ΣΤΡΑΤΙΩΤΙΚΗ
ΣΧΟΛΗ ΕΥΕΛΠΙΔΩΝ
Τμήμα Στρατιωτικών Επιστημών

Ε Λ Η Ν Ι Κ Η Δ Η Μ Ο Κ Ρ Α Τ Ι Α
ΔΙΙΔΡΥΜΑΤΙΚΟ ΔΙΑΤΜΗΜΑΤΙΚΟ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ
ΣΠΟΥΔΩΝ
ΑΚΑΔΗΜΑΪΚΟΥ ΕΤΟΥΣ 2016-17
ΣΧΕΔΙΑΣΗ & ΕΠΕΞΕΡΓΑΣΙΑ ΣΥΣΤΗΜΑΤΩΝ (SYSTEMS ENGINEERING)
(ΠΔ 96 /2015/ΦΕΚ 163Α'/20.08.2014)

ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
Σχολή Μηχανικών Παραγωγής & Διοίκησης

ΣΥΣΤΗΜΑ ΛΗΨΗΣ ΑΠΟΦΑΣΕΩΝ ΚΑΙ ΠΡΟΒΛΕΨΗΣ ΒΛΑΒΩΝ ΠΛΟΙΩΝ ΔΙΟΙΚΗΣΗΣ ΕΠΙΤΗΡΗΣΗΣ ΠΟΛΕΜΙΚΟΥ ΝΑΥΤΙΚΟΥ ΜΕΣΩ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ

Μεταπτυχιακή Διατριβή

ΧΡΗΣΤΟΣ ΠΑΠΑΣΠΥΡΟΣ

Επιβλέπων Καθηγητής:
Καραδήμας Νικόλαος

Αθήνα, Μάιος 2020

Η Μεταπτυχιακή Διατριβή του Παπασπύρου Χρήστου εγκρίνεται:

ΤΡΙΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

Καραδήμας Νικόλαος (Επιβλέπων) ,.....

Επίκουρος Καθηγητής (ΣΣΕ)

Καρανάσιου Ειρήνη ,.....

Αναπληρώτρια Καθηγήτρια (ΣΣΕ)

Τσαφαράκης Στέλιος ,..... .....

Επίκουρος Καθηγητής (ΠΚ)

Παπασπύρος Χρήστος

Μάιος, 2020

"You will never forget a person
who came to you with a torch in the dark"

M Rose

ΕΥΧΑΡΙΣΤΙΕΣ

Η παρούσα εργασία αποτελεί διπλωματική εργασία στα πλαίσια του μεταπτυχιακού προγράμματος "Σχεδίαση & Επεξεργασία Συστημάτων (Systems Engineering)" Πριν την παρουσίαση των αποτελεσμάτων της παρούσας διπλωματικής εργασίας, αισθάνομαι την υποχρέωση να ευχαριστήσω ορισμένους από τους ανθρώπους που έπαιξαν πολύ σημαντικό ρόλο στην πραγματοποίησή της.

Πρώτους από όλους θέλω να ευχαριστήσω τους γονείς μου Μαρία Τατά και Σπυρίδων Παπασπύρο για την ψυχική και οικονομική ενίσχυση καθώς και για την αγάπη που μου έχουν εμφυσήσει για τη μόρφωση. Εν συνεχεία τον επιβλέποντα καθηγητή της διπλωματικής εργασίας, κ. Καραδήμα Νικόλαο για την πολύτιμη καθοδήγηση του την εμπιστοσύνη και εκτίμηση που μου έδειξε. Δε θα μπορούσα να παραβλέψω τον κ. Τσάμη Κωνσταντίνο για την επιστημονική του καθοδήγηση σε θέματα που άπτονται της συλλογής και επεξεργασίας δεδομένων, τον κ. Παπαδήμα Νικόλαο για τη βοήθεια και τις συμβουλές που παρείχε σε μεγάλο αριθμό μαθημάτων, την Διοίκηση Πλοίων Επιτήρησης για την πρόσβαση στα δεδομένα βλαβών που χρησιμοποιήθηκαν στην εργασία, τον Κυβερνήτη μου Αντχο Αναστάσιο Παπαλεοντή ΠΝ για την κατανόηση που επέδειξε στη χορήγηση αδειών ώστε να μπορώ να συμμετέχω σε εξεταστικές καθώς και άτομα του στενού μου περιβάλλοντος που με το τρόπο τους με στήριξαν ηθικά ώστε παρά τις δυσκολίες που αντιμετώπιζα μου υπενθύμιζαν τις ικανότητές μου ώστε να μην τα παρατήσω.

Τις ευχαριστίες μου εκφράζω επίσης και στους καθηγητές Καρανάσιου Ειρήνη και Τσαφαράκη Στέλιο, οι οποίοι δέχτηκαν να είναι μέλη της τριμελούς επιτροπής αξιολόγησης της μεταπτυχιακής εργασίας.

ΑΚΡΩΝΥΜΙΑ

ΠΝ	Πολεμικό Ναυτικό

BI	Business Intelligence
CDA	Confirmatory Data Analysis
EDA	Exploratory Data Analysis
ER	Entity Relationship
HN	Hellenic Navy
LASSO	Least Absolute Shrinkage and Selection Operator
NA	Not Available
PCA	Principal Component Analysis
RSS	Root Sum Square
SVD	Singular Value Decomposition

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΛΗΨΗ

Η παρούσα μεταπτυχιακή διατριβή έχει σκοπό να αναλύσει τον όρο "Ανάλυση Δεδομένων" δίνοντας μια συνοπτική αναφορά ως προς το τι είναι και σε ποιους τομείς υπάρχει ήδη και υπήρχε τόσα χρόνια καθώς και παραδείγματα μεθόδων εκμετάλλευσης δεδομένων με σκοπό να βγάλουν ποικιλόμορφα συμπεράσματα - αν και εφόσον υπάρχουν.

Αρχικά, θα γίνει αναφορά σε συστήματα και σε βάσεις δεδομένων που μπορούν να δεχτούν μεγάλο όγκο δεδομένων αλλά και όχι μόνο. Επιπλέον, η μεταπτυχιακή εργασία αναφέρεται σε συστήματα διαχείρισης δεδομένων, είτε είναι γλώσσες προγραμματισμού είτε είναι έτοιμα τεχνολογικά εργαλεία. Φυσικά, αφού θα έχει γίνει αναφορά στα παραπάνω δεν μπορεί να παραλειφθεί η περιγραφή των πιο διαδεδομένων μεθόδων, τεχνικών και τροπών Ανάλυσης Δεδομένων μεγάλης ποσότητας.

Έπειτα, στόχος της διπλωματικής είναι να επικεντρωθεί στις Ένοπλες Δυνάμεις και ειδικότερα στο Πολεμικό Ναυτικό (ΠΝ). Το ΠΝ αποτελείται από πληθώρα πλοίων όπου όπως είναι φυσικό παρουσιάζουν βλάβες, οι οποίες ποικίλουν ανάλογα με το τύπο του πλοίου αλλά και την παλαιότητα του.

Άρα, θα γίνει σύνδεση της χρησιμότητας της Ανάλυσης Δεδομένων με τη περίπτωση του ΠΝ όσον αφορά την αντιμετώπιση βλαβών ώστε να καταλήξουμε σε ένα σύστημα λήψης απόφασης. Το σύνολο του ελληνικού πολεμικού στόλου συχνά είτε λόγω παλαιότητας, είτε πολυπλοκότητας συστημάτων, τα οποία χρησιμοποιούνται διαρκώς λόγω των αυξημένων επιχειρησιακών απαιτήσεων παρουσιάζουν βλάβες που συχνά και γίνονται αιτία για την μείωση της μαχητικής τους ικανότητας ή ακόμη και για τη καθήλωση τους. Η παρούσα μεταπτυχιακή διατριβή θα ασχοληθεί με τη συλλογή δεδομένων που αφορούν τις βλάβες πλοίων της Διοίκησης Επιτήρησης, τα οποία εφόσον ομαδοποιηθούν και ταξινομηθούν κατάλληλα θα αναλυθούν ωστε να προκύψουν συμπεράσματα και τυχόν προτάσεις.

Τέλος, οποιαδήποτε πληροφορία συλλεχθεί από αυτήν την μεταπτυχιακή διατριβή θα μπορέσει να χρησιμοποιηθεί όχι μόνο για την ορθή και γρήγορη πρόβλεψη του αποθεματικού των υλικών για τις βλάβες αλλά και για την ετοιμότητα των αντίστοιχων συνεργείων που ανήκουν είτε στο ΠΝ είτε σε ιδιώτες αλλά και των πληρωμάτων των πλοίων. Με αυτό τον τρόπο, τα πλοία θα είναι πάντα σε ετοιμότητα για οποιαδήποτε κίνδυνο ή απειλή.

ABSTRACT

This paper is intended to analyze the term "Data Analysis" by giving a summary of what it is and in which areas it has already existed for so many years and some examples of data exploitation methods and the available conclusions.

First, there will be reference to systems and databases that can accommodate a large amount of data. In addition, postgraduate work refers to data management systems whether they are programming languages or technology ready tools. I will describe the most common methods, techniques and ways of analyzing large quantities of data.

Next goal of this paper is to focus on the Armed Forces and in particularly Hellenic Navy (HN). The HN is made up of a large number of ships which naturally suffer from damage, which varies according to the type of each ship.

Therefore, the usefulness of data analysis will be linked to the case of HWN ships in relation to fault management. The Greek fleet as a whole often, either because of its age or the complexity of systems, which are used continuously due to increased operational requirements, is damaged which often and becomes the cause of a reduction in their combat capacity.

Furthermore, this paper will deal with the collection of data, concerning the damage of some types of warships, which grouped and classified accordingly in order to draw conclusions and suggestions.

Finally, any information collected can be used not only to correctly and quickly predict the stock of materials for damage repair but also for the readiness of the crew, of the workshops belonging to either the HN or to individuals. This way, the ships will always be ready for any danger or threat.

1 ΕΙΣΑΓΩΓΗ

Είναι γεγονός ότι το Πολεμικό Ναυτικό (ΠΝ) αποτελείται απο πάρα πολλές μονάδες και η κάθε μία απο αυτές απο πληθώρα διαφορετικών συστημάτων: μηχανικών, ηλεκτρονικών και ηλεκτρολογικών μα και συνδυασμό αυτών. Όπως γίνεται κατανοητό λοιπόν το πολεμικό πλοίο αποτελεί ένα πολύπλοκο σύστημα που για να λειτουργήσει άρτια και επιχειρησιακά όντας αξιόμαχο και αξιόπλοο κάθε υποσύστημα αυτού πρέπει να αποδίδει ικανοποιητικά. Το περιβάλλον λειτουργίας οι υψηλές επιχειρησιακές απαιτήσεις που οδηγούν σε εκτεταμένη χρήση των πλοίων, η παλαιότητα των συστημάτων, η έλλειψη ποσότητας και ποικιλίας ανταλλακτικών και το γεγονός των πολλών μη προβλέψιμων διαφορετικής φύσης βλαβών που δύναται να επηρεάσουν το πλοίο, μας δίνουν το έναυσμα για καταγραφή των μέχρι τούδε βλαβών ωστε εν συνεχεία να μπορέσουμε να τις μελετήσουμε. Είναι πολύ σημαντικό οι βλάβες αυτές να ομαδοποιηθούν ανά κατηγορία, να γίνει καταγραφή των αιτιών αυτών καθώς και άλλων παραμέτρων, όπως μέθοδος επίλυσης και χρόνος αποκατάστασης. Η ανωτέρω διαδικασία είναι πολύ δύσκολη λόγω μη πρότερης καταγραφής ,ομαδοποίησης και επεξεργασίας των δεδομένων συνολικά απο το ΠΝ σε κάποιο ενιαίο πρόγραμμα ώστε να έχουν μια κοινή μορφή εύκολη προς επεξεργασία.

Αρχικά, θα γίνει επικοινωνία με τα αρμόδια τμήματα με σκοπό να συλλεχθούν όλα τα δεδομένα στην εκάστοτε μορφή αρχείου που τα συγκεντρώνουν. Αφού έρθουν στην κατοχή μας, θα αφιερωθεί λίγος χρόνος να αξιολογηθεί η μορφή τους και οι ενέργειες που θα πρέπει να ακολουθήσουν για να έρθουν στην κατάλληλη μορφή. Τα δεδομένα αυτά αφορούν βάθος ετών όπως: είδος βλαβών, αίτιο, τρόπος αντιμετώπισης, απαιτούμενα ανταλλακτικά, χρόνος αποκατάστασης βλάβης, χρόνος μέχρι επανεμφάνιση αυτής, ώρες λειτουργίας συσκευής μέχρις ότου να προκύψει η βλάβη κ.α. μπορεί να δημιουργήσει μια σπουδαία βάση δεδομένων. Εάν ξεκινήσει ενιαία καταγραφή για όλα τα πλοία σε βάθος πολλών ετών και λαμβάνοντας υπόψιν πολλές επιπλέον των προαναφερθέντων παραμέτρους όπως π.χ. απαιτούμενα ανταλλακτικά, περιοδικότητα εμφάνισης βλάβης, μήνας εμφάνισης βλάβης, ώρες απο την τελευταία συντήρηση, ελλείψεις σε υλικά συντήρησης, υπερβάσεις ωρών λειτουργίας, συνολικός αριθμός πλεύσιμων ημερών πλοίου κα (τα οποία δεν βρίσκονταν στη παρούσα χρονική στιγμή στη διάθεσή μας) θα μπορούσε να δημιουργηθεί μια βάση Μεγάλων Δεδομένων. . Με αυτόν τον τρόπο μέσω ενός συστήματος διαχείρισης δεδομένων που θα περιλαμβάνει τα ανωτέρω θα καταγράφονται οι συχνότερες, σημαντικότερες και πιο κρίσιμες βλάβες ώστε να οργανωθεί απόθεμα ανταλλακτικών

που θα ανταποκρίνεται σε αυτές για να ελαχιστοποιηθεί ο χρόνος προμήθειας και συνεπώς επισκευής αυξάνοντας έτσι την επιχειρησιακή διαθεσιμότητα των πλοίων.

Για να μπορεί να εφαρμοστεί οποιαδήποτε τεχνική ανάλυση θα πρέπει να γίνει αρχικά μια επισκόπηση των δεδομένων με σκοπό να γίνει αντιληπτή η σημασία τους και η μορφή τους. Για να γίνουν όμως πλήρως αντιληπτά τα δεδομένα, είναι πολύ πιθανό να χρειαστεί κάποιο Καθάρισμα Δεδομένων (Data Cleansing), κρατώντας έτσι τις παρατηρήσεις που χρειάζονται για την Περιγραφική Ανάλυση (Descriptive Analysis). Αφού τελειώσει η προετοιμασία των δεδομένων, θα μπορεί να εφαρμοστεί η πιο κατάλληλη ανάλυση με σκοπό να μπορεί να γίνει κάποια πρόβλεψη στο μέλλον στα υλικά που θα πρέπει να έχει σαν απόθεμα το ΠΝ αλλά και να δεσμευτεί ο χρόνος των τεχνικών και των συνεργαζόμενων συνεργείων χωρίς να υπάρχουν οι αντίστοιχες καθυστερήσεις στην συνεννόηση και στην επικοινωνία.

Επιπλέον, χρησιμοποιώντας τα μοντέλα πρόβλεψης (predictive analytics) θα γίνει προσπάθεια πρόβλεψης σε ποιο πλοίο θα γίνει η επόμενη βλάβη και τι απαιτείται για να επισκευαστεί μειώνοντας έτσι τους χρόνους που απαιτούνται για να διορθώνονται οι βλάβες. Αν επιτευχθεί κάτι τέτοιο θα έχει γίνει δημιουργώντας κάποιο μοντέλο στατιστικής παλινδρόμησης.

Φυσικά όλα τα ευρήματα της παρούσας διπλωματικής θα απεικονιστούν και με την μορφή γραφημάτων (graphs) που θα είναι απόλυτα κατανοητά και ευανάγνωστα στο ευρύ κοινό παρέχοντας την ευκολία στον οποιοδήποτε αναγνώστη αλλά και στελέχους που ανήκει στο ΠΝ να καταλάβει και να αντιληφθεί την χρησιμότητα παρόμοιων γραφημάτων στην καθημερινότητα της εργασίας τους. Τα γραφήματα (reports) αυτά θα είναι περιγραφικά και θα έχουν παραχθεί είτε από το Microsoft Excel είτε από την γλώσσα προγραμματισμού R που θα χρησιμοποιήσουμε.

Στο 2ο κεφάλαιο, που θα ακολουθήσει, ο αναγνώστης θα μπορεί να έρθει σε επαφή με τον όρο Μεγάλα Δεδομένα (Big Data) και την Ανάλυση Δεδομένων (Data Analysis) διαβάζοντας την ιστορία τους όλα αυτά τα χρόνια, την αναγκαιότητα τους στην σύγχρονη κοινωνία αλλά και κάποια χαρακτηριστικά τους. Έπειτα, προχωρώντας στο 3ο κεφάλαιο, θα παρουσιαστούν τα δημοφιλέστερα εργαλεία που χρησιμοποιούνται για την Ανάλυση των Δεδομένων, αλλά και ειδικότερα τα εργαλεία που χρησιμοποιήθηκαν για την Ανάλυση των Δεδομένων της παρούσας μεταπτυχιακής εργασίας. Στο ίδιο κεφάλαιο γίνεται αναφορά στις μεθοδολογίες που ακολουθούνται συνήθως για την Ανάλυση Δεδομένων και ποιες από αυτές επιλέχθηκαν για την παρούσα μεταπτυχιακή εργασία. Επιπλέον, θα ακολουθήσει το κεφάλαιο 4, όπου και θα γίνει καταγραφή όλων των συμπερασμάτων και των αποτελεσμάτων που προέκυψαν από τις αναλύσεις

που πραγματοποιήθηκαν. Τέλος, ακολουθεί το παράρτημα Α, όπου παρουσιάζεται μέρος του κώδικας που χρησιμοποιήθηκε για να παραχθούν τα αποτελέσματα του κεφαλαίου 4.

2 Η Συλλογη Μεγαλης Ποσοτητας Δεδομενων

2.1 Τι εννοούμε με τον όρο Μεγάλα Δεδομένα

Τα Μεγάλα Δεδομένα (Big Data) είναι μια αφηρημένη έννοια διότι πέραν του ποσοτικού χαρακτήρα των δεδομένων υπάρχουν και κάποια άλλα χαρακτηριστικά τα οποία καθορίζουν τη διαφορά μεταξύ της έννοιας των μαζικών δεδομένων και των πολύ μεγάλων δεδομένων. Πιο αναλυτικά, ο όρος Μεγάλα Δεδομένα (Big Data) χρησιμοποιείται για να περιγράψει δεδομένα τα οποία έχουν κύριο χαρακτηριστικό τον μεγάλο τους όγκο, ο οποίος καθίσταται ιδιαίτερα δύσκολος στην αποθήκευση, στην διαχείριση και στην ανάλυση τους από τις παραδοσιακές εφαρμογές διαχείρισης βάσεων δεδομένων και τα εργαλεία ανάλυσης που υπάρχουν μέχρι σήμερα. Ωστόσο, ο ορισμός αυτός εμπεριέχει μια υποκειμενικότητα όσον αφορά τον ελάχιστο όγκο που πρέπει να έχουν τα δεδομένα, ώστε να μπορούν να θεωρηθούν "Μεγάλα Δεδομένα". Υπάρχουν πολλοί ορισμοί κατά περιόδους που έχουν δοθεί για τον όρο Big Data.

Οι Manyika et al. (2011) αναφέρουν ως Big Data «ένα σύνολο στοιχείων και δεδομένων τα οποία δεν μπορούν να συγκεντρωθούν, να αποθηκευτούν και να επεξεργαστούν τα παραδοσιακά λογισμικά βάσεων δεδομένων». Ομοίως, οι Davis & Patterson (2012) και οι Emani et al. (2015) αναφέρουν ότι «τα Big Data είναι δεδομένα πολύ μεγάλου όγκου ώστε να τα διαχειριστούν και να τα αναλύσουν παραδοσιακά πρωτόκολλα βάσεων δεδομένων όπως η SQL». Οι Mayer, Schonberger, Cukier (2013) από την πλευρά τους ορίζουν τα Big Data ως «τεράστια σύνολα δεδομένων και πληροφοριών απ' όπου μπορούν να διεξαχθούν χρήσιμα συμπεράσματα όπως ο μεγάλος όγκος, η ποικιλία, η ταχύτητα και η αξία τους». Μαζί τους όσον αφορά τον ορισμό συμφωνούν και οι Krishnan (2013) και Reeve (2013) οι οποίοι σημειώνουν ότι η επεξεργασία τους καθίσταται αδύνατη από τα παραδοσιακά λογισμικά.

Πρέπει να ληφθεί υπόψιν του εκάστοτε αναγνώστη ότι η τεχνολογία εξελίσσεται συνεχώς μέσα στο χρόνο και συγχρόνως αλλά και γρηγορότερα αυξάνεται και ο όγκος των δεδομένων πράγμα που καθιστά και τον χαρακτηρισμό τους ως μεγάλα δεδομένα.

Λέγοντας ότι η τεχνολογία εξελίσσεται μέσα στον χρόνο μπορεί να εννοηθεί ότι από το 1950 για την διαχείριση και επεξεργασία μεμονωμένων αρχείων χρησιμοποιούνταν οι γνωστές για την εποχή εκείνη ταινίες και κάρτες. Οι τεχνολογικές όμως εξελίξεις στις συσκευές μαζικής αποθήκευσης και η αύξηση της υπολογιστικής ισχύος, θέτουν τις προϋποθέσεις για την ανάπτυξη

των συστημάτων διαχείρισης δεδομένων ώστε να αντικαταστήσουν τα συστήματα διαχείρισης αρχείων.

Αργότερα στο 1960 τα πρώτα συστήματα διαχείρισης βάσεων δεδομένων ξεκίνησαν να κάνουν την εμφάνισή τους με σκοπό ένα κοινό οργανωτικό πλαίσιο με στόχο την διαχείριση τους, τα οποία δεδομένα μέχρι τότε αποθηκεύονταν σε μεμονωμένα αρχεία. Το 1964, ο Charles Bachman στέλεχος της General Electric πρότεινε ένα δικτυωτό μοντέλο δεδομένων στο οποίο οι εγγραφές των δεδομένων ήταν συνδεδεμένες μεταξύ τους με τέτοιο τρόπο ώστε να σχηματίζουν τεμνόμενα σύνολα δεδομένων. Τα πρώτα λοιπόν συστήματα διαχείρισης βάσεως δεδομένων στηρίχθηκαν σε αυτό το δικτυωτό μοντέλο. Το 1965 η εταιρία IBM και η διεύθυνση διαστήματος της North American Aviation ανέπτυξαν από κοινού το ιεραρχικό μοντέλο δεδομένων. Σε αυτό το μοντέλο, τα δεδομένα παριστάνονταν ως δενδροειδής δομές μέσα σε μια ιεραρχία εγγράφων. Το Σύστημα Διαχείρισης Πληροφοριών (Information Management System-IMS) της IBM που κυκλοφόρησε το 1969 ήταν βασιζόμενο στο ιεραρχικό μοντέλο δεδομένων όπως αναφέρεται από το IBM Rebook (200). Από τα δικτυωτά και ιεραρχικά συστήματα βέβαια, μόνο τα IMS παραμένουν σε χρήση έως σήμερα.

Στη δεκαετία 1970 - 1980, εισήχθη η αρχική πρόκληση με τη μετάβαση από τα megabyte σε gigabyte για τις επιχειρήσεις. Η ανάλυση στοιχείων και εκθέσεων ουσιαστικά οδήγησε στη δημιουργία των βάσεων δεδομένων που χαρακτηρίστηκε ως ιδανική για την εποχή επιλογή στην επίλυση προβλημάτων. Ο ορισμός του σχεσιακού μοντέλου δεδομένων έγινε για πρώτη φορά εν έτη 1970 από τον Edgar Codd της IBM με τίτλο «System R4 Relational». Στην αρχή δεν ήταν ξεκάθαρο κατά πόσο ένα σχεσιακό σύστημα που θα βασιζόταν στο σχεσιακό μοντέλο θα μπορούσε να πετύχει εμπορικά. Το 1976 το Μοντέλο Οντοτήτων-Σχέσεων (ER-Entity Relationship) προτάθηκε από τον Ταϊβανέζο επιστήμονα Η/Υ P.P. Chen για να περιγράψει με γραφικά σύμβολα τα δεδομένα ως συσχετίσεις (σχέσεις), οντότητες και γνωρίσματα. Αναπτύχθηκαν οι έννοιες της Διαχείρισης Συναλλαγών (Transaction Management) από τον Jim Gray όπως περιγράφει και ο Barker Richard (1990). Οι τάσεις που άρχιζαν να εμφανίζονται εκείνη την περίοδο, αφορούσαν τα αντικειμενοστραφή συστήματα, την αρχιτεκτονική πελάτη - διακομιστή και τις κατανεμημένες βάσεις. Οι εγκαταστάσεις των σχεσιακών συστημάτων αυξάνονται με γοργούς ρυθμούς, με πρώτα τα συστήματα της Oracle. Εμφανίζονται επιτυχώς τα σχεσιακά συστήματα διαχείρισης βάσεων δεδομένων και στους γνωστούς τότε, προσωπικούς υπολογιστές. Η Dbase εξελίχθηκε μέχρι τις μέρες μας σε Paradox και στη πιο γνωστή στο κοινό,

την Microsoft Access. Έτσι μέχρι και το 1979 όλες οι εμπορικές υλοποιήσεις βάσεων δεδομένων βασίζονταν είτε στην δικτυωτή είτε στην ιεραρχική προσέγγιση.

Έπειτα, το 1991 το διαδίκτυο, ο παγκόσμιος ιστός όπως τον ξέρουν οι περισσότεροι, γεννιέται. Το Πρωτόκολλο Μεταφοράς Υπερκειμένων (HTTP) γίνεται το βασικό μέσον διαμοιρασμού πληροφοριών. Το 1995 η SUN βγάζει στην κυκλοφορία την πλατφόρμα Java. Η Java, που ανακαλύφθηκε το 1991, γίνεται η δεύτερη πιο διαδεδομένη γλώσσα μετά την γνωστή C. Κυριαρχεί στις εφαρμογές διαδικτύου και καθιερώνεται στις μεσαίου επιπέδου εφαρμογές. Αυτές οι εφαρμογές είναι η πηγή καταγραφής και αποθήκευσης της κίνησης του διαδικτύου. Το παγκόσμιο σύστημα εντοπισμού (GPS) γίνεται πλήρως λειτουργικό. Το GPS είχε αναπτυχθεί αρχικά από την DAPRA (υπηρεσία προγραμμάτων προηγμένης έρευνας και άμυνας του Αμερικανικού στρατού) για στρατιωτικές εφαρμογές στις αρχές της δεκαετίας του ’70. Σήμερα η τεχνολογία αυτή είναι παρούσα, από εφαρμογές πλοήγησης αυτοκινήτων, πλοίων και αεροπλάνων μέχρι την στόχευση πυραύλων με εξαιρετικά μεγάλη ευκρίνεια. Το 1998 ο Carlo Strozzi αναπτύσσει μια ανοιχτού κώδικα βάση δεδομένων και την αποκαλεί No SQL. Δέκα χρόνια αργότερα, η πρωτοβουλία ανάπτυξης βάσεων δεδομένων NoSQL που θα μπορεί να επεξεργάζεται μεγάλα και αδόμητα σύνολα δεδομένων, κερδίζει έδαφος. Ιδρύεται η Google από τους Larry Page και Sergey Brin οι οποίοι ξεκίνησαν σε μια εργασία μιας μηχανής αναζήτησης του πανεπιστημίου Stanford με την ονομασία Back Rub.

