ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΟΡΓΑΝΩΣΗΣ ΚΑΙ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Θέμα:

«Επιχειρηματική Αναλυτική : Ανάλυση Μεγάλων Δεδομένων με χρήση της Γλώσσας Προγραμματισμού R»

ΟΝΟΜΑΤΕΠΩΝΥΜΟ: Ουρανία Βλάχου
ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: Μαραβελάκης Πέτρος

ΑΘΗΝΑ 2021

ΒΕΒΑΙΩΣΗ ΕΚΠΟΝΗΣΗΣ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

«Δηλώνω υπεύθυνα ότι η διπλωματική εργασία για τη λήψη του μεταπτυχιακού τίτλου σπουδών, του Πανεπιστημίου Πειραιώς, στη Διοίκηση Επιχειρήσεων : MBA» με τίτλο:

«Επιχειρηματική Αναλυτική : Ανάλυση Μεγάλων Δεδομένων με χρήση της Γλώσσας Προγραμματισμού R

έχει συγγραφεί από εμένα αποκλειστικά και στο σύνολό της. Δεν έχει υποβληθεί ούτε έχει εγκριθεί στο πλαίσιο κάποιου άλλου μεταπτυχιακού προγράμματος ή προπτυχιακού τίτλου σπουδών, στην Ελλάδα ή στο εξωτερικό, ούτε είναι εργασία ή τμήμα εργασίας ακαδημαϊκού ή επαγγελματικού χαρακτήρα.

Δηλώνω επίσης υπεύθυνα ότι οι πηγές στις οποίες ανέτρεξα για την εκπόνηση της συγκεκριμένης εργασίας, αναφέρονται στο σύνολό τους, κάνοντας πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου»

Υπογραφή Μεταπτυχιακού Φοιτητή Ονοματεπώνυμο

Βλάχου Ουρανία

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου κ. Μαραβελάκη Πέτρο που δέχτηκε να με βοηθήσει στην εκπόνηση της Διπλωματικής μου εργασίας, για το χρόνο που αφιέρωσε και για την άψογη συνεργασία που είχαμε για τη υλοποίησή της.

Παράλληλα ένα μεγάλο ευχαριστώ στην οικογένεια και τους φίλους μου για την αγάπη και την υποστήριξή τους.

Περίληψη

Στην ταχέως εξελισσόμενη πορεία της τεχνολογίας στη ζωή των ανθρώπων, κάθε οργανισμός θα πρέπει να ακολουθεί αυτή την εξελικτική πορεία προκειμένου να είναι βιώσιμος και λειτουργικός. Τα τελευταία χρόνια ο ορισμός «μεγάλα δεδομένα» κυριαρχεί στον κόσμο της αναλυτικής των επιχειρήσεων, προκειμένου να αντιμετωπιστεί η αβεβαιότητα στον επιχειρηματικό τομέα, συλλέγονται δεδομένα που μέσω διαφόρων τεχνικών επιτυγχάνεται η εξαγωγή χρήσιμων πληροφοριών. Ωστόσο ο μεγάλος όγκος των δεδομένων αυτών καθιστά τις κλασσικές τεχνικές της επιχειρησιακής έρευνας αδύνατες να εφαρμοστούν. Για αυτό το λόγο οι επιχειρήσεις κατέφυγαν στην εξέλιξη των προ αναφερόντων τεχνικών ώστε να διαχειρίζονται την πολυπλοκότητα και τον μεγάλο όγκο των δεδομένων προκειμένου να αποκτούν ανταγωνιστικό πλεονέκτημα μέσω των χρήσιμων αποτελεσμάτων από την σωστή διαχείριση τους. Έτσι δημιουργήθηκε ο όρος «Επιχειρηματική Ανάλυση» που αναλύεται στην παρούσα διπλωματική εργασία.

Περιεχόμενα

ΚΕΦΑΛΑΙΟ 1

ΑΝΑΛΥΤΙΚΗ ΜΕΓΑΛΩΝ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΕΞΟΡΙΣΜΟΣ ΑΥΤΩΝ (BIG DATA ANALYTICS AND DATA MINING)

1.1 Ορισμός Μεγάλων Δεδομένων (Big Data)

Ο πιο κοινός ορισμός των Big Data αναφέρεται στις καταστάσεις όπου η ποσότητα των δεδομένων γίνεται συντριπτική και δεν μπορεί να αντιμετωπιστεί με τις παραδοσιακές τεχνολογικές βάσεις δεδομένων και υπολογισμών. Ο ορισμός αυτών των «μεγάλων δεδομένων» αναφέρεται στα λεγόμενα 4 Vs : Volume (όγκος), Velocity (ταχύτητα), Variety (ποικιλία), και Veracity (ακρίβεια). *[Dominik Ryzko (2020) - "Modern BD architecture". John Wiley & Sons, Inc, New Jersey.]*

