το έργο, δηλ. το μέγεθος του πελάτη σε αριθμό εργαζομένων. Περιλαμβάνει τρεις κλάσεις: 1-9, 10-49, 50-249, 250+
Project_Type_Category	factor	Ο τύπος του έργου όπως κατηγοριοποιείται από την εταιρεία και περιλαμβάνει 4 κατηγορίες έργων: eshop, landing page, platform, website
Website_Technology_Platform	factor	Αφορά στην τεχνολογία της πλατφόρμας που χρησιμοποιείται για την υλοποίηση του έργου και περιλαμβάνει 6 κατηγορίες: Decoupled Application, Drupal, HTML, Shopify, SquareSpace, WordPress
Duration_months	numeric	Είναι η ημερολογιακή διάρκεια του έργου σε μήνες η οποία υπολογίζεται με το τιμολόγιο έναρξης και λήξης, δηλ. οι μήνες από την έναρξη του έργου ως την παράδοση, αλλά δεν αφορά σε ανθρωποώρες που καταναλώθηκαν για το έργο.
Seniority_Dev1	factor	Αφορά στο επίπεδο και την εμπειρία του κύριου προγραμματιστή που έχει αναλάβει και δουλεύει το έργο και αντανακλά και το ύψος του μισθού του. Διακρίνεται σε 4 κλάσεις: Junior, Mid, Senior, Team Leader.
Spent_Time	numeric	Είναι οι καταγεγραμμένες ώρες που έχουν καταναλωθεί για την υλοποίηση του έργου
Baseline	numeric	Αφορά στην αρχική εκτίμηση του κόστους σε ευρώ για το έργο, που υπολογίζεται με βάση τις ανθρωποώρες που εκτιμούνται πολλαπλασιασμένες με το μεσο κόστος εργατοώρας της εταιρίας.
Cost	numeric	Είναι το παραγματικό κόστος του έργου σε ευρώ, που υπολογίζεται με βάση τις ώρες που καταναλώθηκαν (spent time) πολλαπλασιασμένες με το σταθμισμένο κόστος εργατοώρας, ανάλογα με το επίπεδο του προγραμματιστή.
Profit	numeric	Είναι το κέρδος του έργου σε ευρώ που προκύπτει μετά την αφαίρεση του κόστους από τα έσοδα του έργου.
Profit_Margin	numeric	Είναι το περιθώριο κέρδους κάθε έργου εκφρασμένο σε ποσοστό, το οποίο προκύπτει από τον λόγο κέρδους προς έσοδα.
Profit_Ratio	numeric	Είναι ένας δείκτης της "καθαρής επιστροφής" της επένδυσης (κόστους) και προκύπτει από τον λόγο των εσόδων προς το κόστος -1. Εκφράζει την αναλογία κέρδους προς το κόστος.
Hourly_Cost	numeric	Είναι το σταθμισμένο κόστος εργατοώρας του κάθε έργου ανάλογα με το επίπεδο του προγραμματιστή που δουλεύει το έργο.
Estimated_Effort_Hours	numeric	Αφορά την αρχική εκτίμηση του φόρτου εργασίας του έργου σε ανθρωποώρες.
Revenue	numeric	Είναι τα έσοδα που έχει φέρει το κάθε έργο σε ευρώ
Complexity_Status	factor	Είναι ένα μέτρο αξιολόγησης του έργου σε όρους πολυπλοκότητας και μεγεθους, από την εταιρία. Περιλαμβάνει 3 κλάσεις: Low, Medium, High
Change_Requests	factor	Αφορά σε ένα εύρος αλλαγών των απαιτήσεων του έργου από τη μεριά του πελάτη, ενδεικτικό της πιθανής καθυστέρησης του έργου και αξιολόγησης της συνεργασίας του πελάτη. Περιλαμβάνει 4 κλάσεις: 0, 1-4, 5-9, 10+
HTML_Templates	numeric	Είναι ο αριθμός των διαφορετικών html templates που κατασκευάστηκαν για την υλοποίηση του έργου, ως ένας δείκτης μεγέθους του έργου.

Πίν. 1. Οι Μεταβλητές & η Περιγραφή τους

Από τις 17 μεταβλητές, οι 6 είναι κατηγορικές και παρακάτω ακολουθεί περαιτέρω επεξήγηση των μεταβλητών και των επιμέρους τιμών που παίρνει η κάθε μία κατηγορική μεταβλητή.

Project Type Category

E-Shop: Ένα ηλεκτρονικό κατάστημα (e-shop) είναι ένας τύπος διαδικτυακού έργου που εμπίπτει στην κατηγορία των δικτυακών ιστοτόπων ηλεκτρονικού εμπορίου. Ένας δικτυακός ιστότοπος ηλεκτρονικού εμπορίου έχει σχεδιαστεί για να πωλεί προϊόντα ή υπηρεσίες στο διαδίκτυο, και ένα e-shop αναφέρεται συγκεκριμένα σε ένα ηλεκτρονικό κατάστημα που πωλεί προϊόντα απευθείας στους καταναλωτές. Είναι ένας δημοφιλής τύπος έργου στον κλάδο της ανάπτυξης ιστοσελίδων. Τα ηλεκτρονικά καταστήματα περιλαμβάνουν συνήθως χαρακτηριστικά όπως λίστες προϊόντων, καλάθια αγορών, πύλες πληρωμών και συστήματα παρακολούθησης παραγγελιών. Τα e-shops μπορούν να κατασκευαστούν χρησιμοποιώντας διάφορες τεχνολογίες ανάπτυξης ιστού, όπως PHP, Ruby, Python ή Java. Μπορεί επίσης να χρησιμοποιούν ένα σύστημα διαχείρισης περιεχομένου (CMS), όπως το WooCommerce για WordPress ή το Magento, για τη διαχείριση των καταλόγων προϊόντων, των δεδομένων των πελατών και των παραγγελιών τους.

Landing page: Η σελίδα προορισμού είναι ένας τύπος διαδικτυακού έργου που έχει σχεδιαστεί για να ενθαρρύνει τους επισκέπτες να προβούν σε μια συγκεκριμένη ενέργεια, όπως η εγγραφή σε ένα ενημερωτικό δελτίο, η λήψη ενός δωρεάν υλικού ή η αγορά ενός προϊόντος. Οι σελίδες προορισμού χρησιμοποιούνται συχνά ως μέρος εκστρατειών ψηφιακού μάρκετινγκ για την προώθηση της επισκεψιμότητας σε μια συγκεκριμένη προσφορά ή προώθηση. Συχνά σχεδιάζονται με απλή διάταξη, με ελάχιστους περισπασμούς και ένα σαφές μήνυμα που επικοινωνεί τα οφέλη της προσφοράς. Οι σελίδες προορισμού μπορούν να κατασκευαστούν χρησιμοποιώντας διάφορες τεχνολογίες ανάπτυξης ιστοσελίδων, αλλά συνήθως σχεδιάζονται για να είναι ελαφριές και να φορτώνουν γρήγορα, με έμφαση στην εμπειρία χρήστη και τη βελτιστοποίηση της μετατροπής της επίσκεψης σε συγκεκριμένη δράση. Μπορεί να περιλαμβάνουν στοιχεία όπως φόρμες, κουμπιά, εικόνες, βίντεο κ.α.

Platform: Μια πλατφόρμα είναι ένας τύπος διαδικτυακού έργου που συνήθως παρέχει έναν κεντρικό σταθμό στους χρήστες για την πρόσβαση σε μια ποικιλία σχετικών υπηρεσιών ή χαρακτηριστικών. Οι πλατφόρμες συχνά σχεδιάζονται ώστε να είναι ιδιαίτερα επεκτάσιμες και ευέλικτες, με δυνατότητα υποστήριξης μεγάλου αριθμού χρηστών και πολύπλοκων αλληλεπιδράσεων μεταξύ τους.

Οι πλατφόρμες μπορούν να πάρουν πολλές διαφορετικές μορφές ανάλογα με τη συγκεκριμένη εστίαση και τη λειτουργικότητά τους. Ορισμένα παραδείγματα έργων πλατφόρμας περιλαμβάνουν: πλατφόρμες κοινωνικής δικτύωσης, πλατφόρμες αναζήτησης εργασίας, ηλεκτρονικής μάθησης, πλατφόρμες cloud computing.

Website: Ένας απλός ιστότοπος είναι ένας βασικός τύπος διαδικτυακού έργου που συνήθως αποτελείται από μερικές σελίδες με στατικό περιεχόμενο που παρέχουν πληροφορίες σχετικά με μια επιχείρηση ή έναν οργανισμό. Ένας απλός ιστότοπος μπορεί να κατασκευαστεί χρησιμοποιώντας μια ποικιλία τεχνολογιών ανάπτυξης ιστού, αλλά συνήθως σχεδιάζεται έτσι ώστε να είναι ελαφρύς και εύκολος στην πλοήγηση, με έμφαση στην παροχή σαφών και συνοπτικών πληροφοριών στους επισκέπτες. Οι απλοί ιστότοποι μπορεί να περιλαμβάνουν κάποια βασική λειτουργικότητα, αλλά γενικά δεν απαιτούν προηγμένα χαρακτηριστικά, όπως πιστοποίηση ταυτότητας χρήστη ή ενσωμάτωση βάσης δεδομένων.

Website Technology Platform

Decoupled Application: Decoupled Application είναι ένας τύπος τεχνολογίας που διαχωρίζει το επίπεδο παρουσίασης του front-end από τη λειτουργικότητα του back-end και το επίπεδο αποθήκευσης δεδομένων μιας διαδικτυακής εφαρμογής. Αυτό μπορεί να προσφέρει βελτιωμένη ευελιξία, επεκτασιμότητα και απόδοση, καθώς και μεγαλύτερη προσαρμογή και ενσωμάτωση με άλλα συστήματα μέσω API. Για παράδειγμα, μια εφαρμογή decoupled θα μπορούσε να χρησιμοποιήσει ένα σύγχρονο front-end πλαίσιο όπως το React ή το Angular για το επίπεδο παρουσίασης, ενώ θα μπορούσε να χρησιμοποιήσει ένα back-end Content Management System όπως το Drupal ή το WordPress για τη διαχείριση περιεχομένου και την αποθήκευση δεδομένων.

Drupal: Το Drupal είναι ένα ελεύθερο και ανοικτού κώδικα σύστημα διαχείρισης περιεχομένου (Content Management System) και ένα πλαίσιο εφαρμογών ιστού που χρησιμοποιείται για την κατασκευή και διαχείριση ιστότοπων και εφαρμογών ιστού. Είναι κατασκευασμένο με τη χρήση γλώσσας προγραμματισμού PHP και έχει σχεδιαστεί για να είναι ιδιαίτερα ευέλικτο, επεκτάσιμο και προσαρμόσιμο. Το Drupal είναι ιδιαίτερα κατάλληλο για σύνθετα και μεγάλης κλίμακας έργα ιστού, με ισχυρά χαρακτηριστικά ασφαλείας και δυνατότητα διαχείρισης μεγάλου όγκου περιεχομένου και επισκεψιμότητας.

HTML: Η HTML (Hypertext Markup Language) είναι μια τυποποιημένη γλώσσα σήμανσης που χρησιμοποιείται για τη δημιουργία και την προβολή περιεχομένου στον Παγκόσμιο Ιστό. Παρέχει τη δομή και τη μορφή των ιστοσελίδων, επιτρέποντας στους προγραμματιστές να δημιουργούν επικεφαλίδες, παραγράφους, λίστες, συνδέσμους, εικόνες και άλλους τύπους περιεχομένου. Η HTML είναι μια θεμελιώδης τεχνολογία για την ανάπτυξη ιστοσελίδων και χρησιμοποιείται ευρέως, με την τρέχουσα έκδοση, την HTML5, να υποστηρίζει πολλά νέα χαρακτηριστικά και δυνατότητες.

Shopify: Το Shopify είναι μια δημοφιλής και ευρέως χρησιμοποιούμενη πλατφόρμα ηλεκτρονικού εμπορίου που είναι ιδανική για επιχειρήσεις που επιθυμούν να δημιουργήσουν γρήγορα και εύκολα ένα ηλεκτρονικό κατάστημα. Παρέχει μια συλλογή εργαλείων για τον σχεδιασμό ιστότοπων, τη διαχείριση προϊόντων, την επεξεργασία πληρωμών και την εκτέλεση αποστολών, καθώς και μια σειρά από προσαρμόσιμα θέματα και πρότυπα. Η πλήρως φιλοξενούμενη πλατφόρμα του και ο φιλικός προς το χρήστη σχεδιασμός του το καθιστούν μια προσιτή επιλογή για επιχειρήσεις όλων των μεγεθών.

SquareSpace: Το Squarespace είναι μια cloud-based πλατφόρμα κατασκευής και φιλοξενίας ιστότοπων που επιτρέπει σε ιδιώτες και επιχειρήσεις να δημιουργούν και να διαχειρίζονται εύκολα ιστότοπους με επαγγελματική εμφάνιση. Προσφέρει μια σειρά από προσαρμόσιμα πρότυπα και θέματα, καθώς και εργαλεία για τον σχεδιασμό ιστότοπων, το ηλεκτρονικό εμπόριο, το μάρκετινγκ και το Search Engine Optimization (SEO). Το Squarespace είναι φιλικό προς τον χρήστη και μια προσιτή επιλογή για χρήστες όλων των τεχνικών επιπέδων δεξιοτήτων.

WordPress: WordPress είναι ένα σύστημα διαχείρισης περιεχομένου ανοικτού κώδικα (CMS), που επιτρέπει στους χρήστες να δημιουργούν και να διαχειρίζονται ιστότοπους και ιστολόγια (blogs). Είναι ιδιαίτερα ευέλικτο και προσαρμόσιμο, με μια σειρά από θέματα, πρότυπα, πρόσθετα και widgets που προσθέτουν πρόσθετη λειτουργικότητα. Το WordPress είναι μια δημοφιλής επιλογή για την ανάπτυξη ιστότοπων, με μια μεγάλη και ενεργή κοινότητα προγραμματιστών που συνεισφέρουν πρόσθετα χαρακτηριστικά και λειτουργίες στην πλατφόρμα.

Seniority Dev1

Το επίπεδο εμπειρίας των προγραμματιστών διακρίνεται σε “Junior”, “Mid”, “Senior” και “Team Leader”. Η χρονική διάρκεια για τη μετάβαση από ένα επίπεδο στο άλλο κυμαίνεται περίπου στα 3 χρόνια και η εμπειρία του κάθε προγραμματιστή μεταφράζεται

και σε μισθολογική κλίμακα. Δηλαδή, ένας “junior” προγραμματιστής αμείβεται λιγότερο από ένα “mid” και ούτω κάθε εξής. Σε γενικές γραμμές, η αμοιβή ενός “junior” κυμαίνεται μεταξύ 850 – 1000€, ενός “mid” γύρω στα 1300€, ενός “senior” περίπου στα 1600€ και ενός “Team Leader” στα 1900€.

Complexity Status:

Τα έργα διακρίνονται σε 3 κατηγορίες ανάλογα με την πολυπλοκότητα τους: “Low”, “Medium” και “High”, δηλαδή χαμηλής, μεσαίας και υψηλής πολυπλοκότητας.

Low complexity: Ένα έργο ιστού χαμηλής πολυπλοκότητας αναφέρεται συνήθως σε ένα έργο που είναι σχετικά απλό όσον αφορά τις απαιτήσεις και τη λειτουργικότητά του. Ένα τέτοιο έργο μπορεί να περιλαμβάνει μικρό αριθμό ιστοσελίδων, βασικά στοιχεία σχεδιασμού και απλή λειτουργικότητα, όπως ιστότοπος παρουσίασης για μια μικρή επιχείρηση, ή ένα προσωπικό ιστολόγιο. Τα έργα ιστού χαμηλής πολυπλοκότητας αναπτύσσονται συχνά σχετικά γρήγορα και εύκολα και μπορούν να ολοκληρωθούν σε σύντομο χρονικό διάστημα και με σχετικά χαμηλό κόστος. Μπορεί να είναι κατάλληλα για ιδιώτες ή μικρές επιχειρήσεις που επιθυμούν να δημιουργήσουν μια βασική διαδικτυακή παρουσία ή για έργα με περιορισμένο προϋπολογισμό ή χρονοδιάγραμμα.

Medium Complexity: Ένα έργο ιστού μεσαίας πολυπλοκότητας αναφέρεται συνήθως σε ένα έργο που είναι πιο πολύπλοκο από έναν απλό ιστότοπο, αλλά όχι τόσο πολύπλοκο όσο ένας ιστότοπος ή μια εφαρμογή μεγάλης κλίμακας σε επίπεδο επιχείρησης. Ένα τέτοιο έργο μπορεί να περιλαμβάνει περισσότερες ιστοσελίδες, προηγμένα στοιχεία σχεδιασμού και πιο σύνθετη λειτουργικότητα, όπως ηλεκτρονικό εμπόριο με πολλαπλά προϊόντα, σύστημα σύνδεσης χρηστών ή ενσωμάτωση με υπηρεσίες από τρίτα μέρη. Μπορεί να περιλαμβάνει μεγαλύτερη ομάδα προγραμματιστών και να απαιτεί πιο προχωρημένες δεξιότητες προγραμματισμού και γνώσεις τεχνολογιών ανάπτυξης ιστοσελίδων.

High Complexity: Ένα έργο ιστού υψηλής πολυπλοκότητας αναφέρεται συνήθως σε έναν ιστότοπο ή μια εφαρμογή μεγάλης κλίμακας που απαιτεί προηγμένη λειτουργικότητα και χαρακτηριστικά, υψηλό επίπεδο προσαρμογής και εξελιγμένα στοιχεία σχεδιασμού. Ένα τέτοιο έργο μπορεί να περιλαμβάνει μεγάλο αριθμό ιστοσελίδων, σύνθετη αρχιτεκτονική βάσεων δεδομένων, προηγμένα συστήματα ηλεκτρονικού εμπορίου ή χρηματοοικονομικά συστήματα, και πολύπλοκες ροές εργασίας. Τα έργα ιστού υψηλής πολυπλοκότητας απαιτούν σημαντική επένδυση χρόνου, πόρων και εμπειρογνομosύνης

για το σχεδιασμό, τη σχεδίαση και την ανάπτυξη αλλά μπορούν να προσφέρουν μια ισχυρή και ιδιαίτερα προσαρμοσμένη λύση για επιχειρήσεις με συγκεκριμένες ανάγκες και απαιτήσεις. Συχνά περιλαμβάνουν μια μεγάλη ομάδα σχεδιαστών, προγραμματιστών και διαχειριστών έργων, με εξειδικευμένες δεξιότητες και γνώσεις προηγμένων τεχνολογιών και πλαισίων ανάπτυξης ιστού.

Change Requests:

Οι αλλαγές των απαιτήσεων στην ανάπτυξη ιστοσελίδων αφορούν αλλαγές στο αντικείμενο του έργου ή στις απαιτήσεις που ζητούνται αφού το έργο έχει ήδη ξεκινήσει ή βρίσκεται σε εξέλιξη. Οι αλλαγές αυτές μπορεί να περιλαμβάνουν τροποποιήσεις στο χρονοδιάγραμμα, τον προϋπολογισμό ή τη λειτουργικότητα του έργου και μπορεί να απαιτούν πρόσθετες εργασίες ανάπτυξης, δοκιμές ή άλλες προσαρμογές στο σχέδιο του έργου. Τα αιτήματα αλλαγών μπορεί να προκύψουν για διάφορους λόγους, όπως αλλαγές στις επιχειρηματικές απαιτήσεις, ανατροφοδότηση από τους χρήστες ή εξωτερικούς παράγοντες, όπως αλλαγές στην τεχνολογία ή στις συνθήκες της αγοράς. Ο αντίκτυπος των αιτημάτων αλλαγής σε ένα έργο εξαρτάται από το μέγεθος και την πολυπλοκότητα του έργου, τη φύση των αιτούμενων αλλαγών και το χρονοδιάγραμμα των αιτημάτων. Αν και τα αιτήματα αλλαγών μπορούν να εμφανιστούν σε οποιοδήποτε επίπεδο πολυπλοκότητας, είναι συχνά πιο συνηθισμένα σε έργα υψηλότερης πολυπλοκότητας, όπου μπορεί να εμπλέκονται πιο σύνθετες ροές εργασίας, ενσωματώσεις και λειτουργίες. Οι αλλαγές σε ένα έργο ιστού υψηλής πολυπλοκότητας μπορεί να έχουν σημαντικότερο αντίκτυπο στο χρονοδιάγραμμα και τον προϋπολογισμό του έργου από ό,τι οι αλλαγές σε ένα έργο χαμηλότερης πολυπλοκότητας.

Οι αλλαγές των απαιτήσεων των έργων αποτυπώνονται σε ένα εύρος 4 κλάσεων για να δηλώσουν το βαθμό και το μέγεθος των αλλαγών αυτών: 0, 1-4, 5-9, 10+.

HTML Templates:

Τα HTML templates αναφέρονται σε προσχεδιασμένα αρχεία HTML που μπορούν να χρησιμοποιηθούν ως σημείο εκκίνησης για την κατασκευή ιστοσελίδων. Αυτά τα πρότυπα περιλαμβάνουν συνήθως προκατασκευασμένο κώδικα HTML, CSS και μερικές φορές JavaScript, ο οποίος μπορεί να προσαρμοστεί ώστε να ανταποκρίνεται στις ανάγκες ενός συγκεκριμένου έργου. Τα πρότυπα HTML μπορούν να κυμαίνονται από απλά, μονοσέλιδα σχέδια έως σύνθετες, πολυσέλιδες διατάξεις με προηγμένη λειτουργικότητα και στοιχεία σχεδιασμού. Η χρήση προτύπων HTML δεν σχετίζεται

απαραίτητα με το μέγεθος ή την πολυπλοκότητα του έργου. Τα πρότυπα μπορούν να χρησιμοποιηθούν σε έργα οποιουδήποτε μεγέθους ή πολυπλοκότητας, από μικρούς ιστότοπους μίας σελίδας έως εφαρμογές μεγάλης κλίμακας σε επίπεδο επιχείρησης. Ωστόσο, η πολυπλοκότητα και το μέγεθος του έργου μπορεί να επηρεάσει τον τύπο και τον αριθμό των προτύπων που χρησιμοποιούνται, καθώς και το επίπεδο προσαρμογής και τροποποίησης που απαιτείται. Σε μικρότερα ή λιγότερο σύνθετα έργα, ένα μόνο προκατασκευασμένο πρότυπο μπορεί να είναι αρκετό για την κατασκευή ολόκληρου του ιστότοπου ή της εφαρμογής. Σε μεγαλύτερα ή πιο σύνθετα έργα, μπορούν να χρησιμοποιηθούν πολλαπλά πρότυπα για τη δημιουργία ενός πιο ολοκληρωμένου συστήματος σχεδιασμού. Σε κάθε περίπτωση ο τύπος και η έκταση της χρήσης των προτύπων θα εξαρτηθεί από τις συγκεκριμένες ανάγκες του έργου και τις προτιμήσεις της ομάδας ανάπτυξης.

Company_Size	Project_Type_Category	Website_Technology_Platform	Duration_months	Seniority_Devel	Spent_Time	Baseline	Cost	Profit	Profit_Margin	Profit_Ratio	Hourly_Cost	Estimated_Effort_Hours	Revenue	Complexity_Status	Change_Requests	HTML_Templates
1 - 9	website	Drupal	4	Mid	48	900 €	864 €	1,636 €	65.44%	1.894	18 €	50	2,500 €	Low	1 - 4	7
1 - 9	website	WordPress	1	Mid	60	936 €	1,080 €	1,520 €	58.46%	1.407	18 €	52	2,600 €	Low	0	11
1 - 9	website	Drupal	3	Mid	170	4,140 €	3,366 €	8,134 €	70.73%	2.417	20 €	230	11,500 €	Medium	5 - 9	15
50 - 249	website	Drupal	3	Senior	54	1,224 €	1,166 €	2,234 €	65.69%	1.915	22 €	68	3,400 €	Low	0	10
50 - 249	eshop	Decoupled Application	8	Senior	727	7,092 €	14,395 €	5,305 €	26.93%	0.369	20 €	394	19,700 €	High	10 +	21
10 - 49	website	Drupal	2	Mid	20	1,224 €	360 €	3,040 €	89.41%	8.444	18 €	68	3,400 €	Low	0	9
10 - 49	website	WordPress	4	Mid	120	1,897 €	2,160 €	3,110 €	59.01%	1.440	18 €	105.4	5,270 €	Medium	1 - 4	11
10 - 49	website	Drupal	2	Mid	110	2,538 €	1,980 €	5,070 €	71.91%	2.561	18 €	141	7,050 €	Low	1 - 4	12
10 - 49	Landing page	HTML	0	Mid	9.6	144 €	173 €	227 €	56.80%	1.315	18 €	8	400 €	Low	0	1
50 - 249	website	Drupal	16	Senior	300	4,954 €	6,480 €	7,280 €	52.91%	1.123	22 €	275.2	13,760 €	High	10 +	32

***Πίν. 2.** Δείγμα από τα Δεδομένα.*

3.3 Περιγραφική Στατιστική των Δεδομένων

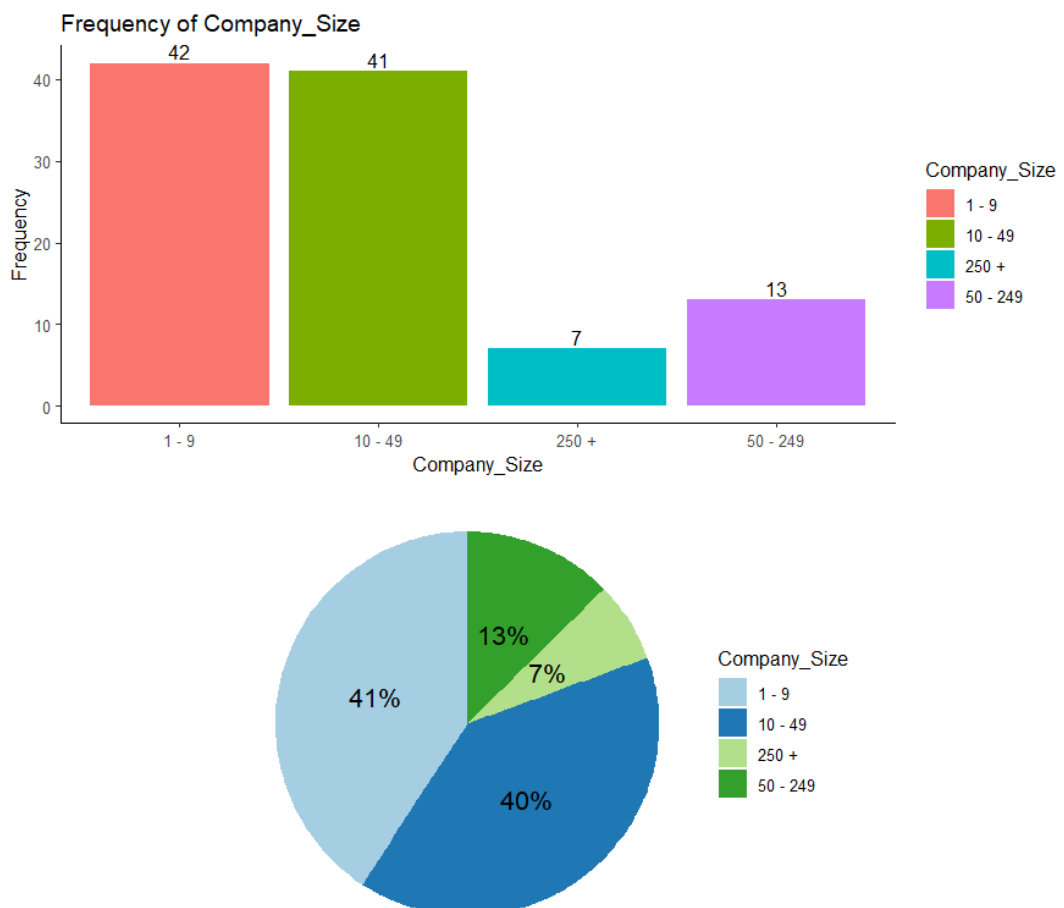
3.3.1 Περιγραφή Συχνοτήτων

Πριν την εφαρμογή αλγορίθμων για τη μελέτη των ερευνητικών ερωτημάτων της έρευνας, πραγματοποιήθηκε περιγραφική στατιστική των δεδομένων, από την οποία σημειώθηκαν οι παρακάτω παρατηρήσεις.

Ποιοτικές Μεταβλητές

Για τις ποιοτικές μεταβλητές κατασκευάστηκαν ραβδογράμματα με τις συχνότητες και διαγράμματα πίτας με τις σχετικές συχνότητες σε ποσοστά.

Company Size

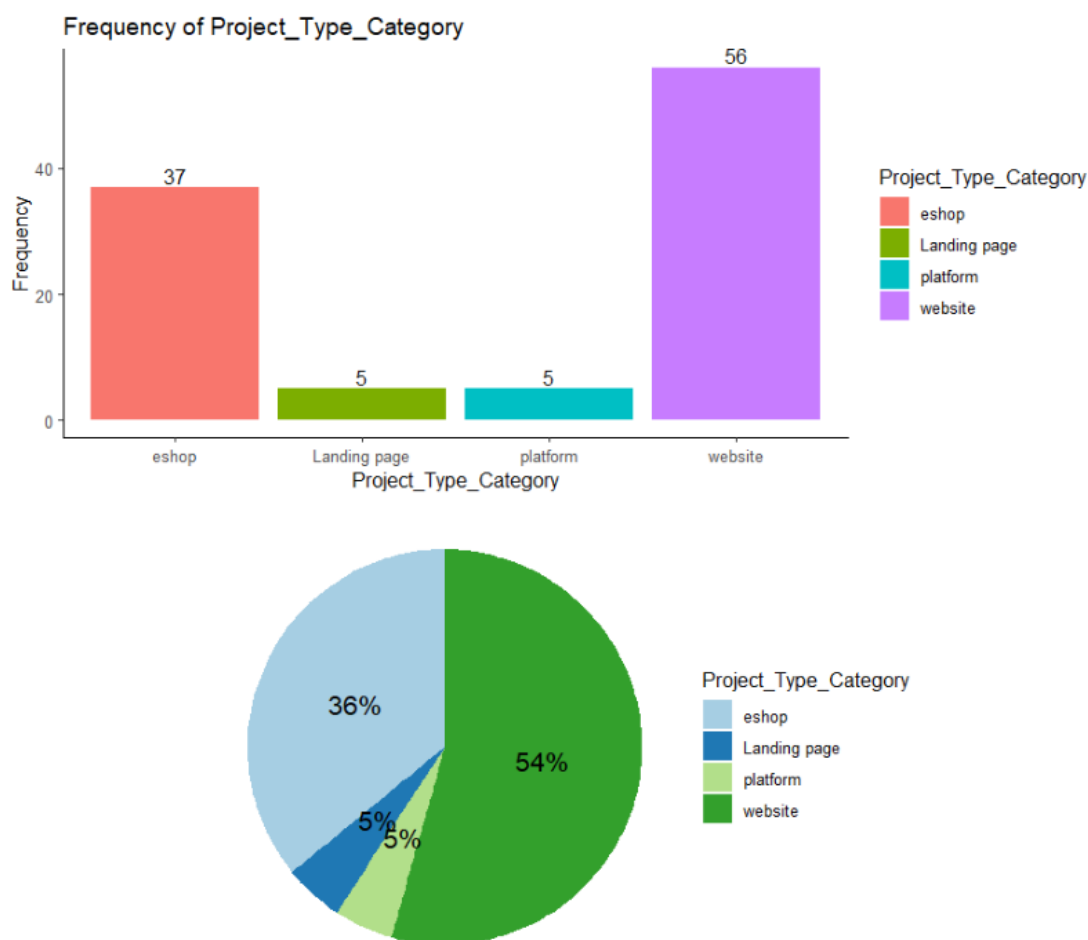


Διάγρ. 1. Συχνότητες Μεταβλητής "Company Size"

Περίπου το 80% των έργων έρχονται από εταιρίες μικρού μεγέθους δηλαδή, 1-9 και 10 – 49 employees, ενώ μόνο 13 projects αφορούν σε εταιρίες μεγαλύτερου μεγέθους 50 -

249 εργαζομένων και μόνο 7 projects αντιστοιχούν σε εταιρίες 250 + εργαζομένων. Προφανώς η εταιρία συνεργάζεται κατά κύριο λόγο με μικρομεσαίες επιχειρήσεις. Αν υποθέσουμε ότι «μεγάλα» και σύνθετα έργα είναι πιθανότερο να προέρχονται και από μεγάλου μεγέθους πελάτες, τότε τα περισσότερα έργα αναμένεται να είναι «μικρά», γεγονός που θα διερευνηθεί και στη συνέχεια.