Μόνο το 2010 εκτιμάται ότι 1.8 zettabytes δεδομένων δημιουργήθηκαν ή αλλιώς 220 δισεκατομμύρια ταινίες υψηλής ευκρίνειας HD, διάρκειας 2 ωρών περίπου. Την ίδια περίοδο ξεκίνησε και η λειτουργία του LinkedIn, του δημοφιλούς μέσου κοινωνικής δια δικτύωσης για επαγγελματίες όπου ήδη το 2014 η ιστοσελίδα είχε περίπου 290 εκατομμύρια χρήστες. Η υπηρεσία κοινωνικής δικτύωσης Facebook, ιδρύθηκε από τον Mark Zuckerberg με άλλους συμφοιτητές του στο Cambridge της Μασαχουσέτης. Έως το 2017, η υπηρεσία αυτή είχε πάνω από 1.8 δισεκατομμύρια χρήστες κάτι το οποίο δείχνει περίτρανα την ραγδαία αύξηση των δεδομένων μέσα από τα σημερινά συστήματα και εφαρμογές.

Το 2005 ο Doug Cutting και ο Mike Caferella δημιούργησαν ένα από τα πιο σημαντικά ερευνητικά έργα που απασχόλησε τα Big Data. Την ίδια περίοδο, το Εθνικό Επιστημονικό Συμβούλιο των ΗΠΑ, πρότεινε στο Εθνικό Ίδρυμα Επιστημών να ιδρυθεί μια νέα επαγγελματική κατηγορία για «έναν επαρκή αριθμό υψηλής ποιότητας καταρτισμένων κατάλληλα επιστημόνων» που θα είναι σε θέση να διαχειριστούν την αυξανόμενη συλλογή των ψηφιακών πληροφοριών.

Ο αριθμός των ηλεκτρονικών συσκευών που είναι συνδεδεμένα στο διαδίκτυο μόλις το 2009 ξεπερνά τον παγκόσμιο πληθυσμό που είναι γύρω στα 6,6 δις. Το πρωτόκολλο IPv4 βασίστηκε σε έναν 33 byte αριθμό που μας δείχνει ότι δύναται να υπάρξουν 4,5 δισεκατομμύρια μοναδικές διευθύνσεις διαθέσιμες. Αυτό μαρτυρά την αλματώδη αύξηση των συσκευών που είναι συνδεδεμένα στο διαδίκτυο. Γι' αυτό το λόγο έχει αναπτυχθεί η νέα έκδοση IPv6 που θα πολλαπλασιάσει τις διαθέσιμες μοναδικές διευθύνσεις.

Το 2019 δημιουργήθηκαν 3.1 zettabytes δεδομένων αλλά μόνο το 4% όσων θα μπορούσαν να χρησιμοποιηθούν για μεγάλα δεδομένα θα έχουν αναλυθεί μέχρι το τέλος του έτους. Προβλέπεται από το IDC ότι μέχρι το 2020 ο ψηφιακός κόσμος θα κατέχει 42 zettabytes, 59 φορές τον συνολικό αριθμό των κόκκων άμμου από όλες τις παραλίες του πλανήτη. Το Harvard αναφέρει το επάγγελμα του αναλυτή δεδομένων ως «την πιο δύσκολη εργασία του 21ου αιώνα».

Επιπρόσθετα, πρέπει να σημειωθεί κιόλας ότι ο ορισμός μπορεί να διαφέρει ανάλογα τον τομέα, ανάλογα με το είδος των λογισμικών, που είναι διαθέσιμα αλλά και ποια είναι τα συνήθη μεγέθη των πακέτων δεδομένων σε κάθε επιστημονικό κλάδο που θέλουμε να αναλύσουμε. Με αυτές τις επισημάνσεις, τα Μεγάλα Δεδομένα σε πολλούς τομείς σήμερα κυμαίνονται από μερικές δεκάδες terabytes έως τα πολλαπλάσια τους petabytes. Ψηφιακά δεδομένα μπορεί να συναντήσει ο καθένας πλέον παντού, σε κάθε οικονομία, σε κάθε οργανισμό και χρήστη της ψηφιακής τεχνολογίας. Τα Μεγάλα Δεδομένα ελκύουν όλο και περισσότερο το ενδιαφέρον των επαγγελματιών από όλους τους τομείς, ενώ οι καταναλωτές προϊόντων και υπηρεσιών αναμένεται και αυτοί να ωφεληθούν από την αξιοποίησή τους όπως και οι ίδιοι οι επιχειρηματίες – επιστήμονες (γιατροί, μηχανικοί κτλ). Η ικανότητα αποθήκευσης, συγκέντρωσης, συνδυασμού δεδομένων και η χρήση των αποτελεσμάτων για την εκπόνηση λεπτομερών αναλύσεων έχει γίνει πολύ πιο προσιτή και εφικτή όπως για παράδειγμα ο Νόμος του Moore (1965)

Ακόμη, τα μέσα άντλησης από τα δεδομένα σημειώνουν σημαντική βελτίωση, καθώς τα διαθέσιμα λογισμικά για την εφαρμογή τεχνικών αυξανόμενης πολυπλοκότητας συνδυάζονται με την αυξανόμενη υπολογιστική ισχύ. Πολλές τεχνολογικές καινοτομίες έχουν οδηγήσει σε δραματική αύξηση των δεδομένων και στη συλλογή αυτών. Αυτός είναι ο λόγος που τα δεδομένα μεγάλης κλίμακας έχουν γίνει περιοχή των στρατηγικών επενδύσεων για τους οργανισμούς με κύριο αντικείμενο την πληροφορική. Αν και είναι σαφές ότι οι νέες τεχνολογίες και νέες μορφές προσωπικής επικοινωνίας οδήγησαν στην τάση των μεγάλης κλίμακας δεδομένων, αν σκεφτούμε ότι ο παγκόσμιος πληθυσμός του διαδικτύου αυξήθηκε κατά 6,5% από το 2010-2011 και τώρα αντιπροσωπεύει πάνω από ένα δισεκατομμύριο ανθρώπους.

Τέλος, ο όγκος των δεδομένων αναμένεται να αυξηθεί 60 φορές μέχρι το 2021. Η Google λαμβάνει πάνω από 2.400.000 ερωτήματα αναζήτησης κάθε λεπτό. Το YouTube με την σειρά του δέχεται 72 καινούργιες ώρες βίντεο κάθε λεπτό, ενώ υπάρχουν 374 νέοι χρήστες του Ιντερνέτ κάθε λεπτό. Το Twitter δεν θα μπορούσε να μείνει πίσω αφού οι χρήστες του στέλνουν πάνω από 130.000 tweets κάθε λεπτό που ισοδυναμεί με πάνω από 170 εκατομμύρια ανά ημέρα. Η IDC προβλέπει ότι η αγορά για την τεχνολογία των μεγάλης κλίμακας δεδομένων και υπηρεσίες θα φτάσει τα $16,9 δισεκατομμύρια μέχρι το 2018 με αύξηση 40% πάνω από τον ορίζοντα της πρόβλεψης.

2.2 Η αναγκαιότητα της συλλογής δεδομένων στην σύγχρονη κοινωνία

Όλο και περισσότερες εταιρείες ψηφιοποιούν και αποθηκεύουν μια αυξανόμενη ποσότητα εξαιρετικά λεπτομερών δεδομένων σχετικά με τις συναλλαγές. Όλο και περισσότεροι αισθητήρες ενσωματώνονται σε φυσικές συσκευές, από τον εξοπλισμό της γραμμής παραγωγής έως και σε αυτοκίνητα αλλά και σε κινητά τηλέφωνα, οι οποίοι μετρούν διαδικασίες, χρήση προϊόντων, και ανθρώπινες συμπεριφορές. Επίσης, οι καταναλωτές ατομικά δημιουργούν και μοιράζονται μια τεράστια ποσότητα δεδομένων μέσω του Blogging, των ενημερώσεων κατάστασης, και την ανάρτηση των φωτογραφιών τους. Μεγάλο μέρος των δεδομένων αυτών μπορεί τώρα να συγκεντρώνεται σε πραγματικό ή σχεδόν πραγματικό χρόνο. Η δυνατότητα πρόσβασης σε όλα τα δεδομένα αυτά και σε ορισμένες περιπτώσεις, η δυνατότητα διαχείρισης των συνθηκών δημιουργίας τους, παρέχουν έναν πολύ διαφορετικό τρόπο λήψης αποφάσεων, τον οποίο εισάγει πιο πολύ η επιστήμη στη Διοίκηση

Η χρήση των δεδομένων προσφέρει τεράστιες ανεκμετάλλευτες δυνατότητες δημιουργικής αξίας. Οι οργανισμοί σε πολλούς κλάδους και πολλές επιχειρηματικές λειτουργίες μπορούν να αξιοποιήσουν τα δεδομένα με σκοπό τη βελτίωση της κατανομής και του συντονισμού των πόρων τους, τον περιορισμό της σπατάλης, την αύξηση της διαφάνειας και της λογοδοσίας, και την ανάδειξη νέων ιδεών και αντιλήψεων. Τα Μεγάλα Δεδομένα δημιουργούν αξία με διάφορους τρόπους. Η δημιουργία διαφάνειας είναι μία από αυτές αφού οι ενδιαφερόμενοι φορείς έχουν εύκολη και έγκαιρη πρόσβαση σε μεγάλα δεδομένα, αποκτώντας έτσι ευκαιρίες δημιουργίας τεράστιας αξίας. Συχνά, τέτοιες ευκαιρίες προκύπτουν σε περιπτώσεις όπου παρατηρείται έλλειψη συμφωνίας κινήτρων για δημιουργία διαφάνειας δεδομένων. Για παράδειγμα, στον δημόσιο τομέα, υπάρχουν περιπτώσεις όπου το προσωπικό διαφόρων υπηρεσιών σπαταλά σημαντικό

ποσοστό του χρόνου του για να εντοπίσει πληροφορίες σε άλλες κυβερνητικές υπηρεσίες, χρησιμοποιώντας μη-ψηφιακά μέσα (π.χ. σε έντυπους καταλόγους ή τηλεφωνώντας) και στη συνέχεια για να πάρουν τις πληροφορίες αυτές θα έπρεπε να επισκεφθούν την πηγή της πληροφορίας για να λάβουν τα στοιχεία με φυσικά μέσα, (πχ. οπτικοί δίσκοι). Αυτή η σπατάλη έχει μειωθεί σημαντικά σε οργανισμούς, που αξιοποιούν τα Μεγάλα Δεδομένα για να ψηφιοποιήσουν την πληροφορία αυτή, χρησιμοποιώντας τα διαθέσιμα δίκτυα, και αναπτύσσοντας εργαλεία ευκολότερης εύρεσης της αναζητούμενης πληροφορίας. Ωστόσο, ακόμη και σε τομείς, που έχουν υιοθετηθεί οι νέες τεχνολογίες και τα Μεγάλα Δεδομένα, υπάρχουν σημαντικά κίνητρα για υψηλότερη απόδοση και περιθώρια αύξησης διαφάνειας και ανταλλαγής Μεγάλων Δεδομένων. Στον τομέα της μεταποίησης, πολλές εταιρείες χρησιμοποιούν τα Μεγάλα Δεδομένα για τη βελτίωση στην απόδοση της Έρευνας και Τεχνολογίας (π.χ. πολύπλοκες προσομοιώσεις) όπως και στη διαχείριση της αλυσίδας εφοδιασμού τους όπως αναφέρεται και από έρευνα του McKinsey (2011).

Αναμενόμενο είναι λοιπόν ένας οργανισμός, που είναι προσανατολισμένος στα δεδομένα να λαμβάνει αποφάσεις με βάση τα εμπειρικά αποτελέσματα και τα οφέλη μιας τέτοιας προσέγγισης που έχουν αποδειχθεί και από την ακαδημαϊκή έρευνα. Οι ηγέτες σε πολλούς τομείς έχουν ήδη αρχίσει να χρησιμοποιούν ελεγχόμενες έρευνες για τη λήψη καλύτερων αποφάσεων. Για παράδειγμα, στον τομέα της υγείας εκπονούνται μελέτες συγκριτικής αξιολόγησης αποτελεσματικότητας σε ολόκληρο τον πληθυσμό, καθώς εντοπίζονται επαρκή κλινικά δεδομένα για τον εντοπισμό και την κατανόηση των πηγών της μεταβλητότητας σε θεραπείες και αποτελέσματα και έτσι βοηθιούνται οι υπεύθυνοι για τη λήψη αποφάσεων στη χάραξη κατευθυντήριων γραμμών, που εξασφαλίζουν ότι οι αποφάσεις για τη θεραπεία βασίζονται στην ορθότερη επιστημονική προσέγγιση. Οι πωλητές, κυρίως εκείνοι που δραστηριοποιούνται διαδικτυακά, προσαρμόζουν τις τιμές και τις προσφορές τους σε μια προσπάθεια εντοπισμού του βέλτιστου συνδυασμού κυκλοφορίας και πωλήσεων. Ωστόσο, δεν είναι πάντα δυνατόν η κατασκευή μιας ελεγχόμενης έρευνας και ο «χειρισμός» μια ανεξάρτητης μεταβλητής. Μια εναλλακτική είναι η εύρεση «φυσικών πειραμάτων», που εντοπίζουν την υπάρχουσα μεταβλητότητα στις μετρήσεις απόδοσης. Η κατανόηση των αιτιών αυτής της μεταβλητότητας μπορεί στη συνέχεια να συμβουλέψει τους υπευθύνους διαχείρισης να λάβουν αποφάσεις και να βελτιώσουν την απόδοση. Στο δημόσιο τομέα, εντοπίζονται υπηρεσίες με τεράστιες αποκλίσεις στην παραγωγικότητα και την ακρίβεια του έργου, οι οποίες εκτελούν σχεδόν πανομοιότυπα καθήκοντα. Η γνωστοποίηση και μόνο αυτής της πληροφορίας μπορεί να έχει ως αποτέλεσμα

σημαντική αύξηση απόδοσης στις υστερούσες υπηρεσίες και χωρίς χρηματικό αντίκρισμα ως κίνητρο.

Η Ανάλυση Δεδομένων και οι εξελιγμένοι αλγόριθμοι τεχνητής νοημοσύνης μπορούν να βελτιώσουν σημαντικά τη λήψη αποφάσεων, την ελαχιστοποίηση της αβεβαιότητας, και την ανάδειξη πολύτιμων πληροφορίων. Τα Μεγάλα Δεδομένα παρέχουν την πρώτη ύλη που απαιτείται είτε για την ανάπτυξη αλγορίθμων, είτε για τη λειτουργία τους. Για παράδειγμα, φορολογικές υπηρεσίες, που εφαρμόζουν και χρησιμοποιούν αυτοματοποιημένες μηχανές αβεβαιότητας που χρησιμοποιούν Μεγάλα Δεδομένα για τον εντοπισμό υποψηφίων, που χρήζουν περαιτέρω διερεύνησης. Οι αλγόριθμοι Μεγάλων Δεδομένων στον τομέα της λιανικής μπορούν να αριστοποιήσουν διαδικασίες λήψης αποφάσεων, επιτρέποντας την αυτόματη ρύθμιση καταλόγων και τιμολογώντας σε πραγματικό χρόνο και σε καταστήματα και σε διαδικτυακές πωλήσεις. Οι κατασκευαστικές εταιρείες μπορούν να προσαρμόσουν τις γραμμές παραγωγής τους αυτόματα, για βελτιστοποίηση αποδοτικότητας, μείωση σπατάλης, και αποφυγή επικίνδυνων συνθηκών. Σε ορισμένες περιπτώσεις, οι εταιρείες δεν αυτοματοποιούν απαραίτητα τις αποφάσεις, αλλά τις διευκολύνουν μέσω της Ανάλυσης των Μεγάλων Δεδομένων που τα περισσότερα από τα δεδομένα είναι πιο προσιτά και εύκολα στη διαχείριση από ένα άτομο χρησιμοποιώντας ένα υπολογιστικό φύλλο. Ορισμένοι οργανισμοί λαμβάνουν ήδη πιο αποτελεσματικές αποφάσεις αναλύοντας ολόκληρα σύνολα δεδομένων από πελάτες και εργαζόμενους ή ακόμα και από αισθητήρες ενσωματωμένους σε προϊόντα.

2.3 Χαρακτηριστικά των Μεγάλων Δεδομένων

Στα χαρακτηριστικά των Big Data οι περισσότεροι συγγραφείς αναφέρουν τα 4V (Volume, Variety, Velocity, Veracity). Αν και τα τρία εξ' αυτών συναντώνται συχνότερα στη βιβλιογραφία (Volume, Variety, Velocity), η ειλικρίνεια (Veracity) και η αξία (Value) αποτελούν επιπλέον χαρακτηριστικά των Big Data.

Ως προς τα χαρακτηριστικά των Big Data σημειώνονται τα εξής:

- Όγκος (Volume). Ο όγκος αναφέρεται στο πλήθος των δεδομένων που δημιουργούνται. Ο τεράστιος όγκος είναι βασική ιδιότητα των Big Data. Γενικά, ακολουθούν εκθετική αύξηση ενώ δημιουργούνται πολύ γρήγορα. Τα Big Data σε επίπεδο όγκου προσφέρουν τη δυνατότητα επεξεργασία πολλών δεδομένων. Το συγκεκριμένο χαρακτηριστικό είναι ιδιαίτερα χρήσιμο ιδιαίτερα τόσο στην αποθήκευση όσο και στην επεξεργασία για είδη δεδομένων από κοινωνικά δίκτυα, υγειονομική περίθαλψη, χρηματοοικονομικά δεδομένα,

βιοχημεία, γενετικά δεδομένα κ.λπ. (Emani et al., 2015; Vitolo, 2015; Singh & Singla, 2015).

- Ποικιλία (Variety). Τα δεδομένα των Big Data δεν έχουν μια σταθερή δομή και σπάνια παρουσιάζονται σε μια απόλυτα επεξεργάσιμη μορφή. Τα δεδομένα αυτά μπορεί να είναι ιδιαίτερα δομημένα (δεδομένα από σχεσιακές βάσεις δεδομένων), ημι-δομημένα (αρχεία καταγραφής ιστού, social media feeds κλπ) ή μη δομημένα (βίντεο, φωτογραφίες).Η ποικιλία είναι μία από τις πιο σημαντικές ιδιότητες των Big Data. Η απουσία συγκεκριμένης δομής δημιουργεί από μόνη της προκλήσεις για τις παραδοσιακές τεχνολογίες και βάσεις δεδομένων. Το εύρος και η ποικιλία των δεδομένων είναι τέτοια ώστε να επηρεάζουν τη σημασιολογία, ή τη μεταβλητότητα του νοήματος (Emani et al., 2015; Hoy, 2014; Wang et al.,2016; Singh & Singla, 2015).
- Ταχύτητα (Velocity). Η ταχύτητα είναι ένα χαρακτηριστικό των Big Data που αφορά το χρόνο δημιουργίας και αποθήκευσης των δεδομένων, τη διαθεσιμότητα και την πρόσβαση. Το ζήτημα της ταχύτητας δεν εστιάζει τόσο στην αποθήκευση όσο στη ταχύτητα με την οποία δημιουργούνται τα δεδομένα. Η ταχύτητα είναι μια άλλη βασική ιδιότητα των Μεγάλων Δεδομένων που παράγονται πολύ γρήγορα καθημερινά στο ψηφιακό κόσμο (Krishnan, 2013; Hoy, 2014; Wang et al.,2016; Singh & Singla, 2015).
- Ειλικρίνεια (Veracity): Η ειλικρίνεια των πληροφοριών βασίζεται στο βαθμό συμφωνίας με την αλήθεια ή την πραγματικότητα. Οι τυχόν αβεβαιότητες στα δεδομένα μπορούν να προκληθούν από λανθασμένες εφαρμογές μοντέλων, παραπλάνηση, επικάλυψη πληροφοριών (Emani et al., 2015; Wang et al., 2016; Yaqoob et al, 2016)

2.4 Χρήση Μεγάλων Δεδομένων στον επαγγελματικό χώρο

Τα Μεγάλα Δεδομένα επιτρέπουν στις επιχειρήσεις όλων των ειδών την ανάπτυξη νέων προϊόντων και υπηρεσιών, την ενίσχυση των υφιστάμενων και την εισαγωγή εντελώς νέων επιχειρηματικών μοντέλων. Στον τομέα της Υγείας, η ανάλυση των κλινικών δεδομένων και δεδομένων τη συμπεριφοράς των ασθενών έχει οδηγήσει σε προγράμματα προληπτικής φροντίδας, στοχευμένα στις κατάλληλες ομάδες ατόμων. Επίσης, στο λιανικό εμπόριο, οι υπηρεσίες σύγκρισης τιμών σε πραγματικό χρόνο δίνουν στους καταναλωτές πλήρη εικόνα των τιμών σε βαθμό, που ποτέ πριν δεν απολάμβαναν και δημιουργούν σημαντικό πλεόνασμα για αυτούς. Άλλες εταιρείες χρησιμοποιούν δεδομένα που λαμβάνονται από αισθητήρες

ενσωματωμένους σε προϊόντα για τη δημιουργία καινοτόμων,μετά την πώληση, προσφορών και υπηρεσιών, όπως η προληπτική συντήρηση και η δημιουργία βάσης για την ανάπτυξη της επόμενης γενιάς προϊόντων.

Υπάρχουν πολλά παραδείγματα περιπτώσεων χρήσης των μεγάλης κλίμακας δεδομένων σε κάθε βιομηχανία που μπορεί να φανταστεί κανείς. Ορισμένες επιχειρήσεις έχουν γίνει πιο δεκτικές στις τεχνολογίες και έχουν ενσωματώσει πιο γρήγορα την Ανάλυση δεδομένων στην καθημερινότητα της επιχείρησης σε σχέση με άλλες. Οι επιχειρήσεις που αγκαλιάζουν την τεχνολογία όχι μόνο θα δουν σημαντικά πρωτοποριακά πλεονεκτήματα, αλλά θα είναι σημαντικά πιο ευέλικτες και πιο προσαρμοστικές στις προσφορές τους. Για παράδειγμα, οι χρηματοπιστωτικές υπηρεσίες υιοθετούν υποδομές Ανάλυσης Μεγάλων Δεδομένων για να βελτιώσουν τις αναλύσεις των πελατών τους για το μετοχικό κεφάλαιο, ασφάλιση, υποθήκη, ή πίστωση. Ακόμη, οι αεροπορικές εταιρείες και εταιρείες οδικών μεταφορών χρησιμοποιούν μεγάλης κλίμακας δεδομένων για να παρακολουθήσουν την κατανάλωση καύσιμων και τα πρότυπα κυκλοφορίας στους στόλους τους σε πραγματικό χρόνο, για να βελτιώσουν την αποτελεσματικότητα και την εξοικονόμηση κόστους. Επιπλέον, οι υγειονομικής περίθαλψης υπηρεσίες διαχειρίζονται και κάνουν κοινή χρήση ηλεκτρονικών μητρώων ασθενών από πολλαπλές πήγες, εικόνες, θεραπείες, και δημογραφικά στοιχεία. Επίσης, οι φαρμακευτικές εταιρείες και οι ρυθμιστικοί οργανισμοί δημιουργούν λύσεις μεγάλης κλίμακας δεδομένων για την παρακολούθηση της αποτελεσματικότητας των φαρμάκων και για να παρέχουν πιο αποτελεσματική και πιο σύντομη ανάπτυξη φαρμάκων.

Τέλος, οι εταιρείες μέσων ενημέρωσης και ψυχαγωγίας αξιοποιούν τις υποδομές της μεγάλης κλίμακας δεδομένων για να βοηθήσουν με την λήψη αποφάσεων γύρω από τον πελάτη και για να παρέχουν πιο εστιασμένο μάρκετινγκ. Υπάρχουν περιπτώσεις και συγκεκριμένα παραδείγματα χρήσης των μεγάλης κλίμακας δεδομένων για κάθε βιομηχανία και εταιρεία . Ως εκ τούτου, έστω και αν αυτήν την περίοδο μια επιχείρηση δεν χρησιμοποιεί λύσεις Μεγάλων Δεδομένων, είναι πιθανόν οι ανταγωνιστές της να χρησιμοποιούν. Το πραγματικό ερώτημα είναι πως μπορεί να βελτιστοποιηθεί καλυτέρα το περιβάλλον της κάθε εταιρείας ώστε να δημιουργηθεί μια πιο γρήγορη αποτελεσματική λύση που δίνει ανταγωνιστικό πλεονέκτημα.

3 ΕΡΓΑΛΕΙΑ ΑΝΑΛΥΣΗΣ ΚΑΙ ΜΕΘΟΔΟΛΟΓΙΕΣ

3.1 Εργαλεία για Ανάλυση Δεδομένων

Μέχρι τώρα πολλές εταιρίες έχουν αποφασίσει ότι η Ανάλυση Δεδομένων δεν είναι κάτι το οποίο απλά αιωρείται στον αέρα αλλά ένα γεγονός στον επαγγελματικό χώρο το οποίο απαιτεί να δημιουργήσουν κάποιες στρατηγικές και αλλαγές στην κουλτούρα τους για να διαχειριστούν τον μεγάλο όγκο των δομημένων και αδόμητων δεδομένων. Αντιμετωπίζοντας την Ανάλυση των Δεδομένων σαν πραγματικότητα πλέον, οι εταιρίες πρέπει να βρουν νέο τρόπο να αντιμετωπίσουν μια νέα πρόκληση, αυτήν της Ανάλυσης όλων αυτών των Δεδομένων. Οι χρήστες και οι διευθυντές πληροφορικής που αρχικά εντόπισαν τα πρώτα προβλήματα στην διαχείριση μεγάλου όγκου δεδομένων έρχονται τώρα αντιμέτωποι με το να βρουν τρόπους να χειριστούν την Ανάλυση αυτών των Δεδομένων με σκοπό να αναγνωρίσουν τάσεις, μοτίβα και την συλλογή σημαντικών πληροφοριών μέσα από αυτό τον ωκεανό δεδομένων. Τα εργαλεία και τα λογισμικά της Ανάλυσης Δεδομένων χρησιμοποιούνται με σκοπό να βρεθούν και να ομαδοποιηθούν οι σχέσεις, τα μοτίβα και οι τάσεις. Με την βοήθεια των τεχνικών της Ανάλυσης και της Εξόρυξης Δεδομένων η κάθε επιχείρηση μπορεί να κερδίσει, αποκτώντας γνώση ακόμη και για τα εσωτερικά ζητήματα ενός οργανισμού ή μιας βιομηχανίας αλλά και τις τάσεις των πελατών.

Είναι πλέον αποδεκτό ότι κάθε μήνα θα ακουστεί μια νέα πληροφορία για ένα καινούργιο εργαλείο που αφορά την Επιχειρησιακή Έρευνα (Operational Research) και την Επιχειρηματική Αναλυτική (Business Analytics) και θα είναι το ίδιο συναρπαστικά σε λειτουργίες όπως τα άλλα λόγω του ότι θα έχουν κάτι καινούριο ή κάτι διαφορετικό. Η Επιχειρησιακή Έρευνα, ως παραδοσιακή προσέγγιση στην επίλυση προβλημάτων λήψης αποφάσεων σε συνδυασμό με την Επιχειρηματική Αναλυτική ως μία σύγχρονη και διευρυμένη όψη της, συνιστούν μία διαρκώς εξελισσόμενη επιστημονική περιοχή, η οποία μάλιστα βιώνει μία εκ νέου ακμή κατά την τελευταία δεκαετία, ακριβώς λόγω του αυξημένου μεγέθους και πολυπλοκότητας των σύγχρονων προβλημάτων λήψης αποφάσεων.