- **Volume (Όγκος)**: Ο όρος "Big" ορίζεται από τον όγκο των δεδομένων τα οποία είναι σε petabytes (1,024 terabytes) και αναμένεται στο κοντινό μέλλον να αυξηθούν σε zettabytes (1.000.000 terabytes). Τα μέσα κοινωνικής δικτύωσης παράγουν τα δεδομένα κατά σειρά terabytes καθημερινά και αυτή η ποσότητα δεδομένων είναι σίγουρα δύσκολο να αντιμετωπιστεί χρησιμοποιώντας τα υπάρχοντα παραδοσιακά συστήματα. Ένα terabyte αποθηκεύει τόσες πληροφορίες ισοδύναμες με όσες σε 1500 CD ή σε 220 DVD, όσες περίπου θα ήταν 16 εκατομμύρια φωτογραφίες στο Facebook.
- **Velocity (Ταχύτητα):** Αυτό το χαρακτηριστικό αφορά την ταχύτητα κατά την οποία ρέουν και εισέρχονται τα δεδομένα από τις διάφορες πηγές στις βάσεις δεδομένων. Για παράδειγμα, τα δεδομένα από αισθητήρες συσκευών μετακινούνταν συνεχώς στη βάση δεδομένων και αυτό το ποσό δεν είναι μικρό, με αποτέλεσμα τα παραδοσιακά συστήματα επεξεργασίας να μην είναι ικανά στην εκτέλεση των αναλύσεων αυτών των εν συνεχή κίνηση δεδομένων. Ο πολλαπλασιασμός των ψηφιακών συσκευών, όπως τα Smart phones και οι αισθητήρες, έχει οδηγήσει σε έναν πρωτοφανή ρυθμό δημιουργίας μεγάλων δεδομένων και οδηγεί σε αυξανόμενη ανάγκη για αναλύσεις σε πραγματικό χρόνο και σχεδιασμό

βάσει στοιχείων. Οι επιχειρήσεις μπορούν να εκμεταλλευτούν τα δεδομένα που προέρχονται από κινητές συσκευές και διακινούνται μέσω κινητών εφαρμογών για τη δημιουργία εξατομικευμένων προσφορών σε πραγματικό χρόνο για τους καθημερινούς πελάτες. Αυτά τα δεδομένα παρέχουν πληροφορίες σχετικά με τους πελάτες, όπως η γεωγραφική τοποθεσία, τα δημογραφικά χαρακτηριστικά και τις συνήθειες αγορών του παρελθόντος, τα οποία μπορούν να αναλυθούν σε πραγματικό χρόνο για να δημιουργήσουν πραγματική αξία για τον πελάτη.

- **Variety (ποικιλία):** η ποικιλία αναφέρεται στην ετερογένεια των μεγάλων δεδομένων. Η τεχνολογική εξέλιξη επιτρέπει στις επιχειρήσεις την επεξεργασία διαφόρων τύπων *δομημένων*, *ημιδομημένων* και *αδόμητων δεδομένων*. Τα *δομημένα δεδομένα* αναφέρονται σε πίνακες που υπάρχουν στα υπολογιστικά φύλλα ή στις σχεσιακές βάσεις δεδομένων, αποτελούν μόνο το 5% του συνόλου των υπό εξέταση δεδομένων. Ο όρος *μη δομημένα* αναφέρεται στα δεδομένα κειμένων, εικόνων, ήχων και βίντεο, τα οποία στερούνται δομικής οργάνωσης το οποίο καθιστά δύσκολη την ανάλυσή τους από τις μηχανές ανάλυσης. Τα *ημι-δομημένα* είναι μια μορφή δεδομένων που καλύπτει το φάσμα μεταξύ δομημένων και μη δομημένων και δεν έχει αυστηρά πρότυπα. Ένα παράδειγμα ημι-δομημένων δεδομένων είναι η γλώσσα σήμανσης (XML), μια γλώσσα κειμένου για την αποθήκευση και μεταφορά δεδομένων στον Ιστό. Έχει παρόμοια μορφή με τα αρχεία HTML, με την διαφορά ότι τα έγγραφα XML περιέχουν ετικέτες δεδομένων σχεδιασμένες από τον χρήστη, οι οποίες είναι αναγνώσιμες από το εκάστοτε μηχάνημα.
- **Veracity (ακρίβεια):** Κατά την IBM, Το τέταρτο χαρακτηριστικό των Big Data αναφέρεται στον βαθμό αξιοπιστίας των δεδομένων. Για παράδειγμα, τα συναισθήματα των πελατών στα κοινωνικά μέσα δικτύωσης είναι αβέβαια από τη φύση τους, καθώς συνεπάγονται ανθρώπινη κρίση. Ωστόσο, παρέχουν πολύτιμες πληροφορίες. Έτσι, η ανάγκη αντιμετώπισης της ανακρίβειας και αβεβαιότητας δεδομένων είναι μια άλλη πτυχή των Big Data, η οποία αντιμετωπίζεται