Project Type Category

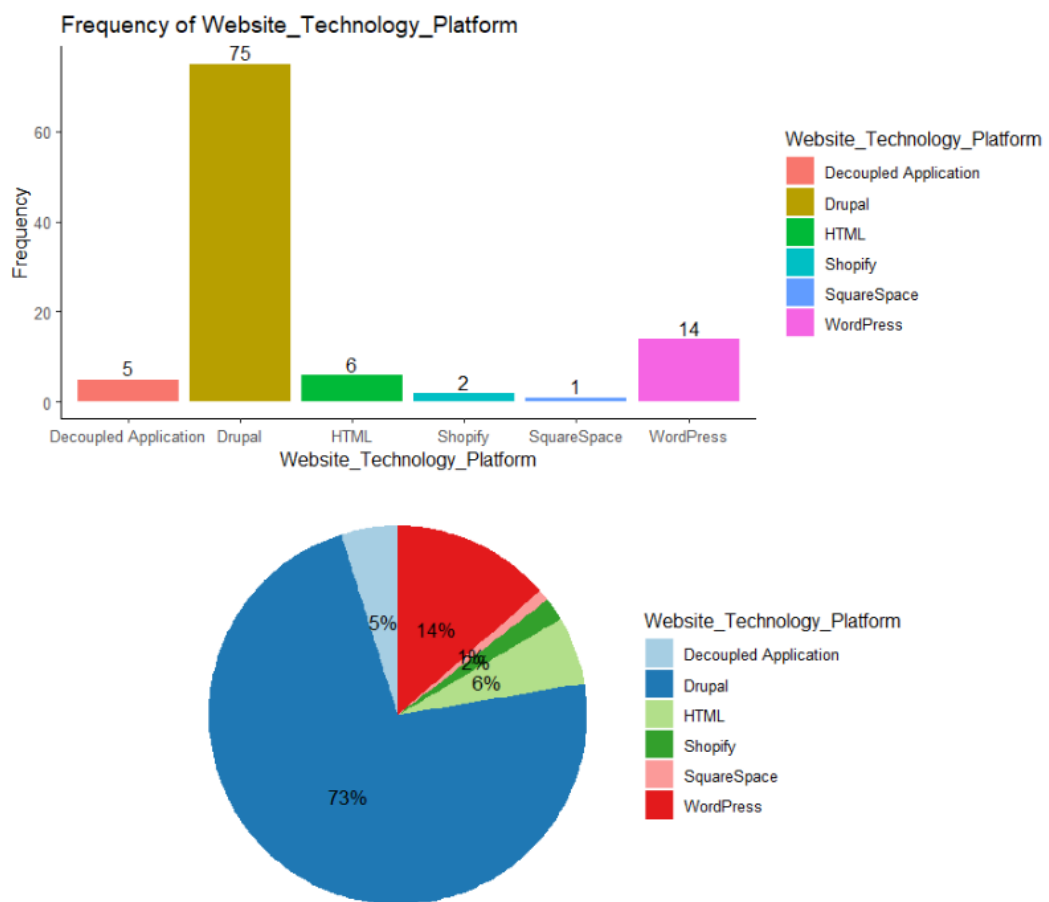


Διάγρ. 2. Συχνότητες Μεταβλητής "Project Type Category"

Το 54% των projects ανήκουν στην κατηγορία Website και το 36% είναι eshop. Στην κατηγορία Landing page και Platform ανήκει το 10% των projects. Όπως αναφέρθηκε κατά την περιγραφή των μεταβλητών, ο τύπος έργου "Website" αναφέρεται σε απλούς ισότοπους παρουσίασης, χωρίς κάποια σύνθετη λειτουργικότητα και επομένως είναι έργα

που απαιτούν λίγους πόρους και χωρίς κάποια εξειδίκευση ή εμπειρία. Προφανώς το γεγονός ότι τα απλά «websites» αντιπροσωπεύουν το 54% των έργων, δικαιολογείται και από το στατιστικό στοιχείο που είδαμε προηγουμένως, ότι δηλαδή το 80% των πελατών είναι μικρομεσαίες επιχειρήσεις που μάλλον δεν έχουν υψηλές απαιτήσεις ή ανάγκες για σύνθετη λειτουργικότητα από τα παραδοτέα έργα.

Website Technology Platform

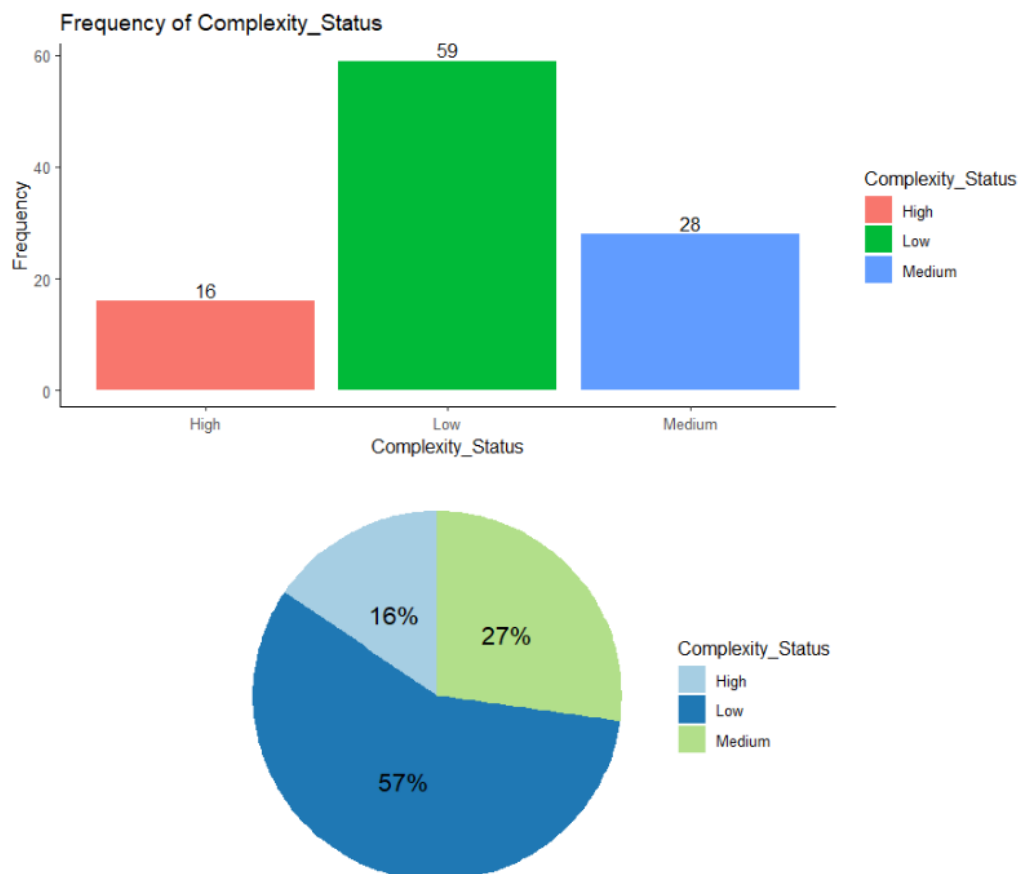


Διάγρ. 3. Συχνότητες Μεταβλητής "Website Technology Platform"

Το 73% των Projects αφορούν στην πλατφόρμα Drupal (75 στο σύνολο) και αμέσως μετά ακολουθεί η πλατφόρμα WordPress με 14 projects συνολικά. Το μεγάλο αυτό ποσοστό υποδεικνύει προφανώς ότι η εταιρία χρησιμοποιεί κατά κύριο λόγο την πλατφόρμα Drupal για την υλοποίηση των web projects. Η απόφαση να χρησιμοποιείται το Drupal για έργα ιστού μπορεί να δικαιολογείται από τις συγκεκριμένες ανάγκες και απαιτήσεις της εταιρείας, καθώς και από τις δεξιότητες και τους πόρους της ομάδας ανάπτυξης.

Ωστόσο, η ευελιξία, η επεκτασιμότητα και οι ισχυρές δυνατότητες διαχείρισης περιεχομένου του Drupal μπορεί να το καταστήσουν ούτως ή άλλως μια ελκυστική επιλογή για πολλά έργα ανάπτυξης ιστού.

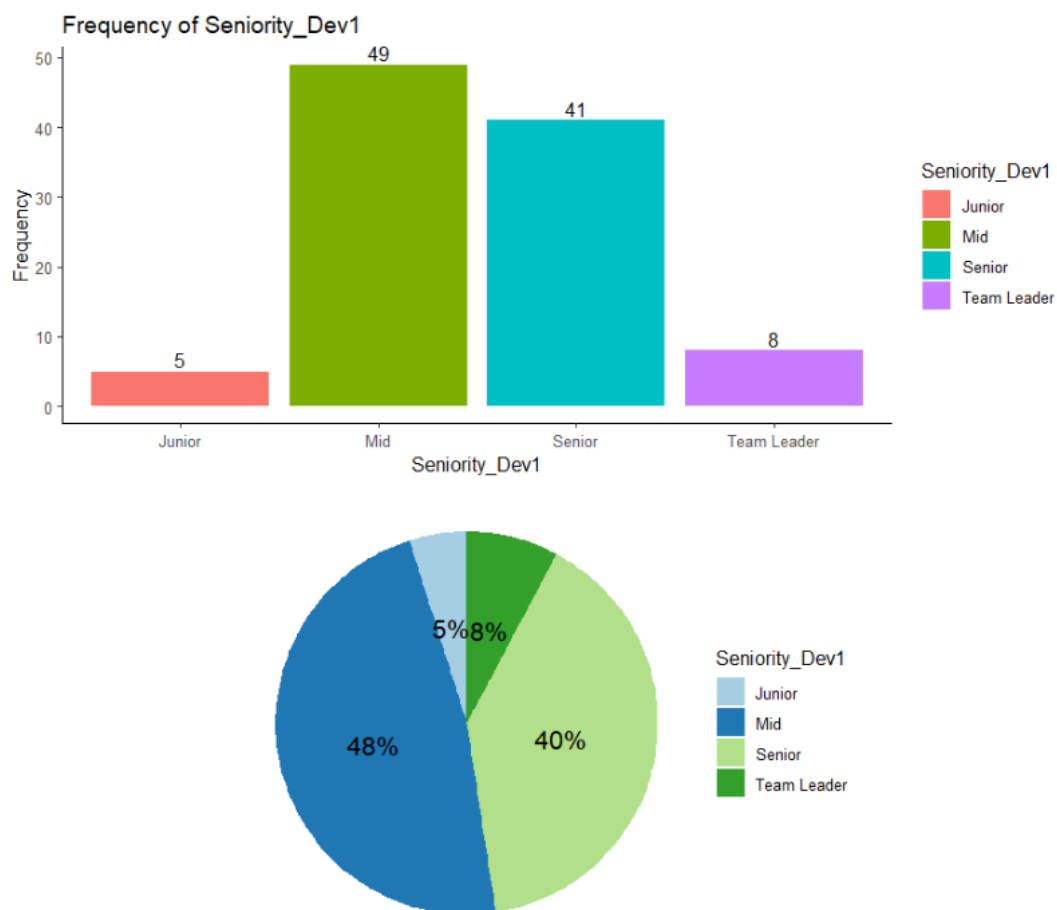
Complexity Status



Διάγρ. 4. Συχνότητες Μεταβλητής "Complexity Status"

Σχετικά με την πολυπλοκότητα των έργων παρατηρείται ότι το 57% των έργων χαρακτηρίζονται ως Low, δηλαδή χαμηλής πολυπλοκότητας και το 27% είναι μεσαίας πολυπλοκότητας (Medium). Το υπόλοιπο 16% έχει χαρακτηριστεί ως υψηλής πολυπλοκότητας (High Complexity) έργα. Το γεγονός ότι η εταιρία έχει κυρίως πελάτες μικρού ή μεσαίου μεγέθους, και αναλαμβάνει έργα που αφορούν σε σημαντικό ποσοστό απλά websites, όπως παρατηρήθηκε προηγουμένως, με τη σειρά του δικαιολογεί και ότι η πλειονότητα των έργων είναι «μικρά» έργα χαμηλής πολυπλοκότητας.

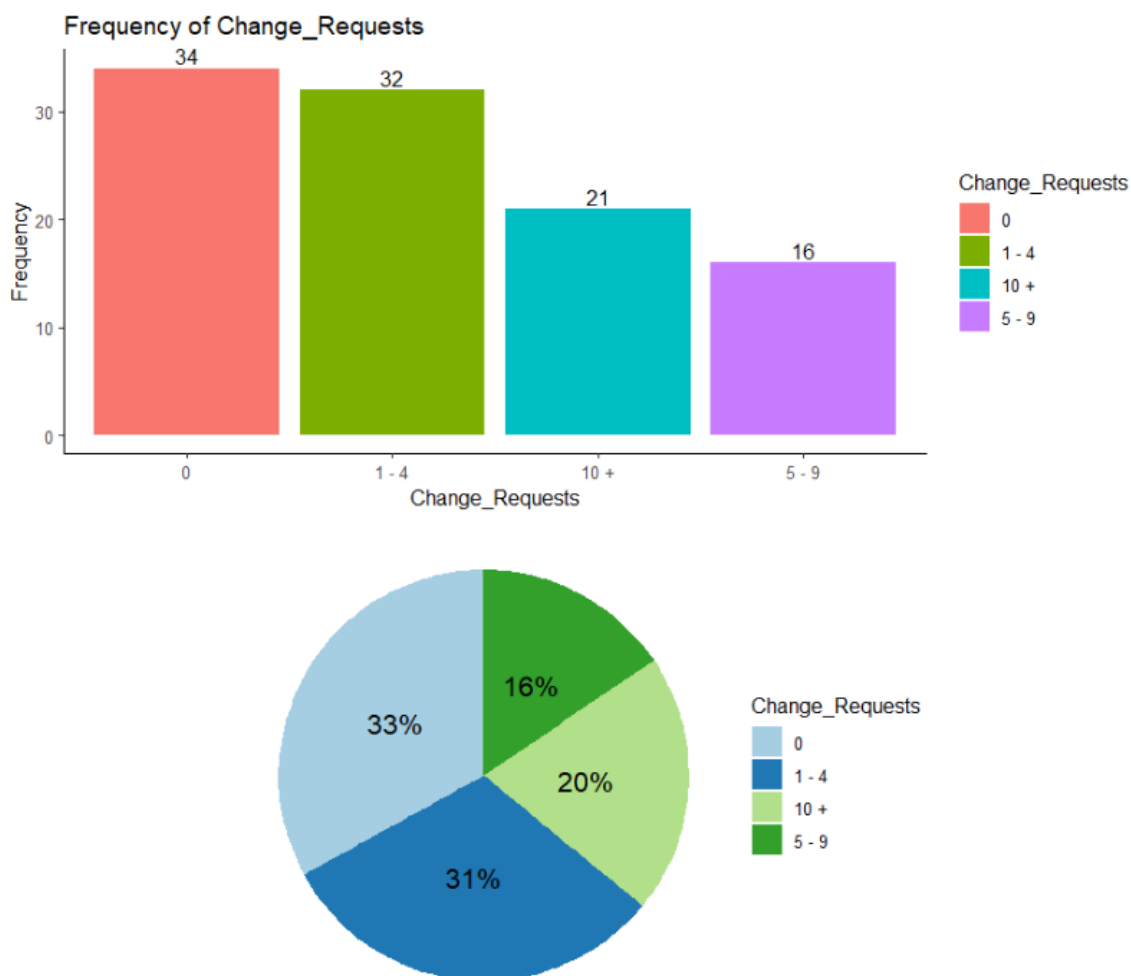
Seniority Dev1



Διάγρ. 5. Συχνότητες Μεταβλητής "Seniority Dev1"

Παρατηρείται ακόμη, ότι τα περισσότερα projects αναλαμβάνονται από Mid και Senior Developers (49 και 41 αντίστοιχα), ενώ μόνο 8 και 5 Projects από Team Leader και Junior αντίστοιχα. Αυτό το στατιστικό στοιχείο υποδεικνύει τη σύνθεση της ομάδας των προγραμματιστών και δικαιολογείται από το γεγονός ότι οι περισσότεροι προγραμματιστές της εταιρίας είναι “Mid” και “Senior” προγραμματιστές, δηλαδή μεσαίας ή και ανώτερης προϋπηρεσίας και εμπειρίας και άρα υπερτερούν σε ποσότητα από τους Junior. Σημειώνεται ότι πάρα το γεγονός ότι τα περισσότερα έργα είναι χαμηλής πολυπλοκότητας, αυτά υλοποιήθηκαν στην πλειονότητα από “Mid” και “Senior” προγραμματιστές.

Change Requests



Διάγρ. 6. Συχνότητες Μεταβλητής "Change Requests"

Τέλος, όπως φαίνεται και στο παραπάνω σχήμα, τα έργα έχουν μια μεγαλύτερη ομοιομορφία ως προς την κατανομή των αλλαγών των απαιτήσεων (change request). Για τα περισσότερα έργα (34) έχουν καταγραφεί 0 αλλαγές. Για 32 projects έχουν καταγραφεί 1 – 4 αλλαγές, και για 21 projects 10+ αλλαγές. Στο εύρος 5 – 9 change requests αντιστοιχούν τα λιγότερα projects (16). Αν υποθέσουμε ότι τα «μεγάλα» και σύνθετα έργα συνοδεύονται μάλλον από περισσότερες αλλαγές απαιτήσεων, αυτό δικαιολογεί και το γεγονός ότι το 64% των έργων σημείωσαν λίγες αλλαγές (0 έως 4), αφού τα περισσότερα έργα είναι χαμηλής πολυπλοκότητας και μικρού μεγέθους.

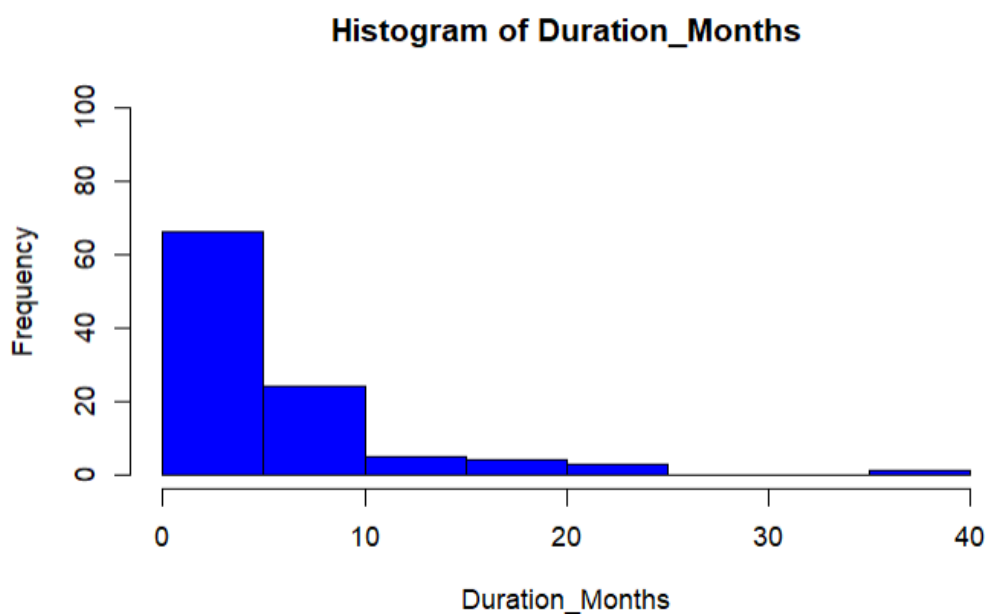
Ποσοτικές Μεταβλητές

Για τις ποσοτικές μεταβλητές υπολογίστηκαν τα περιγραφικά στατιστικά στοιχεία για κάθε μεταβλητή. Αυτά περιλαμβάνουν τη μέση τιμή, την τυπική απόκλιση, τις ελάχιστες και μέγιστες τιμές, όπως φαίνεται στον παρακάτω πίνακα. Επίσης κατασκευάστηκαν τα παρακάτω ιστογράμματα ώστε να διερευνηθεί το σχήμα της κατανομής, αν δηλαδή τα δεδομένα είναι συμμετρικά, λοξά ή με αιχμές, η κεντρική τάση των δεδομένων, η συγκέντρωση των τιμών και ακραίες τιμές.

Numeric.Variables	Mean	Min	Max	sd
<i>Duration_months</i>	5.62	0.00	38.00	6.03
<i>Spent_Time</i>	162.31	9.60	727.00	145.01
<i>Baseline</i>	2791.73	144.00	19800.00	2685.32
<i>Cost</i>	3280.61	172.80	16524.00	3154.19
<i>Profit</i>	4656.24	-5776.00	48070.00	6454.26
<i>Profit_Margin</i>	0.50	-2.89	0.89	0.45
<i>Profit_Ratio</i>	1.78	-0.74	8.44	1.61
<i>Hourly_Cost</i>	19.57	14.40	24.30	2.36
<i>Estimated_Effort_Hours</i>	155.10	8.00	1100.00	149.18
<i>Revenue</i>	7936.84	400.00	55000.00	7942.47
<i>HTML_Templates</i>	13.99	1.00	50.00	7.74

Πίν. 3. Συγκεντρωτικά Στατιστικά για τις Ποσοτικές Μεταβλητές

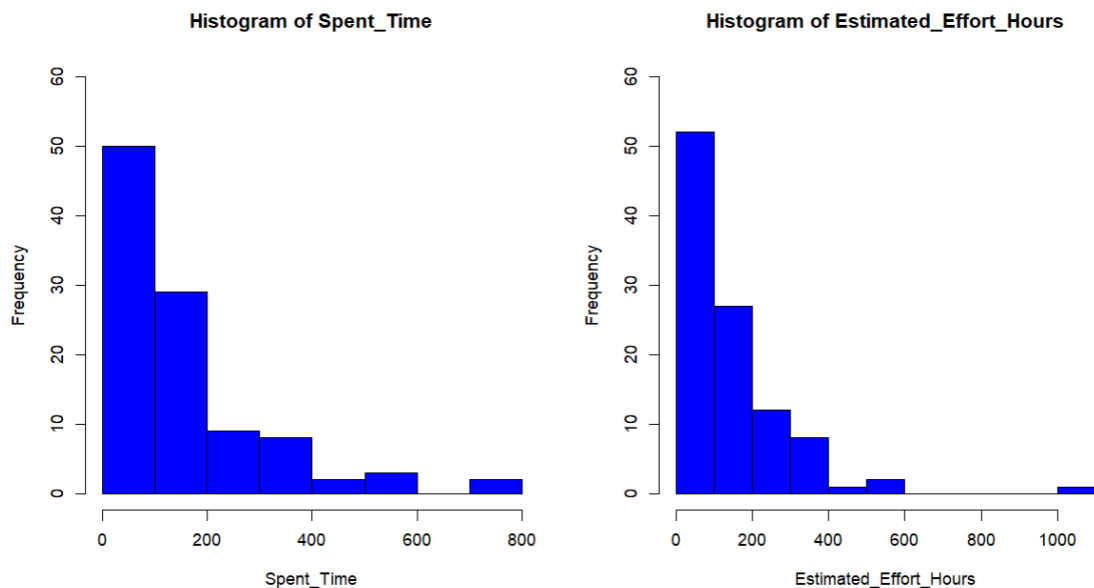
Duration Months



Διάγρ. 7. Ιστόγραμμα Μεταβλητής "Duration Months"

Τα περισσότερα Projects (66) έχουν διάρκεια έως 5 μήνες, ενώ 6 με 10 μήνες διήρκησαν συνολικά 24 Projects. Υπάρχουν ένας μικρός αριθμός έργων (13) που διήρκησαν πάνω από 10 μήνες.

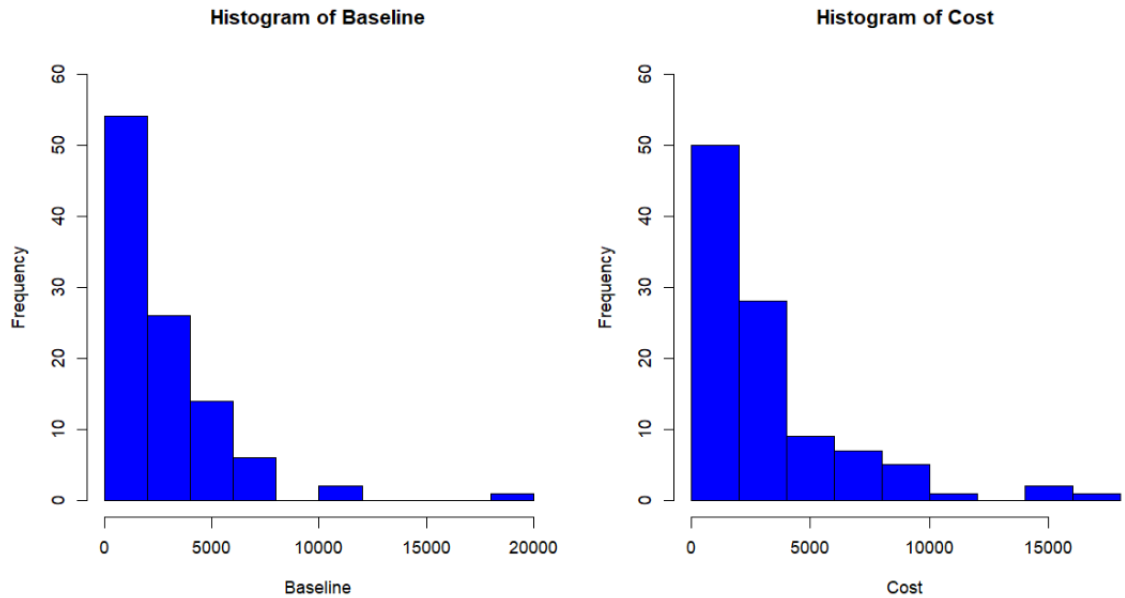
Spent Time & Estimated Effort Hours



Διάγρ. 8. Ιστόγραμμα Μεταβλητής "Spent Time" & "Estimated Effort Hours"

Όσον αφορά στις εργατοώρες που καταναλώθηκαν για τα projects, παρατηρείται ότι για την πλειονότητα των έργων έχουν καταγραφεί έως 200 ώρες, Τα υπόλοιπα Projects (24) ξεπερνούν τις 200 ώρες. Συγκρίνοντας με το αντίστοιχο ιστόγραμμα της εκτίμησης των ανθρωποωρών για τα έργα (Estimated Effort Hours), παρατηρείται ότι ακολουθείται παρόμοια κατανομή, με τα περισσότερα projects να εκτιμάται ότι θα καταναλώσουν έως 200 ανθρωποώρες.

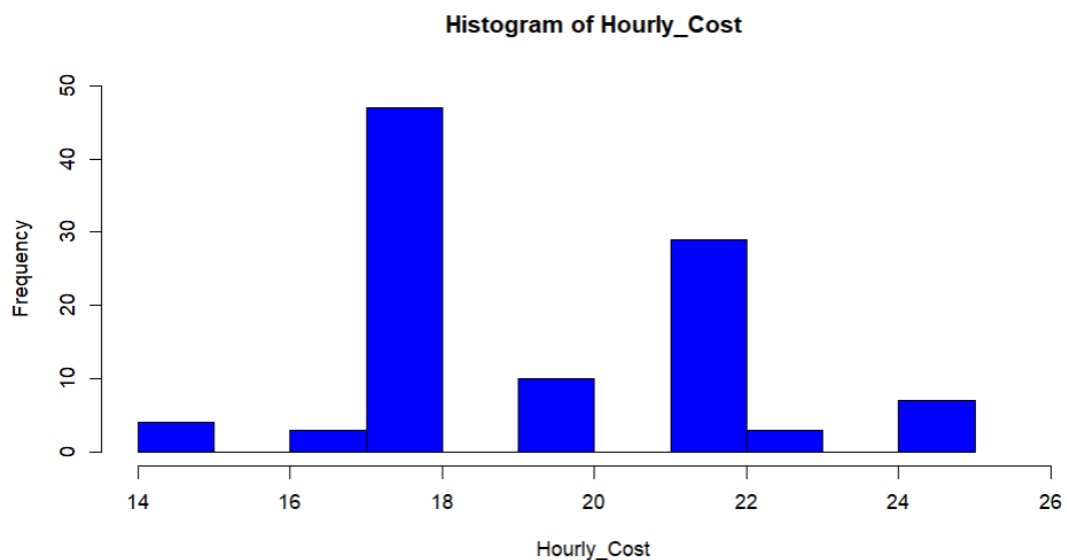
Baseline & Cost



Διάγρ. 9. Ιστόγραμμα Μεταβλητής "Baseline" & "Cost"

Με βάση την εκτίμηση των ανθρωποωρών για κάθε project, παρατηρείται ότι τα περισσότερα έργα εκτιμάται ότι θα έχουν κόστος έως 5000€ (Baseline) ενώ συγκρίνοντας με το πραγματικό κόστος (Cost) που βασίζεται στην πραγματική κατανάλωση χρόνου για το έργο, φαίνεται ότι και σε αυτό ακολουθεί την ίδια κατανομή η συγκέντρωση των τιμών. Ταυτόχρονα όμως υπάρχουν και λίγα projects με αρκετά μεγαλύτερη εκτίμηση κόστους αλλά και μεγαλύτερο πραγματικό κόστος.

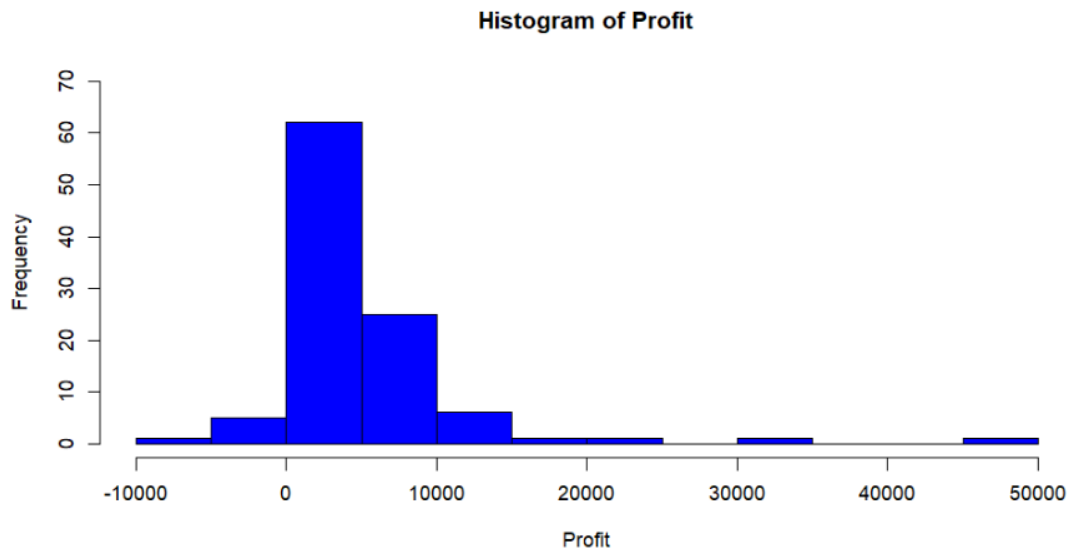
Hourly Cost



Διάγρ. 10. Ιστόγραμμα Μεταβλητής "Hourly Cost"

Από το παραπάνω ιστόγραμμα για την μεταβλητή του μέσου κόστους εργατοώρας (Hourly_Cost), παρατηρείται ότι τα περισσότερα έργα κοστολογούνται με 17-18€ την ανθρωποώρα, ενώ υπάρχουν και μερικά που κοστολογούνται με 21-22€ την ανθρωποώρα, γεγονός που αιτιολογείται από το επίπεδο (seniority) του προγραμματιστή που δουλεύει το έργο.