Υπάρχουν τόσα πολλά εργαλεία εκεί έξω για όλων των ειδών τις αναλύσεις που μπορεί ο καθένας να εφαρμόσει, αλλά και για όλες τις προτιμήσεις του κάθε προγραμματιστή. Με την βοήθεια των εργαλείων και των μεθόδων αυτών δίνεται στον καθένα η δυνατότητα να εμβαθύνει σε μεθόδους Επιχειρησιακής Έρευνας και λήψης αποφάσεων, την εξέταση θεμελιωδών αλλά και εξειδικευμένων θεμάτων Διοίκησης Παραγωγής και λειτουργιών, την ενδελεχή εξέταση μεθόδων Χρηματοοικονομικής Μηχανικής, τον σχεδιασμό μεθόδων Συνδυαστικής Βελτιστοποίησης, την

μοντελοποίηση στοχαστικών προβλημάτων λήψης αποφάσεων μέσω στοχαστικών διαδικασιών ή προσομοίωσης αλλά και την εμβάθυνση σε θέματα Επιχειρηματικής Αναλυτικής και ειδικότερα σε θέματα τεχνολογιών εξατομίκευσης καθώς και Ανάλυσης Δεδομένων στην εφοδιαστική αλυσίδα. Γι' αυτό και κάθε αναλυτής πρέπει να έχει επίγνωση από την αρχή τι είδος δεδομένων θα χρειαστεί να αναλύσει με αποτέλεσμα να διαλέξει το καταλληλότερο εργαλείο.

Επιπρόσθετα, πρέπει να γίνει αποδοχή ότι υπάρχουν δυο τύποι δεδομένων στην αγορά, τα δομημένα και τα αδόμητα. Για το κάθε είδος δεδομένων απαιτείται ανάλυση και το αντίστοιχο εργαλείο που είναι πιο ταιριαστό στον προγραμματιστή - αναλυτή και φυσικά πιο κοντά στο προϋπολογισμό της κάθε επιχείρησης / οργανισμού. Για παράδειγμα, το Microsoft Excel χρησιμοποιείται κατά κόρον από άτομα που δεν είναι συνήθως προγραμματιστές ή αναλυτές. Οι προαναφερθέντες το χρησιμοποιούν γιατί αποθηκεύονται πολλές πληροφορίες στα φύλλα του Microsoft Excel από οργανισμούς και επιχειρήσεις. Είναι απλό σε χρήση για έναν μέσο χρήστη και μπορεί να προσφέρει πολλές δυνατότητες ακόμη και την δημιουργία γραφημάτων. Ακόμη, οι MySQL, SQL, Microsoft SQL Server 2012 είναι οι δημοφιλέστερες σχεσιακές βάσεις δεδομένων που χρησιμοποιούνται ακόμη και για δεδομένα μεγάλης κλίμακας. Αντιθέτως, οι γλώσσες προγραμματισμού Python, Java χρησιμοποιούνται για την επεξεργασία, την ανάλυση ακόμη και την απεικόνιση δεδομένων και των αποτελεσμάτων τους σε μορφή γραφημάτων. Οι R και Matlab είναι γλώσσες προγραμματισμού που η κύρια λειτουργία τους είναι η δημιουργία στατιστικών μοντέλων αλλά και η χειραγώγηση των δεδομένων προσαρμόζοντας κάθε φορά στην επιθυμητή μορφή. Κάτι τέτοιο πραγματοποιείται μέσα από τα ίδια τους τα περιβάλλοντα τα οποία είναι προσιτά προς τον χρήστη προσφέροντας ολοένα και περισσότερες δυνατότητες. Αντίστοιχη γλώσσα προγραμματισμού με δικό της περιβάλλον είναι και η SAS. Οι δυνατότητες που δίνονται στον χρήστη είναι παρόμοιες απλά η χρήση τους απαιτεί πληρωμή κάποιου ποσού και αυτό γιατί σε αντίθεση με την R και την Matlab παρέχεται υποστήριξη από την εταιρία. Επιπρόσθετα, οι αντίστοιχές μη σχεσιακές βάσεις δεδομένων Hadoop, MongoDB, Redis, Spark, Cassandra παρατηρούνται κυρίως σε περιπτώσεις που υπάρχουν αδόμητα δεδομένα και σε χαοτική μορφή συνήθως. Τέτοιου είδους δεδομένα προέρχονται συνήθως από τηλεπικοινωνιακούς κλάδους και από το διαδίκτυο (Facebook, Instagram, twitter κτλ.). Τέλος, αφού έχει καθοριστεί ο χώρος αποθήκευσης των δεδομένων και ο τρόπος ανάλυσης τους θα πρέπει να συνυπολογιστεί και ο τρόπος απεικόνισης των αποτελεσμάτων και της κατάστασης των δεδομένων. Παρόλο που η δυνατότητα απεικόνισης της κατάστασης/μορφής των δεδομένων αλλά και των αποτελεσμάτων υπάρχει και μέσα από τις γλώσσες προγραμματισμού που προαναφέρθηκαν, αν ο χρήστης

χρειαστεί κάποιο εξειδικευμένο εργαλείο που χρησιμοποιείται αποκλειστικά για την δημιουργία γραφημάτων που συνεχώς εξελίσσονται μπορεί να επιλέξει ανάμεσα στα εργαλεία Tableau, Tibco Spotfire, Gephi, IBM Watson, IBM Bluemix, QlikView. Αυτά είναι από τα πιο δημοφιλή εργαλεία που χρησιμοποιούνται για την Ανάλυση Δεδομένων είτε είναι δομημένα είτε αδόμητα.

3.2 Εργαλεία που χρησιμοποιήθηκαν για αυτήν την ανάλυση

Σε πολλές επιχειρήσεις παρατηρούνται κάποιες καθυστερήσεις και κάποιες δυσκολίες στο να αποδεχτούν τις δημοφιλείς τεχνολογίες και τις ακαδημαϊκές τάσεις όχι μόνο για την αποθήκευση των δεδομένων αλλά και για την εκμετάλλευση αυτών. Επίσης, πολλές φορές παρατηρείται ότι οι προτιμήσεις των επιχειρήσεων διαφέρουν ανάλογα με το ακαδημαϊκό και το επαγγελματικό υπόβαθρο του εκάστοτε προσωπικού. Στην περίπτωση του ΠΝ επειδή αποτελείται απο στελέχη που δεν έχουν μεγάλη επαφή με προγραμματιστικό περιβάλλον και ανάλυση χρησιμοποιήθηκαν εργαλεία στα οποία μπορεί να ανταποκριθεί ένας μέσος χρήστης.

3.2.1 Το Microsoft Excel ως Βάση δεδομένων

Στις Ένοπλες Δυνάμεις παρατηρείται να χρησιμοποιείται περισσότερο το Excel για να αποθηκεύονται μικρές και μεγάλες ποσότητες δεδομένων. Το Excel δεν χρησιμοποιείται μόνο σαν κύρια πηγή αποθήκευσης δεδομένων αλλά και για να μεταφερθούν δεδομένα μεταξύ του προσωπικού, λόγω του ότι το γνωρίζουν οι περισσότεροι και ξέρουν να διαχειριστούν ένα μέρος της πληροφορίας που υπάρχει μέσα αν όχι όλη.

Οι σχεσιακές βάσεις δεδομένων όπως η Oracle και η SQL Server είναι λογικότερο να χρησιμοποιούνται σε πιο κεντρικά σημεία για την ομαλοποίηση και την ασφαλέστερη διατήρηση αυτών. Αυτά τα συστήματα είναι καταπληκτικά στην ενοποίηση πολλών πινάκων, αρχείων Excel αλλά και την αναζήτηση της πρωταρχικής δημιουργίας μιας τιμής ή ενός πεδίου. Αυτού του είδους οι βάσεις δεδομένων χρησιμοποιούνται σαν πρωταρχικά εργαλεία αποθήκευσης δεδομένων και ιστορικού κρατώντας την πληροφορία που υπάρχει στις βάσεις δεδομένων και στους πίνακες με ασφάλεια και με ελαχιστοποιημένη την πιθανότητα να καταστραφεί το οτιδήποτε. Στην περίπτωση της παρούσας διπλωματικής εργασίας όμως θα χρησιμοποιηθούν μόνο δεδομένα που προέρχονται από αρχεία Excel. Η αρχική πληροφορία υπήρχε σε πληθώρα αρχείων ασύνδετων μεταξύ τους και μη ομαδοποιημένων και χωρίς κάποιο κοινό τρόπο καταγραφής, συνεπώς έγινε μια μεγάλη προσπάθεια, χρονοβόρα και κοπιαστική , συγκέντρωσης όλων των παρεχόμενων

πληροφοριών και δημιουργήθηκε ένα ενιαίο αρχείο Excel οπού και υπάρχει όλη η απαραίτητη πληροφορία για το σύνολο των βλαβών σε αξιοποιήσιμη μορφή απο τη γλώσσα R.

3.2.2 Η R και οι δυνατότητες της

Στις μέρες μας, η γλώσσα R είναι το πιο συχνό εργαλείο (Open Source) που χρησιμοποιείται για τις αναλύσεις δεδομένων ασχέτως του μεγέθους τους. Είναι ιδανικό για την χρήση που θα γίνει λόγω του ότι έχει μια σειρά από βιβλιοθήκες που περιλαμβάνουν αλγορίθμους από στατιστικούς και μαθηματικούς κανόνες. Η R μπορεί εύκολα και γρήγορα μέσα από το περιβάλλον της να βοηθήσει στην αξιοποίηση της πληροφορίας από ένα αρχείο Excel χωρίς να χρειάζονται γνώσεις προγραμματισμού μεγάλου επιπέδου.

Πιο αναλυτικά, είναι μια γλώσσα προγραμματισμού με δικό της περιβάλλον κυρίως για στατιστικούς υπολογισμούς και για απεικόνιση μέσω γραφημάτων των αποτελεσμάτων. Φυσικά, το περιβάλλον της υποστηρίζεται πλήρως από την R κοινότητα και ας είναι χωρίς επί-πληρωμή το εργαλείο. Επίσης, είναι ευρέως χρησιμοποιούμενο σαν προϊόν ανάμεσα στους στατιστικούς και τους αναλυτές δεδομένων για την ανάπτυξη στατιστικών μοντέλων αλλά και την Ανάλυση Δεδομένων. Η R και οι βιβλιοθήκες της, χρησιμοποιούνται για την υλοποίηση ποικίλων στατιστικών μοντέλων (Γραμμική και Λογιστική Παλινδρόμηση), Χρονικής Διάρκειας Ανάλυση (Time-Series), Ταξινόμηση-Κατηγοριοποίηση (Classification) και Ομαδοποίηση (Clustering).

Ωστόσο, αν και είναι δωρεάν σαν εργαλείο, η R κοινότητα που ασχολείται μένει ενεργή ανανεώνοντας τις συναρτήσεις και δημιουργώντας συνέχεια προεκτάσεις στο περιβάλλον αλλά και στην ίδια την γλώσσα. Πολλές από τις κύριες συναρτήσεις της R έχουν γραφτεί μέσα από την ίδια γλώσσα πράγμα που κάνει ακόμα πιο εύκολο στους χρήστες να ακολουθήσουν τις αλγοριθμικές επιλογές τους. Προφανώς, για μεγάλες υπολογιστικές ανάγκες μπορούν να χρησιμοποιηθούν C, C++ και Fortran κώδικες με τις κατάλληλες συνδέσεις μέσα από το περιβάλλον της R. Ακόμη, δίνεται η δυνατότητα σε χρήστες να μπορούν να γράψουν C, C++, Java ή Python μέσα από το περιβάλλον της R. Η R έχει περισσότερο Αντικειμενοστραφείς Υποδομές (Object-Oriented) προγραμματισμού σε σχέση με τις περισσότερες στατιστικές γλώσσες. Όπως και άλλες στατιστικές γλώσσες προγραμματισμού (APL, MATLAB) έτσι και η R υποστηρίζει ένα σύνολο από αριθμητικών πινάκων, λίστες, πίνακες και αντικείμενα. Επιπρόσθετα, η R περιλαμβάνει τα κατάλληλα αντικείμενα στις βιβλιοθήκες της για Μοντέλα Παλινδρόμησης, Διαγράμματα Χρόνου (Time Series) και χάρτες με γεωγραφικές συντεταγμένες για απεικόνιση

αποτελεσμάτων πάνω σε αυτούς. Στην δική μας περίπτωση αντίστοιχα θα μπορούσε να χρησιμοποιηθεί για να εντοπιστεί ποιο πλοίο και σε ποια περιοχή έπαθε ζημιά. Αν για παράδειγμα μια βλάβη εμφανίζεται συχνότερα όταν πλησιάζει το πλοίο σε μια συγκεκριμένη περιοχή θα πρέπει να καταγραφτεί και να διερευνηθεί παραπάνω.

Τέλος, ένα άλλο αξιοσημείωτο χαρακτηριστικό της R είναι ότι μπορεί να απεικονίσει στατικά γραφήματα περιλαμβάνοντας μαθηματικά σύμβολα. Είτε δυναμικά είτε διαδραστικά γραφήματα είναι άμεσα διαθέσιμα για τον χρήστη κατεβάζοντας τα αντίστοιχα πακέτα.

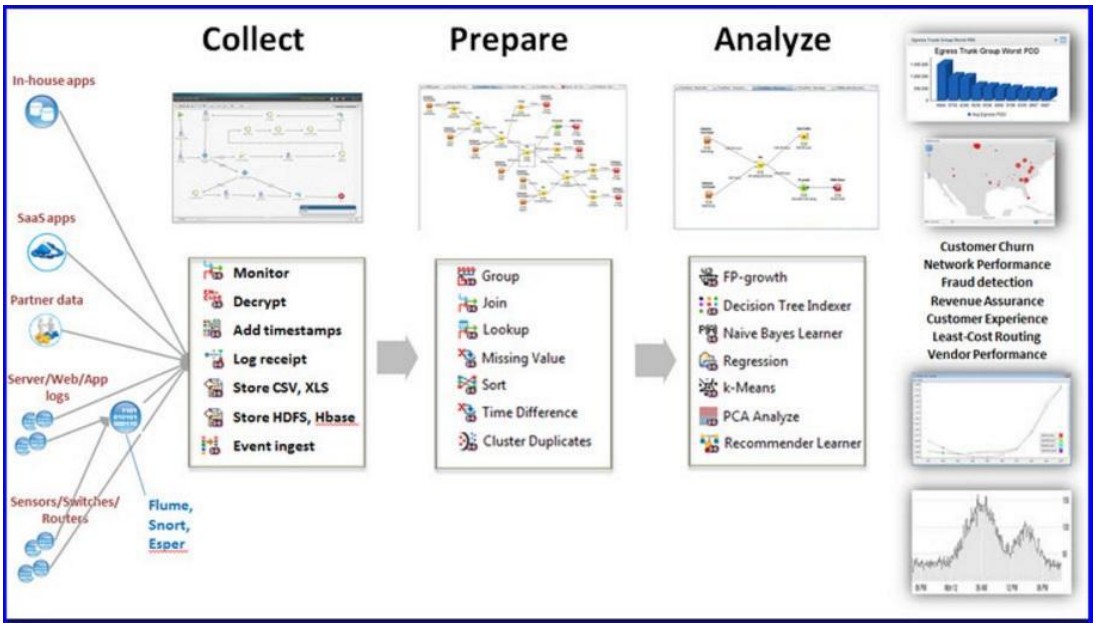
3.3 Μεθοδολογία για την Ανάλυση Δεδομένων

Η Ανάλυση Δεδομένων είναι μία ενοποίηση Διαδικασιών Επιθεώρησης (Investigation), καθαρισμού, μεταμόρφωσης-μετατροπής και μοντελοποίησης των δεδομένων με απώτερο σκοπό την ανακάλυψη και την άντληση χρήσιμης πληροφορίας, προτεινόμενα συμπεράσματα και υποστήριξη στην διαδικασία αποφάσεων. Η Ανάλυση Δεδομένων μπορεί να πραγματοποιηθεί με ποικίλους τρόπους και τεχνικές προσέγγισης, - συνδυάζοντας διάφορες τεχνικές η κάθε μια με διαφορετικά ονόματα σε ξεχωριστούς επιχειρηματικούς κλάδους.

Πριν από την κυρίως χρήση, αυτή της ανάλυσης, στα δεδομένα που θα χρησιμοποιηθούν θα πρέπει ουσιαστικά να υπάρξει κάποια προεργασία με σκοπό να έρθουν στην καταλληλότερη μορφή μέσα από την οποία θα εφαρμοστεί η εκάστοτε ανάλυση που επιθυμούμε. Υπάρχουν πολλές τεχνικές για τα αρχικά στάδια επεξεργασίας που συνήθως ακολουθούν οι αναλυτές. Οι περισσότερες από αυτές είναι απαραίτητες για τον εντοπισμό και την λύση προβλημάτων στα δεδομένα μας, ενώ άλλες είναι χρήσιμες για την μεταμόρφωση προβλημάτων όπου αυτά υπάρχουν για να ετοιμαστούν οι μορφές της ανάλυσης που θα επιλεχθούν από την μεριά του αναλυτή.

Το κυριότερο νόημα όλης αυτής της προεργασίας είναι να καθαριστούν τα δεδομένα και να προετοιμαστούν κατάλληλα για την ανάλυση. Αυτή η διαδικασία αποκαλείται Καθαρισμός Δεδομένων (Data Cleansing). Αρχικά, στο Data Cleansing πραγματοποιείται Διερευνητική Ανάλυση Δεδομένων (Exploratory Data Analysis) η οποία βοηθάει στο να αποκτηθεί μια συνολική εικόνα για τις μεταβλητές μας και την ατομική κατανομή τους. Έπειτα, αν παρατηρηθεί κάποια ανωμαλία κρατείται για το μέρος της ανάλυσης με σκοπό να ελεγχθεί διεξοδικότερα και σε βάθος. Παρατηρείται ότι σε μεγάλα αρχεία δεδομένων με πολλές γραμμές και στήλες υπάρχουν ημιτελή δεδομένα που στις περισσότερες περιπτώσεις πρέπει να τα αποκλειστούν από τις αναλύσεις μας.

Γενικότερα, όλη η μεθοδολογία που ακολουθείται για την Ανάλυση Δεδομένων θα μπορούσε να αποτυπωθεί και μέσω της παρακάτω εικόνας:



Εικόνα 1: Μεθοδολογία Συλλογής, Προετοιμασίας και Ανάλυσης Δεδομένων

Η ανάλυση με σκοπό τις Προβλέψεις και την Ταξινόμηση (Predictive Analytics) πραγματοποιείται κυρίως με την δημιουργία στατιστικών μοντέλων, ενώ η Ανάλυση Κειμένου (Text Analytics) εφαρμόζει περισσότερο στατιστικές, γλωσσικές και δομικές τεχνικές για την εξαγωγή και για την ταξινόμηση πληροφοριών από πηγές κειμένων, ένα είδος αδόμητων δεδομένων. Στις επόμενες σελίδες υπάρχουν λεπτομέρειες για ένα-ένα τα στάδια που περιγράφονται παραπάνω.

3.3.1 Καθαρισμός Δεδομένων

Όπως και να βρει ή να αποκαλέσει κανείς τον Καθαρισμό των Δεδομένων, είτε Data Cleansing, είτε Data Cleaning είτε Data Scrubbing είναι η διαδικασία εντοπισμού και διόρθωσης (ή απαλοιφής) κατεστραμμένων ή ανακριβών εγγραφών στα δεδομένα μας σε ένα πίνακα αλλά ακόμη και σε ολόκληρη βάση δεδομένων. Μπορεί να πραγματοποιηθεί δια-δραστικά με εργαλεία ή μέσω επεξεργασίας συγκεκριμένης παρτίδας ενεργειών.

Όταν ολοκληρωθεί το Καθάρισμα Δεδομένων το Αρχείο Δεδομένων (Dataset) θα πρέπει να έχει μια συνέπεια σε σχέση με τα αντίστοιχα Αρχεία Δεδομένων (Datasets) στο σύστημα. Οι όποιες ανωμαλίες ανιχνεύθηκαν ή μετακινήθηκαν ή διαγράφθηκαν λογικά έχουν προέλθει από

λάθη κατά την εισαγωγή δεδομένων των χρηστών ή από διαφορετική χρήση Data Dictionaries. Το Data Cleansing διαφέρει από το Data Validation το όποιο συνήθως διώχνει τα δεδομένα πριν καν εισαχθούν σε κάποιο αρχείο ή κάποιο πίνακα και πραγματοποιείται είτε κατά την χειροκίνητη εισαγωγή των δεδομένων είτε σε κάποια Batch διαδικασία εισαγωγής δεδομένων.

Η κανονική διαδικασία καθαρισμού δεδομένων μπορεί να εμπεριέχει ακόμη και εργασίες για απαλοιφή τυπογραφικών λαθών και διόρθωση τιμών με βάση κάποια έτοιμη λίστα.

3.3.2 Διερευνητική Ανάλυση Δεδομένων

Αφού τελειώσει το Καθάρισμα Δεδομένων και έρθουν στα χέρια μας στην κατάλληλη επιθυμητή μορφή, τότε ξεκινάει η εργασία μας πάνω σε αυτά από το πιο απλό, δηλαδή κοιτώντας τα δεδομένα πως είναι και αν μπορεί να βγει κάποιο συμπέρασμα στη μορφή που είναι. Αυτό δεν είναι μια τόσο δομημένη και με βήματα διαδικασία αλλά βοηθάει στο να γίνει κατανοητή πλήρως η σημασία και οι μετρικές της κάθε μεταβλητής, αποκτώντας έτσι μια πιο καλή αίσθηση των δεδομένων που έχουμε συλλέξει. Είναι πολύ σημαντικό να αντιληφθεί ο αναλυτής τι απεικονίζει το κάθε στοιχείο στα δεδομένα που υπάρχουν στην κατοχή μας και τι αντιπροσωπεύει. Είναι εξίσου σημαντικό όμως να ελέγξουμε την συνοχή και την συνέχεια των δεδομένων μας κάνοντας μια έρευνα πλοήγησης και απεικόνισης μέσω γραφημάτων.

Μια καλή πρακτική είναι να χρησιμοποιηθεί η συνάρτηση summary()της R στα δεδομένα ή σε ένα σύνολο αυτών για να αποκτηθούν διάφορες μετρικές για την κάθε μεταβλητή όπως μέση τιμή, διαφορά αλλά και μεγαλύτερες ή μικρότερες τιμές. Μερικές φορές είναι αρκετά εύκολο να εντοπιστεί ένα λάθος στην συλλογή των δεδομένων ακόμη και στην ροή εργασιών για την συλλογή μέσω αυτής της διαδικασίας. Επιπλέον, καλή ιδέα θα ήταν να μελετηθεί αν υπάρχουν μεταβλητές με αρκετό θόρυβο. Αυτό θα μπορούσε να οδηγήσει ακόμη και σε διαφορετικές επιλογές ανάλυσης ή θα μπορούσε και να σημαίνει ότι αυτό το στοιχείο πρέπει να αγνοηθεί.

Σημαντικό βήμα στην διαδικασία της Διερευνητικής Ανάλυσης Δεδομένων (Exploratory Data Analysis - EDA) είναι ο αναλυτής να χρησιμοποιήσει γραφήματα με σκοπό την απεικόνιση των δεδομένων. Υπάρχουν μια σειρά από γραφήματα ενσωματωμένα σε μορφή πίνακα που μπορεί να χρησιμοποιηθεί ανάλογα με το περιεχόμενο φυσικά. Για παράδειγμα, μπορεί να χρειαστεί να κατηγοριοποιηθούν τα Γραφήματα Κουτιού (Box Plot) με αριθμητικά στοιχεία με κύριο σκοπό την απεικόνιση διαστημάτων και ποσοτήτων. Τα Γραφήματα Μπάρας (Bar Plots ή Mosaic Plots) είναι χρήσιμα για την απεικόνιση ποσοστών στα δεδομένα με διάφορους

συνδυασμούς των μεταβλητών που υπάρχουν στην κατοχή μας. Δεν θα ειπωθούν παραπάνω λεπτομέρειες για την απεικόνιση μέσω γραφημάτων διότι αυτό ανήκει στην κατηγορία Απεικόνισης (Visualization).

3.3.3 Ελλείπουσες τιμές

Μερικές φορές τυχαίνει στα δεδομένα να υπάρχουν τιμές οι οποίες λείπουν (Missing Values) κατά πάσα πιθανότητα γιατί όταν γινόταν η εισαγωγή των δεδομένων δεν υπήρχε η δυνατότητα συμπλήρωσης και ούτε κάποιος κανόνας που να απαγορεύει στην συνέχεια την συμπλήρωση σε επόμενο στάδιο μέχρι να συμπληρωθεί ορθά όλο το αρχείο. Συνήθως οι κενές τιμές συμβολίζονται με τις λέξεις “NULL” και “NA”. Γι’ αυτό δεν πρέπει να συγχέουμε την τιμή μηδέν στα δεδομένα μας με τις “NULL” τιμές. Πριν παρθεί κάποια απόφαση για το πως θα γίνει ο χειρισμός των δεδομένων που λείπουν, ειδικά αν πρόκειται να ακολουθηθεί η λογική της διαγραφής αυτών των προβληματικών εγγραφών από τα δεδομένα, πρέπει να αναγνωριστούν οι συγκεκριμένες τιμές και αν αυτές ακολουθούν κάποιο μοτίβο. Υπάρχουν πολλές τεχνικές και μέθοδοι για τον χειρισμό των τιμών που λείπουν αλλά δεν θα γίνει παραπάνω αναφορά διότι σε σπάνιες περιπτώσεις υλοποιείται με άλλο τρόπο πέρα από το να διαγραφούν. .

3.3.4 Ανάλυση Δεδομένων

Η Εξόρυξη Δεδομένων (Data Mining) είναι ένα συγκεκριμένο είδος ανάλυσης που στοχεύει στα μοντέλα και την δυνατότητα προβλέψεων παρά για περιγραφικούς λόγους. Πιο αναλυτικά, είναι η διαδικασία ανάλυσης κρυφών μοτίβων στα δεδομένα με σκοπό την άντληση χρήσιμων πληροφορίων, οι οποίες συλλέγονται και ενοποιούνται σε κοινόχρηστες αποθήκες δεδομένων. Αφού γίνει συλλογή όλης της πληροφορίας ακολουθεί η χρήση αυτής για αποτελεσματική ανάλυση, με αλγόριθμους Εξόρυξης Δεδομένων, οι οποίοι κατά συνέπεια διευκολύνουν τη λήψη επιχειρηματικών αποφάσεων.

Τα κύρια βήματα που εμπλέκονται σε μια διαδικασία Εξόρυξης Δεδομένων είναι τα εξής:

- Εξαγωγή, μετατροπή και φόρτωση δεδομένων σε αποθήκη δεδομένων
- Αποθήκευση και διαχείριση δεδομένων σε πολυδιάστατες βάσεις δεδομένων
- Πρόσβαση σε δεδομένα από αναλυτές επιχειρήσεων χρησιμοποιώντας τα κατάλληλα λογισμικά όπως αυτά που αναφέρθηκαν σε προηγούμενες ενότητες
- Παρουσίαση Ανάλυσης Δεδομένων σε εύκολα κατανοητές μορφές, όπως γραφήματα

Επιπλέον, η Ανάλυση Δεδομένων (Data Analysis) καλύπτεται και από την Επιχειρηματική Ευφυία (Business Intelligence) η οποία εξαρτάται σε μεγάλο βαθμό από την συγκέντρωση των δεδομένων εστιάζοντας στις επιχειρηματικές πληροφορίες. Στα στατιστικά εργαλεία μερικοί αναλυτές χρησιμοποιούν τις Περιγραφικές και Επιβεβαιωτικές Αναλύσεις Δεδομένων (Descriptive Exploratory Analysis (DEA) και Confirmatory Data Analysis (CDA). Το DEA εμβαθύνει περισσότερο στην ανακάλυψη νέων στοιχείων μέσα στα δεδομένα μας και το CDA στην επιβεβαίωση ή την παραποίηση ήδη υπάρχοντών υποθέσεων.