χρησιμοποιώντας εργαλεία και αναλυτικά στοιχεία που αναπτύχθηκαν για τη διαχείριση και την εξόρυξη αβέβαιων δεδομένων.

Η *τεχνητή νοημοσύνη (Artificial Intelligence AI*), τα κινητά, τα κοινωνικά και το Διαδίκτυο των πραγμάτων (Internet of Things IoT) οδηγούν την πολυπλοκότητα των δεδομένων μέσω νέων μορφών και πηγών δεδομένων. Για παράδειγμα, τα Big Data προέρχονται από αισθητήρες, συσκευές, βίντεο / ήχο, δίκτυα, αρχεία καταγραφής, εφαρμογές συναλλαγών, ιστούς και μέσα κοινωνικής δικτύωσης - μεγάλο μέρος των οποίων δημιουργήθηκε σε πραγματικό χρόνο και σε πολύ μεγάλη κλίμακα. *[https://www.ibm.com/gr-en]*

Η ανάλυση των Big Data επιτρέπει σε αναλυτές, ερευνητές και επιχειρηματικούς χρήστες να λαμβάνουν καλύτερες και ταχύτερες αποφάσεις χρησιμοποιώντας δεδομένα που προηγουμένως δεν ήταν προσβάσιμα ή αχρησιμοποίητα. Οι επιχειρήσεις μπορούν να χρησιμοποιήσουν προηγμένες τεχνικές ανάλυσης, όπως αναλυτικά κείμενα, μηχανική εκμάθηση (machine learning), αναλυτικά στοιχεία πρόβλεψης, εξόρυξη δεδομένων (data mining), στατιστικά στοιχεία και επεξεργασία φυσικής γλώσσας για να αποκτήσουν νέες πληροφορίες από πηγές δεδομένων που δεν έχουν αξιοποιηθεί προηγουμένως ανεξάρτητα ή μαζί με υπάρχοντα εταιρικά δεδομένα. *[https://www.ibm.com/gr-en]*

Σκεφτείτε όλα τα δεδομένα που συλλέγονται καθημερινά από όλους τους τομείς της επιχείρησης. Τεράστιος όγκος δεδομένων. Θα ήταν μεγάλο επίτευγμα για την επιχείρηση να μπορούσε να κάνει αυτά τα δεδομένα λειτουργικά για εκείνη. Για παράδειγμα, να μπορούσε να προβλέψει με μεγάλη ακρίβεια τις πωλήσεις του προϊόντος Χ το επόμενο τρίμηνο. Με αποτέλεσμα την υψηλότερη ικανοποίηση των πελατών της που θα οδηγούσε σε υψηλότερα κέρδη.