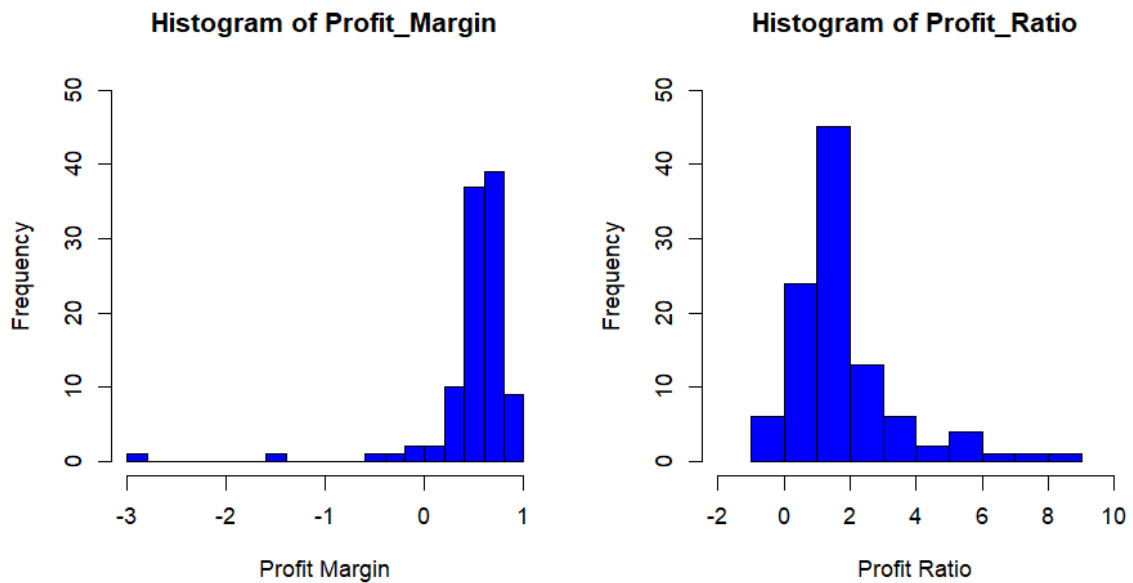
Profit



Διάγρ. 11. Ιστόγραμμα Μεταβλητής "Profit"

Από το παραπάνω ιστόγραμμα για τη μεταβλητή κέρδος (Profit), παρατηρείται ότι τα περισσότερα έργα φέρνουν έως 5000€ κέρδος, ενώ υπάρχουν μερικά για τα οποία το κέρδος φτάνει ως τις 10000€. Αξίζει επίσης να σημειωθεί ότι υπάρχουν και Projects (6) που έχουν φέρει ζημία για την εταιρεία. Τέλος, φαίνεται ακόμη ότι υπάρχουν ελάχιστα έργα που έχουν σημειώσει πολύ μεγάλη κερδοφορία με διαφορά.

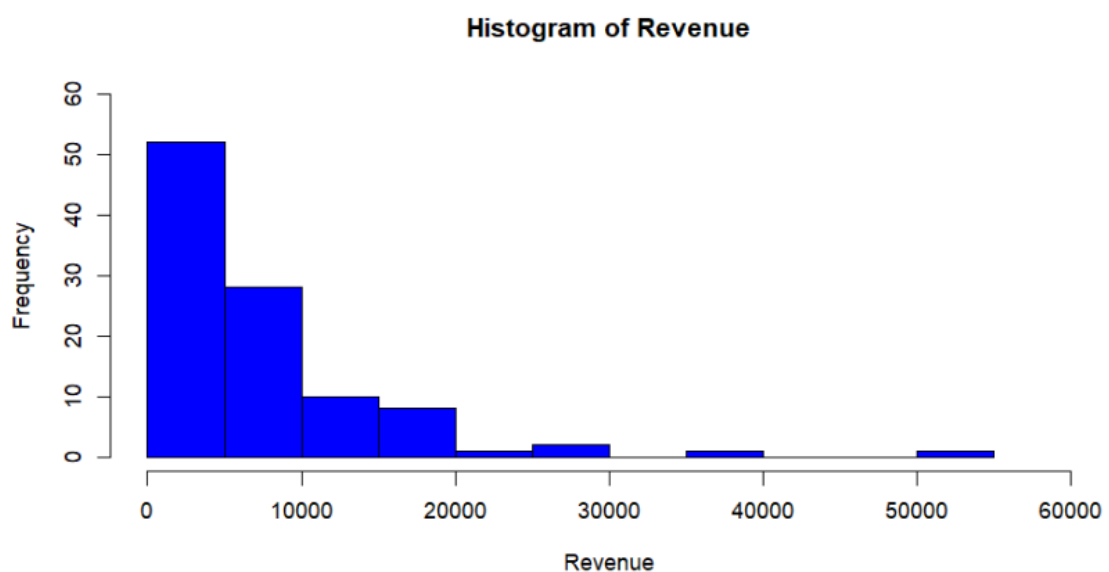
Profit Margin & Profit Ratio



Διάγρ. 12. Ιστόγραμμα Μεταβλητής "Profit Margin" & "Profit Ratio"

Αναφορικά με το περιθώριο κέρδους, φαίνεται ότι τα περισσότερα έργα έχουν περιθώριο κέρδους περίπου 50% με 60% και επομένως συγκριτικά και με τη μεταβλητή Profit Ratio, η εταιρεία τιμολογεί τόσο ώστε να έχει κέρδος μία με δύο φορές πάνω από το κόστος της, όπως ακριβώς φαίνεται και στο ιστόγραμμα.

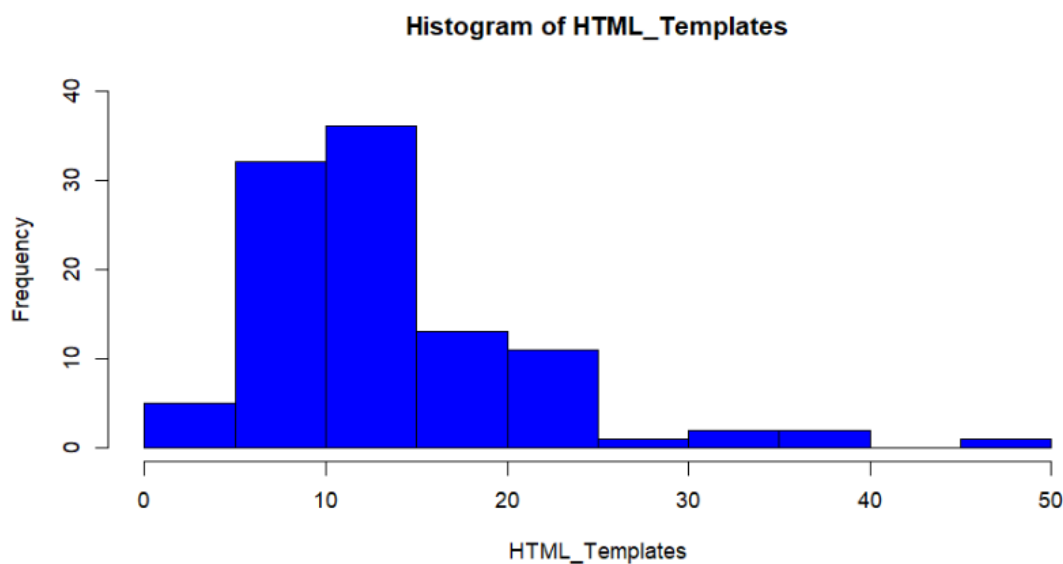
Revenue



Διάγρ. 13. Ιστόγραμμα Μεταβλητής "Revenue"

Αναφορικά με τα έσοδα των έργων, παρατηρείται ότι τα περισσότερα Projects φέρνουν έσοδα έως 10000€. Ενώ υπάρχουν και ακραίες τιμές με πολύ μεγαλύτερα έσοδα.

Html Templates



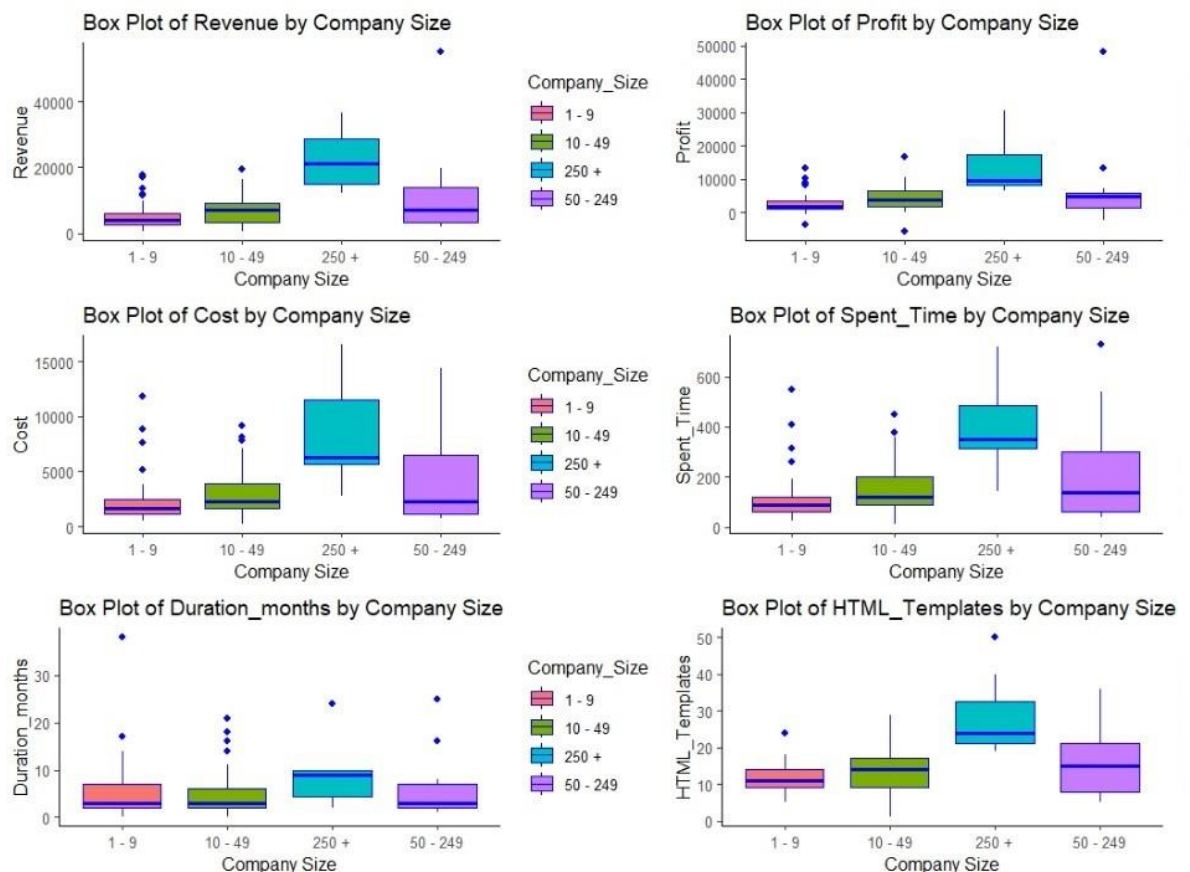
Διάγρ. 14. Ιστόγραμμα Μεταβλητής "HTML_Templates"

Τέλος, τα περισσότερα projects έχουν 5 με 15 html templates.

3.3.2 Περιγραφή Σχέσεων

Company Size

Σχετικά με την ποιοτική μεταβλητή "Company Size" παρατηρείται ότι τα έργα τα οποία προέρχονται από εταιρίες μεγάλου μεγέθους (250+) φέρουν τα περισσότερα έσοδα (Revenue) και κέρδη (Profit). Ταυτόχρονα όμως φαίνεται ότι καταναλώνουν το περισσότερο χρόνο (Spent Time) και ως εκ τούτου σημειώνουν και το μεγαλύτερο κόστος.

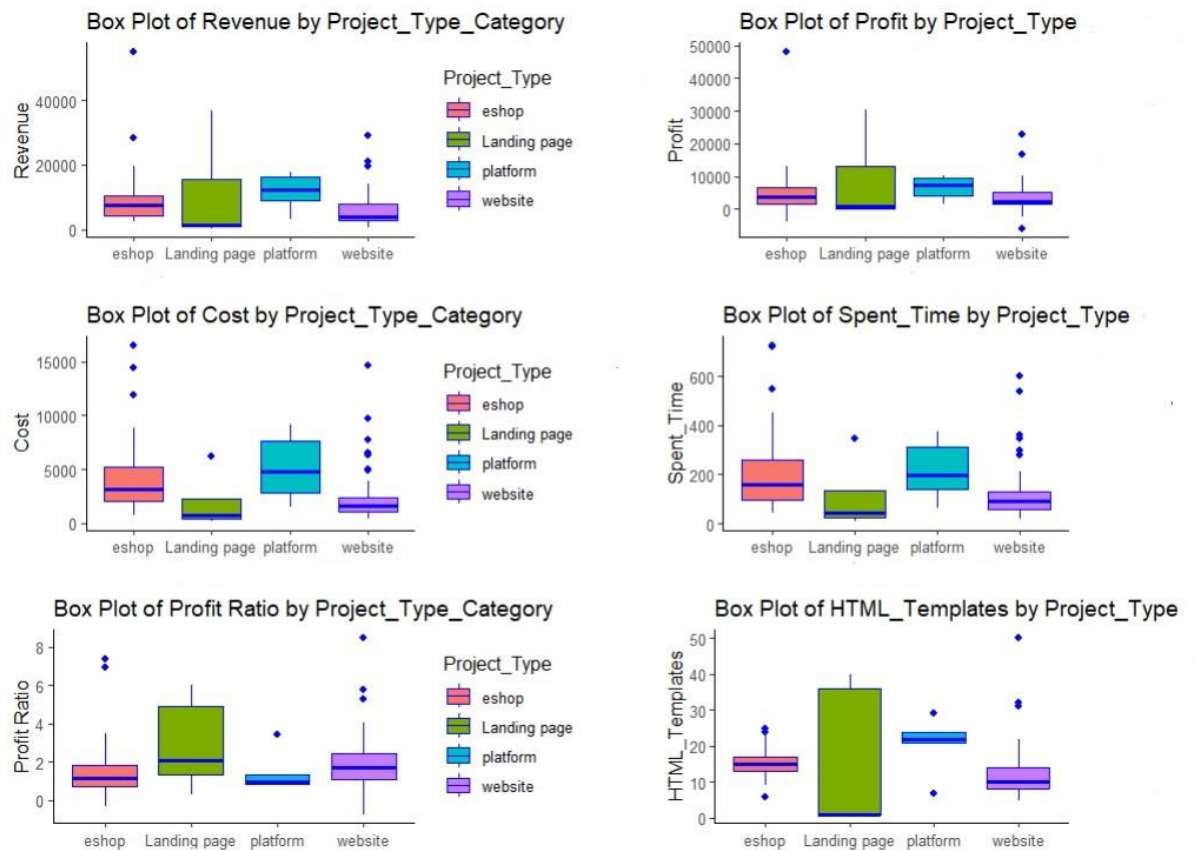


Διάγρ. 15. Boxplots Ποσοτικών Μεταβλητών ανά Company Size

Τα στατιστικά αυτά ερμηνεύονται προφανώς από το γεγονός ότι οι μεγαλύτεροι πελάτες είναι συχνά σε ευνοϊκότερη θέση να επενδύσουν περισσότερους πόρους σε μεγαλύτερα και πιο σύνθετα έργα, γεγονός που μπορεί να οδηγήσει σε υψηλότερα έσοδα, κέρδη και κόστος υλοποίησης. Είναι πιθανόν να έχουν πιο καθιερωμένες δραστηριότητες και εμπορική καταξίωση, πιο σύνθετες απαιτήσεις και ανάγκες και πιο καθιερωμένες διαδικασίες διαχείρισης έργων, τα οποία μπορούν να αυξήσουν τα δυνητικά έσοδα από ένα μεγαλύτερο έργο και να συμβάλουν στην πολυπλοκότητα και το εύρος ενός έργου.

Project Type Category

Αναφορικά με τον τύπο έργων (Project Type Category), παρατηρείται ότι τα έργα που εμπίπτουν στην κατηγορία “platform” ξεχωρίζουν σε ύψος εσόδων, αλλά και κόστους, σε αντίθεση με τα “websites”.



Αιάγρ. 16. Boxplots Ποσοτικών Μεταβλητών ανά Project Type

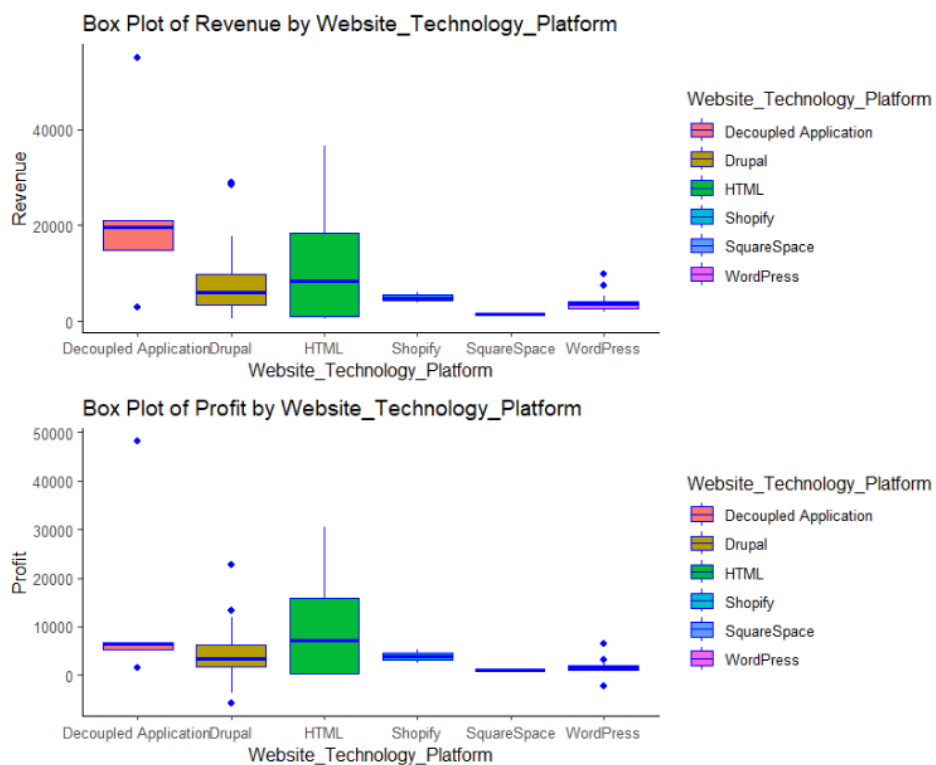
Ο τύπος έργου “platform” μπορεί να έχει τα μεγαλύτερα έσοδα και κόστος λόγω των πολλαπλών χαρακτηριστικών και δυνατοτήτων, και της ανάγκης για ενσωμάτωση με άλλα συστήματα. Τα έργα αυτά είναι μεγάλα και σύνθετα. Η ανάπτυξη και η υλοποίηση αυτών των χαρακτηριστικών απαιτεί σημαντικούς πόρους, όπως χρόνο, χρήματα και προσωπικό, γεγονός που μπορεί να συμβάλει στο υψηλότερο κόστος ενός έργου πλατφόρμας.

Η κατηγορία έργων “landing page” σημειώνει τη μεγαλύτερη αναλογία κέρδους σε σχέση με κόστος (Profit Ratio). Αυτό δικαιολογείται και από το γεγονός ότι αυτά τα έργα υλοποιούνται στο λιγότερο χρόνο (spent time), αφού είναι πολύ απλά έργα και άρα σημειώνουν το λιγότερο κόστος.

Website Technology Platform

Για την μεταβλητή “Website Technology Platform” δηλαδή την τεχνολογία πλατφόρμας των έργων, φαίνεται ότι η “Decoupled Application” ξεχωρίζει σε ύψος εσόδων, κόστους

και χρόνου κατανάλωσης, χωρίς όμως να ξεχωρίζει και σε κέρδος. Λόγω του γεγονότος ότι είναι μεγάλα και σύνθετα έργα, είναι λογικό να φέρνουν υψηλά έσοδα αλλά έχουν και υψηλά κόστη για την εταιρία. Κοστολογούνται με το υψηλότερο κόστος ανθρωποώρας, που δικαιολογείται γιατί απαιτούν μεγάλη εξειδίκευση και εμπειρία και επομένως τα αναλαμβάνουν προγραμματιστές ανωτέρου επιπέδου. Γι' αυτό το λόγο, όσον αφορά στην αναλογία κέρδους προς κόστος, η Decoupled Application φέρνει τη χαμηλότερη ενώ αντίθετα η πλατφόρμα “Shopify” φέρνει την υψηλότερη και αμέσως μετά η τεχνολογία “HTML” που υποδηλώνει και το χαμηλό κόστος αυτών των έργων.



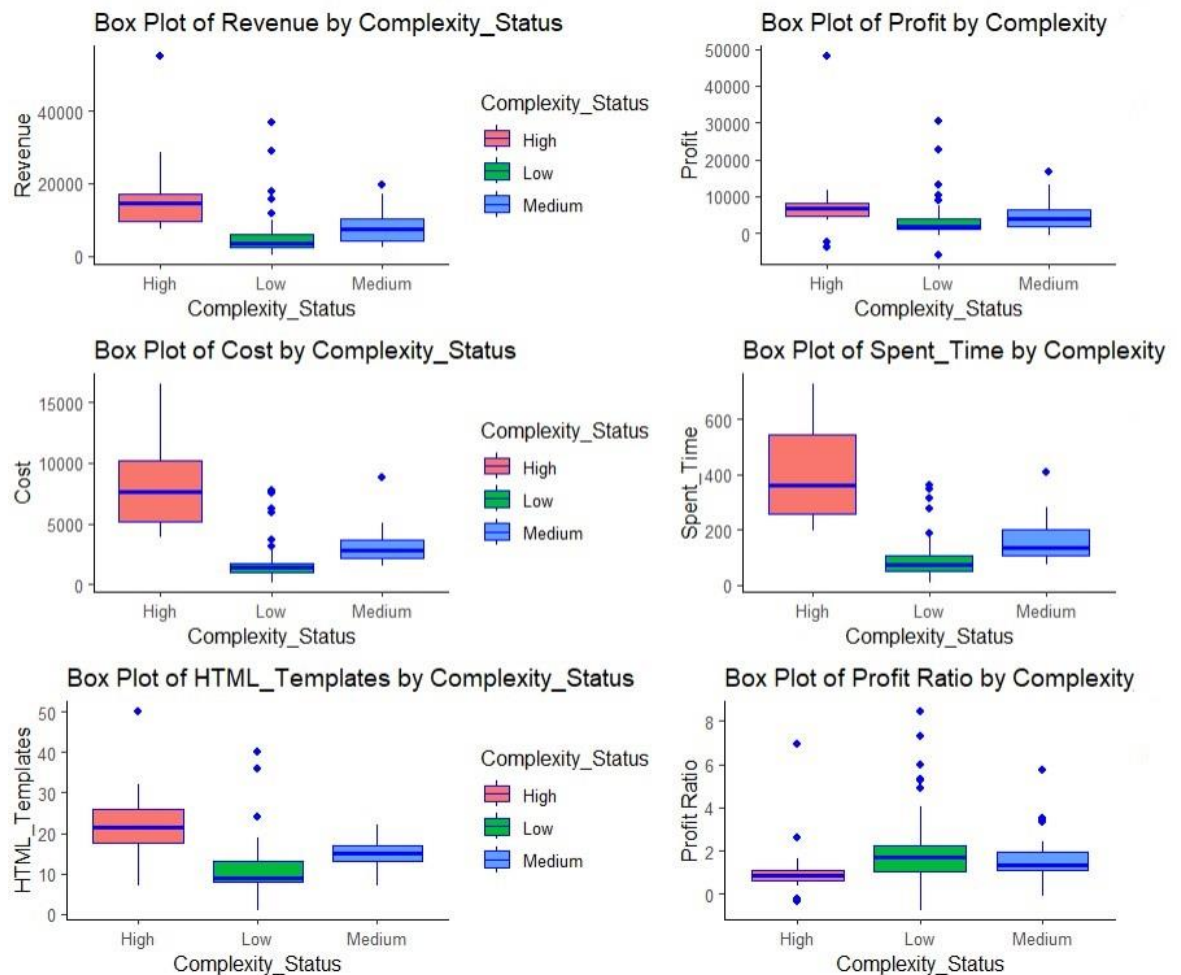


Αίμαρ. 17. Boxplots Ποσοτικών Μεταβλητών ανά Web Technology Platform

Complexity Status

Στο παρακάτω διάγραμμα (Διάγρ. 18) παρατηρούμε τη μεταβλητή της πολυπλοκότητας των έργων (complexity status) σε σχέση με άλλες μεταβλητές. Όπως είναι αναμενόμενο, και σε συνέχεια των όσων αναλυθήκαν παραπάνω, παρατηρείται ότι το υψηλό complexity status (high) παρατηρείται σε έργα με υψηλότερα έσοδα, μεγαλύτερη διάρκεια χρόνου, υψηλό κόστος, μεγάλη κατανάλωση χρόνου (spent time) και με μεγάλο αριθμό html templates.

Επίσης τα έργα με υψηλή πολυπλοκότητα φαίνεται ότι φέρουν μεγαλύτερα κέρδη (Profit) σε απόλυτο μέγεθος αλλά και μικρότερη αναλογία κέρδους προς κόστους (profit ratio), λόγω του μεγάλου κόστους τους.



Διάγρ. 18. Boxplots Ποσοτικών Μεταβλητών ανά Complexity Status

3.4 Μέθοδος Ανάλυσης των Δεδομένων

Όπως αναφέρθηκε αρχικά, η μέθοδος που θα εφαρμοστεί για την έρευνα και τη προσέγγιση του προβλήματος είναι η μέθοδος της συσταδοποίησης (clustering). Στην παρούσα ενότητα ακολουθεί η περιγραφή των μεθόδων - αλγορίθμων συσταδοποίησης που εφαρμόστηκαν και οι λόγοι που επιλέχθηκαν, πριν παρουσιαστούν τα αποτελέσματά τους.

3.4.1 Ανάλυση Συστάδων

Η ανάλυση συστάδων είναι ένας τύπος μάθησης χωρίς επίβλεψη στη μηχανική μάθηση και τη στατιστική. Στόχος της ανάλυσης συστάδων είναι η ομαδοποίηση παρόμοιων αντικειμένων σε συστάδες. Τα αντικείμενα μέσα σε μια συστάδα είναι πιο όμοια μεταξύ τους απ' ό,τι εκείνα σε άλλες συστάδες. Η ανάλυση συστάδων χρησιμοποιείται σε διάφορους τομείς, όπως η τμηματοποίηση της αγοράς, η τμηματοποίηση εικόνων και η ανάλυση γονιδιακής έκφρασης.

Οι αλγόριθμοι ανάλυσης συστάδων βασίζονται είτε σε μέτρα απόστασης μεταξύ σημείων δεδομένων είτε σε κατανομές πιθανοτήτων. Οι κύριοι τύποι μεθόδων ανάλυσης συστάδων είναι οι εξής:

- Συσταδοποίηση με βάση το κεντροειδές (centroid-based) (π.χ. k-means, PAM)
- Ιεραρχική συσταδοποίηση (Hierarchical clustering) (π.χ. Agglomerative, Divisive)
- Συσταδοποίηση με βάση την πυκνότητα (density-based) (π.χ. DBSCAN)
- Συσταδοποίηση με βάση την κατανομή (distribution-based) (π.χ. Gaussian Mixture Model)

Η ποιότητα των συστάδων αξιολογείται συνήθως με τη χρήση μετρικών όπως το silhouette score, ο δείκτης calinski-harabasz ή ο δείκτης Davies-Bouldin. Ο αριθμός των συστάδων που πρέπει να σχηματιστούν καθορίζεται συνήθως με τεχνικές όπως η μέθοδος του αγκώνα (elbow method) ή η στατιστική του χάσματος (gap statistic).

3.4.3 Επιλογή & Καταλληλότητα Αλγορίθμων Συσταδοποίησης

Για την ανάλυση συστάδων, στην παρούσα έρευνα, επιλέχθηκαν οι αλγόριθμοι k-means, το ιεραρχικό clustering (Hierarchical clustering) και ο αλγόριθμος Partitioning Around

Medoids (PAM). Ακολουθεί αναλυτική περιγραφή των συγκεκριμένων αλγορίθμων και η αιτιολόγηση της καταλληλότητάς τους στην συγκεκριμένη περίπτωση εργασίας.

K-means

Ο αλγόριθμος K-Means είναι ένας ευρέως χρησιμοποιούμενος αλγόριθμος συσταδοποίησης που χωρίζει ένα σύνολο n σημείων δεδομένων σε k συστάδες, όπου k είναι ένας αριθμός που καθορίζεται από τον χρήστη. Στόχος του αλγορίθμου είναι η ελαχιστοποίηση του αθροίσματος των τετραγωνικών αποστάσεων μεταξύ των σημείων δεδομένων και του κέντρου συστάδας που τους έχει ανατεθεί. Παρακάτω ακολουθεί η αναλυτική περιγραφή του αλγορίθμου, των πλεονεκτημάτων και μειονεκτημάτων του.

Ο αλγόριθμος K-Means ξεκινά με την τυχαία επιλογή k αρχικών κέντρων συστάδων, που ονομάζονται κεντροειδή. Στη συνέχεια, ο αλγόριθμος αναθέτει κάθε σημείο δεδομένων στο πλησιέστερο κεντροειδές με βάση την ευκλείδεια απόσταση μεταξύ του σημείου δεδομένων και του κεντροειδούς. Μετά την αρχική ανάθεση, ο αλγόριθμος υπολογίζει εκ νέου τα κέντρα των συστάδων ως τον μέσο όρο των σημείων δεδομένων που ανατίθενται σε κάθε συστάδα. Η διαδικασία αυτή επαναλαμβάνεται μέχρι τη σύγκλιση, η οποία επέρχεται όταν οι αναθέσεις συστάδων δεν αλλάζουν πλέον ή όταν επιτυγχάνεται ένας μέγιστος αριθμός επαναλήψεων.

Ο στόχος του αλγορίθμου K-Means είναι η ελαχιστοποίηση του αθροίσματος των τετραγωνικών αποστάσεων εντός της συστάδας (Within-Cluster Sum of Squared), το οποίο μετρά τη συνολική απόσταση μεταξύ των σημείων δεδομένων και του κέντρου της συστάδας που τους έχει ανατεθεί. Το WCSS είναι ένα μέτρο του πόσο καλά τα δεδομένα είναι καταναμημένα στις συστάδες και μια μικρότερη τιμή WCSS υποδηλώνει καλύτερη κατανομή.

Ο αλγόριθμος K-Means έχει πολλά πλεονεκτήματα, όπως η απλότητα, η επεκτασιμότητα και η γρήγορη σύγκλιση. Ο αλγόριθμος είναι εύκολος στην υλοποίηση και μπορεί να χειριστεί μεγάλα σύνολα δεδομένων σε εύλογο χρονικό διάστημα. Επιπλέον, ο αλγόριθμος K-Means είναι μια ευέλικτη μέθοδος που μπορεί να χρησιμοποιηθεί για ένα ευρύ φάσμα προβλημάτων ομαδοποίησης, όπως η κατάτμηση εικόνων, η ταξινόμηση εγγράφων και η κατάτμηση της αγοράς.

Ωστόσο, ο αλγόριθμος K-Means έχει επίσης ορισμένους περιορισμούς. Ένα σημαντικό μειονέκτημα είναι ότι είναι ευαίσθητος στην αρχική επιλογή του κεντροειδούς. Οι τελικές

συστάδες μπορεί να διαφέρουν σημαντικά ανάλογα με την αρχική επιλογή του κεντροειδούς, γεγονός που μπορεί να οδηγήσει σε διαφορετικά αποτελέσματα και ερμηνείες. Επιπλέον, ο αλγόριθμος K-Means υποθέτει ότι οι συστάδες είναι σφαιρικές και ίσου μεγέθους, κάτι που μπορεί να μην ισχύει πάντα στα δεδομένα του πραγματικού κόσμου. Τέλος, ο αλγόριθμος ενδέχεται να μην είναι κατάλληλος για δεδομένα με πολύπλοκες δομές, όπως δεδομένα με μη γραμμικές σχέσεις ή δεδομένα με πολλαπλές συστάδες διαφορετικού μεγέθους και σχήματος.

Συμπερασματικά, ο αλγόριθμος K-Means είναι μια δημοφιλής και αποτελεσματική μέθοδος ομαδοποίησης που χρησιμοποιείται ευρέως σε διάφορες εφαρμογές. Αν και ο αλγόριθμος έχει ορισμένους περιορισμούς, τα πλεονεκτήματά του τον καθιστούν χρήσιμο εργαλείο για την κατανόηση και τον χαρακτηρισμό της δομής μεγάλων συνόλων δεδομένων.

Hierarchical clustering

Η ιεραρχική ομαδοποίηση είναι ένας τύπος αλγορίθμου ομαδοποίησης που δημιουργεί μια ιεραρχική αναπαράσταση των δεδομένων, όπου κάθε σημείο δεδομένων διαιρείται σε όλο και μικρότερες ομάδες, έως ότου κάθε σημείο δεδομένων σχηματίσει τη δική του ομάδα. Παρακάτω ακολουθεί η αναλυτική περιγραφή του αλγορίθμου Ιεραρχικής Συσταδοποίησης και θα συζητηθούν τα πλεονεκτήματα και οι περιορισμοί του.

Υπάρχουν δύο κύριοι τύποι αλγορίθμων ιεραρχικής ομαδοποίησης: Συσσωρευτική (Agglomerative) και διαχωριστική (Divisive). Η συσσωρευτική ιεραρχική ομαδοποίηση ξεκινά αντιμετωπίζοντας κάθε σημείο δεδομένων ως ξεχωριστή συστάδα και στη συνέχεια συγχωνεύει επαναληπτικά τις πλησιέστερες συστάδες έως ότου όλα τα σημεία δεδομένων ομαδοποιηθούν σε μία μεγάλη συστάδα. Από την άλλη πλευρά, η διαχωριστική ιεραρχική ομαδοποίηση ξεκινά αντιμετωπίζοντας όλα τα σημεία δεδομένων ως μία μεγάλη συστάδα και στη συνέχεια χωρίζει επαναληπτικά τη συστάδα σε μικρότερες συστάδες έως ότου κάθε σημείο δεδομένων σχηματίσει τη δική του συστάδα.