Επιπρόσθετα, η Ανάλυση Δεδομένων είναι η διαδικασία που συστηματικά εφαρμόζεται στατιστική και/ή λογικές τεχνικές για να περιγράψουν και να απεικονίσουν, να συμπυκνώσουν και να ανακεφαλαιώσουν τα δεδομένα. Σύμφωνα με την Shamoo και την Resnik (2003), οι αναλυτικές μέθοδοι ποικίλουν «παρέχοντας έτσι έναν τρόπο άντλησης επαγωγικών συμπερασμάτων από τα δεδομένα και διάκρισης του σήματος (δηλαδή του φαινομένου του ενδιαφέροντος) από τον θόρυβο (δηλαδή των στατιστικών διακυμάνσεων) που υπάρχουν στα δεδομένα. Όσον αφορά την Ανάλυση Δεδομένων σε ποσοτική έρευνα μπορεί να συμπεριλάβει μεθόδους στατιστικής όπου πολλές φορές σημαίνει μια συνεχή σε εξέλιξη επαναληπτικής διαδικασίας όπου τα δεδομένα συνεχώς συλλέγονται και αναλύονται σχεδόν ταυτόχρονα. Ισχύει ότι οι επιστημονικοί ερευνητές γενικά αναλύουν τα μοτίβα που ίσως κρύβονται πίσω από το σύνολο των δεδομένων κατά την φάση της συλλογής τους, όπως αναφέρεται από τους Savene και Robinson (2004). Η μορφή ανάλυσης καθορίζεται από συγκεκριμένα ποσοτικά στοιχεία (μελέτη πεδίου, εθνογραφική ανάλυση περιεχομένου, βιογραφία, διακριτική έρευνα) και από την μορφή δεδομένων (σημειώσεις, έγγραφα, κασέτες ακουστικής και βιντεοκασέτες). Το πιο σημαντικό συστατικό για την διαβεβαίωση της ακεραιότητας των δεδομένων είναι η ακρίβεια και τα σωστά ερευνητικά ευρήματα. Οι λανθασμένες στατιστικές αναλύσεις παραμορφώνουν τα επιστημονικά ευρήματα και παραπλανούν τους περιστασιακούς αναγνώστες όπως αναφέρεται και από τον Shepard (2002) και ενδέχεται να επηρεάσουν αρνητικά την αντίληψη της έρευνας για το κοινό. Τα ζητήματα ακεραιότητας έχουν εξίσου μεγάλη σημασία και για την ανάλυση μη στατιστικών δεδομένων. Με άλλα λόγια, ο κύριος σκοπός της Ανάλυσης Δεδομένων είναι να εξετασθεί τι προσπαθούν να πουν τα δεδομένα.

Για να πραγματοποιηθεί Ανάλυση Δεδομένων πρέπει να επιλέξουμε ποιο αρχείο δεδομένων θα χρησιμοποιηθεί, αν υπάρχει παραπάνω από ένα και μετά ποιες μεταβλητές θα χρειαστεί να αναλυθούν. Για να το επιτευχθεί αυτό, υπάρχουν πολλοί τρόποι όπως στοχαστικά ή με κάποια έτοιμα πακέτα της R όπως την Ανάλυση Κύριων Στοιχείων (Principal Component Analysis PCA)

ή την κατηγοριοποίηση των σημαντικών μεταβλητών (Singular Value Decomposition-SVD) ή από οποιοδήποτε εργαλείο γίνει η εργασία της ανάλυσης. Αν δημιουργηθεί ένα στατιστικό μοντέλο με γραμμική ή λογική παλινδρόμηση, η καλύτερη μέθοδος για να επιλεχθούν τα καταλληλότερα χαρακτηριστικά είναι κάποια στατιστική μέθοδος επιλογής μεταβλητών όπως η Lasso (που θα αναλυθεί στα επόμενα υπό-κεφάλαια). Στις περισσότερες περιπτώσεις, έτσι και στις Ένοπλες Δυνάμεις είναι απαραίτητο να επιτευχθεί ο κατακερματισμός των βλαβών. Για να επιτευχθεί αυτό, οι περισσότεροι αναλυτές επιλέγουν τη μέθοδο ομαδοποίησης. Τέλος, ο πλέον προτιμώμενος τρόπος για την δημιουργία κανόνων ή για την εύρεση μοτίβων, είναι η ανάλυση των κανόνων σύνδεσης. Ωστόσο, τις περισσότερες φορές αυτό το είδος ανάλυσης θα μπορούσε να γίνει μέχρι τώρα απλά κοιτώντας τα δεδομένα προηγούμενων ετών μέσω απλής παρατήρησης.

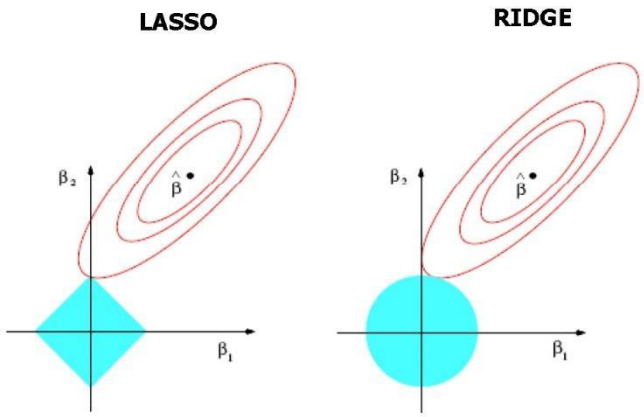
3.3.4.1 Ανάλυση Κύριων Στοιχείων

Ο αλγόριθμος Ανάλυσης Κύριων Στοιχείων (Principal Component Analysis – PCA) δημιουργεί ένα νέο σύνολο μεταβλητών εισόδου που είναι οι γραμμικοί συνδυασμοί των αρχικών μεταβλητών εισόδου. Για το πρώτο βασικό στοιχείο, επιλέγονται τα γραμμικά βάρη συνδυασμού προκειμένου να ληφθεί υπόψη η μέγιστη ποσότητα διακύμανσης στα δεδομένα. Εάν μπορεί να απεικονιστεί η πρώτη κύρια συνιστώσα ως γραμμή στον αρχικό χώρο μεταβλητών, αυτή θα ήταν η γραμμή στην οποία τα δεδομένα ποικίλλουν περισσότερο. Τυχαίνει επίσης να είναι η γραμμή που είναι πιο κοντά σε όλα τα σημεία δεδομένων στον αρχικό χώρο χαρακτηριστικών. Κάθε επόμενη βασική συνιστώσα επιχειρεί να συλλάβει μια γραμμή μέγιστης διακύμανσης, αλλά με τέτοιο τρόπο ώστε το νέο κύριο στοιχείο να μην είναι συνδεδεμένο με τα προηγούμενα που έχουν ήδη υπολογιστεί. Έτσι, το δεύτερο βασικό στοιχείο επιλέγει τον γραμμικό συνδυασμό των αρχικών χαρακτηριστικών εισόδου που έχουν τον υψηλότερο βαθμό διακύμανσης στα δεδομένα, ενώ δεν συνδέονται με το πρώτο κύριο συστατικό.

Τα κύρια συστατικά απεικονίζονται φυσικά σε φθίνουσα σειρά ανάλογα με την ποσότητα διακύμανσης που παρατηρείται. Αυτό επιτρέπει στον αναλυτή να εκτελέσει τη μείωση των διαστάσεων διατηρώντας με έναν απλό τρόπο τα πρώτα στοιχεία N. Επιλέγοντας το N έτσι ώστε τα επιλεγμένα στοιχεία να ενσωματώσουν ένα ελάχιστο ποσό της διακύμανσης από το αρχικό σύνολο δεδομένων.

3.3.4.2 Στατιστική Μέθοδος Επιλογής Μεταβλητών

Η Στατιστική Μέθοδος Επιλογής Μεταβλητών (Least Absolute Shrinkage and Selection Operator – Lasso) είναι μια εναλλακτική μέθοδος κανονικοποίησης για την αύξηση της παλινδρόμησης. Η διαφορά με την Ridge εμφανίζεται μόνο στον όρο της «ποινής», ο οποίος συνεπάγεται την ελαχιστοποίηση του αθροίσματος των απόλυτων τιμών των συντελεστών. Αποδεικνύεται ότι αυτή η διαφορά στον όρο της «ποινής» είναι πολύ σημαντική, καθώς η Lasso συνδυάζει τόσο τη συρρίκνωση όσο και την επιλογή, επειδή συρρικνώνει ορισμένους συντελεστές σε ακριβώς μηδέν, πράγμα που δεν συμβαίνει με την παλινδρόμηση των κορυφών. Παρ 'όλα αυτά, δεν υπάρχει σαφής νικητής μεταξύ αυτών των δύο. Τα μοντέλα που εξαρτώνται από ένα υποσύνολο των χαρακτηριστικών εισόδου θα τείνουν να λειτουργούν καλύτερα με Lasso. Τα μοντέλα που έχουν μια μεγάλη εξάπλωση σε συντελεστές σε πολλές διαφορετικές μεταβλητές θα τείνουν να λειτουργούν καλύτερα με την παλινδρόμηση κορυφογραμμών (Ridge). Συνήθως αξίζει να δοκιμαστούν και τα δύο. Η Lasso είναι η πιο δημοφιλής προσέγγιση που βασίζεται στην ποινικοποίηση. Με αρκετά μεγάλο λ ορίζονται συγκεκριμένοι συντελεστές ακριβώς ίσοι με μηδέν, οπότε η Lasso θα κάνει επιλογή μοντέλου για τους αναλυτές. Η παρακάτω εικόνα 2 δείχνει τον λιγότερο απόλυτο χειριστή συρρίκνωσης και επιλογής.



Εικόνα 2: Ο χειριστής συρρίκνωσης και επιλογής μεταβλητών (Lasso & Ridge)

3.3.4.3 Η αυτόματη μέθοδος επιλογής μεταβλητών για μοντέλα πρόβλεψης (Stepwise)

Ένας ευρέως χρησιμοποιούμενος αλγόριθμος προτάθηκε αρχικά από τον Efroymson (1960). Πρόκειται για μια αυτόματη διαδικασία για την επιλογή στατιστικού μοντέλου σε περιπτώσεις

όπου υπάρχει ένας μεγάλος αριθμός πιθανών επεξηγηματικών μεταβλητών και καμία υποκείμενη θεωρία επί της οποίας θα βασιστεί η επιλογή του μοντέλου. Η διαδικασία χρησιμοποιείται κυρίως στην Ανάλυση Παλινδρόμησης, αν και η βασική προσέγγιση είναι εφαρμόσιμη σε πολλές μορφές επιλογής μοντέλου. Αυτή είναι μια παραλλαγή στην επιλογή προς τα εμπρός. Σε κάθε στάδιο της διαδικασίας, αφού προστεθεί μια νέα μεταβλητή, γίνεται μια δοκιμή για να εξακριβωθεί αν μπορούν να διαγραφούν ορισμένες μεταβλητές χωρίς να αυξηθεί αισθητά το άθροισμα των τετραγώνων (R-Summary Square-RSS). Η διαδικασία τερματίζεται όταν το μέτρο μεγιστοποιείται (τοπικά) ή όταν η διαθέσιμη βελτίωση πέφτει κάτω από κάποια κρίσιμη τιμή.

Ένα από τα βασικά ζητήματα με την σταδιακή παλινδρόμηση είναι ότι αναζητά έναν μεγάλο χώρο πιθανών μοντέλων. Ως εκ τούτου, είναι επιρρεπής στην υπερφόρτωση των δεδομένων. Με άλλα λόγια, η σταδιακή παλινδρόμηση ταιριάζει συχνότερα και πολύ καλύτερα στο δείγμα από ότι στα νέα δεδομένα εκτός δείγματος.

3.3.4.4 Γραμμική και Λογιστική Παλινδρόμηση

Στην Γραμμική Παλινδρόμηση (Linear Regression), η μεταβλητή εξόδου προβλέπεται από ένα γραμμικά σταθμισμένο συνδυασμό χαρακτηριστικών εισόδου. Σε αυτή την εξίσωση, όπως και στην προηγούμενη, μπορεί να παρατηρηθεί ότι υπάρχει ένας ακόμα συντελεστής από τον αριθμό των χαρακτηριστικών. Αυτός ο πρόσθετος συντελεστής, β0, είναι γνωστός ως η τομή και είναι η αναμενόμενη τιμή του μοντέλου όταν η τιμή όλων των χαρακτηριστικών εισόδου είναι μηδέν. Οι άλλοι συντελεστές β μπορούν να ερμηνευθούν ως η αναμενόμενη μεταβολή της τιμής της παραγωγής ανά μονάδα αύξησης ενός χαρακτηριστικού. Σε ένα απλό πρόβλημα μονοδιάστατης παλινδρόμησης, μπορεί να σχεδιαστεί η έξοδος στον άξονα y ενός γραφήματος και η λειτουργία εισαγωγής στον άξονα x. Σε αυτή την περίπτωση, το μοντέλο προβλέπει μια ευθεία σχέση μεταξύ αυτών των δύο, όπου β0 αντιπροσωπεύει το σημείο στο οποίο η ευθεία γραμμή διασχίζει ή παρεμποδίζει τον άξονα y και β1 αντιπροσωπεύει την κλίση της γραμμής. Συχνά γίνεται αναφορά στην περίπτωση ενός μόνο χαρακτηριστικού (ως εκ τούτου δύο συντελεστές παλινδρόμησης) ως Απλή Γραμμική Παλινδρόμηση και την περίπτωση δύο ή περισσοτέρων χαρακτηριστικών ως Πολλαπλή Γραμμική Παλινδρόμηση.

Παρόλο που είναι γνωστό ότι τα προβλήματα ταξινόμησης περιλαμβάνουν ποιοτικά αποτελέσματα, είναι φυσικό να αναρωτιέται κανείς αν θα μπορούσε να χρησιμοποιήσει τις

υπάρχουσες γνώσεις γραμμικής παλινδρόμησης και να τις εφαρμόσει στη ρύθμιση ταξινόμησης. Κάτι τέτοιο θα μπορούσε να γίνει, εκπαιδεύοντας ένα μοντέλο γραμμικής παλινδρόμησης για να γίνει πρόβλεψη μιας τιμής στο διάστημα [0,1], απλά κρατώντας ότι έγινε επιλογή του χαρακτηρισμού με τις δύο τάξεις μας ως 0 και 1. Κάθε φορά που υπάρχουν περισσότερα από ένα χαρακτηριστικά εισαγωγής και υπάρχει επιθυμία για δημιουργία ενός μοντέλου γραμμικής παλινδρόμησης, θα πρέπει να γίνεται επιλογή της Πολλαπλής Γραμμικής Παλινδρόμησης. Η γενική εξίσωση για ένα μοντέλο Πολλαπλής γραμμικής με χαρακτηριστικά εισόδου k είναι:

$$y = \beta_k x_k + \beta_{k-1} x_{k-1} + \cdots + \beta_1 x_1 + \beta_0 + \varepsilon$$

Εικόνα 3: Τύπος γραμμικής παλινδρόμησης

Η Γραμμική Παλινδρόμηση είναι το κατάλληλο μοντέλο για την παροχή μιας γραμμικής σχέσης μεταξύ εξαρτημένων και επεξηγηματικών μεταβλητών, της κανονικότητας και της ομοσκεδαστικότητας των υπολειμμάτων, αλλά δυστυχώς δεν είναι τόσο βολικό μοντέλο για το σύνολο δεδομένων μας, αλλά για το σκοπό αυτής της εργασίας θα γίνει προσπάθεια να βρεθεί το καλύτερο πρότυπο πρόβλεψης με τη βοήθεια αυτής.

Η Λογιστική Παλινδρόμηση (Logistic Regression)υπερτερεί γιατί σε αντίθεση με τη γραμμική παλινδρόμηση ή άλλα στατιστικά μοντέλα, δεν χρειάζεται να γίνουν υποθέσεις για το μοντέλο μας. Η Λογιστική Παλινδρόμηση αντιμετωπίζει όλα αυτά τα σημεία παρέχοντας μια έξοδο που περιορίζεται από το διάστημα [0,1] και εκπαιδεύεται χρησιμοποιώντας ένα εντελώς διαφορετικό κριτήριο βελτιστοποίησης από την γραμμική παλινδρόμηση, οπότε δεν μπορεί πλέον να εφαρμοστεί μια λειτουργία ελαχιστοποιώντας το μέσο τετραγωνικό σφάλμα Median Square Error-MSE. Η μορφή της λειτουργίας της Λογιστικής Παλινδρόμησης έχει ως εξής:

$$f(x) = \frac{e^x}{e^x + 1} = \frac{e^{-x}}{e^{-x}} \cdot \frac{e^x}{\left(e^x + 1\right)} = \frac{1}{1 + e^{-x}}$$

Εικόνα 4: Τύπος λογιστικής παλινδρόμησης

3.3.4.5 Ομαδοποίηση

Η Ομαδοποίηση (Clustering) είναι μια μέθοδος Εξόρυξης Δεδομένων που αναλύει ένα συγκεκριμένο σύνολο δεδομένων και τα οργανώνει με βάση παρόμοια χαρακτηριστικά. Η ομαδοποίηση μπορεί να πραγματοποιηθεί με σχεδόν οποιοδήποτε είδος οργανωμένου ή ημι-

οργανωμένου συνόλου δεδομένων, συμπεριλαμβανομένων κειμένων, εγγράφων, συνόλων αριθμών, απογραφών ή δημογραφικών δεδομένων κλπ. Η βασική ιδέα είναι το σύμπλεγμα, το οποίο είναι μια ομάδα παρόμοιων αντικειμένων. Οι συστοιχίες μπορούν να είναι οποιουδήποτε μεγέθους θεωρητικά, ένα σύμπλεγμα μπορεί να έχει μηδενικά αντικείμενα εντός αυτού ή ολόκληρο το σύνολο δεδομένων μπορεί να είναι τόσο παρόμοιο ώστε κάθε αντικείμενο να πέφτει στο ίδιο σύμπλεγμα. Αυτό θα ήταν σπάνιο διότι είναι πιο συχνό, τα αντικείμενα να συσσωρεύονται λόγω μαθηματικών και στατιστικών ομοιοτήτων. Στην περίπτωση της Ανάλυσης Κειμένου, τα αντικείμενα συχνά συσσωρεύονται λόγω των λέξεων-κλειδιών, της φράσης και του θέματος / πλαισίου. Η Ανάλυση Συμπλέγματος ομαδοποιεί αντικείμενα δεδομένων που βασίζονται μόνο σε πληροφορίες που βρίσκονται στα δεδομένα που περιγράφουν τα αντικείμενα και τις σχέσεις τους. Ο στόχος της είναι τα αντικείμενα μιας ομάδας να σχετίζονται μεταξύ τους και να διαφέρουν από τα αντικείμενα σε άλλες ομάδες. Όσο μεγαλύτερη είναι η ομοιότητα μέσα σε μια ομάδα και τόσο μεγαλύτερη είναι η διαφορά μεταξύ των ομάδων και τόσο καλύτερη ή πιο διακριτή είναι η Ομαδοποίηση. Υπάρχουν πολλοί αλγόριθμοι για την ομαδοποίηση και κάθε αναλυτής πρέπει να επιλέξει ποιο είναι κατάλληλο κάθε φορά για τους σκοπούς του. Ακόμη, ο στόχος είναι να βρεθούν συστάδες / ομάδες ή παρατηρήσεις με την ιδιότητα που μέσα στο σύμπλεγμα οι παρατηρήσεις "μοιάζουν όμοιες", αλλά οι παρατηρήσεις από διαφορετικά συμπλέγματα είναι "διαφορετικές". Υπάρχουν δύο μέθοδοι που χρησιμοποιούνται ευρέως σε άλλους κλάδους όπως η εκμάθηση μηχανών και η εξόρυξη δεδομένων. Η μη επιτηρούμενη μάθηση όπου δεν είναι γνωστή η επισήμανση των παρατηρήσεων και γίνεται προσπάθεια προσδιορισμού και της εποπτευόμενης μάθησης που ανήκουν σε μεθόδους ταξινόμησης. Υπάρχουν πολλοί τύποι ομαδοποίησης και αλγόριθμοι που χρησιμοποιούνται για διαφορετικούς τύπους δεδομένων όπως:

- Ιεραρχική Ομαδοποίηση (Hierarchical clustering)
- Ομαδοποίηση βασισμένη σε κεντροειδή - (Centroid-based clustering (K-Means))
- Ομαδοποίηση βάσει διανομής (Distribution-based Clustering)
- Ομαδοποίηση με βάση την Πυκνότητα (Density-based Clustering)
- Ομαδοποίηση με Μερική Συσσώρευση (Partitional Clustering)
- Ομαδοποίηση Φάσματος (Spectral Clustering)
- Ομαδοποίηση με βάση το Δίκτυο (Grid-based Clustering)
- Ομαδοποίηση Συσχέτισης (Correlation Clustering)
- Ομαδοποίηση Βάρους (Gravitational Clustering)
- Ομαδοποίηση Γεωγραφικού Τμήματος (Herd Clustering)

Οι περισσότεροι αλγόριθμοι είναι ιεραρχικοί. Στους συγκεντρωτικούς ιεραρχικούς αλγόριθμους τα δυο σύνολα που ταιριάζουν περισσότερο ενώνονται για να δημιουργήσουν ένα μεγαλύτερο Cluster σε κάθε βήμα και συνεχίζεται μέχρι να ολοκληρωθεί ο επιθυμητός αριθμός των Clusters. Στους διαχωριστικούς ιεραρχικούς αλγόριθμους η διαδικασία αντιστρέφεται, διότι ξεκινάει με δυο Data Points σε ένα Cluster και τα διαιρεί σε μικρότερα Clusters. Και στις δυο περιπτώσεις πρέπει να μετρήσουμε την απόσταση μεταξύ ενός αντικειμένου και της ομάδας(Cluster) και την απόσταση δυο ομάδων(Clusters).

Τα βασικά πλεονεκτήματα των Ιεραρχικών Μεθόδων είναι τα ακόλουθα:

- Οι ιεραρχικές μέθοδοι παρουσιάζουν καλή προσαρμοστικότητα. Μπορούν να εντοπίσουν καλά διαχωρισμένες, επιμήκεις και ομόκεντρες συστάδες.
- Δημιουργούν πολλαπλά επίπεδα φωλιασμένων συστάδων και επιτρέπουν στον χρήστη να επιλέξει το επίπεδο που αυτός επιθυμεί.

Τα μειονεκτήματα των Ιεραρχικών μεθόδων είναι τα εξής:

- Κάθε ενέργεια, η οποία πραγματοποιείται σε ένα στάδιο, δεν είναι αντιστρέψιμη. Από τη στιγμή που δύο αντικείμενα ενταχθούν στην ίδια ομάδα, θα παραμείνουν στην ίδια ομάδα, και δεν υπάρχει δυνατότητα να διαχωριστούν αργότερα και να ενταχθούν σε διαφορετικές ομάδες.
- Οι ιεραρχικές μέθοδοι χρειάζεται να ελέγξουν πολλές αποστάσεις, και για τον λόγο αυτό καθυστερούν όταν χρειάζεται να επεξεργαστούν μεγάλο αριθμό αντικειμένων.

Σε κάθε συστάδα τα σημεία που περιέχονται σε αυτή παρουσιάζουν ομοιότητα μεταξύ τους. Έτσι για όλες τις τεχνικές της Ανάλυσης Συστάδων είναι σημαντικό να ορίζεται ένα μέτρο ομοιότητας μεταξύ δύο αντικειμένων από το χώρο δεδομένων. Με τη μεγάλη ποικιλία στα χαρακτηριστικά γνωρίσματα η επιλογή του μέτρου ομοιότητας θα πρέπει να είναι προσεγμένη. Σε πολλές περιπτώσεις αυτό το μέτρο ομοιότητας που συνήθως μετράται δεν είναι η ομοιότητα αλλά η διαφορετικότητα δυο σημείων.

Οι Τεχνικές αλγορίθμων Ανάλυσης Συστάδων είναι οι εξής:

- Συγκεντρωτικοί και Διαχωριστικοί (Agglomerative and Divisive): Η διαφοροποίηση των ειδών αυτών σχετίζεται με την λειτουργία και τις δομές του αλγορίθμου. Στην πρώτη

περίπτωση ο αλγόριθμος ξεκινά θεωρώντας κάθε στοιχείο σαν μια ξεχωριστή συστάδα και προχωρά συγχωνεύοντας στοιχεία και συστάδες μέχρις ότου να ικανοποιηθεί μια συνθήκη. Στην περίπτωση ενός Διαχωριστικού Αλγορίθμου, όλα τα στοιχεία θεωρούνται ότι ανήκουν σε μια συστάδα και ακολουθείται μια συνεχής διάσπαση της συστάδας αυτής σε μικρότερες μέχρις ότου να ικανοποιηθεί η συνθήκη τερματισμού.

- Μονοθετικοί και Πολυθετικοί (Monothetic and Polythetic): Η διαφορά αυτών χαρακτηρίζει την σειριακή ή ταυτόχρονη χρήση των χαρακτηριστικών των στοιχείων κατά την διαδικασία της Ανάλυσης Συστάδων. Οι περισσότεροι αλγόριθμοι είναι Πολυθετικοί, κάτι που σημαίνει ότι όλα τα χαρακτηριστικά των στοιχείων συμμετέχουν κάθε φορά στον καθορισμό της απόστασης του στοιχείου από κάποιο άλλο. Ένας Μονοθετικός Αλγόριθμος λαμβάνει υπόψη του μονό ένα χαρακτηριστικό τη φορά και πραγματοποιεί ομαδοποιήσεις με βάση αυτό το χαρακτηριστικό. Σε επόμενη επανάληψη χρησιμοποιεί άλλο χαρακτηριστικό και διαχωρίζει τις ήδη υπάρχουσες ομάδες.
- Σκληροί και Ασαφείς (Hard and Fuzzy): Ένας Σκληρός Αλγόριθμος τοποθετεί κάθε στοιχείο σε ένα και μόνο Cluster, σε αντίθεση με τους Fuzzy αλγορίθμους οι οποίοι δίνουν σε κάθε στοιχείο για κάθε Cluster έναν βαθμό που εκφράζει κατά πόσο το στοιχείο αυτό ανήκει στο cluster αυτό.
- Ντετερμινιστικοί και Στοχαστικοί (Deterministic and Stochastic): Αυτοί οι αλγόριθμοι είναι κυρίως διαιρετικοί και σχετίζονται με την βελτιστοποίηση της ομαδοποίησης.
- Αυξητικοί και μη Αυξητικοί (Incremental and non-Incremental): Η διαφορά αυτών των αλγορίθμων εμφανίζεται όταν το σύνολο των δεδομένων προς ομαδοποίηση είναι πολύ μεγάλο και περιορισμοί που υπάρχουν στον χρόνο εκτέλεσης και τον διαθέσιμο χώρο μνήμης επηρεάζουν την αρχιτεκτονική του αλγορίθμου. Στα πρώτα βήματα της θεωρίας περί Ανάλυσης Συστάδων τα δεδομένα δεν ήταν ιδιαιτέρα πολλά άρα δεν υπήρχαν και προβλήματα με το μέγεθος της πληροφορίας. Με την αύξηση όμως της πληροφορίας υπήρχε η ανάγκη για εύρεση αλγορίθμων οι οποίοι ελαχιστοποιούν τον αριθμό σαρώσεων των δεδομένων, μειώνουν τον αριθμό των στοιχείων που εξετάζονται η μειώνουν το μέγεθος των δομών που χρησιμοποιούνται κατά την εκτέλεση του αλγορίθμου.