Ένα από τα κύρια πλεονεκτήματα της Ιεραρχικής Συσταδοποίησης είναι ότι παρέχει μια οπτική αναπαράσταση της δομής των δεδομένων, με τη μορφή δενδρογράμματος. Το δενδρόγραμμα είναι ένα δενδροειδές διάγραμμα που δείχνει τις σχέσεις μεταξύ των σημείων δεδομένων και των συστάδων και μπορεί να χρησιμοποιηθεί για τον προσδιορισμό του βέλτιστου αριθμού συστάδων για τα δεδομένα.

Ένα άλλο πλεονέκτημα της Ιεραρχικής Συσταδοποίησης είναι ότι δεν είναι ευαίσθητη στις αρχικές συνθήκες, σε αντίθεση με την K-Means Clustering, η οποία μπορεί να παράγει διαφορετικά αποτελέσματα ανάλογα με την αρχική επιλογή των κεντροειδών. Επιπλέον, η Ιεραρχική Συσταδοποίηση μπορεί να χειριστεί μη σφαιρικές και άνισου μεγέθους συστάδες, κάτι που μπορεί να αποτελεί περιορισμό της Συσταδοποίησης K-Means.

Ωστόσο, η Ιεραρχική Συσταδοποίηση έχει επίσης ορισμένους περιορισμούς. Ένας σημαντικός περιορισμός είναι ότι μπορεί να είναι υπολογιστικά δαπανηρή για μεγάλα σύνολα δεδομένων, καθώς ο αλγόριθμος απαιτεί τον υπολογισμό αποστάσεων ανά ζεύγη μεταξύ όλων των σημείων δεδομένων. Επιπλέον, ο αλγόριθμος δεν είναι κατάλληλος για μεγάλα σύνολα δεδομένων με πολυδιαστατικότητα, καθώς το δένδρογραμμα γίνεται δύσκολο να ερμηνευτεί και οι υπολογισμοί των αποστάσεων ανά ζεύγη γίνονται υπολογιστικά δαπανηροί.

Συμπερασματικά, η ιεραρχική ομαδοποίηση είναι ένας ισχυρός και ευέλικτος αλγόριθμος ομαδοποίησης που παρέχει οπτική αναπαράσταση της δομής των δεδομένων. Ο αλγόριθμος είναι κατάλληλος για σύνολα δεδομένων μικρού έως μεσαίου μεγέθους και για δεδομένα με πολύπλοκες δομές. Ωστόσο, οι περιορισμοί του τον καθιστούν λιγότερο κατάλληλο για μεγάλα σύνολα δεδομένων με υψηλές διαστάσεις. Σε κάθε περίπτωση ένα πολύτιμο εργαλείο για την κατανόηση και τον χαρακτηρισμό της δομής πολύπλοκων δεδομένων.

Partitioning Around Medoids (PAM)

Ο PAM (Partitioning Around Medoids) είναι ένας αλγόριθμος συσταδοποίησης που χωρίζει ένα σύνολο δεδομένων σε k συστάδες ελαχιστοποιώντας το άθροισμα των ανομοιοτήτων μεταξύ των σημείων δεδομένων και των πλησιέστερων μεσοειδών τους. Ένα medoid είναι ένα αντιπροσωπευτικό σημείο δεδομένων που βρίσκεται πλησιέστερα σε όλα τα άλλα σημεία δεδομένων εντός της ίδιας συστάδας. Παρακάτω ακολουθεί η περιγραφή του αλγορίθμου PAM, των πλεονεκτημάτων και των περιορισμών του.

Ο αλγόριθμος PAM ξεκινά με την επιλογή k αρχικών medoids, τα οποία συνήθως επιλέγονται τυχαία από τα σημεία δεδομένων. Στη συνέχεια, ο αλγόριθμος αναθέτει κάθε σημείο δεδομένων στο πλησιέστερο medoid με βάση την επιλεγμένη μετρική ανομοιοτήτας. Μετά την αρχική ανάθεση, ο αλγόριθμος εκτελεί μια διαδικασία βελτιστοποίησης δύο βημάτων για την εύρεση των καλύτερων medoids για κάθε

συστάδα. Το πρώτο βήμα είναι η ανταλλαγή του τρέχοντος medoid με ένα μη medoid σημείο δεδομένων στην ίδια συστάδα και το δεύτερο βήμα είναι ο επανυπολογισμός του αθροίσματος των ανομοιοτήτων για να προσδιοριστεί εάν η ανταλλαγή βελτίωσε τη συνολική λύση. Αυτή η διαδικασία επαναλαμβάνεται μέχρι το άθροισμα των ανομοιοτήτων να μην βελτιώνεται πλέον ή να επιτευχθεί ένας μέγιστος αριθμός επαναλήψεων.

Ένα από τα κύρια πλεονεκτήματα του αλγορίθμου PAM είναι ότι είναι πιο ανθεκτικός στις ακραίες τιμές από τον αλγόριθμο K-Means, καθώς χρησιμοποιεί μεσοδιαστήματα αντί για μέσους όρους για την αναπαράσταση κάθε συστάδας. Επιπλέον, ο αλγόριθμος PAM μπορεί να χειριστεί δεδομένα με αυθαίρετα μέτρα ανομοιότητας, συμπεριλαμβανομένων δεδομένων με μη ευκλείδειες μετρικές απόστασης, καθιστώντας τον πιο ευέλικτο από τον αλγόριθμο K-Means. Επιπλέον, ο αλγόριθμος PAM είναι επίσης πιο κατάλληλος για δεδομένα με μη σφαιρικές και άνισου μεγέθους συστάδες, κάτι που μπορεί να αποτελεί περιορισμό του αλγορίθμου K-Means.

Ωστόσο, ο αλγόριθμος PAM έχει επίσης ορισμένους περιορισμούς. Ένα σημαντικό μειονέκτημα είναι ότι είναι υπολογιστικά δαπανηρός, καθώς απαιτεί τον υπολογισμό των ανομοιοτήτων μεταξύ όλων των σημείων δεδομένων σε κάθε επανάληψη. Επιπλέον, ο αλγόριθμος ενδέχεται να μην συγκλίνει στη βέλτιστη λύση για ορισμένα σύνολα δεδομένων, ιδίως για σύνολα δεδομένων με πολλά παρόμοια σημεία δεδομένων ή για σύνολα δεδομένων με πολύπλοκες δομές.

Συμπερασματικά, ο αλγόριθμος PAM είναι μια πολύτιμη μέθοδος ομαδοποίησης. Η ικανότητα του αλγορίθμου να χειρίζεται μη ευκλείδειες αποστάσεις και μη σφαιρικές και άνισου μεγέθους συστάδες τον καθιστά χρήσιμο εργαλείο για την κατανόηση και τον χαρακτηρισμό σύνθετων δεδομένων. Ωστόσο, το υπολογιστικό κόστος και οι περιορισμοί του αλγορίθμου τον καθιστούν λιγότερο κατάλληλο για μεγάλα σύνολα δεδομένων ή για σύνολα δεδομένων με πολλά παρόμοια σημεία δεδομένων.

3.4.4 Καταλληλότητα Αλγορίθμων Συσταδοποίησης

Η αρχική επιλογή για την συσταδοποίηση των δεδομένων ήταν οι αλγόριθμοι με βάση το κεντροειδές (centroid-based), δηλαδή k-means, που είναι δημοφιλείς και αποτελεσματικές μέθοδοι για την ομαδοποίηση δεδομένων. Όπως αναφέρθηκε και

παραπάνω, η βασική ιδέα πίσω από αυτούς τους αλγορίθμους είναι η ελαχιστοποίηση του αθροίσματος των αποστάσεων μεταξύ των σημείων δεδομένων και του πλησιέστερου κεντροειδούς τους, προκειμένου να κατανεμηθούν τα δεδομένα σε συστάδες. Ο αλγόριθμος k-means είναι ευέλικτος αλγόριθμος ομαδοποίησης, κατάλληλος για πολλά σύνολα δεδομένων του πραγματικού κόσμου με παρόμοια σημεία δεδομένων και παράγει ακριβή αποτελέσματα. Επομένως η πρώτη συσταδοποίηση έγινε με αυτόν τον αλγόριθμο.

Ο αλγόριθμος ιεραρχικής ομαδοποίησης (hierarchical clustering) θεωρήθηκε ότι είναι μια καλή επιλογή για ομαδοποίηση για την συγκεκριμένη περίπτωση για τους παρακάτω λόγους:

1. Διερευνητική ανάλυση δεδομένων: Η ιεραρχική ομαδοποίηση είναι χρήσιμη για τη διερεύνηση της δομής ενός συνόλου δεδομένων, καθώς μπορεί να παρέχει μια οπτική αναπαράσταση των σχέσεων μεταξύ των παρατηρήσεων με τη μορφή δενδρογράμματος.
2. Μη παραμετρική προσέγγιση: Η ιεραρχική ομαδοποίηση είναι μια μη παραμετρική προσέγγιση ομαδοποίησης, που σημαίνει ότι δεν κάνει υποθέσεις σχετικά με την κατανομή των δεδομένων. Αυτό την καθιστά κατάλληλη για σύνολα δεδομένων με σύνθετες ή ανομοιόμορφες κατανομές, όπως είναι το συγκεκριμένο σύνολο δεδομένων.
3. Διατήρηση της δομής των συστάδων: Η ιεραρχική συσταδοποίηση είναι μια καλή επιλογή για την διατήρηση της υποκείμενης δομής των δεδομένων, καθώς δεν απαιτεί να καθοριστεί εκ των προτέρων ο αριθμός των συστάδων (όπως ο k-means) και επομένως, η αξιολόγηση ενός τέτοιου αλγορίθμου συμπληρώνει και βελτιώνει την εγκυρότητα της παρούσας έρευνας.

Ο αλγόριθμος Partitioning Around Medoids (PAM) επιλέχθηκε διότι είναι ο καλύτερος για την ομαδοποίηση όταν τα δεδομένα έχουν ανομοιόμορφη κατανομή, όπως το συγκεκριμένο σύνολο δεδομένων, και οι ομάδες έχουν διαφορετικό μέγεθος και σχήμα. Δεδομένου ότι ο PAM χρησιμοποιεί μεσοδιαστήματα αντί για μέσους όρους ως κέντρο των συστάδων, τον καθιστά πιο ανθεκτικό στις ακραίες τιμές και σε θορυβώδη δεδομένα και μπορεί να χειριστεί μη σφαιρικές συστάδες. Επειδή λοιπόν, το συγκεκριμένο σύνολο δεδομένων, έχει ένα μείγμα κατηγορικών και ποσοτικών μεταβλητών και ταυτόχρονα και ακραίες τιμές, θεωρήθηκε πολύ καλή εναλλακτική επιλογή συσταδοποίησης. Ταυτόχρονα δημιουργεί μια πολύ καλή βάση σύγκρισης με τον αλγόριθμο k-means.

Οι αλγόριθμοι K-Means και PAM (Partitioning Around Medoids) είναι δύο δημοφιλείς αλγόριθμοι ομαδοποίησης που χρησιμοποιούνται για την κατάτμηση δεδομένων σε ομάδες. Ωστόσο, υπάρχουν ορισμένες βασικές διαφορές μεταξύ των δύο αλγορίθμων, οι οποίες είναι ακριβώς και οι λόγοι επιλογής των δύο αυτών αλγορίθμων.

1. Κεντροειδή έναντι μεσοειδών: Ο K-Means χρησιμοποιεί κεντροειδή για την αναπαράσταση της συστάδας, τα οποία υπολογίζονται ως ο μέσος όρος των σημείων της συστάδας. Από την άλλη πλευρά, ο PAM χρησιμοποιεί medoids, τα οποία είναι πραγματικά σημεία δεδομένων που αντιπροσωπεύουν τη συστάδα.

2. Μετρική απόστασης: Η K-Means χρησιμοποιεί τη μετρική της ευκλείδειας απόστασης για να μετρήσει την ομοιότητα μεταξύ σημείων και κεντροειδών, ενώ η PAM χρησιμοποιεί το άθροισμα των αποστάσεων μεταξύ κάθε σημείου στη συστάδα και του medoid.

3. Ευαισθησία σε ακραίες τιμές: Η μέθοδος K-Means μπορεί να είναι ευαίσθητη στις ακραίες τιμές, καθώς ο μέσος όρος των δεδομένων επηρεάζεται εύκολα από την παρουσία ακραίων τιμών. Η PAM, από την άλλη πλευρά, είναι λιγότερο ευαίσθητη στις ακραίες τιμές, καθώς βασίζεται στο άθροισμα των αποστάσεων και όχι στον μέσο όρο.

Αντίθετα με τους παραπάνω λόγους επιλογής αλγορίθμων, διερευνήθηκαν και οι λόγοι απόρριψης άλλων μοντέλων συσταδοποίησης όπως το μοντέλο DBSCAN (density-based) και Gaussian Mixture (distribution-based).

Ο μέθοδος DBSCAN δεν επιλέχθηκε για τους παρακάτω λόγους:

1. Η DBSCAN δεν είναι κατάλληλη για αραιά δεδομένα, καθώς απαιτεί έναν ελάχιστο αριθμό παρατηρήσεων εντός μιας ορισμένης ακτίνας για τον σχηματισμό μιας συστάδας. Σε περιπτώσεις όπου τα δεδομένα είναι αραιά, άλλοι αλγόριθμοι όπως ο K-Means μπορεί να είναι καλύτερη επιλογή.

2. Συστάδες με διαφορετικές πυκνότητες: Ο DBSCAN υποθέτει ότι όλες οι συστάδες έχουν παρόμοιες πυκνότητες και ενδέχεται να μην αποδίδει καλά σε περιπτώσεις όπου οι συστάδες έχουν διαφορετικές πυκνότητες.

3. Καθορισμός των σωστών παραμέτρων: Η απόδοση του DBSCAN εξαρτάται από τον προσδιορισμό δύο σημαντικών παραμέτρων: της ακτίνας (eps) και του ελάχιστου αριθμού παρατηρήσεων (minPts) που απαιτούνται για το σχηματισμό μιας συστάδας. Ο

σωστός προσδιορισμός αυτών των παραμέτρων μπορεί να αποτελέσει πρόκληση και να επηρεάσει την απόδοση του αλγορίθμου.

Συνοψίζοντας, ο αλγόριθμος DBSCAN δεν είναι κατάλληλος για αραιά δεδομένα, συστάδες με ποικίλες πυκνότητες και περιπτώσεις όπου είναι δύσκολο να καθοριστούν οι σωστές παράμετροι.

Το μοντέλο Gaussian Mixture δεν επιλέχθηκε για τους παρακάτω λόγους:

1. Το GMM υποθέτει ότι τα δεδομένα είναι γκαουσιανής κατανομής και ενδέχεται να μην αποδίδει καλά σε σύνολα δεδομένων με μη γκαουσιανή κατανομή, όπως το συγκεκριμένο σύνολο δεδομένων.
2. Ο προσδιορισμός του βέλτιστου αριθμού συστάδων για το GMM μπορεί να αποτελέσει πρόκληση, καθώς απαιτεί τον εκ των προτέρων προσδιορισμό του αριθμού των συνιστωσών και τη χρήση τεχνικών επιλογής μοντέλου για τον προσδιορισμό του βέλτιστου αριθμού συστάδων.

Επομένως, δεν είναι κατάλληλη μέθοδος για σύνολα δεδομένων με μη γκαουσιανές κατανομές, δεδομένα υψηλής διάστασης και περιπτώσεις όπου ο αριθμός των συστάδων είναι δύσκολο να προσδιοριστεί.

3.5 Προετοιμασία και Προ-επεξεργασία των Δεδομένων

Μετά την περιγραφική στατιστική των δεδομένων, της οποίας μερικά διαγράμματα παρατέθηκαν σε προηγούμενη ενότητα, το επόμενο βήμα είναι η προ-επεξεργασία δεδομένων, δηλαδή η προετοιμασία τους για στατιστική ανάλυση και αφορά τυχόν βήματα καθαρισμού (cleaning), κανονικοποίησης (normalization) και κλιμάκωσης (scaling). Όπως και για την περιγραφική στατιστική, η επεξεργασία και στη συνέχεια όλη η ανάλυση πραγματοποιήθηκε στο περιβάλλον προγραμματισμού της R, εξειδικευμένο για υπολογισμούς στατιστικής φύσεως και οπτικοποίησης.

Η προεπεξεργασία των δεδομένων είναι ένα κρίσιμο βήμα στην διαδικασία ανάλυσης δεδομένων, καθώς συμβάλλει στη διασφάλιση ότι τα δεδομένα είναι κατάλληλα για ανάλυση και ερμηνεία. Παρακάτω ακολουθεί μια επισκόπηση των κύριων βημάτων που ενεπλάκησαν στην προεπεξεργασία δεδομένων και η σημασία τους.

Το πρώτο βήμα στην προεπεξεργασία δεδομένων είναι ο καθαρισμός των δεδομένων, ο οποίος περιλαμβάνει τον εντοπισμό και την αφαίρεση τυχόν ελλειπών, εσφαλμένων ή ασυνεπών δεδομένων. Αυτό είναι σημαντικό, διότι τα δεδομένα αυτά μπορεί να προκαλέσουν στρεβλώσεις και παραμορφώσεις στα αποτελέσματα της ανάλυσης και να οδηγήσουν σε εσφαλμένα συμπεράσματα. Αφού εισήχθησαν και προβλήθηκαν τα ακατέργαστα δεδομένα, έγινε έλεγχος για ελλείπουσες τιμές. Παράλληλα και σε συνδυασμό με την περιγραφή των δεδομένων και των κατανομών των τιμών των μεταβλητών, προσδιορίστηκαν οι στήλες – μεταβλητές προς αφαίρεση, ώστε να μην επηρεάσουν την έρευνα μεταβλητές που δεν έχουν σημαντικότητα ή οι τιμές έχουν μεγάλη ομοιογένεια. Μετά την αφαίρεση αυτών των στηλών-μεταβλητών, έγινε μετονομασία των υπόλοιπων στηλών για καλύτερη σαφήνεια. Επομένως, οι τελικές μεταβλητές που χρησιμοποιήθηκαν στην ανάλυση είναι οι ακόλουθες 17:

Variable Name
<i>Company_Size</i>
<i>Project_Type_Category</i>
<i>Website_Technology_Platform</i>
<i>Duration_months</i>
<i>Seniority_Dev1</i>
<i>Spent_Time</i>
<i>Baseline</i>
<i>Cost</i>
<i>Profit</i>
<i>Profit_Margin</i>
<i>Profit_Ratio</i>
<i>Hourly_Cost</i>
<i>Estimated_Effort_Hours</i>
<i>Revenue</i>
<i>Complexity_Status</i>
<i>Change_Requests</i>
<i>HTML_Templates</i>

Πίν. 4. Οι Επιλεγθείσες Μεταβλητές

Το επόμενο βήμα για την προετοιμασία των δεδομένων περιλαμβάνει τη μετατροπή των κατηγορικών μεταβλητών σε αριθμητικές αναπαραστάσεις. Αυτό είναι σημαντικό για τη μετέπειτα μοντελοποίηση, δεδομένου ότι πολλοί αλγόριθμοι μηχανικής μάθησης απαιτούν αριθμητικές εισόδους. Το τελικό βήμα είναι η κλιμάκωση των

χαρακτηριστικών, η οποία περιλαμβάνει το μετασχηματισμό των δεδομένων έτσι ώστε κάθε χαρακτηριστικό να έχει την ίδια κλίμακα (scaling). Φέροντας όλες τις μεταβλητές σε κοινή κλίμακα αποτρέπει την κυριαρχία κάποιας μεταβλητής και σε βάρος άλλων.

Συνοπτικά, για την προετοιμασία των δεδομένων εκτελέστηκαν βασικές εργασίες προεπεξεργασίας δεδομένων, όπως ο καθαρισμός δεδομένων, ο μετασχηματισμός δεδομένων και η κανονικοποίηση δεδομένων, για να προετοιμάσει το σύνολο δεδομένων για περαιτέρω ανάλυση και μοντελοποίηση. Αυτά τα βήματα είναι κρίσιμα για να διασφαλιστεί ότι τα δεδομένα είναι σε μορφή κατάλληλη για τους αλγορίθμους μηχανικής μάθησης, για να αποφευχθούν πιθανές μεροληψίες στα αποτελέσματα και να αυξήσουμε τις πιθανότητες λήψης ακριβών και ουσιαστικών αποτελεσμάτων.

3.5.1 Κλιμάκωση των Δεδομένων πριν τη Συσταδοποίηση

Στο σημείο αυτό αξίζει να σημειωθεί ότι η κλιμάκωση των δεδομένων (scaling), ώστε να διασφαλιστεί ότι τα χαρακτηριστικά με μεγαλύτερη κλίμακα δεν κυριαρχούν στα αποτελέσματα, πραγματοποιήθηκε πριν τη συσταδοποίηση σε όλους τους αλγορίθμους. Ωστόσο στα πλαίσια της διερεύνησης, η συσταδοποίηση εφαρμόστηκε και χωρίς κλιμάκωση των δεδομένων. Σε αυτή τη περίπτωση τα αποτελέσματα ήταν καλύτερα, δηλαδή χωρίς την κλιμάκωση (scaling).

Η κλιμάκωση δεδομένων είναι ένα κοινό βήμα προεπεξεργασίας δεδομένων στη μηχανική μάθηση και την ανάλυση δεδομένων. Περιλαμβάνει τον μετασχηματισμό των χαρακτηριστικών ενός συνόλου δεδομένων σε μια κοινή κλίμακα, συνήθως για να βελτιωθεί η απόδοση των αλγορίθμων μηχανικής μάθησης ή για να γίνουν τα δεδομένα πιο προσιτά στη στατιστική ανάλυση. Ωστόσο, ενώ η κλιμάκωση δεδομένων μπορεί να είναι χρήσιμη σε ορισμένες περιπτώσεις, γνωρίζουμε όμως ότι μπορεί επίσης να έχει ανεπιθύμητες συνέπειες, συμπεριλαμβανομένης της διαστρέβλωσης των σχέσεων μεταξύ των σημείων δεδομένων. Στρέβλωση δεδομένων είναι η απόκλιση των δεδομένων από την πραγματική ή ακριβέστερη αναπαράσταση της πλήρους εικόνας.

Το γεγονός ότι το πλάτος της σιλουέτας (μέτρο επικύρωσης της ανάλυσης) προέκυψε καλύτερο πριν από την κλιμάκωση από ό,τι μετά την κλιμάκωση, αυτό υποδεικνύει ότι η διαδικασία κλιμάκωσης ενδεχομένως παραμόρφωσε τις σχέσεις μεταξύ των σημείων δεδομένων. Σε αυτή την περίπτωση, θεωρήθηκε ότι είναι καλύτερο να χρησιμοποιηθούν τα αποτελέσματα χωρίς κλιμάκωση, καθώς η κλιμάκωση δε βελτίωσε τη λύση ομαδοποίησης, αντίθετα μάλλον πρόκυψε στρέβλωση.

Ένας λόγος για τον οποίο η κλιμάκωση δεδομένων μπορεί να στρεβλώσει τις σχέσεις μεταξύ σημείων δεδομένων είναι ότι μπορεί να εντείνει την επίδραση των ακραίων τιμών στα δεδομένα. Οι ακραίες τιμές είναι σημεία δεδομένων που βρίσκονται μακριά από την πλειονότητα των δεδομένων και μπορούν να έχουν δυσανάλογο αντίκτυπο στα αποτελέσματα των στατιστικών αναλύσεων και των αλγορίθμων μηχανικής μάθησης. Όταν τα δεδομένα κλιμακώνονται, το εύρος των δεδομένων συχνά συμπιέζεται, γεγονός που μπορεί να ενισχύσει τον αντίκτυπο των ακραίων σημείων και να τα καταστήσει ακόμη πιο επιδραστικά. Ως αποτέλεσμα, οι σχέσεις μεταξύ των σημείων δεδομένων μπορεί να παραμορφωθούν.

Ένας δεύτερος λόγος για τον οποίο η κλιμάκωση δεδομένων μπορεί να στρεβλώσει τις σχέσεις μεταξύ σημείων δεδομένων είναι ότι μπορεί να εισαγάγει μη γραμμικές σχέσεις μεταξύ μεταβλητών. Ορισμένες μέθοδοι κλιμάκωσης, όπως η λογαριθμική κλιμάκωση (logarithmic scaling) μπορούν να εισάγουν μη γραμμικές σχέσεις μεταξύ των μεταβλητών, γεγονός που μπορεί να οδηγήσει σε στρεβλά αποτελέσματα στις στατιστικές αναλύσεις και στους αλγορίθμους μηχανικής μάθησης. Έχουν βρεθεί μελέτες που επιβεβαιώνουν τον ισχυρισμό, όπως για παράδειγμα, μια μελέτη των Bodoor και Shujaat διαπίστωσε ότι η χρήση λογαριθμικής κλιμάκωσης στην προεπεξεργασία δεδομένων οδήγησε σε μη γραμμικές σχέσεις μεταξύ μεταβλητών, οι οποίες παραμόρφωσαν τα αποτελέσματα μιας ανάλυσης παλινδρόμησης (Bodoor and Shujaat, 2020).

Ένας τρίτος λόγος για τον οποίο η κλιμάκωση των δεδομένων μπορεί να στρεβλώσει τις σχέσεις μεταξύ των σημείων δεδομένων είναι ότι μπορεί να αλλάξει την κατανομή των δεδομένων. Οι μέθοδοι κλιμάκωσης που αλλάζουν το εύρος των δεδομένων μπορούν επίσης να αλλάξουν την κατανομή των δεδομένων, γεγονός που μπορεί να οδηγήσει σε στρεβλές σχέσεις μεταξύ σημείων δεδομένων. Για παράδειγμα, μια μελέτη των Geladi και Kowalski διαπίστωσε ότι οι μέθοδοι κλιμάκωσης δεδομένων που αλλάζουν το εύρος των δεδομένων, όπως η κλιμάκωση min-max (min-max scaling), που μετατρέπει τα χαρακτηριστικά ενός συνόλου δεδομένων σε ένα σταθερό εύρος, συνήθως μεταξύ 0 και 1, επίσης γνωστή και ως κανονικοποίηση, μπορούν να οδηγήσουν σε στρεβλές σχέσεις μεταξύ των μεταβλητών και μπορούν να αποκρύψουν την πραγματική δομή των δεδομένων (Geladi και Kowalski, 1986).

Τέλος μία ακόμη σημαντική αιτία στρέβλωσης μετά από κλιμάκωση προέρχεται και από το είδος και τη μορφή των δεδομένων, και συγκεκριμένα από τα μη συνεχή δεδομένα: Η

κλιμάκωση min-max έχει σχεδιαστεί για συνεχή δεδομένα και μπορεί να μην είναι κατάλληλη για διακριτά ή κατηγορικά δεδομένα. Σε αυτές τις περιπτώσεις, άλλες μέθοδοι κλιμάκωσης, όπως η κωδικοποίηση ενός σημείου (one-hot encoding), η κωδικοποίηση ετικέτας (label encoding) ή η κωδικοποίηση κατά σειρά (ordinal encoding) μπορεί να είναι πιο κατάλληλες.

Η κωδικοποίηση one-hot είναι μια κοινή τεχνική που χρησιμοποιείται για την αναπαράσταση κατηγορικών δεδομένων ως αριθμητικών δεδομένων στη μηχανική μάθηση και την ανάλυση δεδομένων. Τα κατηγορικά δεδομένα αναφέρονται σε δεδομένα που αποτελούνται από κατηγορίες ή ετικέτες, όπως χρώματα, σχήματα ή ονόματα αντικειμένων. Περιλαμβάνει τη δημιουργία ενός δυαδικού διανύσματος για κάθε κατηγορία στο σύνολο δεδομένων, όπου το διάνυσμα είναι τόσο μεγάλο όσο ο συνολικός αριθμός των κατηγοριών στο σύνολο δεδομένων. Το διάνυσμα περιέχει την τιμή 1 στη θέση που αντιστοιχεί στην κατηγορία και 0 σε όλες τις άλλες θέσεις. Για τη συγκεκριμένη περίπτωση, όταν χρειάστηκε να ποσοτικοποιηθούν τα ποιοτικά δεδομένα για χάριν της ανάλυσης, χρησιμοποιήθηκε η μέθοδος one-hot encoding.

Συμπερασματικά, μπορούμε να υποθέσουμε ότι τα χαρακτηριστικά των δεδομένων του δείγματός μας κατέστησαν την κλιμάκωση μάλλον ακατάλληλη επιλογή με αποτέλεσμα την στρέβλωση. Οι λόγοι για αυτό περιλαμβάνουν την ενίσχυση της επίδρασης των ακραίων τιμών, την εισαγωγή μη γραμμικών σχέσεων μεταξύ μεταβλητών, την αλλαγή της κατανομής των δεδομένων, και τα μη συνεχή δεδομένα. Στα δεδομένα που ελήφθησαν, παρατηρήθηκαν ακραίες τιμές, λοξές κατανομές και μη συνεχή δεδομένα, δηλαδή πλήθος ποιοτικών μεταβλητών. Ως εκ τούτου, η απόπειρα για κλιμάκωση διαστρέβλωσε τις σχέσεις μεταξύ των σημείων δεδομένων. Παρ' όλα αυτά, η διαφορά στο πλάτος της σιλουέτας πριν και μετά την κλιμάκωση παρείχε πληροφορίες σχετικά με την επίδραση της κλιμάκωσης στη λύση ομαδοποίησης και βοήθησε στον προσδιορισμό της χρήσης των αποτελεσμάτων με ή χωρίς κλιμάκωση.

4. ΑΠΟΤΕΛΕΣΜΑΤΑ

4.1 K-means

Η ομαδοποίηση K-means είναι μια ευρέως χρησιμοποιούμενη μέθοδος για την ομαδοποίηση παρόμοιων αντικειμένων σε συστάδες. Αυτή η διαδικασία περιγράφει μια ολοκληρωμένη προσέγγιση για την εκτέλεση της συσταδοποίησης k-means σε ένα σύνολο δεδομένων, συμπεριλαμβανομένου του υπολογισμού των μέτρων απόστασης, του προσδιορισμού του βέλτιστου αριθμού συστάδων και της οπτικοποίησης των αποτελεσμάτων της συσταδοποίησης.

Ακολουθεί η περιγραφή της διαδικασίας που ακολουθήθηκε για την εκτέλεση της ομαδοποίησης k-means στο σύνολο των δεδομένων των projects. Συνοπτικά, η διαδικασία αυτή περιλαμβάνει τον υπολογισμό των μέτρων απόστασης, τον προσδιορισμό του βέλτιστου αριθμού συστάδων και την οπτικοποίηση των αποτελεσμάτων της συσταδοποίησης.

Αρχικά, τα δεδομένα τυποποιήθηκαν για να εξασφαλιστεί ότι όλες οι μεταβλητές έχουν τον ίδιο αντίκτυπο στα αποτελέσματα της συσταδοποίησης. Στη συνέχεια, υπολογίζονται τα μέτρα απόστασης μεταξύ των σημείων δεδομένων. Αυτό επιτρέπει τον υπολογισμό των ομοιοτήτων και των διαφορών μεταξύ των σημείων δεδομένων, το οποίο αποτελεί βασικό βήμα στην ομαδοποίηση.

4.1.1 Προσδιορισμός & Αξιολόγηση Αριθμού Συστάδων

Προκειμένου να προσδιοριστεί ο βέλτιστος αριθμός συστάδων, χρησιμοποιήθηκαν διάφορες μέθοδοι για την αξιολόγηση της καταλληλότητας της λύσης συστάδων. Αυτές οι μέθοδοι παρέχουν πληροφορίες σχετικά με το πόσο καλά οι συστάδες ταιριάζουν στα δεδομένα και πόσο παρόμοια είναι τα σημεία δεδομένων μέσα σε κάθε συστάδα. Χρησιμοποιούνται για την αξιολόγηση της απόδοσης των αλγορίθμων ομαδοποίησης και τον προσδιορισμό του βέλτιστου αριθμού ομάδων για ένα δεδομένο σύνολο δεδομένων.

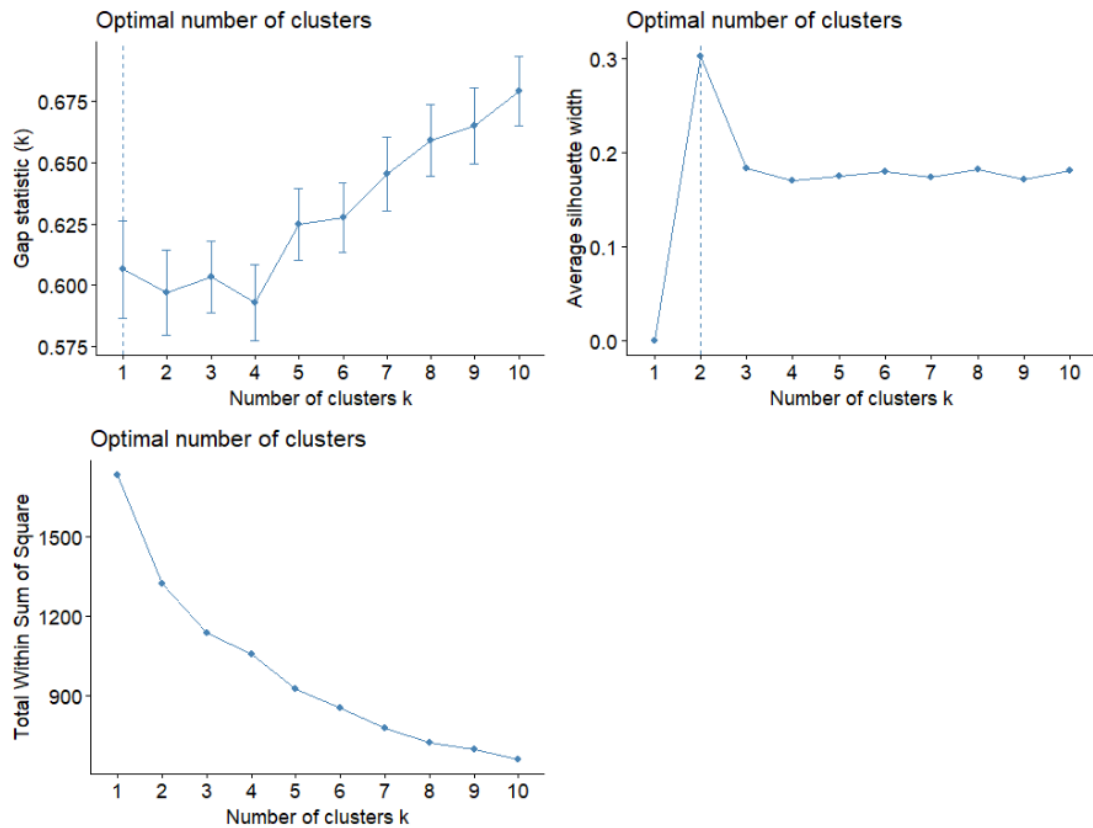
Συγκεκριμένα, για τον προσδιορισμό του βέλτιστου αριθμού συστάδων χρησιμοποιήθηκαν οι μέθοδοι:

- **Gap statistic:** είναι μια μέθοδος που συγκρίνει το άθροισμα τετραγώνων εντός της συστάδας ενός αλγορίθμου ομαδοποίησης με αυτό που αναμένεται βάσει μιας

μηδενικής κατανομής αναφοράς και μπορεί να εφαρμοστεί σε οποιονδήποτε αλγόριθμο ομαδοποίησης.

- **Silhouette width:** το πλάτος σιλουέτας είναι ένα μέτρο της ομοιότητας μεταξύ ενός δείγματος και των άλλων δειγμάτων της ίδιας συστάδας, σε σύγκριση με εκείνα άλλων συστάδων. Αυτή η μέθοδος είναι ειδική για αλγορίθμους συσταδοποίησης με βάση την κατάτμηση, όπως ο k-means.
- **Elbow method** (total within sum of square): Η μέθοδος Elbow (συνολικό άθροισμα τετραγώνων εντός των συστάδων) είναι μια γραφική προσέγγιση για τον προσδιορισμό του βέλτιστου αριθμού συστάδων μέσω της απεικόνισης του συνολικού αθροίσματος τετραγώνων εντός των συστάδων ως συνάρτηση του αριθμού των συστάδων. Η μέθοδος αυτή μπορεί να εφαρμοστεί σε οποιονδήποτε αλγόριθμο συσταδοποίησης που ελαχιστοποιεί το άθροισμα των τετραγωνικών αποστάσεων μεταξύ των σημείων και των καθορισμένων κεντροειδών των συστάδων τους.

Κάθε μία από τις παραπάνω μεθόδους είναι χρήσιμα εργαλεία για την αξιολόγηση της ποιότητας των λύσεων ομαδοποίησης αλλά δεν είναι αλάνθαστες μέθοδοι και μπορεί να μην δίνουν πάντα τα καλύτερα αποτελέσματα. Σε κάθε περίπτωση χρησιμοποιήθηκαν και οι τρεις και τα αποτελέσματα κάθε μεθόδου απεικονίζονται στα παρακάτω γραφήματα:



Διάγρ. 19. Gap Statistic, Silhouette Width & Elbow Method for K-means

Στο διάγραμμα της μεθόδου Gap Statistic ο βέλτιστος αριθμός συστάδων καθορίζεται αναζητώντας το σημείο όπου η τιμή αρχίζει να μειώνεται. Ο αριθμός συστάδων που αντιστοιχεί στο σημείο αυτό λαμβάνεται ως ο βέλτιστος αριθμός συστάδων και στη συγκεκριμένη περίπτωση είναι οι 2 ή και οι 4 συστάδες.

Για την εφαρμογή του πλάτους Silhouette στον αλγόριθμο K-means, υπολογίζεται το πλάτος Silhouette για κάθε σημείο δεδομένων για έναν δεδομένο αριθμό συστάδων, χρησιμοποιώντας τον αλγόριθμο K-means. Το πλάτος Silhouette κυμαίνεται από -1 έως 1, με τιμές κοντά στο 1 να υποδηλώνουν μια καλή λύση ομαδοποίησης και τιμές κοντά στο -1 να υποδηλώνουν μια κακή λύση ομαδοποίησης. Ο αριθμός συστάδων που οδηγεί στο υψηλότερο μέσο πλάτος Silhouette, καθορίζει τον βέλτιστο αριθμό συστάδων για το σύνολο δεδομένων και στην συγκεκριμένη περίπτωση αυτός ο αριθμός είναι 2 συστάδες.

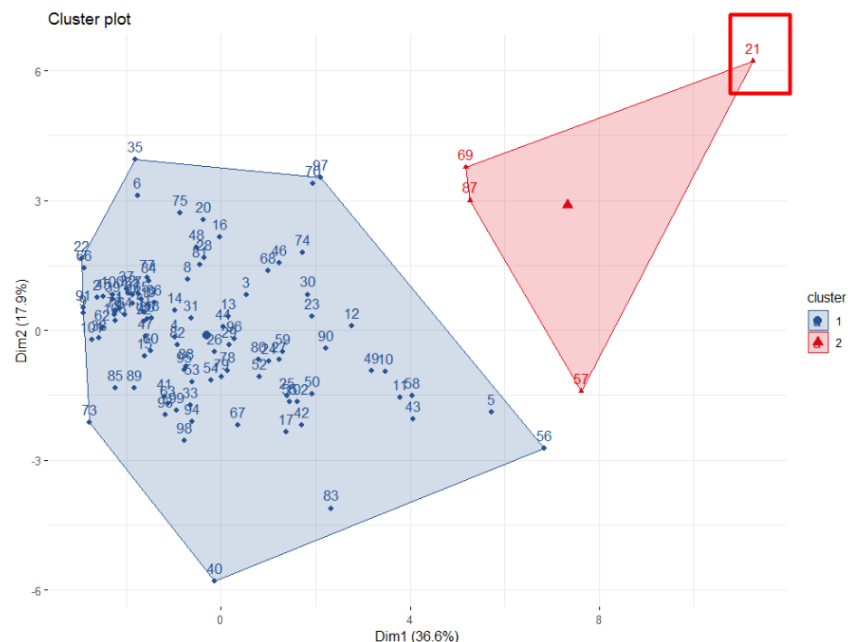
Στη χρήση της μεθόδου Elbow για τον K-means, το συνολικό WCSS (Within-Cluster Sum of Squares) υπολογίζεται για διαφορετικές τιμές του K (αριθμός συστάδων) και απεικονίζεται στο γράφημα. Ο βέλτιστος αριθμός συστάδων προσδιορίζεται με την

εύρεση του "αγκώνα" στο διάγραμμα, ο οποίος είναι το σημείο στο οποίο η προσθήκη άλλης συστάδας οδηγεί σε φθίνουσα οριακή μείωση του WCSS, και στη συγκεκριμένη περίπτωση αυτό γίνεται στις 3 συστάδες. Αυτό το σημείο λαμβάνεται ως ο βέλτιστος αριθμός συστάδων, καθώς η προσθήκη περισσότερων συστάδων πέραν αυτού του σημείου μπορεί να μην οδηγήσει σε σημαντική βελτίωση της λύσης ομαδοποίησης.

Σύμφωνα με τα παραπάνω ο βέλτιστος αριθμός είναι δύο, τρεις ή και τέσσερις συστάδες, ανάλογα τη μέθοδο.

4.1.2 Οπτικοποίηση των Clusters

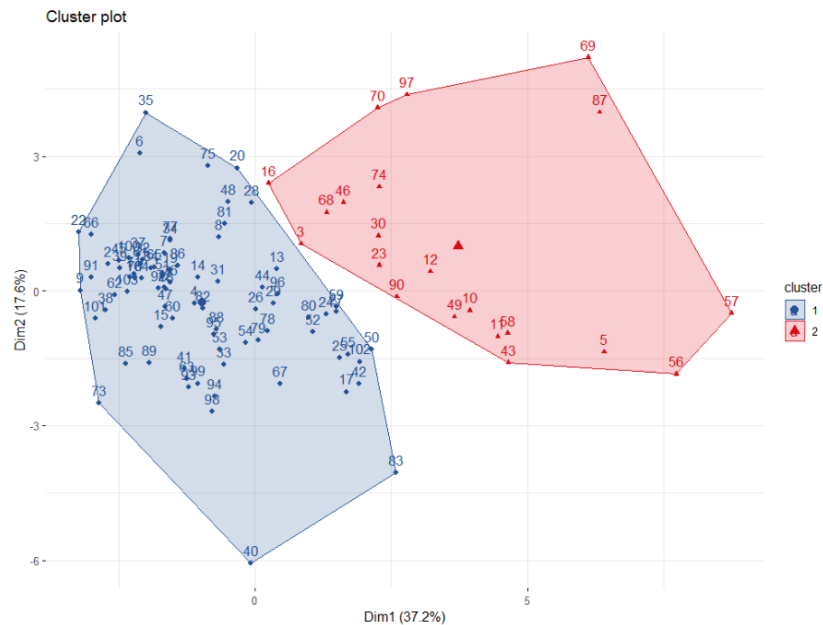
Στη συνέχεια εκτελέστηκε ο αλγόριθμος συσταδοποίησης k-means, όπου τα δεδομένα χωρίζονται στον προκαθορισμένο αριθμό συστάδων, δηλαδή 2 ($k=2$).



Διάγρ. 20. K-means, $k=2$, με το Outlier

Στην πρώτη εφαρμογή του αλγορίθμου, στην οπτικοποίηση των αποτελεσμάτων, παρατηρήθηκε ότι θα πρέπει να αφαιρεθεί η παρατήρηση No21, διότι θεωρείται ακραίο έκτοπο σημείο (outlier) και επηρεάζει σημαντικά την ομαδοποίηση και το μέγεθος των clusters. Επομένως η ανάλυση συνεχίστηκε με την αφαίρεση αυτού του έργου και το σύνολο των δεδομένων διαμορφώθηκε σε 102 έργα.

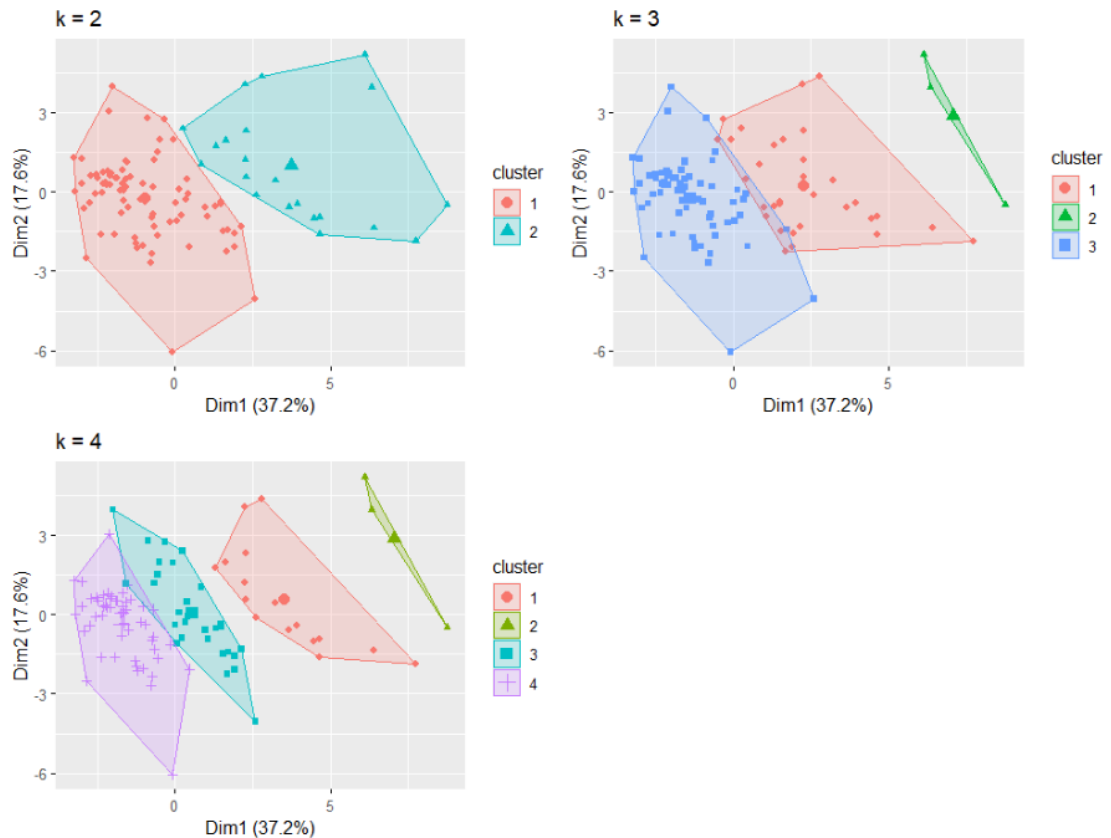
Μετά την αφαίρεση της παρατήρησης 21, τα αποτελέσματα της συσταδοποίησης για $k=2$, οπτικοποιήθηκαν με τρόπο που επιτρέπει τη σαφή ερμηνεία της δομής των δεδομένων, όπως φαίνεται στο σχήμα παρακάτω:



Διάγρ. 21. K-means, $k=2$, χωρίς το Outlier

Παρατηρείται ότι τα δεδομένα χωρίστηκαν σε 2 ομάδες μεγέθους 81 και 21 παρατηρήσεων (projects) αντίστοιχα.

Επιπλέον, ο αλγόριθμος εκτελέστηκε και για 3 και 4 συστάδες, προκειμένου να έχουμε και μια συγκριτική απεικόνιση μεταξύ των δυο, όπως φαίνεται στο σχήμα παρακάτω.



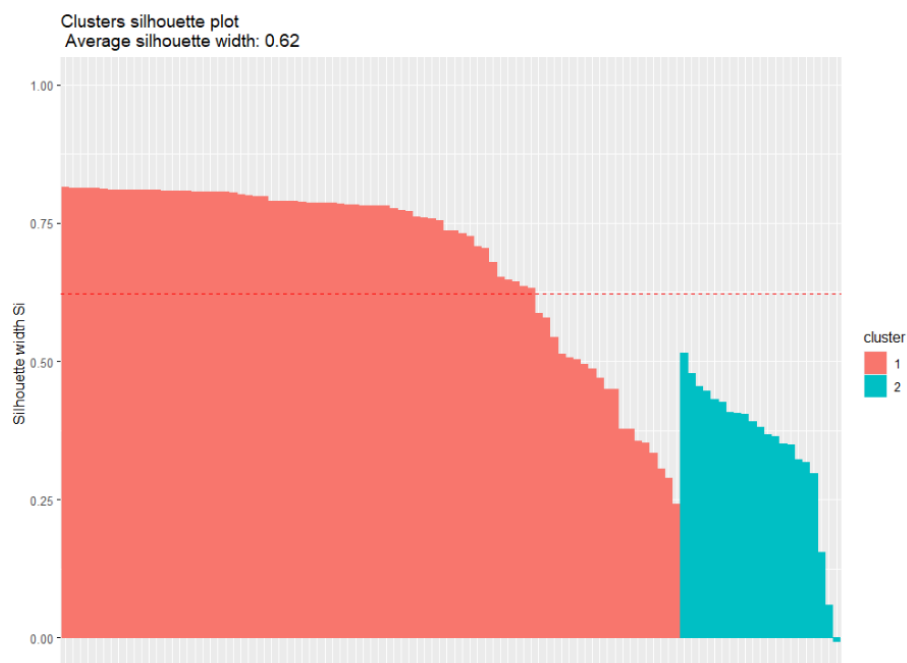
Διάγρ. 22. K-means, $k = 2$, $k = 3$, $k = 4$

4.1.3 Αξιολόγηση Λύσεων

Τέλος για τη διαδικασία της αξιολόγησης της ποιότητας της λύσης ομαδοποίησης, σε 2, 3 και 4 συστάδες, ακολουθήθηκαν τα παρακάτω βήματα.

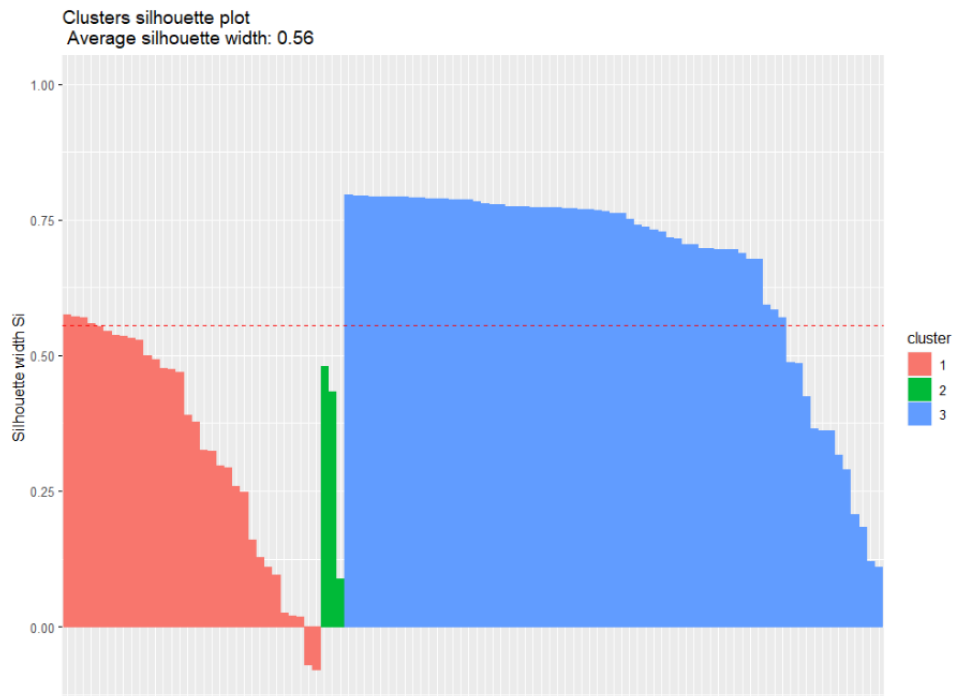
Το πρώτο βήμα είναι ο υπολογισμός της πληροφορίας silhouette, η οποία είναι ένα μέτρο του πόσο παρόμοιο είναι ένα αντικείμενο με τη δική του συστάδα σε σύγκριση με άλλες συστάδες.

Το δεύτερο βήμα είναι η επιθεώρηση του διαγράμματος silhouette, το οποίο παρέχει μια γραφική αναπαράσταση της πληροφορίας silhouette. Το διάγραμμα είναι ένα ραβδόγραμμα που δείχνει το πλάτος της σιλουέτας για κάθε παρατήρηση. Εάν το πλάτος της σιλουέτας (silhouette width) είναι κοντά στο 1, υποδηλώνει ότι η παρατήρηση είναι καλά συγκεντρωμένη σε συστάδες. Εάν το πλάτος της σιλουέτας είναι κοντά στο 0, υποδεικνύει ότι η παρατήρηση δεν είναι καλά συγκεντρωμένη σε συστάδες και μπορεί να έχει ανατεθεί σε λάθος συστάδα. Εάν το silhouette width είναι αρνητικό, υποδηλώνει ότι η παρατήρηση έχει ανατεθεί σε λάθος συστάδα. Το silhouette width αποτυπώνεται στο παρακάτω σχήμα:

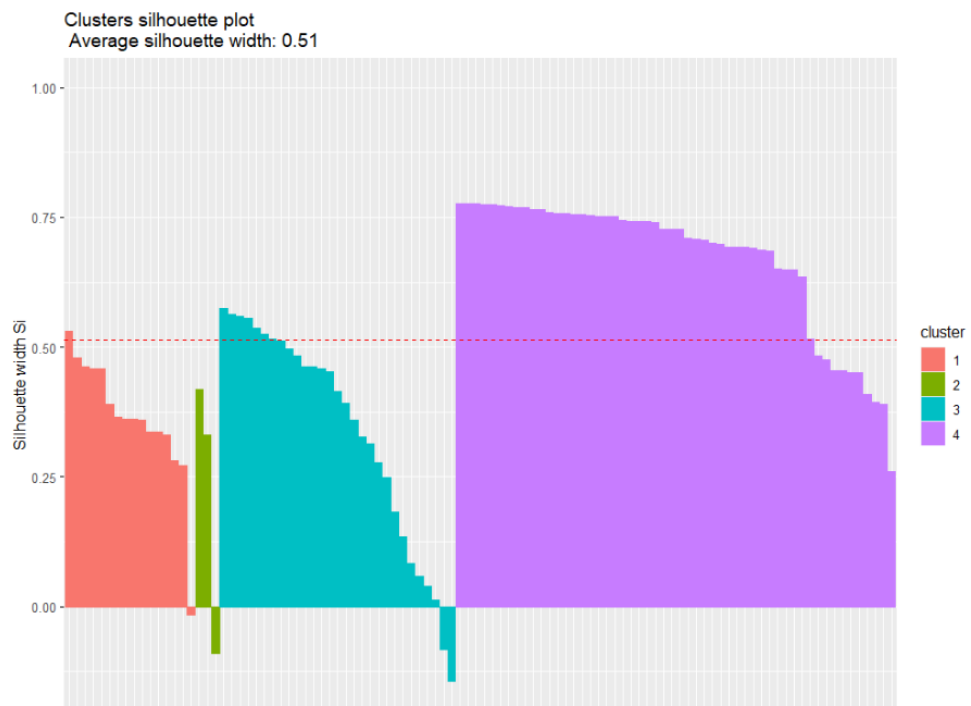


Διάγρ. 23. *K-means Silhouette Width για $k = 2$*

Ένα πλάτος σιλουέτας 0,62 θεωρείται ότι βρίσκεται στο "λογικό" εύρος, υποδεικνύοντας ότι τα δείγματα είναι αρκετά καλά συγκεντρωμένα και διακριτά από τα δείγματα σε άλλες συστάδες. Ωστόσο, δεν είναι απαραίτητα μια εξαιρετική λύση ομαδοποίησης, καθώς η ερμηνεία της βαθμολογίας του πλάτους σιλουέτας εξαρτάται σε μεγάλο βαθμό από τα δεδομένα και το πλαίσιο της ανάλυσης. Το πλάτος σιλουέτας κυμαίνεται από -1 έως 1, με μια τιμή πιο κοντά στο 1 να υποδηλώνει ισχυρή ομοιότητα εντός της εκχωρημένης συστάδας και μια τιμή πιο κοντά στο -1 να υποδηλώνει ισχυρή ομοιότητα με διαφορετική συστάδα.



Διάγρ. 24. K-means Silhouette Width για $k=3$



Διάγρ. 25. K-means Silhouette Width για $k=4$

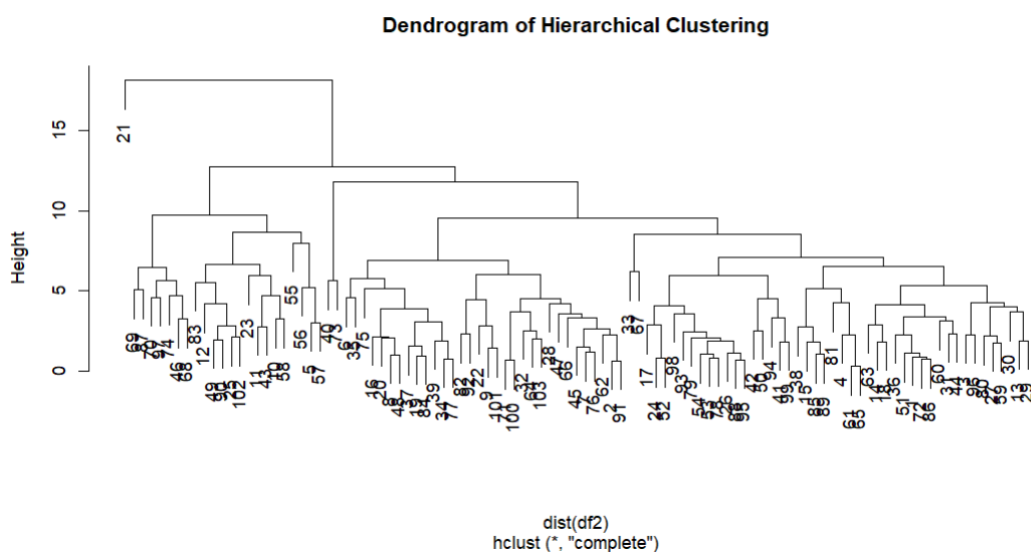
Συγκριτικά με την δοκιμή του αλγορίθμου σε 3 και 4 συστάδες, το silhouette width διαμορφώθηκε σε 0,56 και 0,51 αντίστοιχα (Διάγρ. 24 & 25), δηλαδή ακόμα μικρότερο,

γεγονός που επιβεβαιώνει ότι πράγματι η βέλτιστη λύση για τον αλγόριθμο K-means είναι 2 συστάδες. Ταυτόχρονα οι 3 και οι 4 συστάδες, έχουν μεγάλη ανομοιογένεια στο μέγεθος των clusters. Δηλαδή, οι 3 συστάδες είναι μεγέθους 67, 32, και 3 παρατηρήσεων αντίστοιχα και επομένως αυτή η ομαδοποίηση, δεν είναι η ιδανική για το πλαίσιο της ανάλυσης.

4.2 Hierarchical Clustering

Στα πλαίσια της παρούσας έρευνας, χρησιμοποιήθηκε και μια προσέγγιση ιεραρχικής ομαδοποίησης για την ομαδοποίηση παρόμοιων παρατηρήσεων με βάση την ομοιότητά τους. Εφαρμόστηκε δηλαδή και ιεραρχικό clustering.

Το πρώτο βήμα περιλάμβανε τον υπολογισμό του πίνακα απόστασης μεταξύ κάθε ζεύγους παρατηρήσεων στο σύνολο δεδομένων. Στη συνέχεια εφαρμόστηκε η μέθοδος ιεραρχικής ομαδοποίησης στον πίνακα αποστάσεων, με αποτέλεσμα να προκύψει ένα δένδρογραμμα που έδειχνε τις σχέσεις μεταξύ των παρατηρήσεων, όπως φαίνεται στο παρακάτω σχήμα.

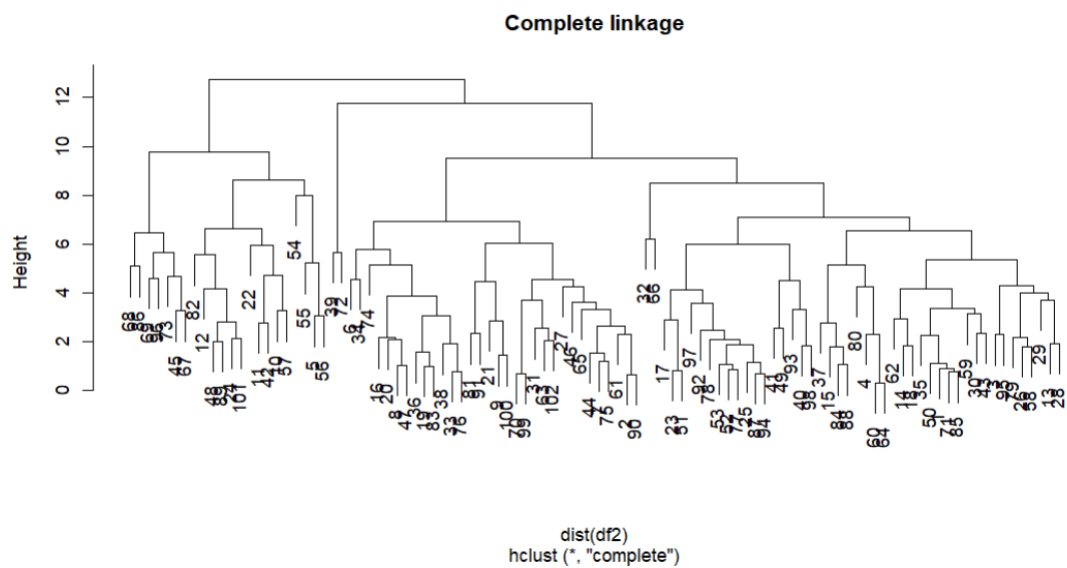


Διάγρ. 26. Hierarchical Clustering with Outlier

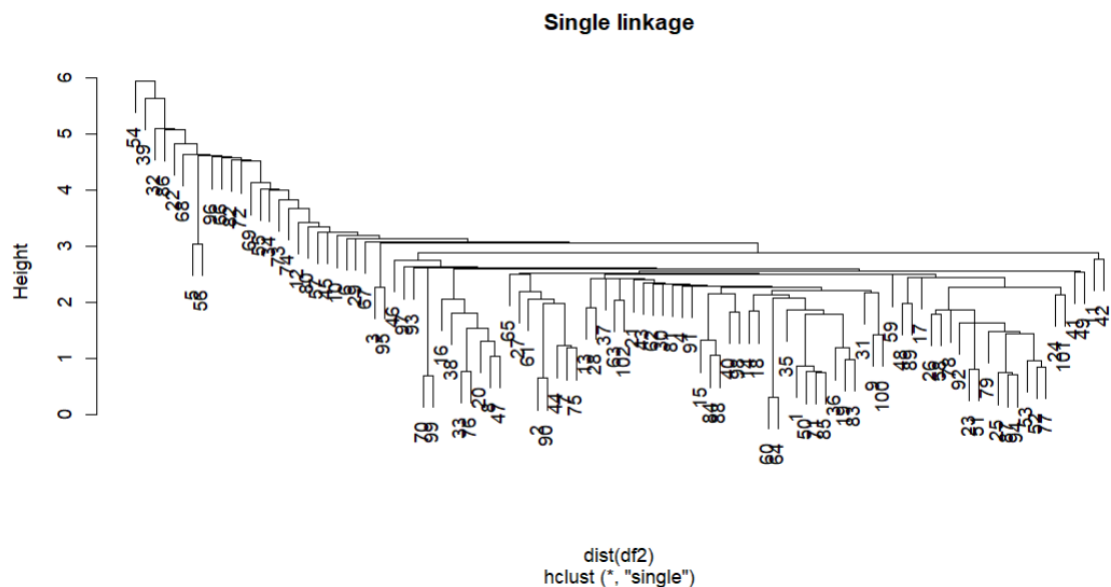
Είναι προφανές ότι το Projects στη γραμμή 21 επηρεάζει αρνητικά το σχήμα του δένδρογράμματος και επομένως αφαιρείται από το πλαίσιο δεδομένων, όπως και στον προηγούμενο αλγόριθμο (k-means), καθώς θεωρείται ακραία τιμή.

Το δενδρόγραμμα χρησιμοποιήθηκε για τον προσδιορισμό του βέλτιστου αριθμού συστάδων. Ο βέλτιστος αριθμός συστάδων αποφασίστηκε με βάση την οπτικοποίηση του δενδρογράμματος και το μέσο πλάτος της σιλουέτας.