3.3.4.6 Κανόνες Συσχέτισης (Association Rules)

Η εκμάθηση κανόνων σύνδεσης είναι μια μέθοδος μάθησης βασισμένη σε κανόνες για την ανακάλυψη ενδιαφερόντων σχέσεων μεταξύ μεταβλητών σε μεγάλες βάσεις δεδομένων. Σκοπός του είναι να εντοπίσει τους ισχυρούς κανόνες που ανακαλύπτονται σε βάσεις δεδομένων χρησιμοποιώντας ορισμένα μέτρα ενδιαφέροντος. Με βάση την έννοια των ισχυρών κανόνων, ο Rakesh Agrawaletal(1993) εισήγαγε κανόνες σύνδεσης για την ανακάλυψη κανονικότητας μεταξύ προϊόντων σε δεδομένα μεγάλης κλίμακας συναλλαγών που καταγράφηκαν από Συστήματα Σημείου Πώλησης (POS) στα σούπερ μάρκετ. Για παράδειγμα, ο κανόνας που υπάρχει στα δεδομένα πωλήσεων ενός σούπερ μάρκετ θα έδειχνε ότι εάν ένας πελάτης αγοράσει μαζί τα κρεμμύδια και τις πατάτες, είναι πιθανό να αγοράσουν και κρέας από χάμπουργκερ. Αυτές οι πληροφορίες μπορούν να χρησιμοποιηθούν ως βάση για πολλές αποφάσεις σχετικά με δραστηριότητες μάρκετινγκ, όπως προωθητικές τιμές ή τοποθετήσεις προϊόντων. Εκτός από το παραπάνω παράδειγμα από την ανάλυση καλαθιού αγοράς, οι κανόνες σύνδεσης χρησιμοποιούνται σήμερα σε πολλούς τομείς εφαρμογής, όπως εξόρυξη χρήσης ιστού, ανίχνευση εισβολών, συνεχής παραγωγή και βίο-πληροφορική. Σε αντίθεση με την εξόρυξη αλληλουχιών, η μάθηση κανόνων σύνδεσης συνήθως δεν λαμβάνει υπόψη τη σειρά των αντικειμένων είτε μέσα σε μια συναλλαγή είτε σε διάφορες συναλλαγές. Υπάρχουν πολλοί αλγόριθμοι από τους Association rules, αλλά οι πιο σημαντικοί είναι οι εξής:

- Apriori αλγόριθμος
- Elcat αλγόριθμος
- FP-Growth αλγόριθμος

Ο πιο συνηθισμένος και προτιμότερος είναι ο Apriori. Πιο πρακτικά, για να γίνει καλύτερη κατανόηση στον τρόπο με τον οποίο λειτουργεί ο αλγόριθμος πρέπει να γίνει αντιληπτό ότι η δύναμη της σύνδεσης μετριέται από την υποστήριξη και την εμπιστοσύνη του κανόνα. Η υποστήριξη για τον κανόνα A□B είναι η πιθανότητα να εμφανιστούν τα δύο σύνολα στοιχείων μαζί. Η υποστήριξη του κανόνα A □ B εκτιμάται από τα ακόλουθα:

$$\frac{transactions\ that\ contain\ every\ item\ in\ A\ and\ B}{all\ transactions}$$

Εικόνα 5: Τύπος Υποστήριξης για Apriori

Η Υποστήριξη (Support) είναι συμμετρική, πράγμα που σημαίνει ότι η Υποστήριξη(Support) του κανόνα A□B είναι η ίδια με την υποστήριξη του κανόνα B□A. Η εμπιστοσύνη ενός κανόνα σύνδεσης είναι A□B είναι η υπό όρους πιθανότητα ενός στοιχείου που περιέχει τα στοιχεία B δεδομένου ότι περιέχει σύνολο στοιχείων A. Η Εμπιστοσύνη υπολογίζεται από τα ακόλουθα:

$$\frac{transactions\ that\ contain\ every\ item\ in\ A\ and\ B}{transactions\ that\ contain\ the\ items\ in\ A}$$

Εικόνα 6: Τύπος Εμπιστοσύνης για Apriori

Η ερμηνεία της επίπτωσης (□) των κανόνων σύνδεσης είναι επισφαλής. Η υψηλή Εμπιστοσύνη και Υποστήριξη δεν συνεπάγεται αιτία και αποτέλεσμα. Τα δύο στοιχεία ίσως να μην συσχετιστούν. Ο όρος εμπιστοσύνης δεν σχετίζεται με τη στατιστική χρήση, συνεπώς, δεν υπάρχει επαναλαμβανόμενη ερμηνεία δειγματοληψίας.

3.3.5 Απεικόνιση Αποτελεσμάτων

Η οπτικοποίηση δεδομένων είναι η διαδικασία μετατροπής των δεδομένων σε εύκολα κατανοητές εικόνες πληροφοριών που επιτρέπουν γρήγορες και αποτελεσματικές αποφάσεις. Στις αρχές του 20ου αιώνα, οι ψυχολόγοι παρατήρησαν ότι όταν τα στοιχεία συγκεντρώνονταν σε μια εικόνα, ο αριθμός πήρε μια αντιληπτική σημασία που υπερέβαινε το άθροισμα των τμημάτων του. Η αποτελεσματική απεικόνιση βοηθά τους χρήστες στην ανάλυση, τη συλλογιστική και την τεκμηρίωση σχετικά με τα δεδομένα και τα αποδεικτικά στοιχεία, καθιστά πολύπλοκα δεδομένα πιο προσιτά, κατανοητά και χρησιμοποιήσιμα. Οι πίνακες χρησιμοποιούνται γενικά όπου οι χρήστες θα αναζητήσουν ένα συγκεκριμένο μέτρο μιας μεταβλητής, ενώ διαγράμματα διαφόρων τύπων χρησιμοποιούνται για την εμφάνιση μοτίβων και σχέσεων στα δεδομένα για μία ή περισσότερες μεταβλητές.

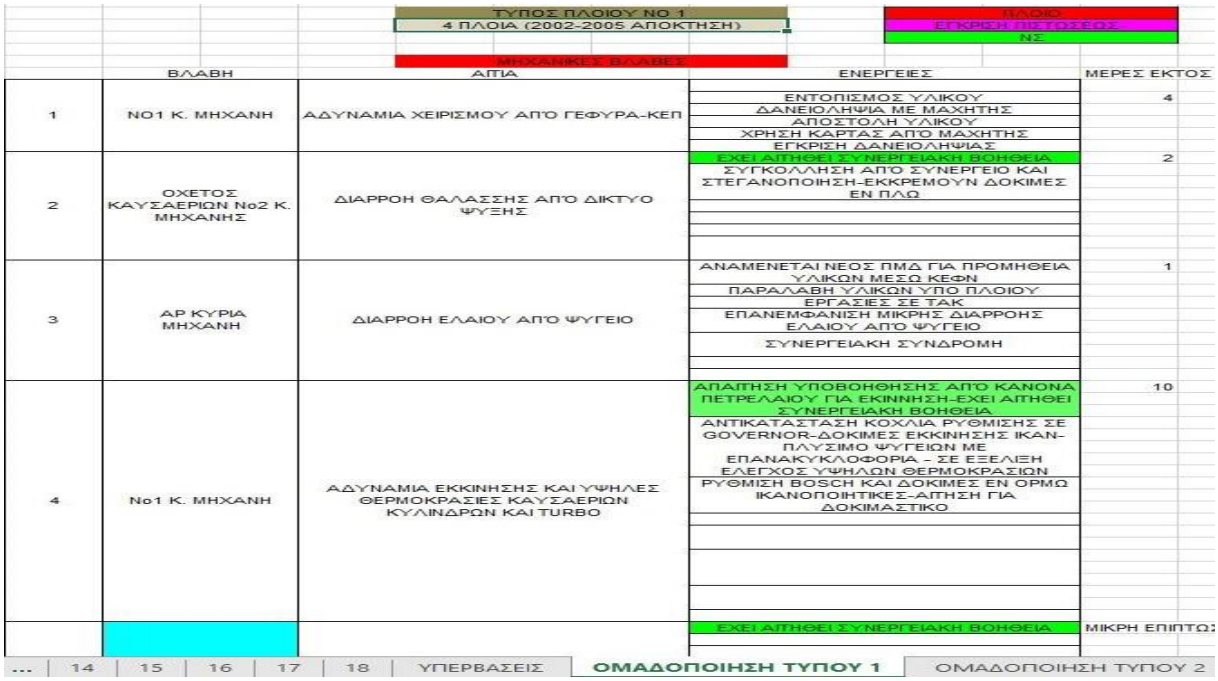
Ένας πρωταρχικός στόχος της απεικόνισης δεδομένων είναι η επικοινωνία των πληροφοριών με σαφήνεια και αποτελεσματικότητα μέσω στατιστικών γραφημάτων,

διαγραμμάτων και γραφικών πληροφοριών. Τα αριθμητικά δεδομένα μπορούν να κωδικοποιούνται χρησιμοποιώντας κουκκίδες, γραμμές ή ράβδους, για την οπτική επικοινωνία ενός ποσοτικού μηνύματος. Κανονικά η R παρέχει και αυτήν την δυνατότητα γι' αυτό και θα χρησιμοποιηθεί στην παρούσα διπλωματική εργασία αλλά δεν είναι το καταλληλότερο εργαλείο για την απεικόνιση αποτελεσμάτων (Visualization of Results).

3.4 Μεθοδολογία που ακολουθήθηκε για αυτήν την ανάλυση

Στην περίπτωση αυτής της μεταπτυχιακής διατριβής χρησιμοποιήθηκαν δεδομένα από το ΠΝ που αφορούν τις βλάβες μερίδας πλοίων του στόλου της χώρας που ανήκουν στη Διοίκηση Πλοίων Επιτήρησης, φυσικά αποκρύπτοντας στοιχεία που θα μπορούσαν να χαρακτηριστούν απόρρητα και η διαρροή τους θα έβλαπτε την ακεραιότητα την άμυνα και την εθνική ασφάλεια της χώρας. Λόγω των ευαίσθητων και διαβαθμισμένων δεδομένων έγινε μια τροποποίηση σε αυτά ώστε να είναι μεν αληθοφανή με βάση τις πραγματικές συνθήκες μα τροποποιημένα κατάλληλα ποιοτικά και ποσοτικά ώστε να μην εκτίθενται μετρήσιμα μεγέθη που μπορούν να χρησιμοποιηθούν για εξαγωγή συμπερασμάτων προς την ακριβή κατάσταση της άμυνας της χώρας . Ακόμη και οι ονομασίες των πλοίων, αλλάξαν με κεφαλαία γράμματα του αγγλικού αλφάβητου με σκοπό να μην αποκαλυφθούν πληροφορίες επειδή αυτές θεωρούνται διαβαθμισμένες. Η ίδια τροποποίηση ίσχυσε και με την αλλαγή στον τύπο των πλοίων που ονομάστηκαν «Ομαδοποίηση 1 έως 5» αντίστοιχα. Συνεπώς τα δεδομένα αυτά αφορούν πλοία τα όποια είναι μέρος μιας συγκεκριμένης διοίκησης με βάση το τύπο και την αποστολή τους. Η πληροφορία αυτή συλλέγεται σε μορφή Excel από τα αρμόδια τμήματα, φυσικά με τις όποιες γνώσεις κατέχουν τα άτομα που συμπεριλαμβάνονται σε αυτά αγνοώντας συχνά την πραγματική σημασία που κρύβει μια σωστή καταγραφή των βλαβών αλλά και την αξία που θα κέρδιζαν αν γνώριζαν ποιο πλοίο εμφανίζει μεγαλύτερη συχνότητα βλαβών αλλά και κάθε πότε. Σωστή καταγραφή σημαίνει να αποτυπώνεται όσον το δυνατόν περισσότερη πληροφορία μπορεί να συλλεχθεί στο Excel ή σε κάποια άλλη βάση δεδομένων έτσι ώστε να μπορεί να αξιοποιείται από αναλυτές ευκολά και γρήγορα με σκοπό την εγρήγορση του ΠΝ. Αξίζει να σημειωθεί ότι η αρχική πληροφορία υπάρχει σε ένα σύνολο από αρχεία Excel αλλά για λόγους ασφαλείας στην παρακάτω εικόνα απεικονίζεται ένα από αυτά. Με την συνεργασία των αρμόδιων τμημάτων του ΠΝ συλλέχθηκε η πληροφορία και από άλλα αρχεία καταγραφής, ενώθηκε σε ένα τελικό Excel το οποίο χρησιμοποιήθηκε και για όλες τις αναλύσεις της συγκεκριμένης διπλωματικής εργασίας. Η αρχική πληροφορία υπήρχε σε πληθώρα αρχείων ασύνδετων μεταξύ τους και μη ομαδοποιημένων

και χωρίς κάποιο κοινό τρόπο καταγραφής, συνεπώς έγινε μια μεγάλη προσπάθεια, χρονοβόρα και κοπιαστική , συγκέντρωσης όλων των παρεχόμενων πληροφοριών και δημιουργήθηκε ένα ενιαίο αρχείο Excel οπού και θα υπάρχει όλη η απαραίτητη πληροφορία για το σύνολο των βλαβών σε αξιοποιήσιμη μορφή απο τη γλώσσα R. Δυστυχώς δεν υπάρχει ακόμη κάποιο αρμόδιο τμήμα επιφορτισμένο μα την καταγραφή βλαβών ούτε κάποια γενική οδηγία για τον τρόπο που αυτή θα γίνεται και μέσω αυτής της πτυχιακής εργασίας δίνεται το έναυσμα για να ξεκινήσουν τέτοιες ενέργειες.



Εικόνα 7: Αρχική μορφή αρχείου Excel

Όπως είναι εμφανές απο τις διαφορές των αρχείων Excel των εικόνων 7 και 8, γίνεται κατανοητό πόσο σημαντική είναι η σωστή μορφοποίηση και καταγραφή και πόση προσπάθεια απαιτήθηκε για να φτάσουμε στη τελική μορφή. Όπως είναι φυσικό η τελική μορφή του αρχείου είναι και η ιδανικότερη για την συλλογή των δεδομένων έτσι ώστε να βελτιώνεται συνεχώς και να μπορεί να αποθηκευτεί ολοένα και περισσότερη πληροφορία που θα βοηθήσει να αποκτήσει παραπάνω γνώση προς όφελος του ΠΝ.

1	TYPOS PLOIOY	PLOIOY	APOKTHSH	PROSWPIK	BLABH	AITIA	BLABHS	BLABH	ENERGEI	ENERGEIES	EKTO
614	ΟΜΑΔΟΠΟΙΗΣΗ 3	G	1990	48	No2 K. ΜΗΧΑΝΗ	ΕΞΩΤΕΡΙΚΗ ΔΙΑΡΡΟΗ Α3-Α4 ΚΑΙ Α7-Α8 ΚΥΛΙΝΔΡΩΝ	ΜΗΧΑΝΙΚΕΣ	2006	1	ΕΧΕΙ ΑΙΤΗΘΕΙ ΣΥΝΕΡΓΕΙΑΚΗ ΒΟΗΘΕΙΑ	5
615	ΟΜΑΔΟΠΟΙΗΣΗ 3	G	1990	48	No2 K. ΜΗΧΑΝΗ	ΕΞΩΤΕΡΙΚΗ ΔΙΑΡΡΟΗ Α3-Α4 ΚΑΙ Α7-Α8 ΚΥΛΙΝΔΡΩΝ	ΜΗΧΑΝΙΚΕΣ	2006	2	ΕΓΚΡΙΣΗ ΑΡΣΗΣ ΦΡΑΓΜΟΥ ΥΛΙΚΑ ΑΠΟΚ. ΔΙΑΡΡΟΩΝ	5
616	ΟΜΑΔΟΠΟΙΗΣΗ 3	G	1990	48	No2 K. ΜΗΧΑΝΗ	ΕΞΩΤΕΡΙΚΗ ΔΙΑΡΡΟΗ Α3-Α4 ΚΑΙ Α7-Α8 ΚΥΛΙΝΔΡΩΝ	ΜΗΧΑΝΙΚΕΣ	2006	3	ΑΡΣΗ ΦΡΑΓΜΟΥ ΣΤΕΓΑΝΟΠΟΙΗΤΙΚΑ	5
617	ΟΜΑΔΟΠΟΙΗΣΗ 3	G	1990	48	No2 K. ΜΗΧΑΝΗ	ΕΞΩΤΕΡΙΚΗ ΔΙΑΡΡΟΗ Α3-Α4 ΚΑΙ Α7-Α8 ΚΥΛΙΝΔΡΩΝ	ΜΗΧΑΝΙΚΕΣ	2006	4	ΑΡΣΗ ΦΑΡΤΜΟΥ SEAL-BOLT	5
618	ΟΜΑΔΟΠΟΙΗΣΗ 3	G	1990	48	No2 K. ΜΗΧΑΝΗ	ΕΞΩΤΕΡΙΚΗ ΔΙΑΡΡΟΗ Α3-Α4 ΚΑΙ Α7-Α8 ΚΥΛΙΝΔΡΩΝ	ΜΗΧΑΝΙΚΕΣ	2006	5	ΕΝΑΡΞΗ ΕΡΓΑΣΙΩΝ ΓΙΑ Α4 ΚΥΛΙΝΔΡΟ	5
619	ΟΜΑΔΟΠΟΙΗΣΗ 3	G	1990	48	No2 K. ΜΗΧΑΝΗ	ΕΞΩΤΕΡΙΚΗ ΔΙΑΡΡΟΗ Α3-Α4 ΚΑΙ Α7-Α8 ΚΥΛΙΝΔΡΩΝ	ΜΗΧΑΝΙΚΕΣ	2006	6	ΟΛΟΚΛΗΡΩΣΗ ΕΡΓΑΣΙΩΝ ΑΠΟΚΑΤΑΣΤΑΣΗΣ ΔΙΑΡΡΟΗΣ-ΔΟΚΙΜΕΣ ΕΝ ΟΡΜΩ ΙΚΑΝΟΠΟΙΗΤΙΚΕΣ	5
620	ΟΜΑΔΟΠΟΙΗΣΗ 3	G	1990	48	No2 K. ΜΗΧΑΝΗ	ΕΞΩΤΕΡΙΚΗ ΔΙΑΡΡΟΗ Α3-Α4 ΚΑΙ Α7-Α8 ΚΥΛΙΝΔΡΩΝ	ΜΗΧΑΝΙΚΕΣ	2006	7	ΑΠΟΣΤΟΛΗ ΥΛΙΚΩΝ ΑΠΟ ΛΟΓ.2 ΣΕ ΛΟΓ.1	5
621	ΟΜΑΔΟΠΟΙΗΣΗ 3	G	1990	48	No2 K. ΜΗΧΑΝΗ	ΕΞΩΤΕΡΙΚΗ ΔΙΑΡΡΟΗ Α3-Α4 ΚΑΙ Α7-Α8 ΚΥΛΙΝΔΡΩΝ	ΜΗΧΑΝΙΚΕΣ	2006	8	ΑΠΟΣΤΟΛΗ ΥΛΙΚΩΝ ΑΠΟ ΛΟΓ.2 ΣΕ ΛΟΓ.1	5
622	ΟΜΑΔΟΠΟΙΗΣΗ 3	G	1990	48	No2 K. ΜΗΧΑΝΗ	ΕΞΩΤΕΡΙΚΗ ΔΙΑΡΡΟΗ Α3-Α4 ΚΑΙ Α7-Α8 ΚΥΛΙΝΔΡΩΝ	ΜΗΧΑΝΙΚΕΣ	2006	9	ΤΟΠΟΘΕΤΗΣΗ ΕΠΟΞΙΚΟΥ ΥΛΙΚΟΥ ΣΕ ΠΕΡΙΧΙΤΩΝΙΟ ΧΩΡΟ-ΑΡΜΟΣΗ ΠΑΛΑΙΟΥ ΧΙΤΩΝΙΟΥ ΚΑΙ ΚΑΠΑΚΙΟΥ-ΔΟΚ. ΙΚΑΝΟΠ.	5
623	ΟΜΑΔΟΠΟΙΗΣΗ 3	G	1990	48	No2 K. ΜΗΧΑΝΗ	ΕΞΩΤΕΡΙΚΗ ΔΙΑΡΡΟΗ Α3-Α4 ΚΑΙ Α7-Α8 ΚΥΛΙΝΔΡΩΝ	ΜΗΧΑΝΙΚΕΣ	2006	10	ΕΜΦΑΝΙΣΗ ΔΙΑΡΡΟΗΣ ΚΑΙ Α7-Α8-ΣΥΝΕΡΓΕΙΑΚΗ ΣΥΝΔΡΟΜΗ	5
624	ΟΜΑΔΟΠΟΙΗΣΗ 3	G	1990	48	No2 K. ΜΗΧΑΝΗ	ΕΞΩΤΕΡΙΚΗ ΔΙΑΡΡΟΗ Α3-Α4 ΚΑΙ Α7-Α8 ΚΥΛΙΝΔΡΩΝ	ΜΗΧΑΝΙΚΕΣ	2006	11	ΔΙΑΡΡΟΗ ΑΠΟ Α3 ΠΟΛΥ ΜΙΚΡΗ-ΔΙΑΡΡΟΗ ΑΠΟ Α7-Α8 ΑΠΟ ΔΙΚΤΥΟ-ΑΠΟΚΑΤΑΣΤΑΣΗ ΑΠΟ ΠΛΟΙΟ	5
625	ΟΜΑΔΟΠΟΙΗΣΗ 3	H	1990	48	No2 K. ΜΗΧΑΝΗ	ΔΙΑΡΡΟΗ ΘΑΛΑΣΣΗΣ ΑΠΟ ΕΞΑΡΤΗΜΕΝΗ	ΜΗΧΑΝΙΚΕΣ	2004	1	ΜΙΚΡΗ ΔΙΑΡΡΟΗ-ΠΑΡΑΚΟΛΟΥΘΗΣΗ	2
626	ΟΜΑΔΟΠΟΙΗΣΗ 3	H	1990	48	No2 K. ΜΗΧΑΝΗ	ΔΙΑΡΡΟΗ ΘΑΛΑΣΣΗΣ ΑΠΟ ΕΞΑΡΤΗΜΕΝΗ	ΜΗΧΑΝΙΚΕΣ	2004	2	ΑΙΤΗΣΗ ΕΝΕΡΓΟΥΣ ΔΑΝΕΙΟΛΗΨΙΑ ΑΠΟ ΝΑΥΜΑΧΟ	2
627	ΟΜΑΔΟΠΟΙΗΣΗ 3	H	1990	48	No2 K. ΜΗΧΑΝΗ	ΔΙΑΡΡΟΗ ΘΑΛΑΣΣΗΣ ΑΠΟ ΕΞΑΡΤΗΜΕΝΗ	ΜΗΧΑΝΙΚΕΣ	2004	3	ΕΓΚΡΙΣΗ ΕΝΕΡΓΟΥΣ ΔΑΝΕΙΟΛΗΨΙΑΣ	2
628	ΟΜΑΔΟΠΟΙΗΣΗ 3	H	1990	48	No2 K. ΜΗΧΑΝΗ	ΔΙΑΡΡΟΗ ΘΑΛΑΣΣΗΣ ΑΠΟ ΕΞΑΡΤΗΜΕΝΗ	ΜΗΧΑΝΙΚΕΣ	2004	4	ΠΑΡΑΔΟΣΗ ΥΛΙΚΩΝ ΕΠΙΣΚΕΥΗΣ ΑΝΤΛΙΑΣ	2
629	ΟΜΑΔΟΠΟΙΗΣΗ 3	H	1990	48	No2 K. ΜΗΧΑΝΗ	ΔΙΑΡΡΟΗ ΘΑΛΑΣΣΗΣ ΑΠΟ ΕΞΑΡΤΗΜΕΝΗ	ΜΗΧΑΝΙΚΕΣ	2004	5	ΜΙΚΡΗ ΔΙΑΡΡΟΗ-ΠΑΡΑΚΟΛΟΥΘΗΣΗ	2
630	ΟΜΑΔΟΠΟΙΗΣΗ 3	H	1990	48	No2 K. ΜΗΧΑΝΗ	ΔΙΑΡΡΟΗ ΘΑΛΑΣΣΗΣ ΑΠΟ ΕΞΑΡΤΗΜΕΝΗ	ΜΗΧΑΝΙΚΕΣ	2004	6	ΑΙΤΗΣΗ ΕΝΕΡΓΟΥΣ ΔΑΝΕΙΟΛΗΨΙΑ ΑΠΟ ΝΑΥΜΑΧΟ	2
631	ΟΜΑΔΟΠΟΙΗΣΗ 3	H	1990	48	No2 K. ΜΗΧΑΝΗ	ΔΙΑΡΡΟΗ ΘΑΛΑΣΣΗΣ ΑΠΟ ΕΞΑΡΤΗΜΕΝΗ	ΜΗΧΑΝΙΚΕΣ	2004	7	ΕΓΚΡΙΣΗ ΕΝΕΡΓΟΥΣ ΔΑΝΕΙΟΛΗΨΙΑΣ	2
632	ΟΜΑΔΟΠΟΙΗΣΗ 3	H	1990	48	No2 K. ΜΗΧΑΝΗ	ΔΙΑΡΡΟΗ ΘΑΛΑΣΣΗΣ ΑΠΟ ΕΞΑΡΤΗΜΕΝΗ	ΜΗΧΑΝΙΚΕΣ	2004	8	ΠΑΡΑΔΟΣΗ ΥΛΙΚΩΝ ΕΠΙΣΚΕΥΗΣ ΑΝΤΛΙΑΣ	2