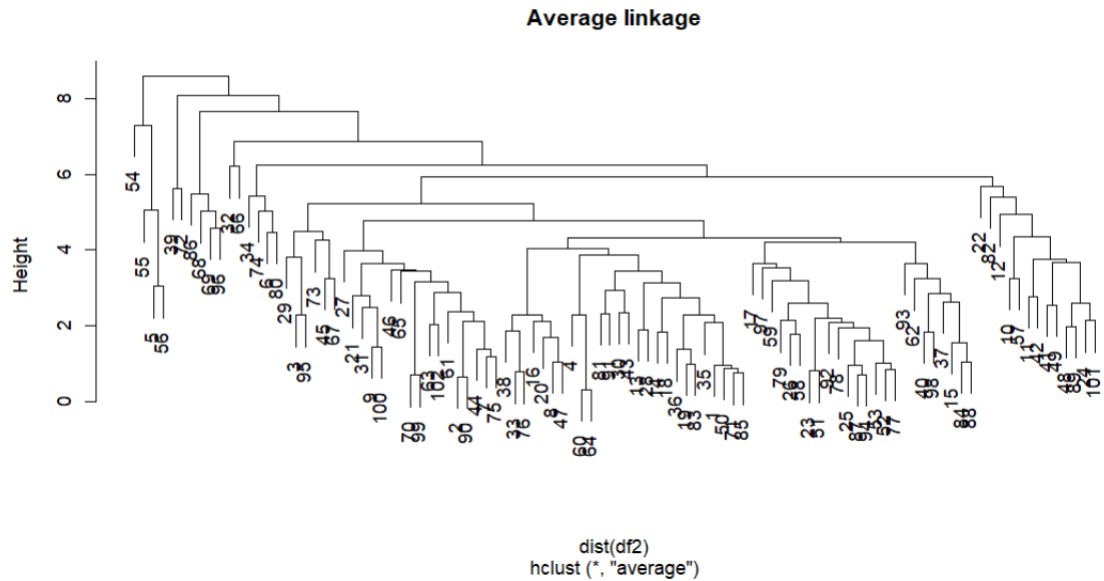
Η μέθοδος πλήρους σύνδεσης (complete linkage) επιλέχθηκε ως κριτήριο σύνδεσης για τον αλγόριθμο ιεραρχικής συσταδοποίησης, αλλά συγκρίθηκαν επίσης τα αποτελέσματα των μεθόδων απλής (single) και μέσης σύνδεσης (average) και τα δενδρογράμματα που προκύπτουν απεικονίζονται για να συγκριθούν οι διαφορές μεταξύ των τριών μεθόδων.



Διάγρ. 27. Hierarchical Clustering Complete Method



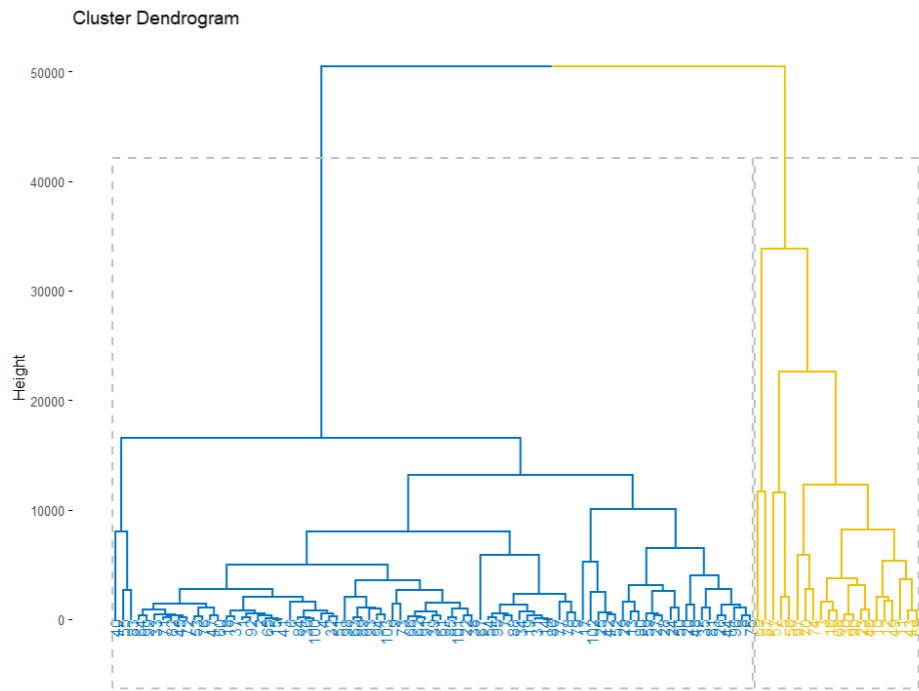
Διάγρ. 28. Hierarchical Clustering Single Method



Διάγρ. 29. Hierarchical Clustering Average Method

4.2.1 Οπτικοποίηση των Λύσεων

Στη συνέχεια η διαδικασία προσδιορισμού του βέλτιστου αριθμού συστάδων και η αξιολόγηση των αποτελεσμάτων της ομαδοποίησης επαναλαμβάνεται για $k=2$, $k=3$ και $k=4$ και τα αποτελέσματα απεικονίστηκαν με τη χρήση δένδρογραμμάτων. Το πλάτος της σιλουέτας (silhouette width) υπολογίστηκε και απεικονίστηκε για κάθε ένα από αυτά τα μοντέλα για να αξιολογηθεί η απόδοση της μεθόδου συσταδοποίησης, όπως φαίνεται στα παρακάτω σχήματα.

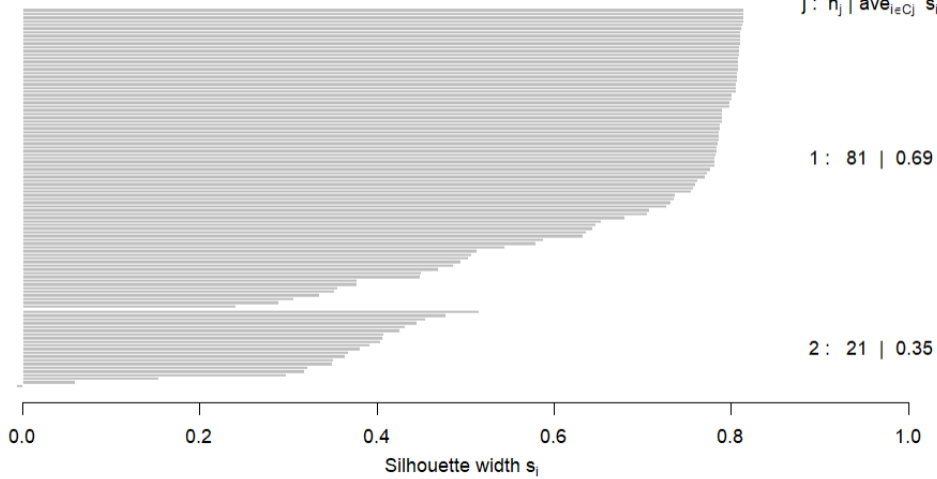


Διάγρ. 30. Hierarchical Clustering, $k=2$

Silhouette width plot

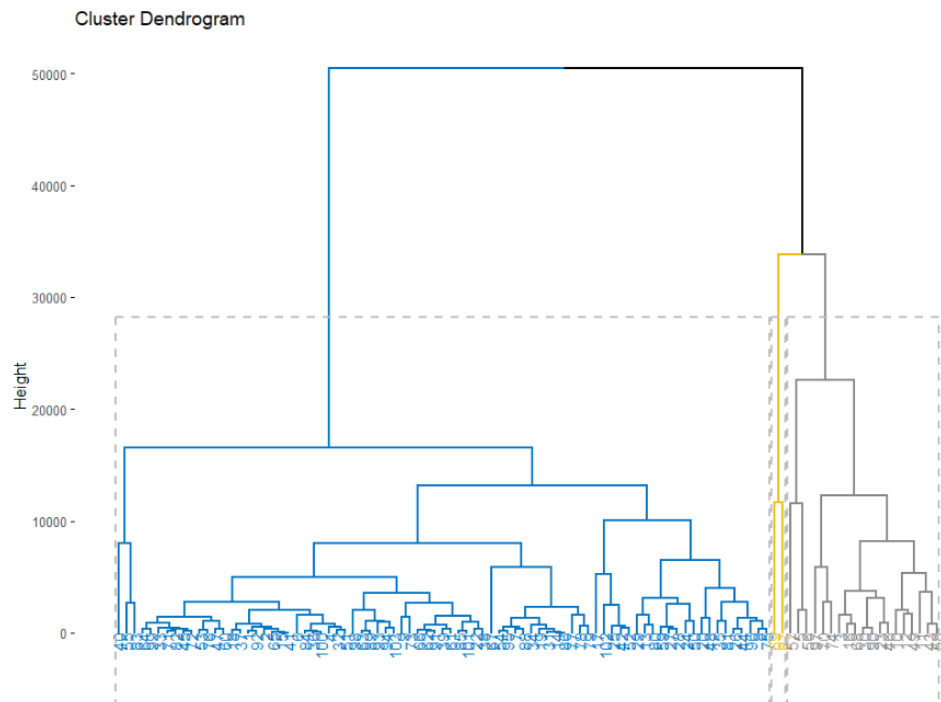
$n = 102$

2 clusters C_j
 $j : n_j \mid \text{ave}_{i \in C_j} s_i$

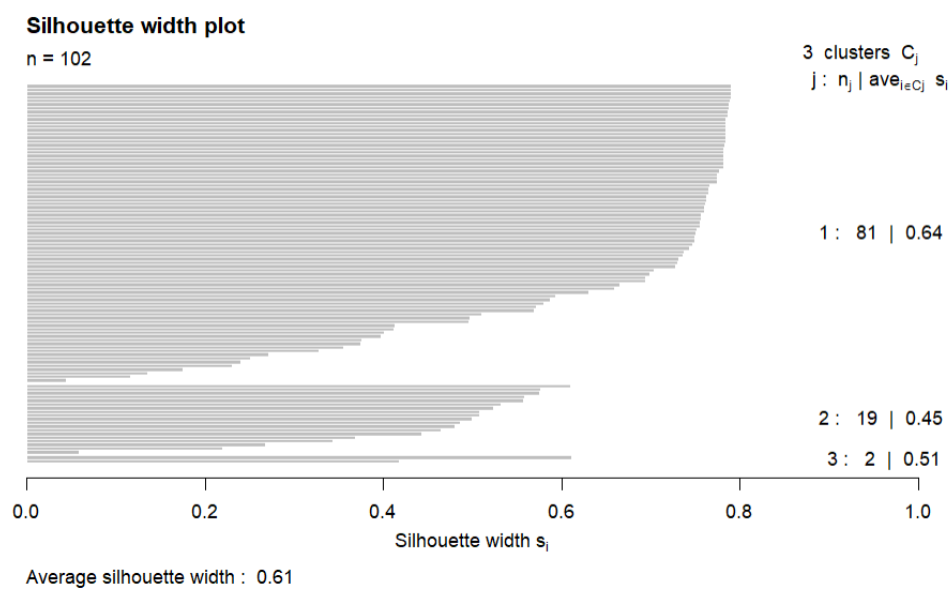


Average silhouette width : 0.62

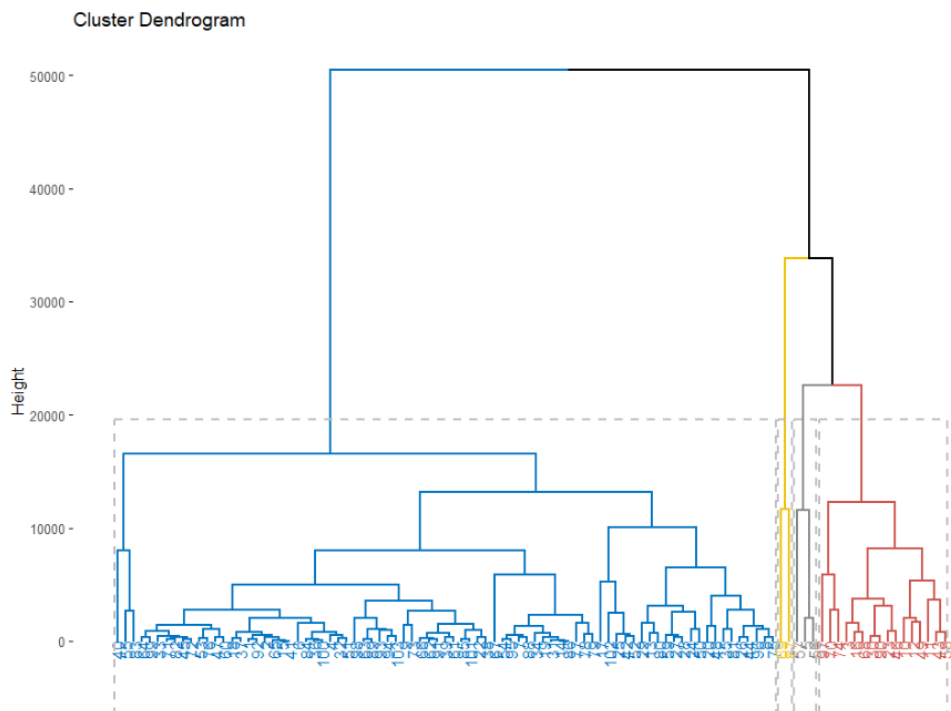
Διάγρ. 31. Silhouette Width Hier. Cl., $k=2$



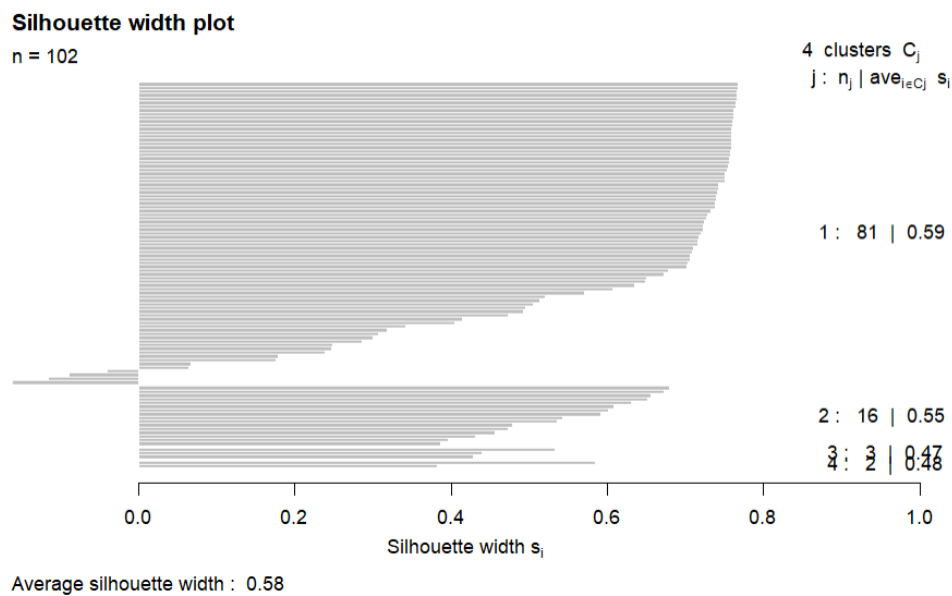
Διάγρ. 32. Hierarchical Clustering, $k=3$



Διάγρ. 33. Silhouette Width Hier. Cl., $k=3$



Διάγρ. 34. Hierarchical Clustering, $k=4$



Διάγρ. 35. Silhouette Width Hier. Cl., $k=4$

4.2.2 Αξιολόγηση των Λύσεων

Όπως αναφέρθηκε και προηγουμένως το μέσο πλάτος σιλουέτας παρέχει ένα συνολικό μέτρο της ποιότητας της λύσης ομαδοποίησης. Ένα υψηλότερο μέσο πλάτος σιλουέτας υποδηλώνει μια καλύτερη λύση.

Με βάση τα αποτελέσματα της ανάλυσης, ο βέλτιστος αριθμός συστάδων καθορίστηκε ως η τιμή που οδηγεί στο υψηλότερο μέσο πλάτος σιλουέτας και αυτή είναι με ελάχιστη διαφορά μεταξύ τους, οι 2 ή οι 3 συστάδες, μεγέθους 81 και 21 παρατηρήσεων και 81, 19 και 2 παρατηρήσεων αντίστοιχα. Το μέσο πλάτος σιλουέτας ισούται με 0.62 για τις 2 συστάδες και 0.61 για τις 3 συστάδες. Ένα μέσο πλάτος σιλουέτας 0,61 υποδεικνύει ότι τα έργα είναι αρκετά καλά συγκεντρωμένα σε συστάδες και αρκετά διακριτά από τα έργα σε άλλες συστάδες, αλλά δεν είναι απαραίτητα μια εξαιρετική λύση ομαδοποίησης.

Δεδομένου λοιπόν ότι το μέσο πλάτος σιλουέτας είναι σχεδόν ίδιο, και το μέγεθος των 3 συστάδων είναι αρκετά δυσανάλογο μεταξύ τους, όπως και στην περίπτωση του αλγορίθμου kmeans για $k=3$, θεωρούμε ότι η βέλτιστη λύση από το ιεραρχικό clustering, είναι η λύση των 2 συστάδων.

Συγκριτικά με τον πρώτο αλγόριθμο Kmeans, παρατηρείται, ότι λαμβάνουμε παρόμοια αποτελέσματα και ότι βέλτιστος αριθμός συστάδων και στις δύο περιπτώσεις αλγορίθμων είναι οι 2 συστάδες. Αν και μικρή διαφορά στο πλάτος σιλουέτας, οι 3 συστάδες στο ιεραρχικό clustering και στον αλγόριθμο k-means, έχουν σημαντική διαφορά στο μέγεθος μεταξύ τους και επομένως η βέλτιστη λύση σε κάθε περίπτωση είναι τα 2 clusters.

4.3 PAM Clustering

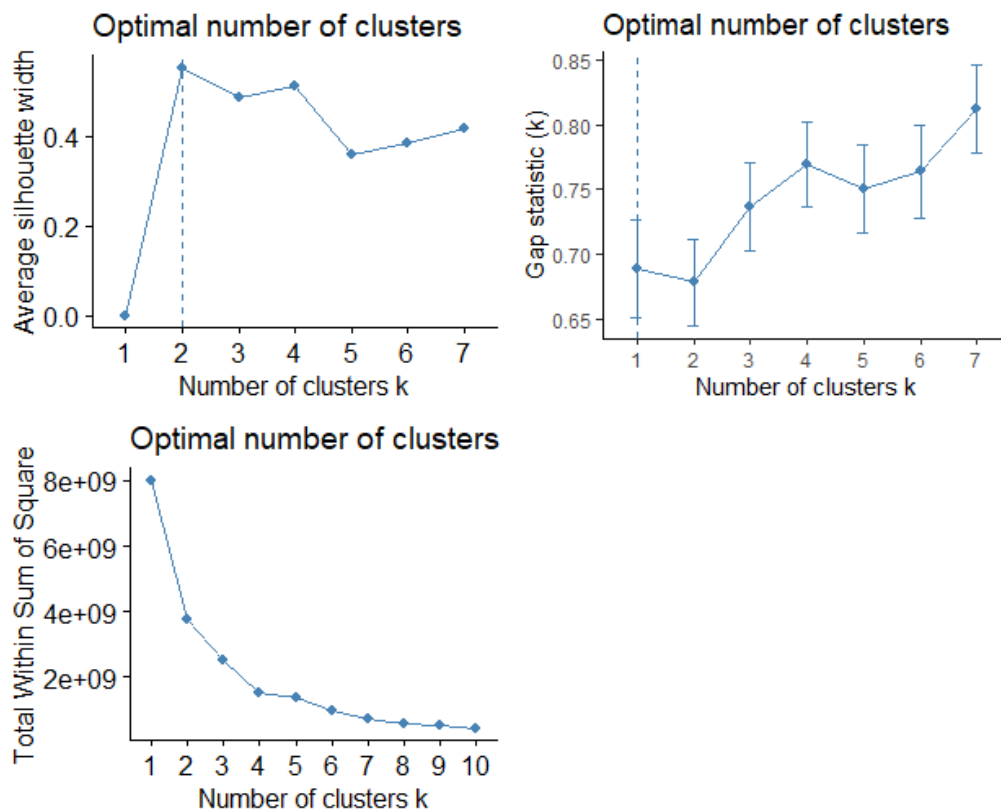
4.3.1 Ο Βέλτιστος Αριθμός Συστάδων

Στα πλαίσια της αναζήτησης βέλτιστης λύσης για τη συσταδοποίηση των δεδομένων, προχωρήσαμε στην εφαρμογή τρίτου αλγορίθμου, την ομαδοποίηση PAM. Η PAM (Partitioning Around Medoids) είναι μια τεχνική που χρησιμοποιείται για την ομαδοποίηση δεδομένων σε ομάδες με βάση την ομοιότητά τους. Οι λόγοι για τους οποίους επιλέχθηκε αυτή η τεχνική στην έρευνα είναι γιατί η PAM είναι ανθεκτική στις ακραίες τιμές και είναι λιγότερο ευαίσθητη στην παρουσία θορύβου στα δεδομένα. Αυτό την καθιστά καλή επιλογή όταν τα δεδομένα περιέχουν μεγάλο αριθμό ακραίων τιμών ή όταν τα δεδομένα είναι θορυβώδη. Επιπλέον Η PAM είναι ικανή να χειρίζεται

κατηγορικά δεδομένα, ή σύνολα δεδομένων με συνδυασμό αριθμητικών και κατηγορικών δεδομένων, ενώ άλλες τεχνικές ομαδοποίησης, όπως η k-means, περιορίζονται συνήθως σε αριθμητικά δεδομένα. Ακόμη, η PAM παράγει αποτελέσματα που είναι εύκολα ερμηνεύσιμα και κατανοητά, καθώς αναθέτει κάθε σημείο δεδομένων στο πλησιέστερο medoid. Επομένως για τους παραπάνω λόγους εκτιμήθηκε ότι η ομαδοποίηση PAM είναι κατάλληλη μέθοδος για το περίπτωσι των δεδομένων της παρούσας έρευνας και ενδεχομένως θα προσφέρει αρκετά πλεονεκτήματα έναντι των άλλων τεχνικών ομαδοποίησης.

Παρακάτω ακολουθεί η περιγραφή της διαδικασίας για την εκτέλεση της ομαδοποίησης PAM στο σύνολο των δεδομένων μας. Συνοπτικά, η διαδικασία εκτέλεσης της ομαδοποίησης PAM περιλαμβάνει, όπως συνήθως, τον προσδιορισμό του βέλτιστου αριθμού συστάδων, την ομαδοποίηση των σημείων δεδομένων σε συστάδες με βάση την ομοιότητά τους, την οπτικοποίηση των αποτελεσμάτων και την αξιολόγηση της ποιότητας των αποτελεσμάτων της ομαδοποίησης.

Το πρώτο βήμα ήταν ο προσδιορισμός του βέλτιστου αριθμού συστάδων. Αυτό έγινε, όπως και στους προηγούμενους αλγόριθμους, κατασκευάζοντας διαγράμματα για τις μεθόδους Gap statistic, Silhouette width, Elbow method (total within sum of square), οι οποίες βοηθούν στην αξιολόγηση της καταλληλότητας της λύσης συστάδων.



Διάγρ. 36. Gap Statistic, Silhouette Width & Elbow Method for PAM

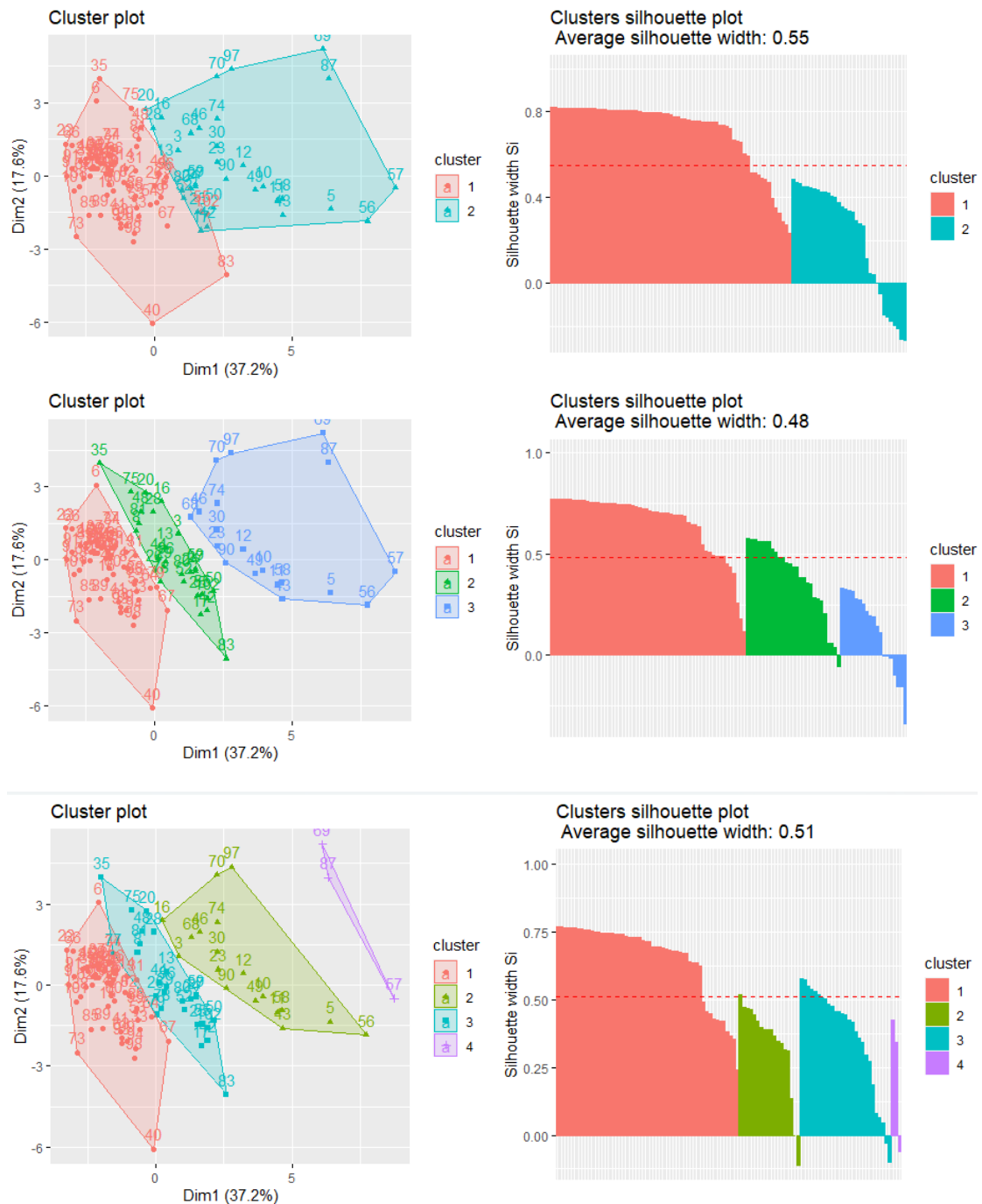
Από τα παραπάνω σχήματα, φαίνεται ότι βέλτιστες είναι 2, 3 ή και 4 συστάδες. Επομένως η ανάλυση προχώρησε με τη δοκιμή διαφόρων τιμών του k, όπου το k αντιπροσωπεύει τον αριθμό των συστάδων. Σε αυτή την περίπτωση, δοκιμάσαμε τρεις διαφορετικές τιμές του k: 2, 3 και 4, δηλαδή τιμές που έχουν νόημα για τα δεδομένα. Μόλις καθοριστεί η τιμή του k, το επόμενο βήμα είναι η τελική εκτέλεση της ομαδοποίησης PAM. Αυτό περιλαμβάνει την ομαδοποίηση των σημείων δεδομένων σε k συστάδες με βάση την ομοιότητά τους. Το αποτέλεσμα αυτού του βήματος είναι ένα σύνολο k συστάδων, καθεμία από τις οποίες περιέχει έναν αριθμό σημείων δεδομένων.

Σημειώνεται ότι οι δοκιμές έγιναν και μετά την αφαίρεση του Project στη γραμμή 21 από τα δεδομένα, όπως και προηγουμένως στο ιεραρχικό clustering, καθώς θεωρείται εξαιρετικά ακραία τιμή και πράγματι η αφαίρεση βελτίωσε ελαφρώς την απόδοση των αποτελεσμάτων συσταδοποίησης.

4.3.2 Οπτικοποίηση των Συστάδων

Στη συνέχεια ακολούθησε η οπτικοποίηση των αποτελεσμάτων της ομαδοποίησης PAM. Αυτό περιλαμβάνει τη δημιουργία ενός γραφήματος που παρουσιάζει τα σημεία

δεδομένων, χρωματισμένα ανάλογα με τη συστάδα στην οποία ανήκουν, όπως και στους προηγούμενους αλγορίθμους. Στα παρακάτω σχήματα φαίνεται η συνολική δομή των αποτελεσμάτων της συσταδοποίησης για όλες τις περιπτώσεις που αξιολογήθηκαν.



Διάγρ. 37. PAM clustering for $k=2$, $k=3$, $k=4$

4.3.3 Αξιολόγηση των Συστάδων

Τέλος, η ποιότητα των αποτελεσμάτων της συσταδοποίησης αξιολογήθηκε με τη χρήση του διαγράμματος σιλουέτας. Όπως και προηγουμένως, το διάγραμμα σιλουέτας μας βοηθά να προσδιορίσουμε αν τα σημεία δεδομένων είναι καλά διαχωρισμένα σε διαφορετικές συστάδες και αν υπάρχει επικάλυψη μεταξύ των συστάδων. Στο Διάγρ. 37 παραπάνω, φαίνεται το μέσο πλάτος σιλουέτας για κάθε περίπτωση αντίστοιχα.

Με βάση τα παραπάνω γραφήματα παρατηρείται ότι και στην περίπτωση της PAM, η βέλτιστη λύση είναι ο σχηματισμός 2 συστάδων, μεγέθους 69 και 33 παρατηρήσεων, καθώς στις 2 συστάδες το μέσο πλάτος σιλουέτας λαμβάνει τη μέγιστη τιμή και ισούται με 0.55, ενώ μια εξίσου καλή λύση είναι και οι 3 συστάδες, μεγέθους 56, 27, 19 έργων αντίστοιχα, αφού το μέσο πλάτος σιλουέτας δεν μειώνεται σημαντικά. Η λύση των 4 συστάδων αν και με μέσο πλάτος σιλουέτας μεγαλύτερο από τις 3 συστάδες, δημιουργεί μία συστάδα από τις 4 με πολύ μικρό μέγεθος, δηλαδή με 3 έργα και επομένως απορρίπτεται διότι δεν εξυπηρετεί το σκοπό της έρευνας.

Με δεδομένο ότι στη παρούσα έρευνα και για το σκοπό που γίνεται μας ενδιαφέρει η μελέτη των έκτοπων σημείων (outliers), η λύση των 3 συστάδων με μέσο πλάτος σιλουέτας 0.48 περιλαμβάνει ακριβώς μια συστάδα από τις 3 με όλες τις ακραίες τιμές. Αυτή η λύση να προσεγγίζει ικανοποιητικά την τμηματοποίηση των έργων, γεγονός που θα αξιολογηθεί στη συνέχεια στην ανάλυση και ερμηνεία μεταξύ των παρατηρήσεων.

Σε κάθε περίπτωση η τεχνική PAM, έφερε τα καλύτερα αποτελέσματα στη συσταδοποίηση των έργων αξιολογώντας συνδυαστικά κριτήρια που αφορούν στο μέγεθος των επιμέρους clusters και το πλάτος σιλουέτας και επομένως, θα επιλεγεί αυτή για την περαιτέρω ανάλυση και ερμηνεία των χαρακτηριστικών των συστάδων.

4.4 Διερεύνηση & Προσδιορισμός των Χαρακτηριστικών των Clusters

4.4.1 Περιγραφή των Χαρακτηριστικών των Clusters

Δεδομένου του ερευνητικού σκοπού σχετικά με την τμηματοποίηση των έργων με βάση την κερδοφορία τους, και επειδή θεωρούμε σκόπιμο να συμπεριλάβουμε παρατηρήσεις που θεωρούνται ακραίες τιμές σε ξεχωριστή συστάδα, αποφασίζουμε να συνεχίσουμε την ανάλυση με 3 συστάδες. Η επιλογή αυτή έγινε λαμβάνοντας υπόψη την ανάλυση και

τη σύγκριση μεταξύ των λύσεων που προηγήθηκε η οποία υπέδειξε ότι οι 3 συστάδες είναι η καλύτερη λύση για τη συγκεκριμένη περίπτωση.

Το επόμενο βήμα είναι η κατανόηση και η εξαγωγή πληροφοριών σχετικά με σχετικά με τη δομή και τις σχέσεις των δεδομένων εντός του κάθε cluster. Η διερεύνηση των χαρακτηριστικών των συστάδων είναι ένα σημαντικό βήμα για την κατανόηση της δομής των δεδομένων. Ακολουθεί η περιγραφή της διαδικασίας που ακολουθήθηκε για τους δύο τύπους μεταβλητών που υπάρχουν στα δεδομένα, ποιοτικές και ποσοτικές.

Αριθμητικές μεταβλητές

Για τις αριθμητικές μεταβλητές, υπολογίστηκαν τα περιγραφικά στατιστικά για κάθε συστάδα. Αυτά τα στατιστικά στοιχεία περιλαμβάνουν τη μέση τιμή, τη διάμεσο, τις ελάχιστες και μέγιστες τιμές και τα κβαντικά ποσοστημόρια (όπως το 25ο, 50ο και 75ο εκατοστημόριο). Αυτά τα στατιστικά, χρησιμοποιήθηκαν για να διερευνηθεί η κεντρική τάση και η διασπορά των τιμών σε κάθε συστάδα. Έπειτα δημιουργήθηκαν γραφήματα (box plots) για να οπτικοποιηθεί η κατανομή των τιμών σε κάθε συστάδα και ραβδογράμματα (bar plots) των μέσων όρων. Με αυτές τις οπτικοποιήσεις εντοπίστηκαν τα σχετικά μοτίβα στα δεδομένα των επιμέρους clusters.

Κατηγορικές μεταβλητές

Για τις κατηγορικές μεταβλητές, υπολογίστηκαν οι συχνότητες κάθε κατηγορίας μέσα σε κάθε cluster αλλά και τα ποσοστά (σχετικές συχνότητες) κάθε κατηγορίας. Έπειτα δημιουργήθηκαν οπτικοποιήσεις, δηλαδή ραβδογράμματα, για να εντοπιστούν οι αναλογίες κάθε κατηγορίας για κάθε cluster. Αυτό έδωσε μια αίσθηση του πώς κατανέμονται οι κατηγορίες μέσα στις συστάδες και αν ορισμένες κατηγορίες υπερεκπροσωπούνται ή υποεκπροσωπούνται σε ορισμένες συστάδες για τον εντοπισμό των μοτίβων στα δεδομένα, και πώς σχετίζονται οι κατηγορίες μεταξύ τους.

Όπως αναφέρθηκε και προηγουμένως δημιουργήθηκαν 3 clusters, μεγέθους 56, 27 και 19 παρατηρήσεων. Παρακάτω ακολουθεί η περιγραφή των clusters και οι σχετικές γραφικές απεικονίσεις για τη διερεύνηση των σχέσεων και χαρακτηριστικών.

Cluster 1: Στην συστάδα 1 παρατηρείται ότι ο μέσος όρος διάρκειας των projects είναι περίπου 5 μήνες, οι ανθρωποώρες που χρειάστηκαν κυμαίνονται στις 82 ώρες, ενώ η αρχική εκτίμηση των ανθρωποωρών είναι 64 ώρες. Ο αριθμός των html templates είναι περίπου 10. Η εκτίμηση του κόστους ήταν περίπου στα 1157€ ενώ το πραγματικό κόστος κυμαίνεται στα 1597€. Το μέσο κέρδος αυτών των έργων είναι 1617€, το μέσο περιθώριο

κέρδους 43% ενώ η αναλογία κέρδους σε σχέση με το κόστος είναι 1.3 φορά πάνω. Το σταθμισμένο κόστος εργατοώρας είναι 19€. Τέλος τα μέσα έσοδα των έργων στην συστάδα αυτή είναι 3.214€

	Numeric_Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
cluster 1	Duration_months	0.00	2.00	3.00	5.18	6.00	38.00
	Spent_Time	9.60	54.90	73.50	82.25	96.00	360.00
	Baseline	144.00	830.25	1134.00	1157.11	1440.00	2196.00
	Cost	172.80	1046.52	1528.74	1597.42	1762.20	7776.00
	Profit	-5776.00	1000.20	1747.52	1616.78	2451.40	4180.00
	Profit_Margin	-2.89	0.46	0.58	0.43	0.65	0.89
	Profit_Ratio	-0.74	0.85	1.36	1.39	1.83	8.44
	Hourly_Cost	14.40	18.00	18.00	19.21	21.60	24.30
	Estimated_Effort_Hours	8.00	46.13	63.00	64.28	80.00	122.00
	Revenue	400.00	2306.25	3150.00	3214.20	4000.00	6100.00
	HTML_Templates	1.00	8.00	9.00	9.61	12.25	18.00

Πίν. 5. Συγκεντρωτικά Στατιστικά των Ποσοτικών Μεταβλητών του Cluster 1

Αναφορικά με τις ποιοτικές μεταβλητές, παρατηρείται ότι τα έργα αυτά έρχονται περισσότερο από μικρομεσαίες εταιρείες, ο τύπος κατηγορίας των έργων είναι «website», δηλαδή απλοί ιστότοποι παρουσίασης, τα δουλεύουν κυρίως μεσαίου επιπέδου προγραμματιστές, ενώ το cluster περιλαμβάνει και 4 έργα που έχουν φέρει ζημιά. Τέλος η πλειονότητα αυτών είναι χαμηλής πολυπλοκότητας έργα.

Cluster 2: Στην συστάδα 2 παρατηρείται ότι ο μέσος όρος διάρκειας των projects είναι περίπου 4.5 μήνες, οι ανθρωποώρες που χρειάστηκαν κυμαίνονται στις 197 ώρες, ενώ η αρχική εκτίμηση των ανθρωποωρών είναι 176 ώρες. Ο αριθμός των html templates των έργων είναι περίπου 16. Η εκτίμηση του κόστους ήταν περίπου στα 3170€ ενώ το πραγματικό κόστος κυμαίνεται στα 4000€. Το μέσο κέρδος αυτών των έργων είναι 4700€, το μέσο περιθώριο κέρδους 54% ενώ η αναλογία κέρδους σε σχέση με το κόστος είναι 2 φορές πάνω. Το σταθμισμένο κόστος εργατοώρας είναι 20€. Τέλος τα μέσα έσοδα των έργων στην συστάδα αυτή είναι 8.660€

	Numeric_Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
cluster 2	Duration_months	0.00	2.00	3.00	4.33	6.00	17.00
	Spent_Time	50.00	120.00	176.00	197.53	219.60	550.00
	Baseline	2160.00	2682.00	3204.00	3169.20	3535.20	4500.00
	Cost	720.00	2430.00	3366.00	3982.20	4885.92	11880.00
	Profit	-3680.00	4137.86	5070.00	4676.69	6132.00	8848.80
	Profit_Margin	-0.45	0.48	0.62	0.54	0.71	0.88
	Profit_Ratio	-0.31	0.94	1.65	2.04	2.50	7.33
	Hourly_Cost	14.40	18.00	21.60	19.82	21.60	24.30
	Estimated_Effort_Hours	120.00	149.00	178.00	176.07	196.40	250.00
	Revenue	6000.00	7450.00	8840.00	8658.89	9750.00	11700.00
	HTML_Templates	5.00	14.00	15.00	15.70	17.00	31.00

***Πίν. 6.** Συγκεντρωτικά Στατιστικά των Ποσοτικών Μεταβλητών του Cluster 2*

Αναφορικά με τις ποιοτικές μεταβλητές, παρατηρείται ότι τα έργα αυτά έρχονται περισσότερο από μεσαίες εταιρείες, ο τύπος κατηγορίας των έργων είναι «eshop» και «website» ισόποσα και τα δουλεύουν κυρίως ανώτερου επιπέδου προγραμματιστές (senior) και λιγότερο μεσαίου (mid). Το cluster περιλαμβάνει τα υπόλοιπα 2 έργα που έχουν φέρει ζημιά. Τέλος η πλειονότητα αυτών είναι χαμηλής και μεσαίας πολυπλοκότητας έργα ισόποσα.

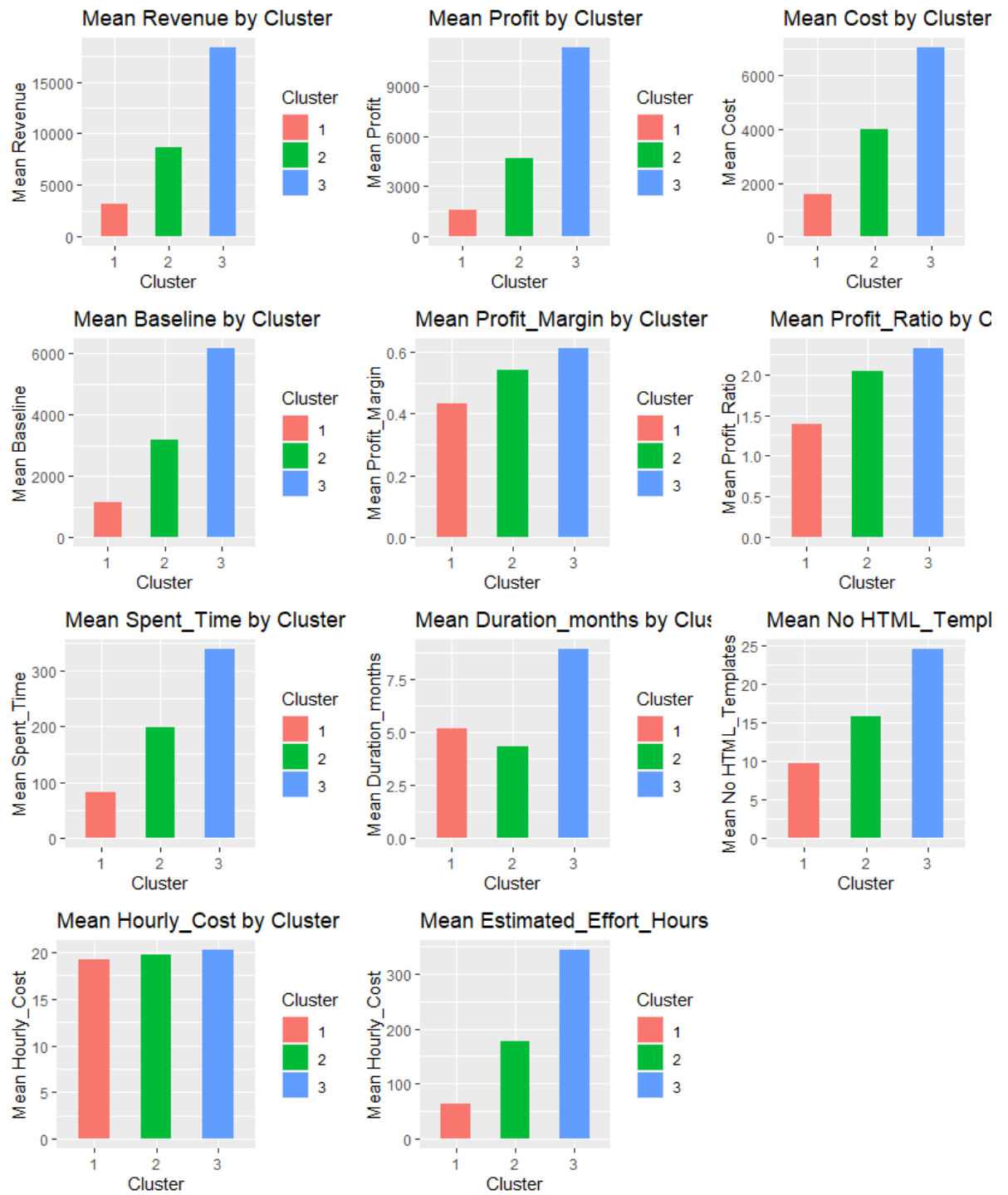
Cluster 3: Στην συστάδα 3 παρατηρείται ότι ο μέσος όρος διάρκειας των projects είναι περίπου 9 μήνες, οι ανθρωποώρες που χρειάστηκαν κυμαίνονται στις 338 ώρες, ενώ η αρχική εκτίμηση των ανθρωποωρών είναι 343 ώρες. Ο αριθμός των html templates των έργων είναι περίπου 25. Η εκτίμηση του κόστους ήταν περίπου στις 6178€ ενώ το πραγματικό κόστος κυμαίνεται στα 7053€. Το μέσος κέρδος αυτών των έργων είναι 11300€, το μέσο περιθώριο κέρδους 61% ενώ η αναλογία κέρδους σε σχέση με το κόστος είναι 2.3 φορές πάνω. Το σταθμισμένο κόστος εργατοώρας είναι 20.3€. Τέλος τα μέσα έσοδα των έργων στην συστάδα αυτή είναι 18.350€.

cluster 3	Numeric_Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	Duration_months	1.00	3.50	8.00	8.89	10.00	24.00
	Spent_Time	137.00	194.50	312.00	338.36	374.40	727.00
	Baseline	4392.00	5004.00	5832.00	6177.98	6714.00	10440.00
	Cost	2219.40	3851.10	6264.00	7052.51	8318.70	16524.00
	Profit	5305.40	7471.30	9428.00	11300.64	12547.10	30440.00
	Profit_Margin	0.27	0.46	0.62	0.61	0.78	0.86
	Profit_Ratio	0.37	0.86	1.63	2.32	3.45	5.98
	Hourly_Cost	16.20	18.00	19.80	20.27	22.28	24.30
	Estimated_Effort_Hours	244.00	278.00	324.00	343.22	373.00	580.00
	Revenue	12200.00	14050.00	16200.00	18353.16	19600.00	36650.00
	HTML_Templates	7.00	20.00	21.00	24.47	27.00	50.00

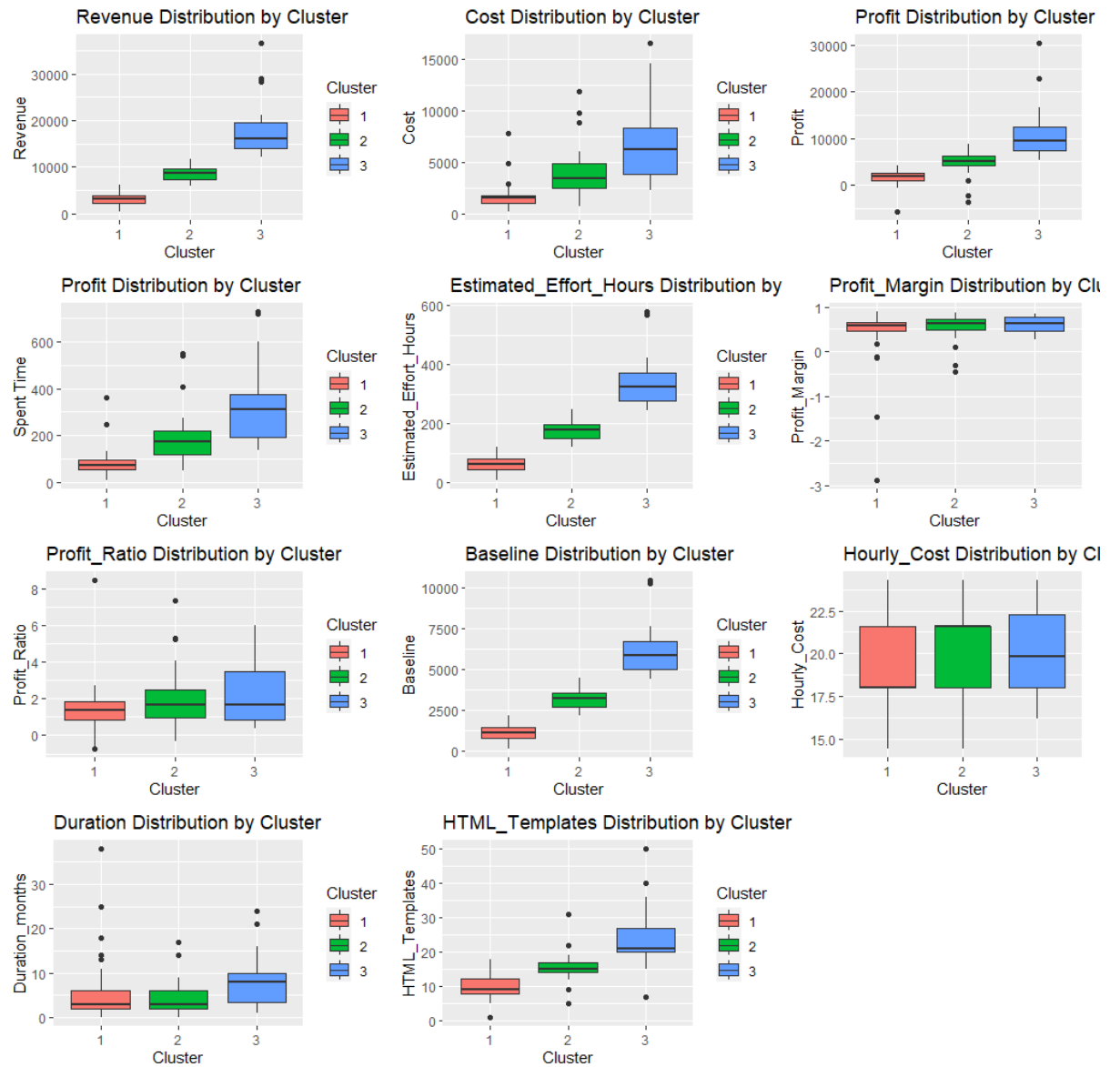
Πίν. 7. Συγκεντρωτικά Στατιστικά των Ποσοτικών Μεταβλητών του Cluster 3

Αναφορικά με τις ποιοτικές μεταβλητές, παρατηρείται ότι τα έργα αυτά έρχονται περισσότερο από μεγάλες εταιρείες, ο τύπος κατηγορίας των έργων είναι στην πλειονότητα «eshop» και λιγότερο «website», και τα δουλεύουν κυρίως ανώτερου επιπέδου προγραμματιστές (senior) και λιγότερο μεσαίου (mid). Τέλος η πλειονότητα αυτών είναι υψηλής πολυπλοκότητας έργα.

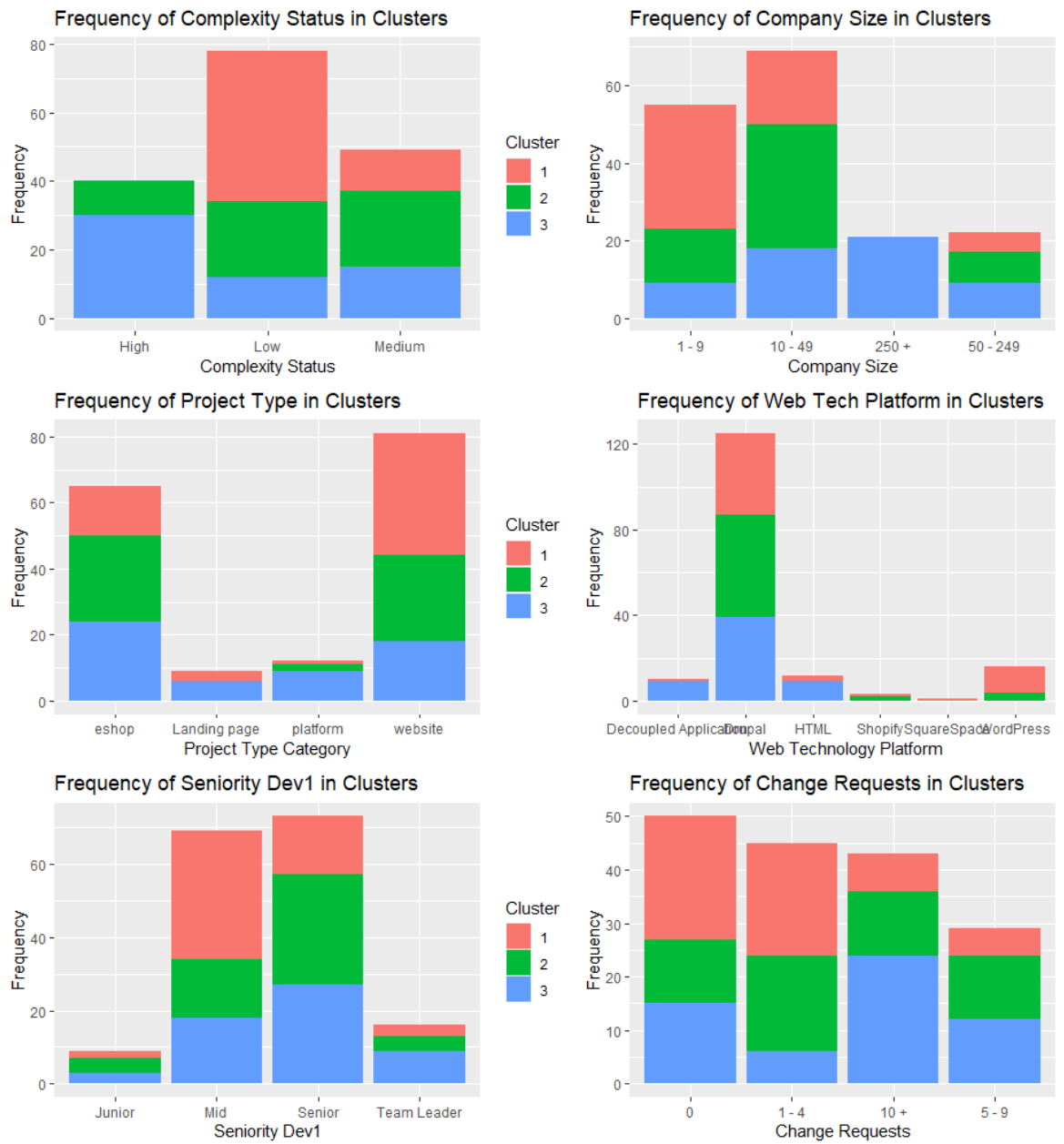
4.4.2 Οπτικοποίηση των Χαρακτηριστικών των Clusters



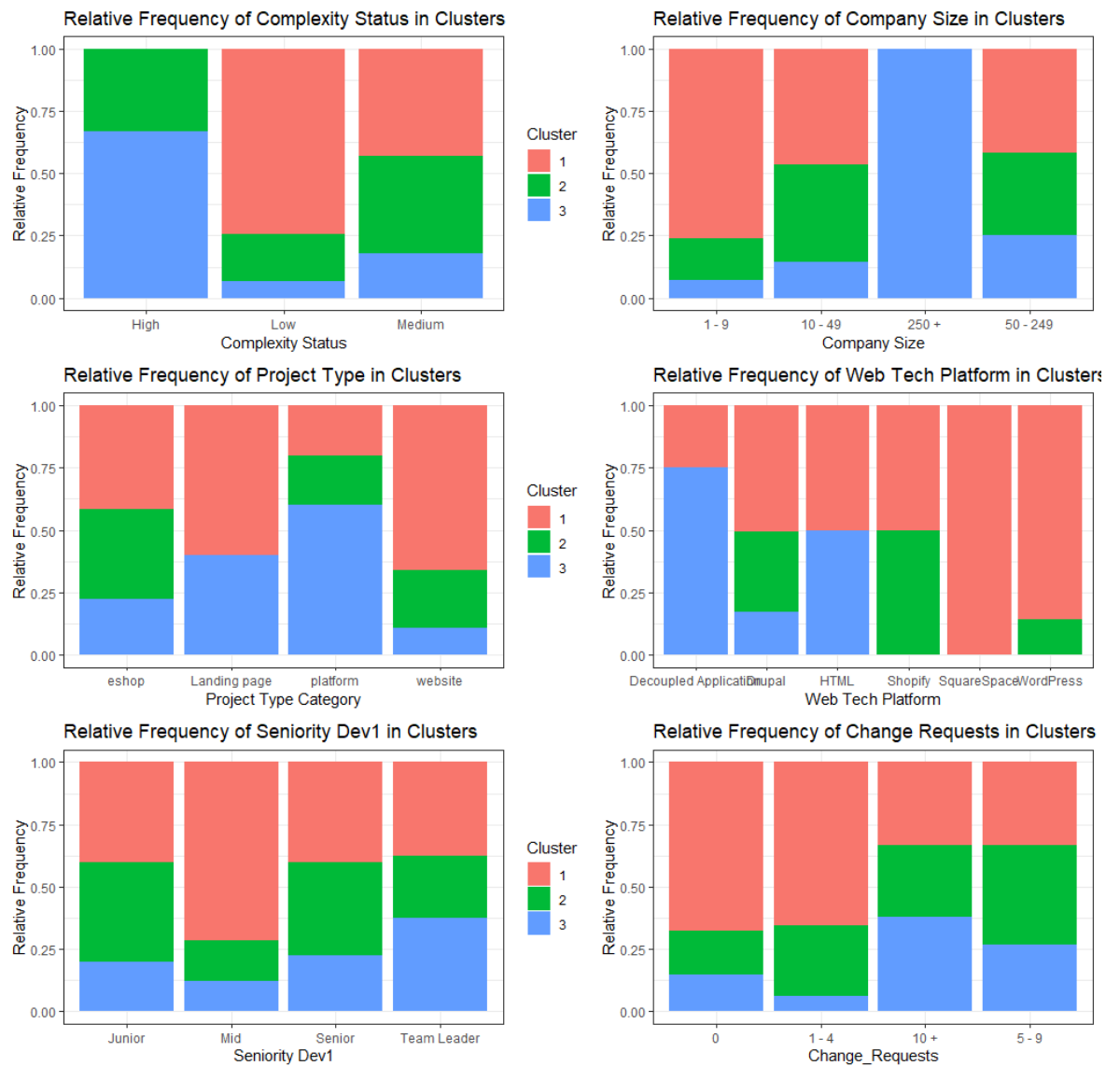
Διάγρ. 38. Μέσοι Όροι Ποσοτικών Μεταβλητών ανά Cluster



Διάγρ. 39. Κατανομή Ποσοτικών Μεταβλητών ανά Cluster



Διάγρ. 40. Συχνότητες Ποιοτικών Μεταβλητών ανά Cluster



Διάγρ. 41. Σχετικές Συχνότητες Ποιοτικών Μεταβλητών ανά Cluster

4.5 Συγκριτική Αξιολόγηση των Clusters

Είναι προφανές ότι σε σχέση με τον αντικειμενικό σκοπό της παρούσας έρευνας, τα 3 clusters, τμηματοποίησαν ικανοποιητικά τα έργα σε σχέση με την οικονομική αξία που δίνουν στην επιχείρηση και δίνουν σημαντικές πληροφορίες ως προς τα χαρακτηριστικά τους.

Το Cluster 1 συγκεντρώνει τα έργα με τα λιγότερα έσοδα και κέρδη σε απόλυτο μέγεθος. Αυτά τα έργα αντιπροσωπεύουν περίπου το 55% των έργων που αναλαμβάνει η εταιρεία, καταναλώνουν τις λιγότερες ανθρωποώρες (spent time) και είναι κυρίως χαμηλής πολυπλοκότητας, με τον μικρότερο αριθμό html templates και ο τύπος τους είναι απλά websites. Έρχονται από μικρού μεγέθους πελάτες.

Αντίθετα το Cluster 2 είναι το 26.4% των έργων και φαίνεται ότι είναι μεσαίας οικονομικής αξίας για την εταιρεία. Έχουν αυξημένα έσοδα και κέρδη σε σχέση με το Cluster 1 και είναι λίγο πιο σύνθετα έργα ως προς την υλοποίηση. Αξίζει να σημειωθεί ότι σε σχέση με το Cluster 1, στο Cluster 2, το πλήθος των απλών «websites» σαν τύπος έργων μειώνεται σημαντικά, ενώ περιλαμβάνονται ισόποσα και έργα eshop. Τα έργα στο cluster 3 έρχονται κυρίως από μεσαίου μεγέθους πελάτες.

Τέλος, το τελευταίο Cluster 3, είναι προφανές ότι έχει ομαδοποιήσει όλα τα έργα που φαίνεται να λειτουργούν σαν εξαίρεση στο πλήθος και περιλαμβάνει το 18.6% των έργων. Αφορά στα έργα με την μεγαλύτερη κερδοφορία και μάλιστα με διαφορά σε απόλυτα μεγέθη στα έσοδα και στα κέρδη. Τα έργα στο cluster 3 έχουν τη μεγαλύτερη διάρκεια υλοποίησης σε ανθρωποώρες. Επίσης, συγκριτικά με τα άλλα clusters, το cluster 3 έχει τα περισσότερα έργα υψηλής πολυπλοκότητας και ακόμη φαίνεται ότι τείνει να συγκεντρώνει τα έργα με τις περισσότερες αλλαγές απαιτήσεων (change requests) και τον μεγαλύτερο αριθμό html templates. Περιλαμβάνει ακόμη και άλλους τύπους έργων, όπως “landing page” και “platform”. Τέλος στο Cluster 3 φαίνεται ότι ανήκουν οι πελάτες μεγάλου μεγέθους.

5. ΣΥΜΠΕΡΑΣΜΑΤΑ & ΠΡΟΤΑΣΕΙΣ

Σύμφωνα με τον αρχικό σκοπό της παρούσας έρευνας, η συσταδοποίηση έδωσε πολύτιμες πληροφορίες σχετικά με την ποιότητα των έργων και με βάση αυτές τις πληροφορίες, μπορούν να δοθούν στρατηγικές κατευθύνσεις που αφορούν στη αποδοτική διαχείριση των έργων και στην κατάλληλη στόχευση έργων – πελατών. Αυτές οι κατευθύνσεις μάλιστα δεν αφορούν μόνο στην συγκεκριμένη εταιρία από την οποία ελήφθησαν τα δεδομένα, αλλά οποιαδήποτε εταιρία που δραστηριοποιείται στο χώρο του “web development” καθώς είναι περισσότερες προσεγγίσεις αποδοτικού και στρατηγικού “management”.

Παρατηρήθηκε ότι και στα 3 clusters η εκτίμηση του κόστους (Baseline) είναι μεν κοντά, αλλά παραμένει μικρότερη από το πραγματικό κόστος (Cost). Το γεγονός αυτό υποδηλώνει μια αδυναμία σε σχέση με τις αρχικές εκτιμήσεις και ενδεχομένως και με την εσωτερική διαχείριση των έργων. Ταυτόχρονα όμως δίνει και μια κατεύθυνση σχετικά με το εύρος τιμολόγησης των έργων. Δηλαδή, η εταιρεία θα πρέπει να τιμολογεί με τέτοια προσαύξηση επί του κόστους ή επί της τιμής ώστε να έχει ένα ασφαλές περιθώριο σφάλματος στις εκτιμήσεις και να αποφεύγει πιθανές αποκλίσεις που θα μειώσουν το περιθώριο κέρδους.

Αναφορικά με το μέσο ωριαίο κόστος εργατοώρας (Hourly Cost), σημειώνεται μια βαθμιαία αύξηση μεταξύ των clusters, αλλά όχι σημαντική. Δηλαδή, στο cluster 1, το μέσο κόστος εργατοώρας είναι το μικρότερο 19.2€, στο cluster 2 λίγο υψηλότερο 19.8€ και στο cluster 3 το υψηλότερο 20.2€. Τα έργα που είναι «μικρά» και χαμηλής πολυπλοκότητας, αναμένεται, να τα αναλαμβάνουν κυρίως «junior» προγραμματιστές και επόμενος το ωριαίο κόστος αυτών θα είναι μικρό, ώστε η εταιρία να έχει μεγαλύτερο περιθώριο κέρδους. Αντίστοιχα τα «μεγάλα» έργα που είναι πολύπλοκα, απαιτούν τη μεγάλη εμπειρία των προγραμματιστών (Senior ή Team leaders) και επομένως αναμένεται το ωριαίο κόστος μεγαλύτερο. Συνεπώς η ανάθεση των έργων στους προγραμματιστές, θα πρέπει να γίνεται έτσι ώστε όπου είναι εφικτό να μειώνεται το ωριαίο κόστος για να αυξάνεται η απόδοση του έργου σε οικονομική αξία. Το μέσο ωριαίο κόστος θα πρέπει να είναι ενδεικτικό της αποδοτικής κατανομής των πόρων, γεγονός που μάλλον δεν αναδεικνύεται στην παρούσα περίπτωση.

Σε συνέχεια του παραπάνω, αν η εταιρία έχει μεγάλη ποσότητα σε «μικρά» έργα, που φέρουν λίγα σχετικά έσοδα και κέρδη στην εταιρία, θα πρέπει να είναι ανάλογη και η

σύνθεση της ομάδας των προγραμματιστών σε ποσότητα και επίπεδο εμπειρίας. Δηλαδή, η ομάδα θα πρέπει να αποτελείται από πολλούς «junior» προγραμματιστές που θα αναλαμβάνουν αυτά τα έργα με χαμηλό κόστος για την επιχείρηση. Με βάση τα χαρακτηριστικά των clusters, αυτό δεν φαίνεται να συμβαίνει.

Σχετικά με την καθαρή επιστροφή της επένδυσης των έργων (Profit ratio), τα έργα που ανήκουν στο cluster 1, έχουν την μικρότερη αναλογία του κέρδους προς το κόστος 1,38. Στο cluster 2 αυτή η αναλογία έρχεται στο 2, και στο cluster 3 στο 2.3. Δεδομένου ότι το cluster 1, αφορά σε «μικρά» έργα και χαμηλών εσόδων σε απόλυτα μεγέθη, τα έργα αυτά θα πρέπει να έχουν μεγαλύτερο ratio, για να έχει νόημα για την εταιρία να τα αναλαμβάνει. Αντίστοιχα για τα «μεγαλύτερα» έργα που φέρουν υψηλά έσοδα σε απόλυτα μεγέθη, αναμένεται μάλλον μικρότερο ratio. Επομένως και αυτή η μεταβλητή είναι ενδεικτική της αποδοτικότητας των έργων, και θα πρέπει να δίνει κατευθύνσεις στην αποδοτική διαχείριση πόρων και τιμολόγηση των έργων.