TYPOS PLOIOY	ONOMA PLOIOY	ETOS APOKTHSHS	ATOMA PROSWPIKOY	BLABH	AITIA	TYPOS BLABHS
ΟΜΑΔΟΠΟΙΗΣΗ 1	A	2002	21	NO1 K. ΜΗΧΑΝΗ	ΑΔΥΝΑΜΙΑ ΧΕΙΡΙΣΜΟΥ ΑΠΟ ΤΕΦΥΡΑ-ΚΕΠ	ΜΗΧΑΝΙΚΕΣ
ΟΜΑΔΟΠΟΙΗΣΗ 1	A	2002	21	NO1 K. ΜΗΧΑΝΗ	ΑΔΥΝΑΜΙΑ ΧΕΙΡΙΣΜΟΥ ΑΠΟ ΤΕΦΥΡΑ-ΚΕΠ	ΜΗΧΑΝΙΚΕΣ
ΟΜΑΔΟΠΟΙΗΣΗ 1	A	2002	21	NO1 K. ΜΗΧΑΝΗ	ΑΔΥΝΑΜΙΑ ΧΕΙΡΙΣΜΟΥ ΑΠΟ ΤΕΦΥΡΑ-ΚΕΠ	ΜΗΧΑΝΙΚΕΣ
ΟΜΑΔΟΠΟΙΗΣΗ 1	A	2002	21	NO1 K. ΜΗΧΑΝΗ	ΑΔΥΝΑΜΙΑ ΧΕΙΡΙΣΜΟΥ ΑΠΟ ΤΕΦΥΡΑ-ΚΕΠ	ΜΗΧΑΝΙΚΕΣ
ΟΜΑΔΟΠΟΙΗΣΗ 1	A	2002	21	NO1 K. ΜΗΧΑΝΗ	ΑΔΥΝΑΜΙΑ ΧΕΙΡΙΣΜΟΥ ΑΠΟ ΤΕΦΥΡΑ-ΚΕΠ	ΜΗΧΑΝΙΚΕΣ
ΟΜΑΔΟΠΟΙΗΣΗ 1	C	2005	21	ΟΙΚΕΤΟΣ ΚΑΥΣΑΕΡΙΩΝ No2 K. ΜΗΧΑΝΗΣ	ΔΙΑΡΡΟΗ ΘΑΛΑΣΣΗΣ ΑΠΟ ΔΙΚΤΥΟ ΨΥΞΗΣ	ΜΗΧΑΝΙΚΕΣ
ΟΜΑΔΟΠΟΙΗΣΗ 1	C	2005	21	ΟΙΚΕΤΟΣ ΚΑΥΣΑΕΡΙΩΝ No2 K. ΜΗΧΑΝΗΣ	ΔΙΑΡΡΟΗ ΘΑΛΑΣΣΗΣ ΑΠΟ ΔΙΚΤΥΟ ΨΥΞΗΣ	ΜΗΧΑΝΙΚΕΣ
ΟΜΑΔΟΠΟΙΗΣΗ 1	A	2002	21	Α.Ρ ΚΥΡΙΑ ΜΗΧΑΝΗ	ΔΙΑΡΡΟΗ ΕΛΑΙΟΥ ΑΠΟ ΨΥΤΕΙΟ	ΜΗΧΑΝΙΚΕΣ
ΟΜΑΔΟΠΟΙΗΣΗ 1	A	2002	21	Α.Ρ ΚΥΡΙΑ ΜΗΧΑΝΗ	ΔΙΑΡΡΟΗ ΕΛΑΙΟΥ ΑΠΟ ΨΥΤΕΙΟ	ΜΗΧΑΝΙΚΕΣ
ΟΜΑΔΟΠΟΙΗΣΗ 1	A	2002	21	Α.Ρ ΚΥΡΙΑ ΜΗΧΑΝΗ	ΔΙΑΡΡΟΗ ΕΛΑΙΟΥ ΑΠΟ ΨΥΤΕΙΟ	ΜΗΧΑΝΙΚΕΣ
ΟΜΑΔΟΠΟΙΗΣΗ 1	A	2002	21	Α.Ρ ΚΥΡΙΑ ΜΗΧΑΝΗ	ΔΙΑΡΡΟΗ ΕΛΑΙΟΥ ΑΠΟ ΨΥΤΕΙΟ	ΜΗΧΑΝΙΚΕΣ
ΟΜΑΔΟΠΟΙΗΣΗ 1	A	2002	21	Α.Ρ ΚΥΡΙΑ ΜΗΧΑΝΗ	ΔΙΑΡΡΟΗ ΕΛΑΙΟΥ ΑΠΟ ΨΥΤΕΙΟ	ΜΗΧΑΝΙΚΕΣ
ΟΜΑΔΟΠΟΙΗΣΗ 1	B	2005	21	No1 K. ΜΗΧΑΝΗ	ΑΔΥΝΑΜΙΑ ΕΚΚΙΝΗΣΗΣ ΚΑΙ ΥΨΗΛΕΣ ΘΕΡΜΟΚΡΑΣΙΕΣ	ΜΗΧΑΝΙΚΕΣ
ΟΜΑΔΟΠΟΙΗΣΗ 1	B	2005	21	No1 K. ΜΗΧΑΝΗ	ΑΔΥΝΑΜΙΑ ΕΚΚΙΝΗΣΗΣ ΚΑΙ ΥΨΗΛΕΣ ΘΕΡΜΟΚΡΑΣΙΕΣ ΚΑΥΣΑΕΡΙΩΝ ΚΥΛΙΝΔΡΩΝ ΚΑΙ TURBO	ΜΗΧΑΝΙΚΕΣ
ΟΜΑΔΟΠΟΙΗΣΗ 1	B	2005	21	No1 K. ΜΗΧΑΝΗ	ΑΔΥΝΑΜΙΑ ΕΚΚΙΝΗΣΗΣ ΚΑΙ ΥΨΗΛΕΣ ΘΕΡΜΟΚΡΑΣΙΕΣ ΚΑΥΣΑΕΡΙΩΝ ΚΥΛΙΝΔΡΩΝ ΚΑΙ TURBO	ΜΗΧΑΝΙΚΕΣ
ΟΜΑΔΟΠΟΙΗΣΗ 1	A	2002	21	No2 ΚΥΡΙΑ ΜΗΧΑΝΗ	ΔΙΑΡΡΟΗ ΕΛΑΙΟΥ ΜΙΚΡΗΣ ΕΚΤΑΣΕΩΣ ΑΠΟ ΨΥΤΕΙΟ ΕΛΑΙ	ΜΗΧΑΝΙΚΕΣ
ΟΜΑΔΟΠΟΙΗΣΗ 1	C	2005	21	Α.Ρ Κ. ΜΕΣΣΤΗΡΑΣ	ΔΙΑΡΡΟΗ ΕΛΑΙΟΥ ΑΠΟ ΚΕΛΥΦΟΣ ΕΞΑΡΤΗΜΕΝΗΣ	ΜΗΧΑΝΙΚΕΣ
ΟΜΑΔΟΠΟΙΗΣΗ 1	D	2002	21	No2 ΚΥΡΙΑ ΜΗΧΑΝΗ	ΒΛΑΒΗ ΑΝΤΙΣΤΑΣΗΣ ΠΡΟΘΕΡΜΑΝΣΕΩΣ	ΜΗΧΑΝΙΚΕΣ
ΟΜΑΔΟΠΟΙΗΣΗ 1	D	2002	21	Α.Ρ Κ. ΜΗΧΑΝΗ	ΔΙΑΡΡΟΗ ΠΕΤΡΕΛΑΙΟΥ ΑΠΟ ΔΙΚΤΥΟ ΕΠΙΣΤΡΟΦΩΝ ΚΑΙ Τ	ΜΗΧΑΝΙΚΕΣ
ΟΜΑΔΟΠΟΙΗΣΗ 1	D	2002	21	Α.Ρ Κ. ΜΗΧΑΝΗ	ΔΙΑΡΡΟΗ ΠΕΤΡΕΛΑΙΟΥ ΑΠΟ ΔΙΚΤΥΟ ΕΠΙΣΤΡΟΦΩΝ ΚΑΙ Τ	ΜΗΧΑΝΙΚΕΣ
ΟΜΑΔΟΠΟΙΗΣΗ 1	D	2002	21	Α.Ρ Κ. ΜΗΧΑΝΗ	ΔΙΑΡΡΟΗ ΠΕΤΡΕΛΑΙΟΥ ΑΠΟ ΔΙΚΤΥΟ ΕΠΙΣΤΡΟΦΩΝ ΚΑΙ Τ	ΜΗΧΑΝΙΚΕΣ
ΟΜΑΔΟΠΟΙΗΣΗ 1	C	2005	21	No1 ΚΥΡΙΑ ΜΗΧΑΝΗ	ΔΙΑΡΡΟΗ ΘΑΛΑΣΣΗΣ	ΜΗΧΑΝΙΚΕΣ
ΟΜΑΔΟΠΟΙΗΣΗ 1	C	2005	21	No1 ΚΥΡΙΑ ΜΗΧΑΝΗ	ΔΙΑΡΡΟΗ ΘΑΛΑΣΣΗΣ	ΜΗΧΑΝΙΚΕΣ

Εικόνες 8 α, β: Τελική μορφή αρχείου Excel που δημιουργήθηκε και χρησιμοποιήθηκε για την ανάλυση

Έπειτα, αφού ήρθε στην επιθυμητή μορφή για ανάλυση το αρχείο Excel το επόμενο βήμα είναι να προσδιοριστεί το εργαλείο που θα γίνει η σχετική Ανάλυση των Δεδομένων. Για τους

σκοπούς αυτής της διπλωματικής εργασίας χρησιμοποιήθηκε η γλώσσα προγραμματισμού R. Με την βοήθεια αυτής της γλώσσας έγινε Καθαρισμός των Δεδομένων, Περιγραφική Ανάλυση των Δεδομένων, αλλά και Απεικόνιση των Αποτελεσμάτων σε διαγράμματα με σκοπό την καταγραφή των συμπερασμάτων που ακολουθούν στο επόμενο κεφάλαιο.

Αρχικά, για να ξεκινήσει η οποιαδήποτε ανάλυση θα πρέπει να γίνει εισαγωγή των δεδομένων στο περιβάλλον της R. Αυτό γίνεται με μία σειρά εντολών που μπορούν να βρεθούν στο παράρτημα. Αφού γίνει η εισαγωγή των δεδομένων θα πρέπει να γίνει μια Περιγραφική Ανάλυση με σκοπό να έρθει σε άμεση επαφή ο αναλυτής και να γνωρίσει τα δεδομένα τα οποία υπάρχουν στην κατοχή του. Περιγραφική Ανάλυση είναι όταν χρησιμοποιούνται διάφορα μέσα (γραφήματα ή σύνολα αριθμών) για την απεικόνιση των δεδομένων. Στην παρούσα μεταπτυχιακή διατριβή το πρώτο βήμα είναι να γίνει εντοπισμός των λανθασμένων τιμών που μπορεί να οφείλονται είτε στην αγνοία για το τρόπο κάποιων εγγραφών είτε σε λάθη κατά την εισαγωγή των δεδομένων από τα αρμόδια άτομα που έχουν αναλάβει αυτό το κομμάτι. Το αρχικό αρχείο Excel είχε περίπου 1550 γραμμές δεδομένων οπότε αναμενόμενο είναι ότι αν βρεθούν λανθασμένες τιμές θα μειωθούν σχετικά. Στην αναζήτηση λανθασμένων δεδομένων διαπιστώθηκε ότι υπάρχουν «προβληματικές εγγραφές» οι οποίες αντί να έχουν αξιοποιήσιμες τιμές είχαν NA τιμές , όπως μπορεί να απεικονιστεί και παρακάτω στην εικόνα 9.

```
sapply(MyData, function(x) sum(is.na(x)))
    TYPOS PLOIOY       ONOMA PLOIOY     ETOS APOKTHSHS   ATOMA PROSWPIKOY            BLABH            AITIA
               0                  0                  0                  0                0               17
    TYPOS BLABHS     ETOS BLABHS ARITHMOS ENERGEIAS         ENERGEIES      MERES EKTOS
               0                  5                  0                  0                2
```

Εικόνα 9: Not Available (NA) τιμές

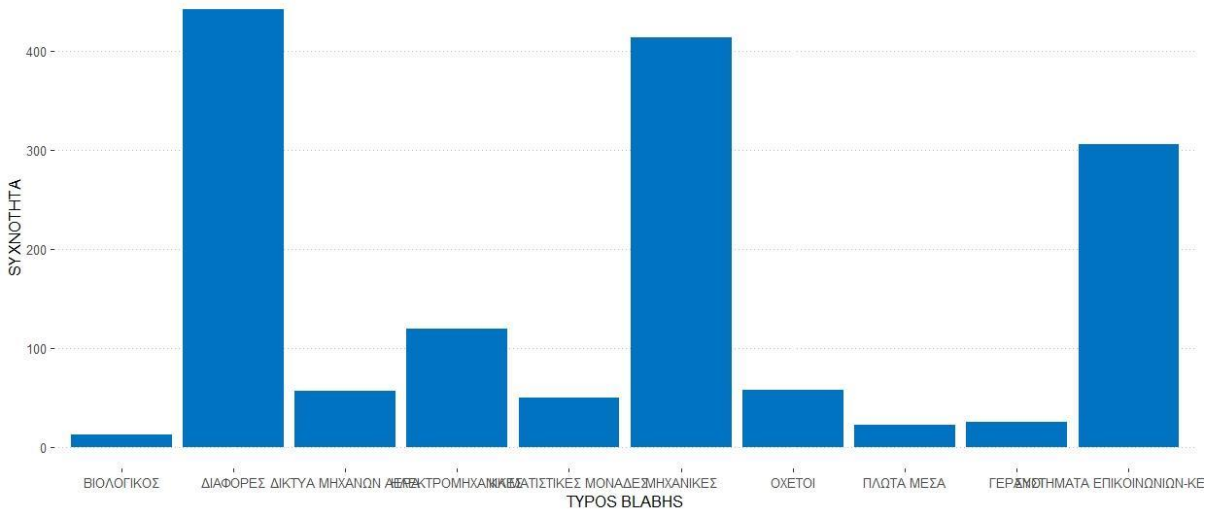
Η εικόνα 9 προειδοποιεί τον αναλυτή ότι υπάρχουν NA τιμές στα δεδομένα. Η πιο συνήθης κίνηση που ακολουθήθηκε και στην παρούσα εργασία είναι η διαγραφή αυτών των εγγραφών με σκοπό των Καθαρισμό των Δεδομένων. Άλλη κίνηση θα μπορούσε να ήταν η διόρθωση των τιμών αυτών σε συνεργασία με τα αρμόδια τμήματα που γνωρίζουν τις εγγραφές καλύτερα από τον καθένα. Αφού γίνει διαγραφή των προβληματικών εγγραφών θα πρέπει να μην υπάρχει καμία NA τιμή στα δεδομένα. Το επόμενο βήμα μετά την απαλλαγή των προβληματικών εγγραφών είναι να συνεχίσει η Περιγραφική Ανάλυση για να γίνει πιο εύπεπτη η πληροφορία που κρύβεται στα δεδομένα. Αφού καθαρίστηκε το τελικό αρχείο της μεταπτυχιακής διατριβής από τα προβληματικά δεδομένα, έχει 1506 γραμμές δεδομένων και 11 στήλες που είναι έτοιμες για ανάλυση. Ο αριθμός των NA τιμών δεν επηρέασε ποσοτικά και ποιοτικά τα δεδομένα μας παρά

τη μείωση των εγγραφών συγκριτικά με το αρχικό αρχείο Excel. Συνεπώς με το συγκεκριμένο όγκο δεδομένων δεν αλλοιώνονται αποτελέσματα και συμπεράσματα. Σίγουρα θα πρέπει να γίνει μια προσπάθεια ορθότερης και προσεκτικότερης καταγραφής ώστε όσο το δυνατόν περισσότερες τιμές να είναι διαθέσιμες καθώς εκτιμάται ότι σε περίπτωση που είχαμε να κάνουμε με μεγάλο όγκο δεδομένων η ύπαρξη NA τιμών θα είχε μεγαλύτερη επίδραση αλλοιώνοντας ίσως αποτελέσματα και συμπεράσματα. Πρέπει να εστιάσουμε στην προσεκτικότερη καταχώρηση δεδομένων ώστε να ομαδοποιούνται ορθά και στην εκπαίδευση προσωπικού πάνω σε αυτό το τομέα.

Αρχικά θα γίνει μια Περιγραφική Ανάλυση για να αναγνωριστούν οι ακραίες τιμές και αν υπάρχει κάπου περαιτέρω κίνητρο για ανάλυση μέσω αυτών. Στα γραφήματα που ασχολούνται με τις ακραίες τιμές όπως θα δούμε και στις αντίστοιχες εικόνες τους δεν μας ενδιαφέρει τόσο πολύ η ομαδοποίηση στον οριζόντιο άξονα καθώς μας αρκεί μόνο η διαπίστωση αν υπάρχουν και ποιες είναι οι ακραίες τιμές και όχι τόσο με ποια βλάβη ή με ποιο πλοίο σχετίζονται. Για να επιτευχθεί αυτό θα πραγματοποιηθούν μια σειρά από γραφήματα που θα απεικονίζουν ποσοτικά τις τιμές της κάθε στήλης. Φυσικά για τις αριθμητικές στήλες κάτι τέτοιο θα είναι πιο εύκολο και ευανάγνωστο από τις στήλες που περιέχουν χαρακτήρες.

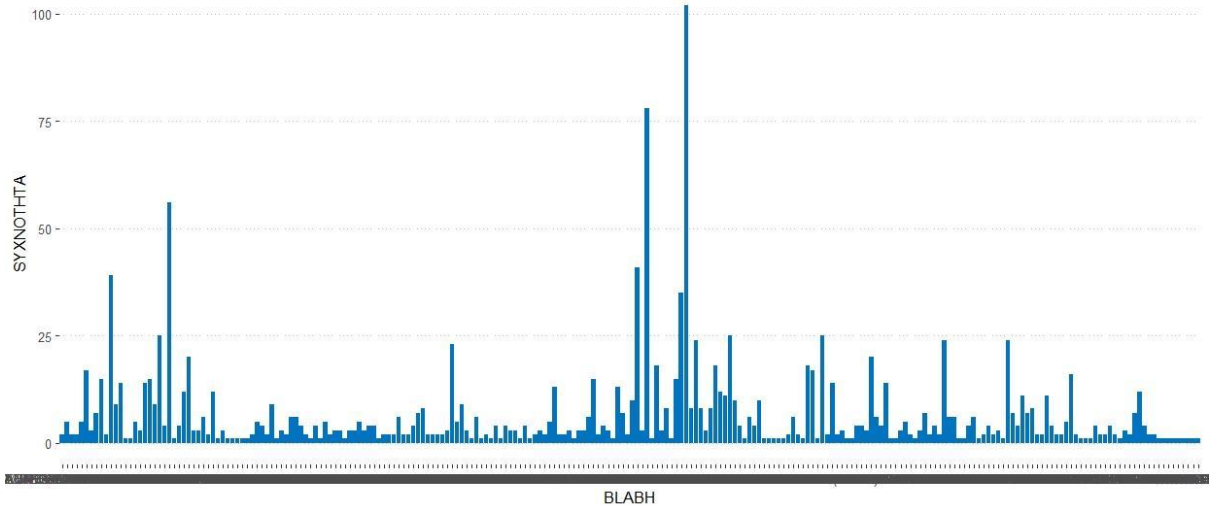
Αφού η κύρια ιδέα της διπλωματικής εργασίας είναι η μελέτη των βλαβών σε μερίδα του στόλου του ΠΝ θα πρέπει πρώτα και κύρια να ερευνηθούν οι στήλες που αφορούν τις βλάβες. Στην εικόνα 10 που ακολουθεί μπορεί κανείς να παρατηρήσει τους διάφορους τύπους βλαβών που υπάρχουν στο εκάστοτε πλοίο. Οι κατηγορίες είναι συνολικά 10. Όπως γίνεται κατανοητό φαίνεται να εμφανίζονται με διαφορά περισσότερο οι κατηγορίες «Διάφορες», «Μηχανές» και «Συστήματα Επικοινωνιών -Κεραίες». Με διαφορά λιγότερες φορές εμφανίζονται βλάβες που ανήκουν στην κατηγορία «Βιολογικός». Οι κατηγορίες σχετίζονται με το τι είδους μηχάνημα χαλάει κάθε φορά οπότε και μπαίνει ο αντίστοιχος τίτλος. Απο το είδος των βλαβών που παρουσιάζονται συχνότερα βγαίνουν συμπεράσματα για το ποια μηχανήματα είναι πιο ευπαθή, ποια χρήζουν προσοχή κατά τη χρήση και συχνότερη συντήρηση, ποιο προσωπικό οφείλει να εκπαιδευτεί περισσότερο και ποια συνεργεία πρέπει να στελεχωθούν και να εξοπλιστούν περισσότερο ώστε να είναι έτοιμα να αντιμετωπίσουν τις βλάβες. Δίνεται επίσης η δυνατότητα για προμήθεια ανταλλακτικών που αφορούν τις συνηθέστερες βλάβες ώστε να υπάρχει επαρκές απόθεμα και να μην χάνεται χρόνος σε παραγγελίες. Επίσης θα μπορούσε να εξεταστεί η δυνατότητα αντικατάστασης εξολοκλήρου κάποιου συστήματος με νεότερης τεχνολογίας αν οι

διαρκείς επισκευές του είναι αύμφορες και επηρεάζουν τη μαχητική ικανότητα του εκάστοτε πλοίου.



Εικόνα 10: Συχνότητα Βλαβών ανά τύπο βλάβης

Παρόλο που αν απεικονιστούν οι βλάβες θα βγει ένα χαοτικό γράφημα, το οποίο δεν θα είναι καθόλου ευανάγνωστο, προτιμήθηκε να πραγματοποιηθεί, και να παρουσιαστεί υπό την κάτωθι μορφή για να γίνουν αντιληπτές οι ακραίες τιμές, (δηλαδή ποιες εμφανίζονται περισσότερο και ποιες λιγότερο)



Εικόνα 11: Ακραίες τιμές συχνότητας βλαβών

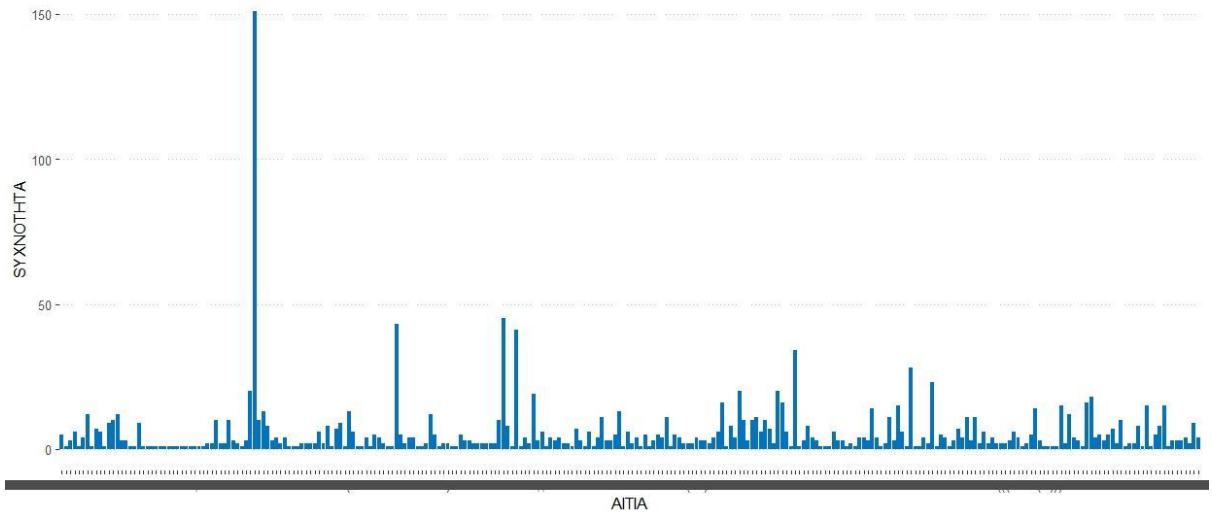
Όπως ειπώθηκε σκοπός αυτού του γραφήματος ήταν να μπορούν να γίνουν αντιληπτές οι ακραίες τιμές ανεξάρτητα απο το είδος της βλάβης. Αφού λοιπόν υπάρχουν το μόνο που μένει είναι να δούμε με κάποιον άλλον τρόπο που θα είναι πιο ευανάγνωστος ποιες είναι οι 10 πιο συνήθεις βλάβες, κάτι το οποίο φαίνεται καθαρά στην εικόνα 12.

```
tail(names(sort(table(MyData$`ΒΛΑΒΗ`))), 10)
[1] "LINK-11"           "ΑΡ Κ. ΜΗΧΑΝΗ"      "ΟΧΕΤΟΙ Κ. ΜΗΧΑΝΩΝ" "ΠΗΔΑΛΙΟ"           "No2 ΗΛΕΚΤΡΟΜΗΧΑΝΗ" "ΑΡ Κ. ΜΕΙΩΤΗΡΑΣ"
[7] "No1 ΗΛΕΚΤΡΟΜΗΧΑΝΗ" "ΑΦΑΛΑΤΩΤΗΣ"       "No1 Κ. ΜΗΧΑΝΗ"     "No2 Κ. ΜΗΧΑΝΗ"
```

Εικόνα 12: Οι Top 10 Βλάβες

Κοιτώντας την εικόνα 12 πρέπει να γίνει ξεκάθαρο ότι η πρώτη βλάβη από τις 10, που εμφανίζεται περισσότερο είναι αυτή στη «No2 Κ. Μηχανή» και λιγότερο η βλάβη στο «LINK-11».

Αφού απεικονίστηκαν οι τύποι βλαβών και οι πιο συχνά εμφανιζόμενες βλάβες επόμενο βήμα είναι να εντοπιστεί αν υπάρχει κάποια πιο συνηθισμένη αιτία βλάβης. Φυσικά όπως και στις βλάβες λόγω του ότι υπάρχουν πολλές αιτίες στα δεδομένα, σκοπός του πρώτου γραφήματος είναι πάντα ο εντοπισμός ακραίων τιμών χωρίς να ασχοληθούμε με το ποια αιτία σχετίζεται με ποια ακραία τιμή κάτι το οποίο ξεκαθαρίζεται μέσω της εικόνας 14 . Όπως παρατηρείται στην εικόνα 13, υπάρχουν αιτίες που ξεχωρίζουν από τις υπόλοιπες.



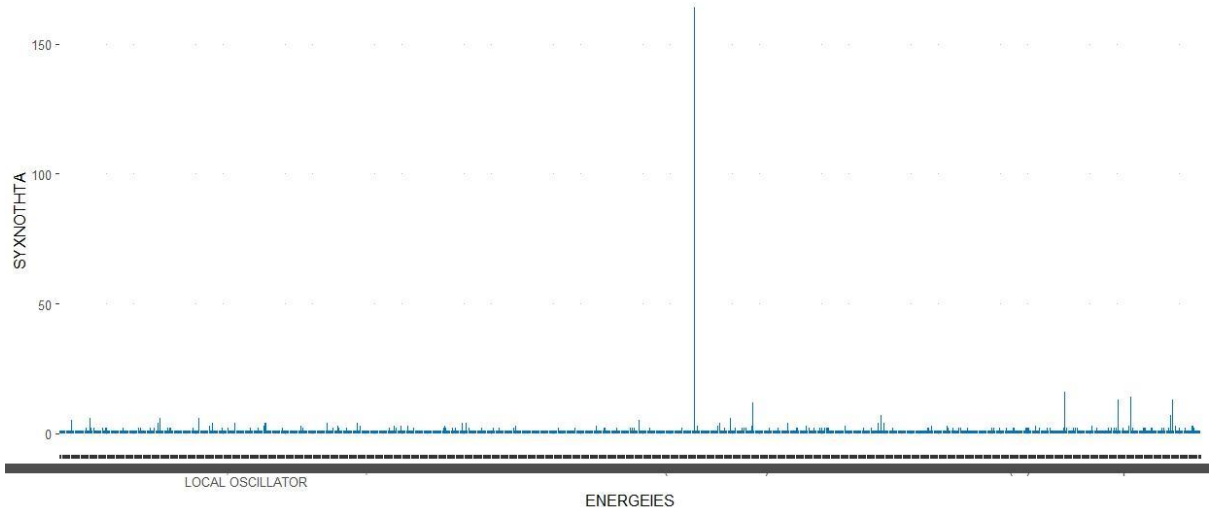
Εικόνα 13: Ακραίες τιμές αιτίας βλάβης

Άρα αυτό που μένει είναι με την βοήθεια της εικόνας 14 να εντοπιστούν οι πιο συνηθισμένες αιτίες βλαβών. Κάτι άλλο το οποίο παρατηρείται εύκολα και γρήγορα από την παρακάτω εικόνα είναι ότι οι χρήστες που καταχωρούν τα δεδομένα δεν έχουν κάνει πάντα ορθή καταγραφή αφού υπάρχουν εγγραφές όπως «Αδυναμία λειτουργία» και «Αδυναμία λειτουργίας» που αναφέρονται ακριβώς στην ίδια αιτία πλην όμως οφείλεται να καταχωρείται επακριβώς με τον ίδιο τρόπο.

```
> tail(names(sort(table(MyData$`ΑΙΤΙΑ`))), 10)
 [1] "ΑΔΥΝΑΜΙΑ ΛΕΙΤΟΥΡΓΙΑ"
 [2] "ΔΙΑΡΡΟΗ ΘΑΛΑΣΣΗΣ"
 [3] "ΔΙΑΡΡΟΗ ΘΑΛΑΣΣΗΣ ΚΑΙ ΚΑΥΣΑΕΡΙΩΝ ΟΧΕΤΟΥ-ΕΜΦΡΑΞΗ ΟΧΕΤΟΥ ΑΝΩ ΙΣΑΛΟΥ"
 [4] "ΚΑΤΑΔΙΚΗ ΤΡΟΦΟΔΟΤΙΚΗΣ ΑΝΤΛΙΑΣ"
 [5] "ΕΡΕΥΣΗ ΡΙΝΙΣΜΑΤΩΝ ΣΕ ΕΞΑΡΤΗΜΕΝΗ ΑΝΤΛΙΑ ΜΕΤΑΦΟΡΑ ΕΛΑΙΟΥ"
 [6] "ΔΥΣΛΕΙΤΟΥΡΓΙΑ"
 [7] "ΑΥΤΟΔΙΑΤΡΗΣΗ"
 [8] "ΑΔΥΝΑΜΙΑ ΣΥΝΤΟΝΙΣΜΟΥ"
 [9] "ΑΥΞΗΣΗ ΘΕΡΜΟΚΡΑΣΙΑΣ ΕΞΟΔΟΥ ΚΑΥΣΑΕΡΙΩΝ ΜΕΤΑ TURBO (Τ2)"
[10] "ΑΔΥΝΑΜΙΑ ΛΕΙΤΟΥΡΓΙΑΣ"
```

Εικόνα 14: Οι Top 10 Αιτίες

Αφού απεικονίστηκαν οι βλάβες και από τι προκλήθηκαν επόμενο βήμα είναι να εντοπιστούν οι ενέργειες που έγιναν για να αντιμετωπιστούν οι βλάβες. Με την έννοια ενέργειες εννοούμε πόσα βήματα χρειάστηκαν να γίνουν απο την εμφάνιση της βλάβης μέχρι την οριστική αποκατάσταση αυτής (π.χ. αίτηση συνεργειακής βοήθειας, εξάρμωση βεβλαμένου υλικού, παράδοση σε αρμόδιο συνεργείο, έλεγχος αυτού, επισκευή ή αντικατάσταση με έτερο όμοιο που είτε υπάρχει σε απόθεμα είτε το παραγγέλνουμε. κ.α.). Αν παρατηρηθεί η εικόνα 15, μπορεί να γίνει αντιληπτό ότι μια ενέργεια εμφανίζεται παραπάνω από τις άλλες.