Επιπλέον, η παραπάνω ανάλυση των clusters υποδεικνύει ότι οι μεγάλοι μεγέθους πελάτες φέρουν «μεγάλα» έργα και αυτά με τη σειρά τους είναι που συνεισφέρουν στο μεγαλύτερο βαθμό στην αύξηση της κερδοφορίας της εταιρίας, παρόλο που, όπως είναι αναμενόμενο, έχουν και το μεγαλύτερο κόστος αλλά και ρίσκο. Τα μεγάλα και σύνθετα έργα, εξαιτίας της πολυπλοκότητας τους περιέχουν και απρόβλεπτους παράγοντες που αυξάνουν τον κίνδυνο αποτυχίας και ζημίας. Αυτό δηλώνεται και από τον αυξημένο αριθμό των αλλαγών των απαιτήσεων (change requests) στο cluster 3. Στην συγκεκριμένη περίπτωση φαίνεται ότι αυτά τα έργα, παρά το ρίσκο, καταφέρνουν να γίνονται τα πιο κερδοφόρα, με το μεγαλύτερο profit ratio, όπως είδαμε παραπάνω. Αυτό ερμηνεύεται πολύ πιθανόν από την εκμετάλλευση της οικονομικής ευκαιρίας που προκύπτει από τα έργα αυτά. Δεδομένου ότι η τιμολόγηση γίνεται εξ' αρχής σε μεγάλα ποσά, η εταιρία βρίσκει έδαφος να τιμολογήσει με μεγάλη προσαύξηση που εύκολα δικαιολογείται από το μέγεθος και την πολυπλοκότητα του έργου, σε συνδυασμό με το μεγάλο μέγεθος του πελάτη, χαρακτηριστικό που υποδηλώνει την οικονομική του άνεση.

Αναφορικά με τον τύπο project, μπορούμε να συμπεράνουμε με βάση τη συσταδοποίηση, ότι τα πιο κερδοφόρα έργα τείνουν να είναι τύπου «platform» και «eshop» που με τη σειρά τους είναι τα έργα με την υψηλότερη πολυπλοκότητα, δηλαδή «μεγάλα» έργα.

Παρατηρείται ότι η πλειοψηφία των έργων που αναλαμβάνει η εταιρία είναι «μικρά» έργα, τα οποία όμως ισορροπούν μεταξύ κόστους και εσόδων και επομένως, δεν

αποδίδουν κέρδος ικανοποιητικό στην εταιρία. Δεδομένου όμως ότι αντιπροσωπεύουν μεγαλύτερη ποσότητα του συνόλου των έργων, η εταιρία καλείται να απασχολεί μεγάλο μέρος των πόρων της για αυτά τα έργα, χωρίς όμως να υπάρχει βιώσιμη επιστροφή κέρδους που θα συνεισφέρει στην ανάπτυξη της. Με δεδομένη αυτή την κατάσταση, η εταιρία έχει δύο επιλογές, είτε να μειώσει το κόστος, εάν είναι δυνατό, διαμορφώνοντας αναλόγως και τη σύνθεση της ομάδας προγραμματιστών, είτε να διαφοροποιήσει την τιμολόγηση της, ή ακόμη και τα δύο, ώστε να είναι ικανή να πετύχει μία μεγαλύτερη καθαρή επιστροφή της επένδυσης.

Διαφορετικά η εταιρία, έχει την επιλογή να αλλάξει τη στρατηγική της, και να στοχεύσει στην απόκτηση «μεγάλων» πελατών. Σε αυτή την περίπτωση ο στόχος της είναι να αναλαμβάνει «μεγάλα» και σύνθετα έργα αλλά σε μικρότερη ποσότητα, τα οποία έχουν μεγάλη κερδοφορία και μεγάλη καθαρή επιστροφή της επένδυσης.

Σε κάθε περίπτωση, αναδεικνύονται δύο διαφορετικές στρατηγικές κατευθύνσεις μεταξύ των οποίων μπορεί να επιλέξει η εταιρία. Οι δύο βασικές στρατηγικές είναι η στόχευση σε μεγάλα και πολύπλοκα έργα και η στόχευση σε μεγάλο αριθμό μικρών και χαμηλής πολυπλοκότητας έργων. Και οι δύο, έχουν τα πλεονεκτήματα, τα μειονεκτήματα και τους περιορισμούς τους που θα πρέπει να λάβει υπόψιν της η διοίκηση μιας εταιρίας. Η υιοθέτηση της κάθε μιας απαιτεί συγκεκριμένες τακτικές κινήσεις, προϋποθέσεις και κατάλληλες συνθήκες εσωτερικές και εξωτερικές για να ευδοκιμήσουν και εξασφαλίσουν την υψηλή κερδοφορία στην επιχείρηση.

5.1 Στρατηγική Α: Στοχεύοντας σε μεγάλα & πολύπλοκα έργα

Η στόχευση σε μεγάλα και πολύπλοκα έργα περιλαμβάνει την αναζήτηση έργων υψηλής αξίας που απαιτούν σημαντική επένδυση πόρων και εμπειρογνομosύνης. Τέτοια έργα περιλαμβάνουν συνήθως πολύπλοκα συστήματα και αρχιτεκτονικές, προηγμένες τεχνολογίες και έμπειρες ομάδες προγραμματιστών. Οι στρατηγικές τακτικές για τη στόχευση μεγάλων και πολύπλοκων έργων υπαγορεύουν:

Δημιουργία ισχυρής φήμης: Για να κερδίσει μια εταιρεία μεγάλα και πολύπλοκα έργα, πρέπει να έχει ισχυρή φήμη στην αγορά. Αυτό περιλαμβάνει την ύπαρξη ιστορικού επιτυχούς ολοκλήρωσης τέτοιων έργων και την ύπαρξη ομάδας έμπειρων

προγραμματιστών και αρχιτεκτόνων που μπορούν να χειριστούν την πολυπλοκότητα του έργου.

Ανάπτυξη εξειδικευμένων δεξιοτήτων: Τα μεγάλα και πολύπλοκα έργα απαιτούν εξειδικευμένες δεξιότητες και τεχνογνωσία. Μια εταιρεία πρέπει να επενδύσει στην ανάπτυξη τέτοιων δεξιοτήτων προσλαμβάνοντας έμπειρους προγραμματιστές ή εκπαιδεύοντας το υπάρχον προσωπικό.

Συνεργασία με εταίρους: Η συνεργασία με συνεργάτες, όπως προμηθευτές τεχνολογίας και συμβούλους, μπορεί να βοηθήσει μια εταιρεία να αποκτήσει πρόσβαση στην απαιτούμενη τεχνογνωσία και τους πόρους για μεγάλα και πολύπλοκα έργα.

Δημιουργία μιας ισχυρής διαδικασίας διαχείρισης έργων: Οι εταιρείες πρέπει να έχουν σαφή κατανόηση των απαιτήσεων του έργου, των χρονοδιαγραμμάτων και της κατανομής των πόρων και πρέπει να διαθέτουν ισχυρά εργαλεία και μεθοδολογίες διαχείρισης έργων που εξασφαλίζουν την έγκαιρη παράδοση, την ποιότητα και τον έλεγχο του κόστους.

Ταυτόχρονα στο μακρο-περιβάλλον, οι εταιρείες που στοχεύουν σε μεγάλα και πολύπλοκα έργα πρέπει να λαμβάνουν υπόψη παράγοντες όπως το επίπεδο του ανταγωνισμού, η ζήτηση της αγοράς και οι αναδυόμενες τεχνολογίες. Στο μικρο-περιβάλλον, πρέπει να εξετάζουν τη διαθεσιμότητα εξειδικευμένων πόρων, το επίπεδο κινδύνου του έργου και τις στρατηγικές τιμολόγησης των ανταγωνιστών.

Τέλος οι περιορισμοί της αγοράς, συμπεριλαμβανομένης της τιμολόγησης, είναι επίσης σημαντικοί παράγοντες που πρέπει να λαμβάνονται υπόψη κατά τη στόχευση μεγάλων και σύνθετων έργων. Οι εταιρίες πρέπει να είναι σε θέση να παραδώσουν το έργο κερδοφόρα και ταυτόχρονα να παραμείνουν ανταγωνιστικές στην αγορά. Η υψηλή τιμολόγηση μπορεί να οδηγήσει σε χαμηλότερη ζήτηση, ενώ η χαμηλή τιμολόγηση μπορεί να οδηγήσει σε χαμηλότερα περιθώρια κέρδους.

Η στόχευση σε μεγάλα και πολύπλοκα έργα έχει διάφορα πλεονεκτήματα, όπως:

Υψηλό δυναμικό εσόδων: Λόγω της υψηλότερης τιμολόγησης και του μεγαλύτερου πεδίου εφαρμογής τους, τα μεγάλα και πολύπλοκα έργα μπορούν να αποφέρουν υψηλά έσοδα, που μπορούν να οδηγήσουν και σε υψηλότερη κερδοφορία.

Δημιουργία φήμης μακροπρόθεσμα: Η επιτυχής ολοκλήρωση μεγάλων και σύνθετων έργων μπορεί να ενισχύσει τη φήμη και την αξιοπιστία μιας εταιρίας και να προσελκύσει νέους πελάτες.

Τα μειονεκτήματα της στόχευσης σε μεγάλα και πολύπλοκα έργα είναι:

Υψηλός κίνδυνος: Τα μεγάλα και πολύπλοκα έργα είναι έργα υψηλού κινδύνου που μπορεί να οδηγήσουν σε σημαντική οικονομική ζημία και ζημία στη φήμη, εάν δεν παραδοθούν με επιτυχία.

Υψηλές απαιτήσεις σε πόρους: Τα μεγάλα και πολύπλοκα έργα απαιτούν σημαντική επένδυση πόρων, η οποία μπορεί να επιβαρύνει τα οικονομικά μιας εταιρίας.

Μεγάλοι κύκλοι έργων: Τα μεγάλα και πολύπλοκα έργα μπορεί να έχουν μεγάλους κύκλους έργων, γεγονός που μπορεί να επηρεάσει την ικανότητα μιας εταιρείας να παράγει έσοδα βραχυπρόθεσμα.

5.2 Στρατηγική B: Στοχεύοντας σε μεγάλο αριθμό μικρών & χαμηλής πολυπλοκότητας έργων

Μία άλλη στρατηγική για τις εταιρίες στον κλάδο της ανάπτυξης λογισμικού είναι να στοχεύουν σε μεγάλο αριθμό μικρών και χαμηλής πολυπλοκότητας έργων. Αυτά τα έργα χαρακτηρίζονται από μικρότερο πεδίο εφαρμογής, χαμηλότερα επίπεδα πολυπλοκότητας, την ανάγκη για λιγότερους πόρους για την ολοκλήρωσή τους με μικρότερο προϋπολογισμό και μικρότερα χρονοδιαγράμματα. Η στόχευση μεγάλου αριθμού μικρών και χαμηλής πολυπλοκότητας έργων είναι μια στρατηγική που επικεντρώνεται στον όγκο και όχι στο μέγεθος, σε αντίθεση με την ανάληψη λιγότερων, αλλά μεγαλύτερων, πολύπλοκων έργων. Η προσέγγιση αυτή περιλαμβάνει την ανάληψη πολλών μικρότερων έργων ταυτόχρονα, παρέχοντας μια σταθερή ροή εσόδων για την εταιρία. Παρόλο που αυτά τα έργα μπορεί να μην είναι τόσο σημαντικού προφίλ όσο τα μεγάλα και πολύπλοκα έργα, μπορούν να αποτελέσουν μια αξιόπιστη πηγή εσόδων για τις εταιρείες ανάπτυξης έργων λογισμικού.

Ωστόσο, για να ακολουθήσει μια εταιρία αυτή τη στρατηγική, υπάρχουν αρκετές προϋποθέσεις που πρέπει να πληροί:

Αποτελεσματικές διαδικασίες: Οι επιχειρήσεις πρέπει να διαθέτουν αποτελεσματικές διαδικασίες για τη διαχείριση μεγάλου όγκου έργων ταυτόχρονα.

Επεκτάσιμη υποδομή και πόροι: Η επιχείρηση πρέπει να διαθέτει κλιμακούμενη υποδομή που να μπορεί να διαχειριστεί την αύξηση του όγκου των έργων. Για την ανάληψη μεγάλου αριθμού μικρών και χαμηλής πολυπλοκότητας έργων, μια εταιρεία πρέπει να διαθέτει κλιμακούμενο εργατικό δυναμικό, υποδομές και τεχνολογία. Η εταιρεία πρέπει να είναι σε θέση να διαχειρίζεται τον αυξημένο όγκο εργασιών χωρίς να υποβαθμίζει την ποιότητα.

Διαφοροποίηση των προσφερόμενων υπηρεσιών: Για να προσελκύσουν ένα ευρύ φάσμα πελατών, οι επιχειρήσεις θα πρέπει να διαφοροποιήσουν τις προσφορές υπηρεσιών τους, ώστε να περιλαμβάνουν μια ποικιλία διαφορετικών τύπων έργων. Αυτό μπορεί να περιλαμβάνει την ανάπτυξη ιστοσελίδων, την ανάπτυξη εφαρμογών για κινητά, την ανάπτυξη προσαρμοσμένου λογισμικού και πολλά άλλα.

Βελτιστοποίηση των διαδικασιών: Δεδομένου ότι ο στόχος είναι η διαχείριση μεγάλου όγκου έργων, είναι σημαντικό να εξορθολογιστούν οι διαδικασίες για να αυξηθεί η αποδοτικότητα και να μειωθεί ο χρόνος που απαιτείται για την ολοκλήρωση κάθε έργου.

Αξιοποίηση της τεχνολογίας: Η αξιοποίηση της τεχνολογίας μπορεί να βοηθήσει στην αυτοματοποίηση επαναλαμβανόμενων εργασιών και στη μείωση του χρόνου που απαιτείται για την ολοκλήρωση κάθε έργου. Μπορεί επίσης να βοηθήσει τις επιχειρήσεις να διαχειριστούν ταυτόχρονα μεγάλο αριθμό έργων.

Ένας πιθανός περιορισμός αυτής της προσέγγισης είναι ότι η αγορά για μικρά και χαμηλής πολυπλοκότητας έργα μπορεί να είναι ιδιαίτερα ανταγωνιστική, γεγονός που μπορεί να καταστήσει δύσκολο να ξεχωρίσει κανείς από τους ανταγωνιστές. Μια άλλη ανησυχία είναι το ενδεχόμενο χαμηλών περιθωρίων κέρδους, καθώς τα μικρότερα έργα μπορεί να μην παρέχουν το ίδιο επίπεδο εσόδων με τα μεγαλύτερα, πιο πολύπλοκα έργα.

Ένα πλεονέκτημα αυτής της προσέγγισης είναι ότι επιτρέπει στις επιχειρήσεις να προσελκύσουν μεγάλο αριθμό πελατών, γεγονός που μπορεί να συμβάλει στη δημιουργία μιας σταθερής ροής εσόδων. Η προσέγγιση αυτή προσφέρει επίσης τη δυνατότητα ανάπτυξης, καθώς οι επιχειρήσεις μπορούν να κλιμακώσουν τις δραστηριότητές τους καθώς αναλαμβάνουν περισσότερους πελάτες και έργα.

Ωστόσο, υπάρχουν και ορισμένα μειονεκτήματα αυτής της προσέγγισης. Μια πρόκληση είναι η ταυτόχρονη διαχείριση μεγάλου αριθμού έργων, η οποία μπορεί να είναι συντριπτική και να απαιτεί σημαντικό αριθμό πόρων. Υπάρχει επίσης ο κίνδυνος να

γενικευτούν υπερβολικά οι προσφερόμενες υπηρεσίες, γεγονός που μπορεί να δυσχεράνει τη διαφοροποίηση από τους ανταγωνιστές και να μειώσει την κερδοφορία.

Συμπερασματικά, η στόχευση σε μεγάλο αριθμό μικρών και χαμηλής πολυπλοκότητας έργων αποτελεί βιώσιμη στρατηγική στον κλάδο της ανάπτυξης λογισμικού αλλά με προϋποθέσεις. Οι επιχειρήσεις μπορούν να προσελκύσουν ένα ευρύ φάσμα πελατών, να δημιουργήσουν σταθερά έσοδα και να κλιμακώσουν τις δραστηριότητές τους καθώς αναλαμβάνουν περισσότερα έργα. Ωστόσο, η προσέγγιση αυτή συνοδεύεται επίσης από τις προκλήσεις της, όπως η διαχείριση μεγάλου όγκου έργων ταυτόχρονα και το ενδεχόμενο χαμηλότερων περιθωρίων κέρδους.

5.3 Η Επιλογή Στρατηγικής

Όταν πρόκειται να αποφασίσει μια εταιρία μεταξύ της στόχευσης σε μεγάλα και πολύπλοκα έργα έναντι ενός μεγάλου αριθμού μικρών και χαμηλής πολυπλοκότητας έργων στη βιομηχανία ανάπτυξης λογισμικού, υπάρχουν διάφοροι περιορισμοί που πρέπει να ληφθούν υπόψη. Αυτοί οι περιορισμοί περιλαμβάνουν:

Περιορισμοί πόρων: Τα μεγάλα και πολύπλοκα έργα συχνά απαιτούν υψηλότερη επένδυση πόρων, συμπεριλαμβανομένου του χρόνου, των χρημάτων και του ανθρώπινου δυναμικού. Εάν μια εταιρεία δεν διαθέτει επαρκείς πόρους για να διαθέσει προς αυτούς τους τύπους έργων, ενδέχεται να μην είναι εφικτή η εφαρμογή αυτής της στρατηγικής. Από την άλλη πλευρά, μια εταιρεία που στοχεύει σε μεγάλο αριθμό μικρών και χαμηλής πολυπλοκότητας έργων μπορεί να απαιτεί λιγότερες επενδύσεις, αλλά μπορεί να αντιμετωπίσει προκλήσεις στην κατανομή και διαχείριση των πόρων σε πολλά έργα ταυτόχρονα.

Τεχνική εμπειρογνωμοσύνη: Τα μεγάλα και πολύπλοκα έργα απαιτούν συχνά εξειδικευμένη τεχνική εμπειρογνωμοσύνη, όπως στους τομείς της τεχνητής νοημοσύνης, της μηχανικής μάθησης ή της αλυσίδας μπλοκ. Μια εταιρεία πρέπει να διαθέτει την απαραίτητη τεχνική εμπειρογνωμοσύνη για να χειρίζεται αποτελεσματικά τέτοια έργα. Από την άλλη πλευρά, τα μικρά και χαμηλής πολυπλοκότητας έργα μπορεί να μην απαιτούν τέτοια εξειδικευμένη εμπειρογνωμοσύνη, καθιστώντας ευκολότερο για μια εταιρεία να βρει και να διατηρήσει το απαραίτητο τεχνικό ταλέντο.

Προτιμήσεις πελατών: Οι πελάτες μπορεί να έχουν διαφορετικές προτιμήσεις για τον τύπο των έργων που θέλουν να εκτελέσουν. Ορισμένοι πελάτες μπορεί να προτιμούν να

συνεργάζονται με εταιρείες που έχουν αποδεδειγμένο ιστορικό στο χειρισμό μεγάλων και πολύπλοκων έργων, ενώ άλλοι μπορεί να προτιμούν να συνεργάζονται με εταιρείες που μπορούν να χειριστούν πολλά μικρά έργα ταυτόχρονα. Είναι σημαντικό για μια εταιρεία να κατανοεί τις προτιμήσεις των πελατών της και να προσαρμόζει τη στρατηγική της ανάλογα.

Διαχείριση κινδύνων: Τα μεγάλα και πολύπλοκα έργα συχνά ενέχουν υψηλότερο βαθμό κινδύνου, όπως τεχνικές προκλήσεις, αλλαγές στις απαιτήσεις ή απρόβλεπτα ζητήματα. Μια εταιρεία πρέπει να διαθέτει ένα ισχυρό σχέδιο διαχείρισης κινδύνων για τον εντοπισμό, τον μετριασμό και την αποτελεσματική διαχείριση των κινδύνων του έργου. Από την άλλη πλευρά, τα μικρά και χαμηλής πολυπλοκότητας έργα μπορεί να είναι λιγότερο επικίνδυνα, αλλά μπορεί επίσης να απαιτούν μια διαφορετική προσέγγιση διαχείρισης κινδύνων.

Κερδοφορία: Και οι δύο στρατηγικές έχουν διαφορετικό προφίλ κερδοφορίας. Τα μεγάλα και πολύπλοκα έργα μπορούν να προσφέρουν υψηλότερα περιθώρια κέρδους, αλλά μπορεί επίσης να συνεπάγονται μεγαλύτερο κύκλο πωλήσεων και υψηλότερο βαθμό ανταγωνισμού. Τα μικρά και χαμηλής πολυπλοκότητας έργα μπορεί να προσφέρουν χαμηλότερα περιθώρια κέρδους, αλλά μπορεί επίσης να είναι ευκολότερο να πωληθούν και να αποφέρουν έσοδα πιο γρήγορα.

Εν κατακλείδι, η απόφαση για τη στόχευση μεγάλων και πολύπλοκων έργων έναντι ενός μεγάλου αριθμού μικρών και χαμηλής πολυπλοκότητας έργων θα πρέπει να βασίζεται σε μια προσεκτική αξιολόγηση των πόρων της εταιρείας, της τεχνικής εμπειρογνομosύνης, των προτιμήσεων των πελατών, των δυνατοτήτων διαχείρισης κινδύνων και των στόχων κερδοφορίας. Με την κατανόηση αυτών των περιορισμών, μια εταιρεία μπορεί να λάβει μια τεκμηριωμένη απόφαση και να υιοθετήσει μια στρατηγική προσέγγιση που ευθυγραμμίζεται με τους μακροπρόθεσμους στόχους της.

5.4 Επίλογος

Ο σκοπός της μελέτης αυτής ήταν να εξερευνήσει τη σχέση μεταξύ της υψηλής κερδοφορίας των έργων και άλλων διαθέσιμων ποιοτικών χαρακτηριστικών τους, ώστε μέσα από τη τμηματοποίηση των έργων να δώσει κατευθύνσεις στην επιχείρηση για το ποια έργα έχει νόημα να αναλαμβάνει. Τα ευρήματα της παρούσας μελέτης έδειξαν ότι

πράγματι τα έργα ομαδοποιήθηκαν σε ομάδες ανάλογα τον βαθμό κερδοφορίας και τα χαρακτηριστικά των επιμέρους ομάδων, δίνουν πολύτιμες πληροφορίες για το είδος των έργων σε κάθε ομάδα. Η στρατηγική τμηματοποίηση των έργων ανάλογα την κερδοφορία τους και η μετέπειτα στόχευση τους, δίνει στις εταιρίες μακροπρόθεσμη αξία, βιωσιμότητα και ανάπτυξη και προφανώς επιχειρεί να υπερκαλύψει τις προκλήσεις και τις δυσκολίες στην επιμέρους διαχείριση τους βραχυπρόθεσμα. Ως εκ τούτου προτείνεται μία εναλλακτική προσέγγιση στο «πρόβλημα» της εκτίμησης των ανθρωποωρών σε επίπεδο project καθώς και στις δύο περιπτώσεις τελικός σκοπός είναι να προκύπτει το ζητούμενο κέρδος.

Η βιομηχανία ανάπτυξης λογισμικού είναι ένας ταχέως αναπτυσσόμενος και ανταγωνιστικός τομέας, με διάφορους τύπους έργων που απαιτούν διαφορετικά επίπεδα πολυπλοκότητας, πόρων και τιμών. Οι εταιρίες πρέπει να λαμβάνουν στρατηγικές αποφάσεις σχετικά με τους τύπους έργων που θέλουν να στοχεύσουν για να παραμείνουν ανταγωνιστικές και κερδοφόρες. Και για να το πετύχουν αυτό, σε αυτόν τον δυναμικό κλάδο, οι εταιρίες πρέπει να υιοθετήσουν μια στρατηγική προσέγγιση που να ευθυγραμμίζεται με τους μακροπρόθεσμους στόχους τους και να εξασφαλίζει την κερδοφορία τους.

5.5 Προτάσεις για έρευνα

Η παρούσα μελέτη μπορεί να βελτιωθεί με τον εμπλουτισμό των μεταβλητών των έργων και με μεγαλύτερο όγκο δεδομένων που ενδεχομένως να δώσουν νέα «σήματα» στον αλγόριθμο και να προκύψουν νέες πληροφορίες που αξίζει να ληφθούν υπόψη από τις επιχειρήσεις. Επιπλέον η έρευνα είναι δυνατόν να συνεχιστεί με μοντέλα πρόβλεψης ταξινόμησης (classification). Δεδομένου ότι μετά την ανάλυση συστάδων μπορούν να δοθούν ετικέτες στις παρατηρήσεις ανάλογα το cluster που ανήκουν, όπως στη συγκεκριμένη περίπτωση «χαμηλής», «μεσαίας» και «υψηλής» κερδοφορίας, το επόμενο βήμα είναι η πρόβλεψη της ετικέτας ενός νέου έργου που θα εισαχθεί στα δεδομένα. Ο στόχος της ταξινόμησης είναι η δημιουργία ενός μοντέλου που μπορεί να προβλέψει με ακρίβεια την ετικέτα κλάσης νέων, αθέατων περιπτώσεων με βάση τα χαρακτηριστικά τους.

Εφόσον η έρευνα επεκταθεί προς την μέθοδο της ταξινόμησης, πρέπει να ληφθούν υπόψη διάφορα πράγματα:

Επιλογή χαρακτηριστικών: Πριν από την έναρξη της ταξινόμησης, είναι σημαντικό να επιλεγούν οι πιο σημαντικές και σχετικές μεταβλητές για την εργασία ταξινόμησης. Αυτό μπορεί να γίνει με την ανάλυση των συστάδων και τον εντοπισμό των χαρακτηριστικών που τις διαφοροποιούν. Μπορούν επίσης να χρησιμοποιηθούν και άλλες μέθοδοι, είτε ανεξάρτητα από τον αλγόριθμο ταξινόμησης, όπως η απλή ανάλυση συσχέτισης ή η ανάλυση κύριων συνιστωσών (Principal Component Analysis), είτε χρησιμοποιώντας τον αλγόριθμο για να βρουν το βέλτιστο σύνολο χαρακτηριστικών που μεγιστοποιεί την απόδοση ταξινόμησης (forward selection, backward elimination, recursive feature elimination).

Επιλογή αλγορίθμου ταξινόμησης: Η επιλογή του αλγορίθμου ταξινόμησης εξαρτάται από τη φύση των δεδομένων και το συγκεκριμένο πρόβλημα. Ορισμένοι δημοφιλείς αλγόριθμοι ταξινόμησης περιλαμβάνουν τη λογιστική παλινδρόμηση, τα δέντρα αποφάσεων, τα τυχαία δάση (Random Forests), τις μηχανές διανυσμάτων υποστήριξης (Support Vector Machines), τα νευρωνικά δίκτυα (Neural Networks), ο αλγόριθμος K-Nearest Neighbors (KNN) και ο αλγόριθμος Naive Bayes.

Εκπαίδευση και δοκιμή του μοντέλου: Αφού προσδιοριστούν τα σχετικά χαρακτηριστικά και ο αλγόριθμος ταξινόμησης, είναι σημαντικό να εκπαιδευτεί το μοντέλο σε ένα σύνολο δεδομένων με ετικέτες. Το σύνολο δεδομένων θα πρέπει να χωριστεί σε σύνολα εκπαίδευσης και δοκιμής για την αξιολόγηση της απόδοσης του μοντέλου. Το μοντέλο θα πρέπει να βελτιστοποιηθεί με τη ρύθμιση των παραμέτρων του για τη βελτίωση της ακρίβειάς του.

Αξιολόγηση της απόδοσης του μοντέλου: Η απόδοση του μοντέλου ταξινόμησης θα πρέπει να αξιολογείται χρησιμοποιώντας μετρικές όπως η ακρίβεια, η ανάκληση και το F1-score. Αυτές οι μετρικές βοηθούν στην αξιολόγηση της απόδοσης του μοντέλου και στη σύγκρισή του με άλλα μοντέλα.

Γενίκευση: Το μοντέλο θα πρέπει να δοκιμαστεί σε ένα νέο σύνολο δεδομένων για να διασφαλιστεί ότι γενικεύεται καλά σε αθέατα δεδομένα. Αυτό βοηθά στην επαλήθευση της αξιοπιστίας του μοντέλου σε πραγματικές εφαρμογές.

Συνολικά, η προσεκτική εξέταση των παραπάνω παραγόντων θα βοηθήσει στη δημιουργία ενός ακριβούς και αξιόπιστου μοντέλου ταξινόμησης.

6. ΒΙΒΛΙΟΓΡΑΦΙΑ

- Ali, A., & Gravino, C. (2019). A systematic literature review of software effort prediction using machine learning methods. *Journal of software: evolution and process*, 31(10), e2211.
- Alsaadi, B., & Saeedi, K. (2022). Data-driven effort estimation techniques of agile user stories: a systematic literature review. *Artificial Intelligence Review*, 1-32.
- Azhar, D., Mendes, E., & Riddle, P. (2012, September). A systematic review of web resource estimation. In *proceedings of the 8th International Conference on Predictive Models in Software Engineering* (pp. 49-58).
- Bodoor, K. H., & Shujaat, A. (2020). Impact of data preprocessing on regression analysis. *International Journal of Engineering Research & Technology*, 9(7), 733-738.
- Corazza, A., Di Martino, S., Ferrucci, F., Gravino, C., & Mendes, E. (2009, October). Applying support vector regression for web effort estimation using a cross-company dataset. In *2009 3rd International Symposium on Empirical Software Engineering and Measurement* (pp. 191-202). IEEE.
- Dave, C. V., Patel, A., & Keshri, U. Estimation Approaches of Machine Learning in Scrum Projects: A Review.
- Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica chimica acta*, 185, 1-17.
- Guide, A. (2001). Project management body of knowledge (pmbok® guide). In *Project Management Institute* (Vol. 11, pp. 7-8).
- Gupta, S. K., Gunasekaran, A., Antony, J., Gupta, S., Bag, S., & Roubaud, D. (2019). Systematic literature review of project failures: Current trends and scope for future research. *Computers & Industrial Engineering*, 127, 274-285.
- Huang, S. J., & Han, W. M. (2008). Exploring the relationship between software project duration and risk exposure: A cluster analysis. *Information & Management*, 45(3), 175-182.
- Idri, A., Hosni, M., & Abran, A. (2016). Systematic literature review of ensemble effort estimation. *Journal of Systems and Software*, 118, 151-175.
- Lee, T. K., Wei, K. T., & Abd Ghani, A. A. (2016, December). Systematic literature review on effort estimation for Open Sources (OSS) web application development. In *2016 Future Technologies Conference (FTC)* (pp. 1158-1167). IEEE.

- Matos, O., Conte, T., & Mendes, E. (2014, May). Is there a place for qualitative studies when identifying effort predictors? a case in web effort estimation. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering* (pp. 1-10).
- Mendes, E. (2007). The use of a bayesian network for web effort estimation. In *Web Engineering: 7th International Conference, ICWE 2007 Como, Italy, July 16-20, 2007 Proceedings 7* (pp. 90-104). Springer Berlin Heidelberg.
- Mendes, E. (2010). An overview of web effort estimation. *Advances in Computers*, 78, 223-270.
- Mendes, E., & Mosley, N. (2008). Bayesian network models for web effort prediction: a comparative study. *IEEE Transactions on Software Engineering*, 34(6), 723-737.
- Priya Varshini, A. G., & Anitha Kumari, K. (2020). Predictive analytics approaches for software effort estimation: A review. *Indian J. Sci. Technol*, 13, 2094-2103.
- Saif, S. M., & Wahid, A. (2017, November). Web complexity factors! A novel approach for predicting size measures for web application development. In *2017 International Conference on Inventive Computing and Informatics (ICICI)* (pp. 897-902). IEEE.
- Satapathy, S. M., & Rath, S. K. (2016, September). Effort estimation of web-based applications using machine learning techniques. In *2016 International conference on advances in computing, communications and informatics (ICACCI)* (pp. 973-979). IEEE.
- Sharma, P., & Singh, J. (2017, December). Systematic literature review on software effort estimation using machine learning approaches. In *2017 International Conference on Next Generation Computing and Information Systems (ICNGCIS)* (pp. 43-47). IEEE.
- Torrecilla-Salinas, C. J., Sedeño, J., Escalona, M. J., & Mejías, M. (2015). Estimating, planning and managing Agile Web development projects under a value-based perspective. *Information and Software Technology*, 61, 124-144.
- Wen, J., Li, S., Lin, Z., Hu, Y., & Huang, C. (2012). Systematic literature review of machine learning based software development effort estimation models. *Information and Software Technology*, 54(1), 41-59.