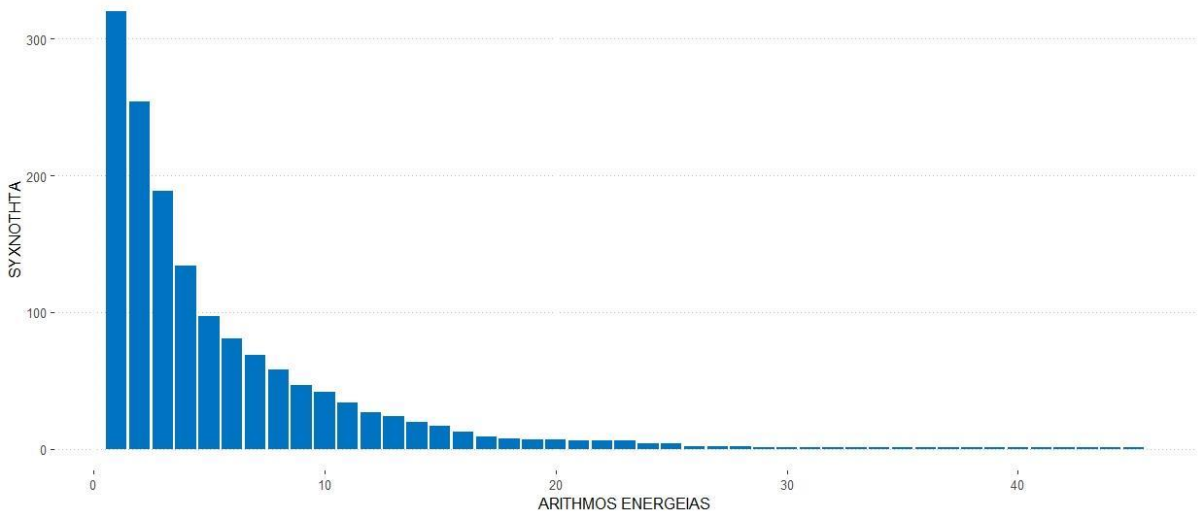
Εικόνα 15: Ακραίες τιμές ενεργειών αντιμετώπισης βλαβών

Αυτή η ενεργεία που εμφανίζεται συχνότερα είναι η «έχει αιτηθεί συνεργειακή βοήθεια» δηλαδή να αναθέτουν την εργασία είτε στα εξουσιοδοτημένα συνεργεία του ΠΝ είτε σε εξωτερικούς συνεργάτες/ιδιωτικούς φορείς, που αναλαμβάνουν έναντι χρηματικού αντιτίμου την αποκατάσταση της βλάβης παρέχοντας είτε ανταλλακτικά είτε εργασία είτε και τα δύο τόσο εντός ναυστάθμων όσο και στα δικά τους συνεργεία. Απο αυτό προκύπτει το συμπέρασμα πως είτε το προσωπικό των πλοίων δεν έχει πάντα την κατάλληλη τεχνογνωσία και εξειδίκευση σε σχέση με τα αρμόδια συνεργεία , είτε τα κατάλληλα εργαλεία είτε ότι οι βλάβες που παρουσιάζονται είναι τέτοιας φύσεως ώστε να μην μπορούν να φτιαχτούν επι τόπου εντός του πλοίου.

```
 tail(names(sort(table(MyData$`ENERGEIES`))), 10)
 [1] "ΑΝΑΛΗΨΗ ΕΝΕΡΓΕΙΩΝ"                "ΕΓΚΡΙΣΗ ΑΡΣΗΣ ΦΡΑΓΜΟΥ"              "ΟΛΟΚΛΗΡΩΣΗ ΕΠΙΣΚΕΥΗΣ ΑΠΟ ΝΣ"
 [4] "ΣΥΝΕΡΓΕΙΑΚΗ ΣΥΝΔΡΟΜΗ"             "ΕΓΚΡΙΣΗ ΕΝΕΡΓΟΥΣ ΔΑΝΕΙΟΛΗΨΙΑΣ"      "ΥΛΙΚΑ ΕΠΙΣΚΕΥΗΣ ΑΝΤΛΙΑΣ"
 [7] "ΣΥΝΕΡΓΕΙΑΚΗ ΣΥΝΔΡΟΜΗ ΑΠΟ ΝΣ"     "ΥΠΟ ΠΑΡΑΚΟΛΟΥΘΗΣΗ"                 "ΠΡΟΕΓΚΡΙΣΗ ΥΛΙΚΑ ΕΠΙΣΚΕΥΗΣ"
10] "ΕΧΕΙ ΑΙΤΗΘΕΙ ΣΥΝΕΡΓΕΙΑΚΗ ΒΟΗΘΕΙΑ"
```

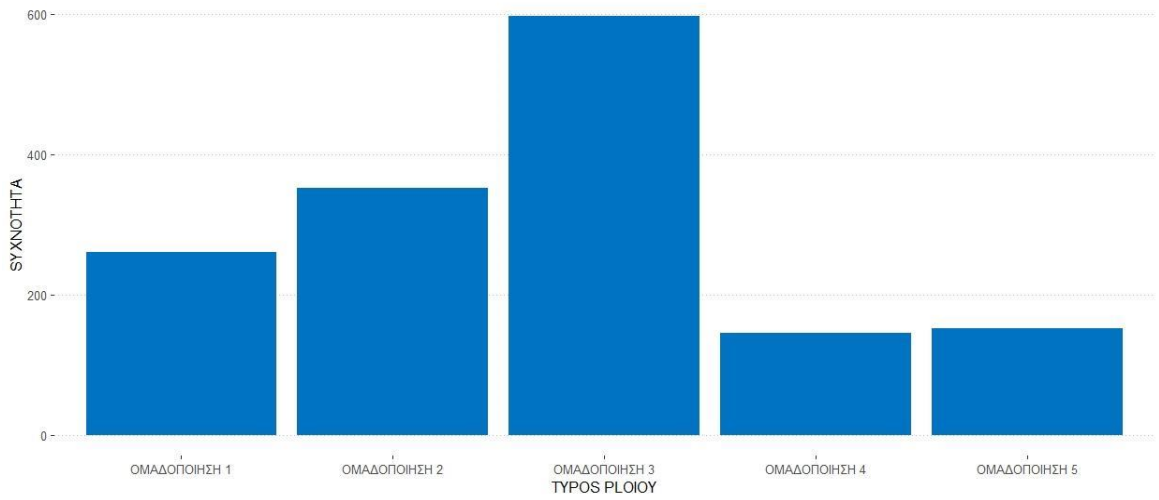
Εικόνα 16: Οι Top 10 Ενέργειες

Όταν εμφανίζεται μια βλάβη ξεκινάει ένας κύκλος εργασιών και ενεργειών και μας ενδιαφέρει να ποσοτικοποιήσουμε τον αριθμό αυτών φτάνοντας έτσι σε χρήσιμα συμπεράσματα. Κάνοντας καταμέτρηση ενεργειών ανά βλάβη φαίνεται να υπερτερούν με διαφορά οι λιγότερες ενέργειες 0-5. Κάτι τέτοιο απεικονίζεται καθαρά στην εικόνα 17 που ακολουθεί και μας δείχνει ότι απο τη στιγμή που η βλάβη έχει εντοπιστεί ως προς την αιτία αυτής, συνήθως απαιτούνται 0-5 ενέργειες (βήματα) για την αποκατάσταση αυτής. Αυτό μας δείχνει ότι συνήθως ο αριθμός των ενεργειών δεν είναι πολύ μεγάλος καθώς αυτές είναι τυποποιημένες συνεπώς δεν υπάρχει χρονοτριβή και μας δίνει την δυνατότητα να τις βελτιώσουμε και εξελίξουμε ακόμη περισσότερο.



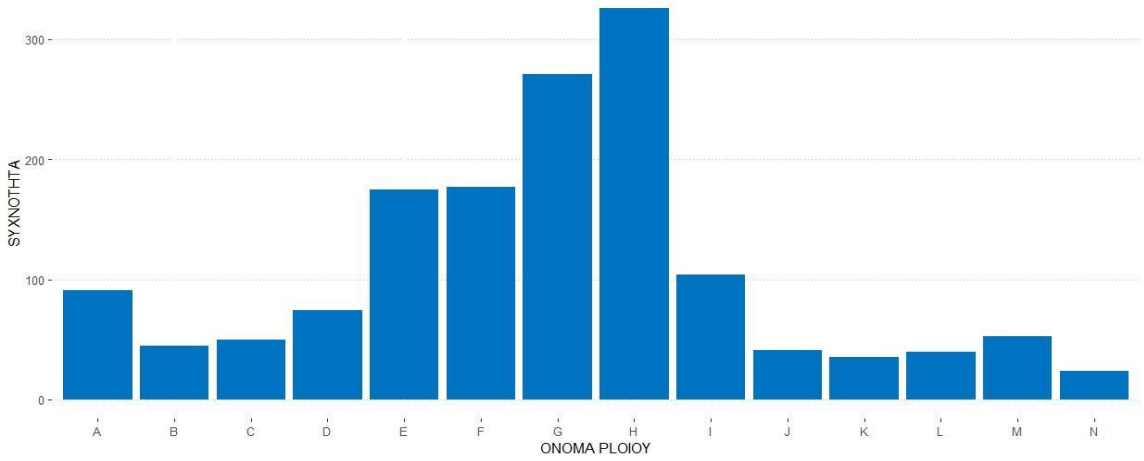
Εικόνα 17: Αριθμός ενεργειών μετά την βλάβη

Από την αρχή της εργασίας γίνεται αναφορά στις βλάβες των πλοίων συνεπώς πρέπει να γίνει ξεκάθαρο πόσες κατηγορίες πλοίων υπάρχουν και πόσα πλοία υπάρχουν στις κατηγορίες αυτές και να γίνει σύνδεση τους με τον αριθμό βλαβών. Κάτι τέτοιο απεικονίζεται στις επόμενες δύο εικόνες 18 & 19. Η ομαδοποίηση των κατηγοριών και ο αριθμός των πλοίων για λόγους ευαισθησίας των δεδομένων ,διαρροή των οποίων θα έπληττε την ασφάλεια της χώρας έγινε με τέτοιο τρόπο ωστε να μην εμφανίζονται οι πραγματικοί τύποι των πλοίων παρά μόνο χρησιμοποιούνται οι γενικοί όροι «ομαδοποίηση 1,2,3» κτλ αντίστοιχα και τα ονόματα των πλοίων τοποθετήθηκαν συμβολικά με ένα αρχικό γράμμα της αγγλικής αλφαβήτου.



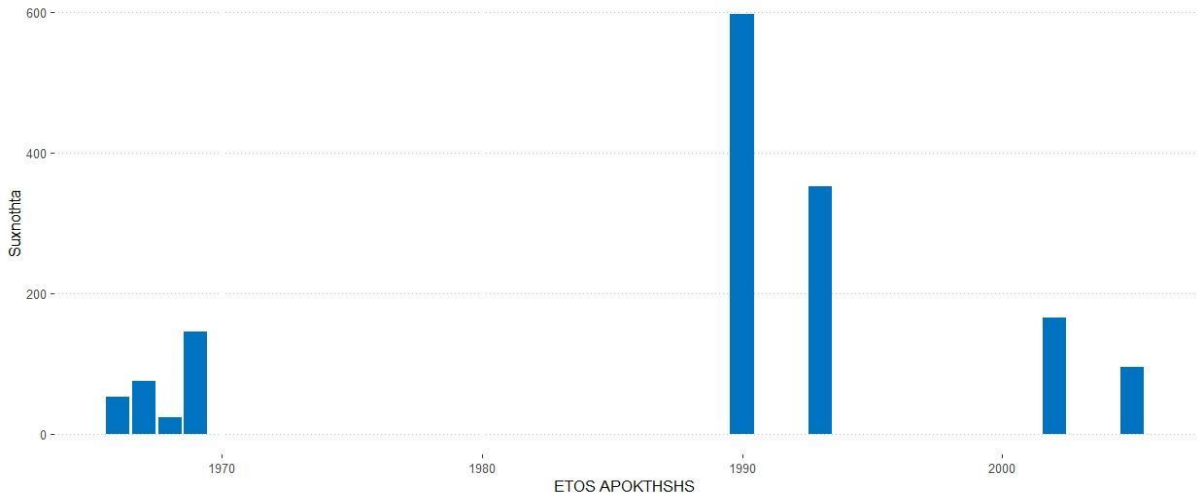
Εικόνα 18: Συχνότητα βλαβών ανά τύπο πλοίου

Υπάρχουν 5 ομαδοποιήσεις πλοίων με μια διαφορά ότι η ομαδοποίηση 3 των πλοίων εμφανίζεται πολύ παραπάνω στα δεδομένα. Εκτός απο το διάγραμμα με τις ομαδοποιήσεις των πλοίων εξετάζουμε ξεχωριστά και το κάθε πλοίο ώστε να προκύψουν περισσότερα συμπεράσματα. Όπως απεικονίζεται και στην παρακάτω εικόνα, τα πλοία «G» και «H» έχουν τις περισσότερες βλάβες με διαφορά. Μέσω αυτού κατανοούμε πως θα πρέπει να γίνει εμπεριστατωμένη έρευνα για αυτά τα δύο πλοία συγκεκριμένα. Έτσι θα γίνει κατανοητός ο λόγος που παρουσιάζουν τις πιο πολλές βλάβες μέσω συλλογής και άλλων επιμέρους δεδομένων για αυτά, ωστε να ξεκινήσουν ενέργειες πρόληψης βλαβών. Θα μπορούσαν να αναζητηθούν δεδομένα ως προς των αριθμό πλεύσιμων ημερών τους, τις επιμέρους γνώσεις του προσωπικού, έρευνα για τυχόν κακοτεχνίες σε αυτά τα δύο συγκεκριμένα, απαίτηση για ολική συντήρηση και επισκευή ακόμη και αντικατάστασή τους.



Εικόνα 19: -Συχνότητα βλαβών ανά πλοίο

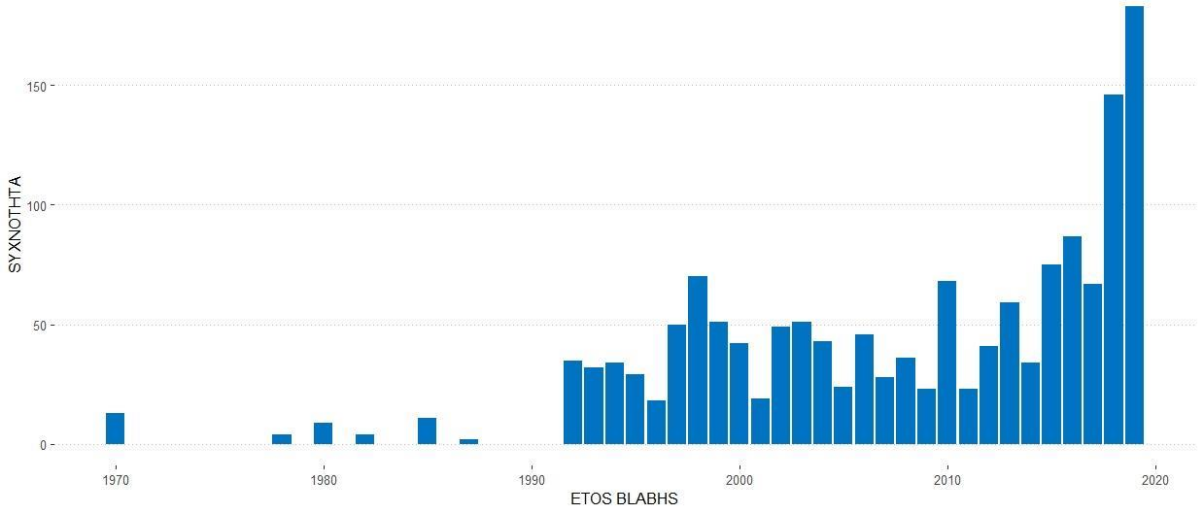
Αφού ερευνώνται οι βλάβες καλό θα ήταν γίνει και ένας έλεγχος με σκοπό να ερευνηθεί πότε αποκτήθηκαν τα πλοία και πότε παρουσιαστήκαν οι βλάβες. Όπως δείχνουν οι εικόνες υπάρχουν πλοία ναυπηγημένα από το 1966, αλλά και άλλα της δεκαετίας 1990-2000. Στο διάγραμμα της εικόνας 20 γίνεται σύνδεση του έτους απόκτησης του πλοίου με τον αριθμό βλαβών για να εξάγουμε συμπεράσματα το κατά πόσον αυτές επηρεάζονται απο την παλαιότητα των πλοίων και κατά πόσο τα νεότερα σε τεχνολογία παρουσιάζουν ή όχι βλάβες ώστε να κριθεί η αξιοπιστία τους.



Εικόνα 20: Συχνότητα βλαβών ανά έτος απόκτησης πλοίου

Όπως παρατηρούμε στο διάγραμμα πλοία που αποκτήθηκαν τη δεκαετία του 90 παρουσιάζουν περισσότερες βλάβες. Αυτό είναι κάτι το οποίο άξιζε να μελετηθεί και προκύπτει

απο το γεγονός πως αυτά τα πλοία είναι επιφορτισμένα με τον κύριο όγκο των αποστολών του ΠΝ συγκριτικά με τα πιο καινούρια που ακόμη δεν έχουν φτάσει σε μεγάλο αριθμό πλεύσιμων ημερών. Όσον αφορά τα παλαιότερα πλοία αυτά έχουν πιο μειωμένες επιχειρησιακές δράσεις αλλά και λιγότερα και πιο απλά συστήματα και ως εκ τούτου εμφανίζονται λιγότερο στις βλάβες πράγμα το οποίο αν και φαντάζει περίεργο λόγω της μεγάλης τους παλαιότητας επιβεβαιώθηκε μέσω του διαγράμματος. Εγείρονται φυσικά ερωτήματα βλέποντας τα διαγράμματα κατά πόσον τα πλοία δεκαετίας 1990-2000 παρουσιάζουν πιο πολλές βλάβες γιατί ενδεχομένως έχουν προβληματικά συστήματα, είτε γίνονται κακές συντηρήσεις και δίνεται το έναυσμα για περαιτέρω έρευνα ώστε να διαπιστωθούν επι μέρους αιτίες και να βρεθούν λύσεις με καλύτερο ίσως διαμοιρασμό επιχειρησιακών δράσεων σε όλα τα πλοία.

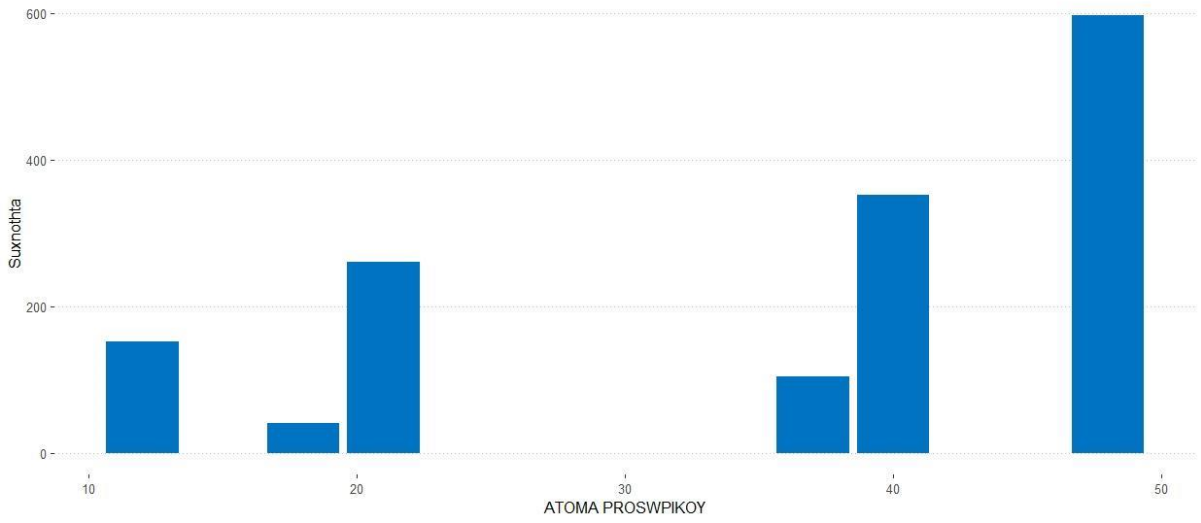


Εικόνα 21: Συχνότητα βλαβών ανά έτος βλάβης

Στο διάγραμμα της εικόνας 21 θα συνδέσουμε τις βλάβες με τις χρονιές εμφάνισης αυτών ωστε να βγάλουμε χρήσιμα συμπεράσματα κατά πόσον όσο περνάνε τα χρόνια και τα πλοία καταπονούνται με αποστολές και μέρες πλεύσης εμφανίζονται περισσότερες βλάβες και αν υπήρξαν χρονιές που λόγω συνθηκών (πχ αυξημένη επιχειρησιακή δράση) παρουσιάστηκαν περισσότερες βλάβες. Οι περισσότερες βλάβες εμφανίζονται τα τελευταία χρόνια μετά το 2010 όπως απεικονίζεται στην εικόνα 21. Η πληροφορία αυτή είναι πολύ σημαντική και θα μπορούσε να ξεκινήσουν ενέργειες για εντατικότερη συντήρηση και μοντέλο πρόβλεψης για τυχόν αύξηση ρυθμού εμφανίσεων βλαβών όσο περνάνε τα χρόνια ωστε να εξεταστεί τυχόν αντικατάσταση πλοίων του στόλου. Αυτό ίσως ερμηνεύεται διότι αρχικά με την πάροδο των ετών αυξάνεται τόσο η παλαιότητα των πλοίων όσο και οι μέρες πλεύσης τους, επίσης η οικονομική κρίση επηρέασε

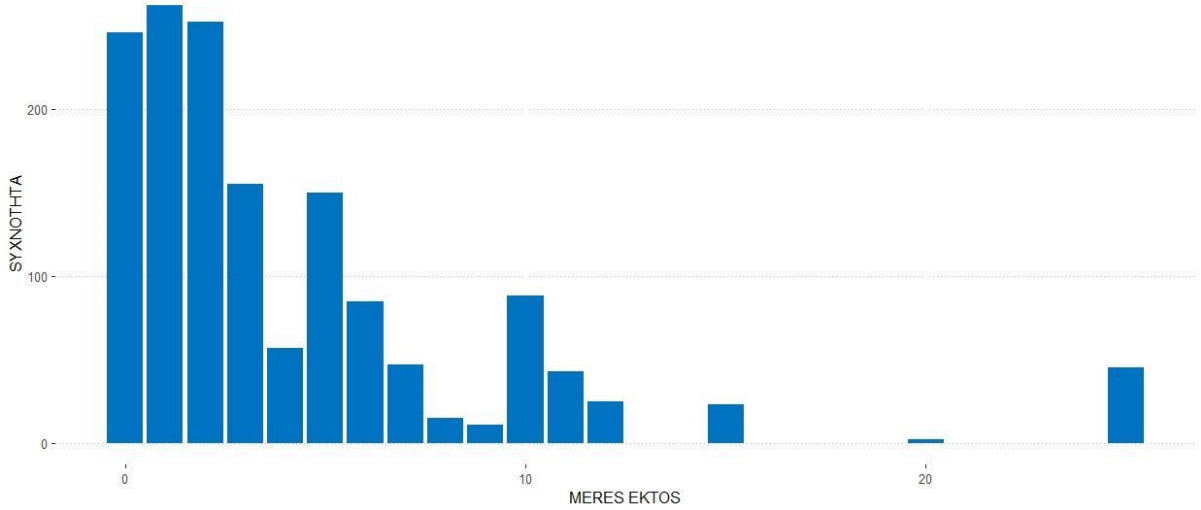
σημαντικά τόσο την προμήθεια υλικών για ορθές συντηρήσεις, όσο και ανταλλακτικών με συνέπεια την καταπόνηση μηχανημάτων, αλλά και παράλληλα η έξαρση του μεταναστευτικού αύξησε κατά πολύ τον αριθμό πλεύσιμων ημερών των πλοίων στο Αιγαίο τα τελευταία χρόνια.

Έγινε μια προσπάθεια συσχετισμού του αριθμού προσωπικού με τη συχνότητα βλαβών όπως φαίνεται στο διάγραμμα της εικόνας 22. Όπως είναι λογικό μεγαλύτερα πλοία έχουν περισσότερο προσωπικό και είναι φορείς περισσότερων συστημάτων άρα παρουσιάζουν μεγαλύτερη συχνότητα βλαβών. Μεγάλο ρόλο παίζει και η μεγαλύτερη ποικιλία στα άτομα που χειρίζονται κάθε συσκευή, συνεπώς συχνά όσο μεγαλύτερος είναι ο αριθμός προσωπικού οδηγεί σε κακή χρήση άρα και βλάβες.

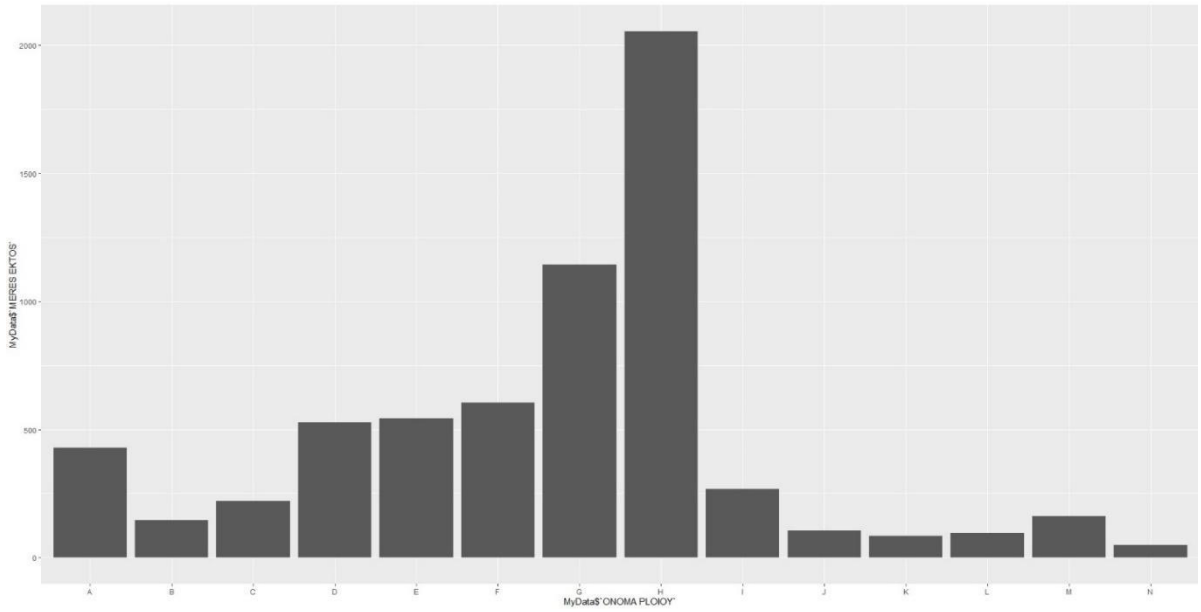


Εικόνα 22: Συχνότητα βλαβών ανά αριθμό προσωπικού πλοίου

Όταν ένα πλοίο παρουσιάζει μια βλάβη το πιο σημαντικό από όλα είναι τι επίπτωση θα έχει για το πλοίο αυτό (π.χ. θα μείνει εκτός πλεύσης, θα χάσει την μαχητική του ικανότητα κτλ.). Στα δεδομένα υπάρχει μόνο η μια πληροφορία, δηλαδή πόσες μέρες μένει εκτός ένα πλοίο. Κοιτώντας τις παρακάτω εικόνες 23 & 24, γίνεται ευκολά αντιληπτό ότι στις περισσότερες βλάβες τα πλοία δεν μένουν εκτός πλεύσης πολλές ημέρες αλλά έχουν υπάρξει και περιπτώσεις που έχουν μείνει εκτός και πάνω από 20 μέρες.



Εικόνα 23: Συχνότητα ημερών εκτός πλεύσης πλοίων



Εικόνα 24: Μέρες εκτός λειτουργίας ανά πλοίο

Αν συνδυαστούν και τα δύο γραφήματα θα μπορούσε να γίνει αντιληπτό ότι τα πλοία που εμφάνιζαν τις περισσότερες βλάβες ήταν και αυτά που έμεναν εκτός πλεύσης πάνω από 20 μέρες, παρέχοντας μας μια ακόμη σημαντική πληροφορία. Πλέον γνωρίζοντας ποια πλοία είναι πιο πολύ εκτός αλλά και πόσο κατά μέσο όρο μένει εκτός το σύνολο των πλοίων θα μπορέσουν να σχεδιαστούν πιο προσεκτικά οι επιχειρησιακές δράσεις (π.χ. μη χρησιμοποίηση πλοίων που παρουσιάζουν πολλές βλάβες, ταυτόχρονα σε αποστολές στο Αιγαίο ώστε να μην ρισκάρουμε να μείνουν ταυτόχρονα εκτός). Επιπρόσθετα θα γνωρίζουμε πόσο ευάλωτοι είμαστε και για πόσο ενώ δίνεται και το έναυσμα για συλλογή πιο πολλών δεδομένων. Τέτοια δεδομένα θα μπορούσε να

είναι ενδεχομένως η χρονική περίοδος που μένει εκτός ένα πλοίο (π.χ. καλοκαίρι ή χειμώνας) για να γίνει τυχόν σύνδεση με τις καιρικές συνθήκες δίνοντας μας έτσι μια ακόμα καλύτερη εικόνα για το πότε και γιατί είμαστε πιο ευάλωτοι.

4 ΣΥΜΠΕΡΑΣΜΑΤΑ

Η Ανάλυση των Δεδομένων που μας παρείχαν παρότι αυτά δεν ήταν πάρα πολλά, οδήγησε σε πολύ σημαντικά αποτελέσματα και συμπεράσματα και δίνει το έναυσμα για συλλογή πολλών περισσότερων δεδομένων και παραμέτρων ώστε να γίνει ακόμη πιο ολοκληρωμένη η συμπερασματολογία. Μέσω της μεθοδολογίας που ακολουθήθηκε έγινε εμφανής η σημαντικότητα της Ανάλυσης Δεδομένων ώστε να οδηγηθούμε σε ένα Σύστημα Λήψης Απόφασης και μας κάνει εμφανές πόσο σημαντικό είναι να συγκεντρώσουμε πολύ μεγαλύτερο όγκο ποικίλων δεδομένων σε βάθος χρόνου ώστε μετά να έχουμε να διαχειριστούμε Βάση Μεγάλων Δεδομένων προς ανάλυση, τα οποία με τη σειρά τους θα μας οδηγήσουν σε ακόμη πιο χρήσιμα συμπεράσματα.

Μέσω της επεξήγησης των διαγραμμάτων που παρουσιάστηκαν ανωτέρω, αναλύθηκαν τα διάφορα συμπεράσματα που προέκυψαν και δίνουν την ευκαιρία για περισσότερη έρευνα και ανάληψη ενεργειών με γνώμονα την αύξηση της μαχητικής ικανότητας του ΠΝ, προλαμβάνοντας ή ελαχιστοποιώντας τις βλάβες ή σε περίπτωση που αυτό δεν είναι εφικτό μειώνοντας ενέργειες διαδικασίες και ημέρες εκτός πλεύσης.

Οι βλάβες σε ένα πλοίο είναι ένα από τα σημαντικότερα πράγματα στα όποια θα πρέπει να επικεντρώνεται ο εκάστοτε κυβερνήτης, διότι είναι εξίσου σημαντικές με τις προμήθειες και τα καύσιμα. Αν για παράδειγμα ο εκάστοτε εχθρός μάθαινε τι βλάβες έχει το κάθε πλοίο και με τι συχνότητα εμφανίζονταν θα μπορούσε να προβλέψει και να επιτεθεί όταν ο στόλος βρίσκεται σε καταστολή. Γι’ αυτό και είναι σημαντικό να γίνεται καταγραφή των βλαβών σε κάθε ένα από τα πλοία έτσι ώστε να μελετηθεί η συχνότητα των βλαβών, που οφείλεται και αν οι βλάβες εμφανίζονται με ομαλότητα σε όλα τα πλοία και όχι μεμονωμένα. Φυσικά κάτι τέτοιο έχει σημασία αν το πλοίο μένει κάποιες μέρες εκτός δράσης ή αν οι βλάβες αυτές κοστίζουν πολύ ή αν επιδρούν σημαντικά στη μαχητική του ικανότητα. Ίσως αν έχει παρατηρηθεί κάποια αυξημένη συχνότητα να πρέπει να γίνει μια έρευνα αντικατάστασης αυτού του προβληματικού τμήματος του στόλου, είτε αλλαγή προβληματικών συστημάτων και ανανέωση με καινούρια.

Στην περίπτωση της παρούσας μεταπτυχιακής διατριβής οι περισσότερες βλάβες με διαφορά εμφανίζονται στα πλοία ομαδοποίησης 2 και 3 τα οποία αποκτήθηκαν μετά το 1990. Οι βλάβες που ξεχωρίζουν με διαφορά είναι αυτές που οφείλονται στην Κ. Μηχανή και συγκεκριμένα όπως αναφέρθηκε παραπάνω στις «No2 Κ. Μηχανή» και «No1 Κ. Μηχανή» (κάθε πλοίο έχει 2 Κ. Μηχανές), αλλά και αυτές που ανήκουν στην ευρύτερη κατηγορία «Διάφορες». Όσον αφορά

αυτήν την δεύτερη κατηγορία είναι κάτι το οποίο ίσως να έπρεπε να αξιολογηθεί διότι είναι μια κατηγορία που δεν αρμόζει σε καλή καταγραφή βλαβών. Ο χρήστης επιλέγει πολλές φορές για συντομία χρόνου ή και άγνοιας ακόμη, να χαρακτηρίζει τις βλάβες ότι ανήκουν στην κατηγορία «Διάφορες». Αυτό δημιουργεί προβλήματα και δυσκολίες στην άντληση συμπερασμάτων και πληροφορίων. Είναι βέβαιο πως για περισσότερες από μια βλάβες θα μπορούσε να είχε επιλεχθεί κανονική κατηγορία από τις υπόλοιπες που υπάρχουν με σκοπό την καλύτερη ανάλυση σε επόμενο στάδιο.

Επίσης, παρατηρείται ότι οι ομαδοποιήσεις που εμφανίζουν με διαφορά το μεγαλύτερο ποσοστό βλαβών είναι αυτές με το μεγαλύτερο αριθμό προσωπικού πράγμα που υποδηλώνει το μέγεθος τους αλλά και ότι κάποιες βλάβες ίσως να προέρχονται λόγω αυτού. Στην σύγχρονη κοινωνία είναι σύνηθες και παρατηρείται ολοένα και περισσότερο ότι όπου υπάρχει μεγάλο σύνολο ατόμων δεν υπάρχει και η κατάλληλη συνεννόηση. Άρα, θα μπορούσε να γίνει και μια υπόθεση ότι μόνο και μόνο γι' αυτό δημιουργούνται περισσότερες βλάβες σε αυτά τα πλοία και ιδίως μηχανικές. Τονίζεται το «Μηχανικές» διότι κάτι τέτοιο θα μπορούσε να οφείλεται στην κακή λειτουργία των μηχανών, αλλά και στην κακή συντήρηση που μπορεί να προέλθει από μια κακή συνεννόηση. Τα πλοία παρόλο που περιβάλλονται από μεγαλύτερο αριθμό ατόμων παρατηρείται ότι δεν είναι άρτια καταρτισμένα αφού η συχνότερη και άμεση ενέργεια διόρθωσης της βλάβης είναι να καλείται κάποια συνεργειακή βοήθεια και λόγω του ότι τα συνεργεία είναι εξειδικευμένα και διαθέτουν κατάλληλα εργαλεία αφήνοντάς το πλοίο δεμένο και στάσιμο πράγμα το οποίο καθιστά ευάλωτη την άμυνα της χώρας.

Μία ακόμη υπόθεση που θα μπορούσε να γίνει, είναι ότι τα πλοία αυτά που αποκτήθηκαν μετά από το 1990 όπως αυτά των ομαδοποιήσεων 2 και 3, δεν ήταν εξαρχής προβληματικά μιας και οι περισσότερες βλάβες παρουσιάζονται με διαφορά τα τελευταία χρόνια.. Φυσικά όμως για να μετρηθεί κάπως αυτό θα έπρεπε να είναι γνωστό και ο συνολικός αριθμός ημερών πλεύσης όλα αυτά τα χρόνια σε σχέση και με τα υπόλοιπα πλοία το οποίο δε βρισκόταν στα δεδομένα μας. Επιπλέον, αφού παρατηρείται ότι αυτές οι συγκεκριμένες ομάδες πλοίων παρουσιάζουν αυξημένα ποσοστά βλαβών ίσως να έπρεπε να αξιολογηθεί κατά ποσό είναι προς το συμφέρον της χώρας να διατηρούνται σε ενέργεια τα πλοία αυτά. Φυσικά κάτι τέτοιο θα μπορούσε να καταμετρηθεί με το κόστος που απαιτείται για να επισκευάζονται όλες αυτές οι βλάβες και το πόσες μέρες μένει εκτός πλεύσης ένα πλοίο καθιστώντας την χώρα ευάλωτη σε προκλήσεις συγκριτικά με το κόστος απόκτησης νέων μονάδων. Αν το κόστος που απαιτείται για να επισκευάζονται όλες αυτές οι βλάβες αποδειχθεί ασύμφορο θα μπορούσε να μελετηθεί μια υποτιθέμενη αγορά πιο σύγχρονων

πλοίων που ίσως λόγω λιγότερων ημερών πλεύσης ως τώρα μα και πιο σύγχρονων συστημάτων να αποδειχθούν καλύτερη επιλογή.

Επιπρόσθετα, τα αρμόδια τμήματα θα πρέπει να ενημερώνονται και εκσυγχρονίζονται με τα νέα συστήματα και να εκπαιδεύονται ανά τακτά χρονικά διαστήματα στην σωστή καταγραφή όλων των κινήσεων, ενεργειών και βλαβών έτσι ώστε στο μέλλον να μπορεί να γίνεται ακόμη καλύτερη αξιοποίηση της αντίστοιχης πληροφορίας. Κάτι τέτοιο θα μπορούσε να είναι η καταγραφή όχι μόνο του εκάστοτε κόστους της εκάστοτε βλάβης αλλά και του κόστους απόκτησης του εκάστοτε πλοίου.

Αποκτώντας πλήρη εικόνα των βλαβών και αποκτώντας καλύτερη εικόνα στα χρονικά δεδομένα και όχι απλά ανά έτος, ίσως να μπορούσε να γίνει προβλέψιμή η εκάστοτε βλάβη με σκοπό να γίνεται καλύτερη αξιοποίηση του αποθεματικού φυσικά με απώτερο σκοπό την ετοιμότητα των πλοίων στον χώρο δράσης, μιας και τα κατάλληλα ανταλλακτικά και θα είχαν προμηθευτεί απο πριν και θα γίνονταν άμεσες αντικαταστάσεις και συντηρήσεις.

Η ανωτέρω μεθοδολογία μας δίνει το έναυσμα για διεύρυνση των συμπερασμάτων και των αποτελεσμάτων απο την μελέτη βλαβών αν και εφόσον διευρυνθεί η Data Base των παρεχόμενων στοιχείων και αποφασιστεί η δημιουργία ενός ενιαίου συστήματος καταγραφής δεδομένων.

Αρχικά θα μπορούσε να γίνει σύνδεση των πλεύσιμων ημερών κάθε πλοίου με τη συχνότητα και τον τύπο των βλαβών που παρουσιάζουν ώστε να καθοριστεί κρίσιμος αριθμός ημερών στην οποία δύναται να γίνει παύση για συντήρηση. Με την ίδια λογική και εφαρμόζοντας το παραπάνω μοντέλο θα μπορούσε να γίνει σύνδεση των ωρών λειτουργίας κάθε μηχανήματος με την εμφάνιση βλάβης αυτού ώστε τυχόν υπερβάσεις λειτουργίας πέραν των προβλεπόμενων συντηρήσεων να αποφευχθούν.

Πολύ σημαντικό εργαλείο για αντίστοιχη έρευνα θα ήταν και η λεπτομερής καταγραφή ανταλλακτικών με κάποια κωδικοποίηση τα οποία και απαιτήθηκαν για την αποκατάσταση βλαβών παράλληλα με το κόστος αυτών. Σε αυτή τη περίπτωση και με τη μεθοδολογία που ακολούθησα θα μπορούσαμε να βγάλουμε συμπεράσματα ώστε να έχουμε σε αποθήκες stock σε ορισμένα κρίσιμα ανταλλακτικά ώστε να πετύχουμε αμεσότητα αποκατάστασης βλαβών και λόγω μαζικών παραγγελιών τυχόν καλύτερες τιμές αγοράς. Παράλληλα λαμβάνοντας υπόψη και το κόστος θα μπορούμε κάθε χρονιά να εξάγουμε έναν προϋπολογισμό με βάση τις ανάγκες του επόμενου έτους ώστε να αντιμετωπίσουμε τυχόν προβλήματα που προκύπτουν απο την ρευστότητα χρημάτων λόγω κρίσης κάνοντας έτσι ορθολογική κατανομή πόρων.

Ένας ακόμη τρόπος να διευρύνουμε τις ανωτέρω μεθόδους για πληρέστερη συμπερασματολογία, είναι η λεπτομερής καταγραφή και σύνδεση με έκαστη βλάβη του συνόλου των ενεργειών που απαιτήθηκαν και αφορούσαν συνεργειακή συνδρομή είτε απο συνεργεία του ΠΝ είτε απο ιδιώτες. Με αυτό το τρόπο θα μπορούσαμε να στελεχώσουμε με μεγαλύτερη επάρκεια τα συνεργεία με σωστό αριθμό ατόμων αφού θα καταφέρουμε να εξάγουμε συμπεράσματα για το ποιες ειδικότητες απαιτούνται πιο πολύ. Έτσι θα εξασφαλιζόταν η ομαλότερη και αποδοτικότερη λειτουργία των συνεργείων και θα δινόταν έμφαση στην στοχευμένη εκπαίδευση προσωπικού πάνω στην αντιμετώπιση και την επίλυση των σημαντικότερων και συχνότερων βλαβών. Παράλληλα με αυτό το τρόπο θα μπορούσε να διευρυνθεί η λίστα των εξωτερικών συνεργατών που θα χρησιμοποιηθούν σε βλάβες ώστε να εξασφαλίζεται η ποικιλία και πιο ανταγωνιστικές τιμές λόγω αυτής.

Τα ανωτέρω δεδομένα δεν ήταν προσβάσιμα και ως τώρα δεν έχει προβλεφθεί ώστε να μπουν σε ενιαία βάση δεδομένων πλην όμως δίνεται το έναυσμα με τη παρούσα εργασία να ξεκινήσει μια προσπάθεια εκπαίδευσης συγκεκριμένου προσωπικού σε συστήματα καταχώρησης και καταγραφής ώστε με τις διαδικασίες PCA, Lasso, Stepwise, Linear-Logistic Regression που περιγράφηκαν προηγουμένως να βρίσκουμε τις εκάστοτε σημαντικές μεταβλητές για τη δημιουργία βέλτιστων μοντέλων ωστε να έχουμε πληρέστερα συμπεράσματα και να οδηγηθούμε στην πλήρη αξιοποίηση της μεθοδολογίας με την οποία ασχολήθηκε η εργασία ώστε το ΠΝ να έχει πολλαπλά οφέλη. Ο τρόπος που περιγράφηκε και η μεθοδολογία που ακολουθήθηκε με χρήση αρχείων Excel και γλώσσας R, μπορεί να αποτελέσει οδηγό για έναρξη τέτοιων εργασιών ώστε σε βάθος ετών και με περισσότερες παραμέτρους δεδομένων όπως περιγράφηκαν και πιο πάνω να καταλήξουμε σε δημιουργία Βάσης Μεγάλων Δεδομένων επωφελούμενοι απο την ανάλυσή τους.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Anderberg, M. (1973). "Cluster analysis for applications". NewYork: AcademicPress.

Barker, Richard (1990). "CASE Method: Entity Relationship Modelling". Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA

Boyd, D. and Crawford, K (2011). "Six Provocations for Big Data". Social Science Research Network: A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society.

Bughin, J., Chui, M., Manyika, J. (2010). "Clouds, big data, and smart assets: Ten tech-enabled business trends to watch". [URL: http://www.mckinsey.com/industries/high-tech/our-insights/clouds-big-data-andsmart-assets-ten-tech-enabled-business-trends-to-watch]

Davis, K., Patterson, D. (2012). "Ethics of Big Data: Balancing risk and innovation". San Francisco: O'Reilly Media.

Efroymson M.A. (1960), "Multiple Regression Analysis," In: A. Ralston and H. S. Wilf, Eds., Mathematical Methods for Digital Computers, John Wiley, New York.

Emani, C., Cullot, N., Nicolle, C. (2015). "Understandable Big Data: A survey". Computer Science Review.

Guojun Gan, Chaoqun Ma, Jianhog Wu, (2007). "Data Clustering Theory, Algorithms and Applications"

Hilbert, Martin. "Big Data for Development: A Review of Promises and Challenges. Development Policy Review." martinhilbert.net.

http://artemis.cslab.ece.ntua.gr:8080/jspui/bitstream/123456789/13486/1/DT2017-0144.pdf

http://data-mining.philippe-fournier-viger.com/introduction-clustering-k-means-java-code/

http://hypatia.teiath.gr/xmlui/bitstream/handle/11400/20185/lb_04174_thanos_thesis.pdf?sequence=1

http://people.revoledu.com/kardi/tutorial/Similarity/EuclideanDistance.html

http://repository.library.teimes.gr/xmlui/bitstream/handle/123456789/4537/%CE%A4%CE%95%CE%A7%CE%9D%CE%99%CE%9A%CE%95%CE%A3%20%CE%91%CE%9D%CE%91%CE%9B%CE%A5%CE%A3%CE%97%CE%A3%20%CE%9C%CE%95%CE%93%CE%91%CE%9B%CE%A9%CE%9D%20%CE%94%CE%95%CE%94%CE%9F%CE%9C%CE%95%CE%9D%CE%A9%CE%9D..pdf?sequence=1&isAllowed=y

http://www.dummies.com/programming/big-data/data-science/data-science-performing-hierarchical-clustering-with-python/

https://docs.tibco.com/products/tibco-analytics-7-7-0

https://en.wikipedia.org/wiki/Association_rule_learning

https://en.wikipedia.org/wiki/Cluster_analysis

https://en.wikipedia.org/wiki/Data_analysis

https://en.wikipedia.org/wiki/Euclidean_distance

https://en.wikipedia.org/wiki/Hierarchical_clustering

https://en.wikipedia.org/wiki/K-means_clustering

https://en.wikipedia.org/wiki/Mahalanobis_distance

https://en.wikipedia.org/wiki/Microsoft_SQL_Server

https://en.wikipedia.org/wiki/R_(programming_language)

https://en.wikipedia.org/wiki/Stepwise_regression

https://en.wikipedia.org/wiki/Taxicab_geometry

https://en.wiktionary.org/wiki/Manhattan_distance

https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/

https://joernhees.de/blog/2015/08/26/scipy-hierarchical-clustering-and-dendrogram-tutorial/

https://towardsdatascience.com/unsupervised-learning-and-data-clustering-eeecb78b422a

https://www.programcreek.com/2014/02/k-means-clustering-in-java/

https://www.quora.com/How-do-banks-store-financial-data

https://www.sas.com/el_gr/insights/big-data/what-is-big-data.html

Jacobs, A. (2009). "The Pathologies of Big Data". ACMQueue.

Jain, A. and Dubes, R. (1988). “Algorithms for Clustering Data”. Englewood Cliffs, NJ: Prentice–Hall.

Journal of Computing in Higher Education (2005). “Using Qualitative research methods in higher education”

Ka-Chun Wong. “A short Survey on Data Clustering Algorithms”

Krishnan, K. (2013). “Data warehousing in the age of big data: A volume in MK series on business intelligence”. Elsevier.

Mántaras, R. (1991). “A distance-based attribute selection measure for decision tree induction. In Machine Learning”. Boston: Kluwer Academic

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A. (2011). “Big data: the next frontier for innovation, competition, and productivity”. McKinsey Global Institute.

Mashey J.R. (1998). "Big Data and the Next Wave of InfraStress". Slides from invited talk. Usenix.

Mayer – Schonberger, V., Cukier, K. (2013). “Big data: a revolution that will transform how we live, work, and think”. New York: Eamon Dolan/Houghton Mifflin Harcourt.

Moore, Gordon E. (1965). “Cramming more components onto integrated circuits”. Retrieved 2016-07-01.

Rakesh Agrawal, Tomasz Imielinski, Arun Swami (1993). “Mining Association Rules between Sets of Items in Large Databases”. IBM Almaden Research. Washington, D.C., US, San Jose, CA

Rui Miguel Forte June (2015). “Mastering Predictive Analytics with R”.

Shamoo & Resnik (2003). “Responsible Conduct of Research”. Oxford University Press, New York

Sheppard, D. C. Tomberlin, J. K. Joyce, J. A. Kiser, B. C. Sumner, S. M. (2002). “Rearing methods for the black soldier fly” (Diptera: Stratiomyidae). J. Med. Entomol

Singh, J., Singla, V. (2015). “Big Data: Tools and technologies in Big Data”. International Journal of Computer Applications.

Steve Lohr (2013). "The Origins of 'Big Data': An Etymological Detective Story". New York Times.

Steve Lohr, S (. (2013). "The Origins of 'Big Data': An Etymological Detective Story". New York Times.

Vitolo, C., Elkhatib, Y., Reusser, D., Macleod, C., Buytaert, W. (2015). “Web technologies for environmental Big Data”. Environmental Modelling & Software.

Wishart, D. (2002). “k-means clustering with outlier detection, mixed variables and missing values”. In Schwaiger, M., and Opitz, O., editors, Exploratory Data Analysis in Empirical Research, New York: Springer.

Yaqoob, I., Hashem, I., Gani, A., et al. (2016). “Big data: From beginning to future”. International Journal of Information Management.

ΠΑΡΑΡΤΗΜΑ Α

```
library(readxl)
Teliko <- read_excel("C:/Users/grckont/OneDrive - SAS/Chris Pap/Excel/Teliko.xlsx")
View(Teliko)
MyData<-Teliko
View(MyData)

# Check for Na's
sapply(MyData, function(x) sum(is.na(x)))
library(ggpubr)
ggplot(MyData, aes(`TYPOS BLABHS`)) +
  geom_bar(fill = "#0073C2FF") +
  theme_pubclean()

library(ggpubr)
ggplot(MyData, aes(`TYPOS PLOIOY`)) +
  geom_bar(fill = "#0073C2FF") +
  theme_pubclean()
library(ggpubr)
ggplot(MyData, aes(`ONOMA PLOIOY`)) +
  geom_bar(fill = "#0073C2FF") +
  theme_pubclean()
library(ggpubr)
ggplot(MyData, aes(`ETOS APOKTHSHS`)) +
  geom_bar(fill = "#0073C2FF") +
  theme_pubclean()
library(ggpubr)
ggplot(MyData, aes(`ATOMA PROSWPIKOY`)) +
  geom_bar(fill = "#0073C2FF") +
  theme_pubclean()
library(ggpubr)
```

```
ggplot(MyData, aes(`BLABH`)) +
  geom_bar(fill = "#0073C2FF") +
  theme_pubclean()
library(ggpubr)
ggplot(MyData, aes(`AITIA`)) +
  geom_bar(fill = "#0073C2FF") +
  theme_pubclean()
library(ggpubr)
ggplot(MyData, aes(`ETOS BLABHS`)) +
  geom_bar(fill = "#0073C2FF") +
  theme_pubclean()
library(ggpubr)
ggplot(MyData, aes(`ARITHMOS ENERGEIAS`)) +
  geom_bar(fill = "#0073C2FF") +
  theme_pubclean()
library(ggpubr)
ggplot(MyData, aes(`ENERGEIES`)) +
  geom_bar(fill = "#0073C2FF") +
  theme_pubclean()
library(ggpubr)
ggplot(MyData, aes(`MERES EKTOS`)) +
  geom_bar(fill = "#0073C2FF") +
  theme_pubclean()

library(dplyr)
df <- MyData %>%
  group_by(`ONOMA PLOIOY`, `TYPOS PLOIOY`) %>%
  summarise(`ENERGEIES` = n())
df
library(tidyverse)
ggplot(MyData) + geom_col(aes(x = MyData$`ONOMA PLOIOY`, y =MyData$`MERES
EKTOS`))
```

```
MyData4 <- MyData
MyData4$`AA` <- NULL
MyData4$`TYPOS PLOIOY` <- NULL
MyData4$`ONOMA PLOIOY` <- NULL
MyData4$`BLABH` <- NULL
MyData4$`AITIA` <- NULL
MyData4$`TYPOS BLABHS` <- NULL
MyData4$`ENERGEIES` <- NULL
MyData4$`KOSTOLOGO` <- NULL

MyData4$`ETOS APOKTHSHS`<- as.numeric(MyData4$`ETOS APOKTHSHS`)
MyData4$`ATOMA PROSWPIKOY`<- as.numeric(MyData4$`ATOMA PROSWPIKOY`)
MyData4$`ETOS BLABHS`<- as.numeric(MyData4$`ETOS BLABHS`)
MyData4$`ARITHMOS ENERGEIAS`<- as.numeric(MyData4$`ARITHMOS ENERGEIAS`)
MyData4$`MERES EKTOS`<- as.numeric(MyData4$`MERES EKTOS`)

MyData4
MyData5 <- na.omit(MyData4)
set.seed(1236)
# we start the lasso procedure
library(glmnet)
mfull <- lm(`MERES EKTOS`~.,data=MyData5)
x2 <-model.matrix(mfull)
lasso4 = glmnet(x2, MyData5$`MERES EKTOS` )
plot(lasso4, xvar='lambda', label=T)
summary(lasso4)
# after we finish lasso we create the model with the variables that lasso inidicates to us
Myfull <- lm(`MERES EKTOS`~ ., data= MyData5)
summary(Myfull)
#library(MASS)
```

```
# stepAIC(Myfull)
# summary(Myfull)
### STEP . . .. .
MyData_null <- lm(`MERES EKTOS`~ 1.,data=MyData5)
Myfull2 <- lm(`MERES EKTOS`~ .,data=MyData5)
model3 <- step(MyData_null, scope = list(lower= MyData_null, upper=Myfull2), direction = "forward")
model4 <- step(Myfull, scope = list(lower= MyData_null, upper=Myfull2), direction = "backward")
summary(model3)
summary(model4)
MyData5
Correlation_matrix<-cor(MyData5)
library(corrplot)
corrplot(Correlation_matrix)
palette = colorRampPalette(c("green", "white", "red")) (20)
heatmap(x = Correlation_matrix, col = palette, symm = TRUE)
par(mfrow=c(1,1))
pie(table(MyData$`BLABH`))
labls <- count(MyData$`BLABH`)
labls[MyData$count <  10] <- NA
pie(MyData$count, paste(labls))
tail(names(sort(table(MyData$`BLABH`))), 10)
tail(names(sort(table(MyData$`ONOMA PLOIOY`,MyData$`ATOMA PROSWPIKOY`))), 3)
tail(names(sort(table(MyData$`AITIA`))), 10)
```