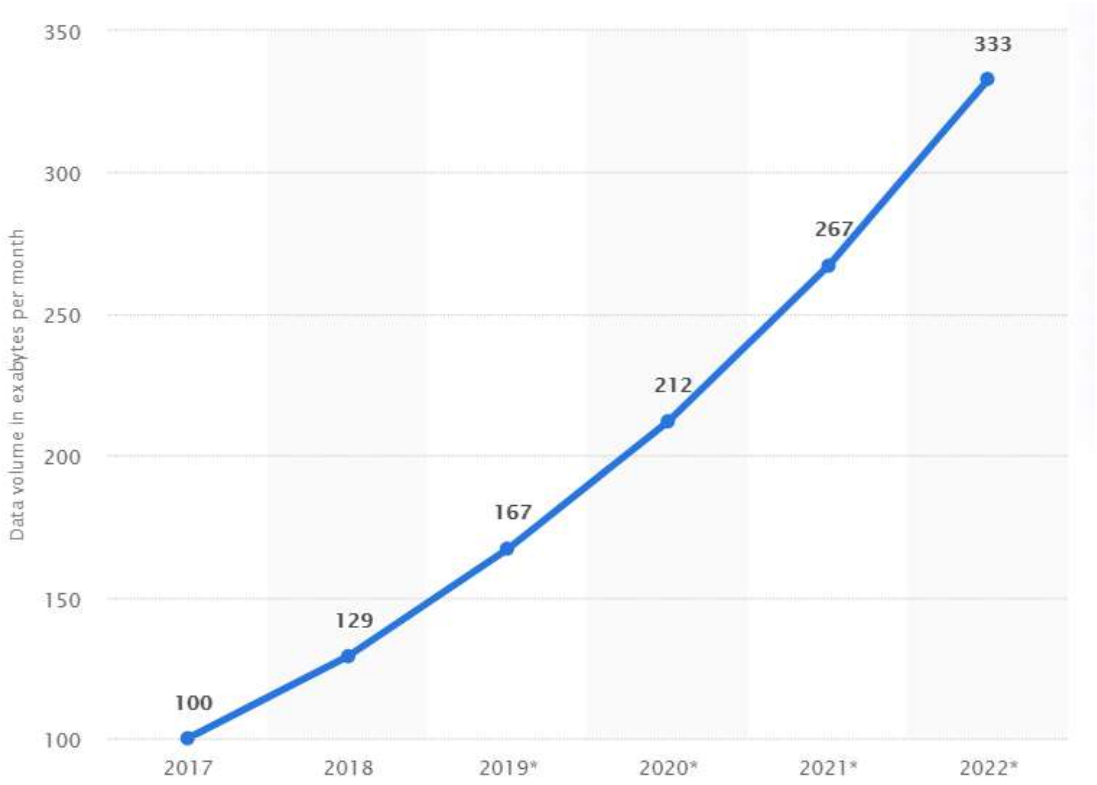
Εάν η επιχείρηση εκμεταλλευτεί την αποτελεσματικότητα των δεδομένων της, μπορεί να επιτύχει τον στόχο των υψηλότερων κερδών. Ωστόσο, πρόσφατη έρευνα από τους New Vantage Partners διαπιστώνει ότι οι περισσότερες μεγάλες εταιρείες αντιμετωπίζουν προβλήματα στην εκμετάλλευση αυτών των δεδομένων για την επίτευξη αυτού του στόχου. Βάση της συγκεκριμένης έρευνας, ένα ποσοστό της

τάξεως 69% των IT professionals δήλωσαν την αποτυχία τους να δημιουργήσουν data-driven επιχειρήσεις. Επιπλέον, το ποσοστό των επιχειρήσεων που αποκαλούν τον εαυτό τους “data-driven” μειώθηκε τα τελευταία χρόνια, από 37,1% το 2017 σε 32,4% το 2018 και κάτω του 31,0% στην τελευταία έρευνα.

Η **μηχανή εκμάθησης** (machine learning), βοηθά τις εταιρείες να εντοπίσουν μοτίβα βάση των αγοραστικών συνηθειών των πελατών, που έχουν κάνει «κλικ», που έχουν περάσει αρκετό χρόνο στις οθόνες τους, τι τους ενδιαφέρει περισσότερο κτλ. Χρησιμοποιώντας αυτές τις πληροφορίες σε συνδυασμό με άλλα δεδομένα, για παράδειγμα την ηλικία, το φύλο, την γεωγραφική τοποθεσία, οι εταιρείες μπορούν να προτείνουν στους πελάτες άλλα προϊόντα που ίσως τους ενδιαφέρουν προσαρμοσμένα στα γούστα τους. Η μηχανή εκμάθησης (machine learning) υπόσχεται να αντιστρέψει αυτή την αρνητική τάση, αυξάνοντας τον εσωτερικό βαθμό απόδοσης (ROI) των επιχειρήσεων που την χρησιμοποιούν (αύξηση ROI κατά 17% σύμφωνα με την Deloitte).

1.1.2 Μεγάλος όγκος δεδομένων (Large Data Sets)

Η ραγδαία εξελισσόμενη τεχνολογία υπολογιστών, επικοινωνιών και ψηφιακής αποθήκευσης, μαζί με την ανάπτυξη τεχνολογιών απόκτησης δεδομένων υψηλής απόδοσης, κατέστησαν δυνατή την συλλογή και αποθήκευση τεράστιου όγκου δεδομένων. Αυτά τα δεδομένα μπορεί να προέρχονται από το μητρώο ταμείου ενός γειτονικού καταστήματος, τις συναλλαγές πιστωτικών καρτών από τις τράπεζες, αρχεία βιολογικών εργαστηρίων, αρχεία μοτίβων τηλεφωνικών κλήσεων και πολλά άλλα. Οι περισσότερες εφαρμογές δημιουργούν ροές ψηφιακών αρχείων που αρχειοθετούνται σε τεράστιες βάσεις δεδομένων επιχειρήσεων. Ηλεκτρονικά ταχυδρομεία, ιστολόγια, δεδομένα συναλλαγών και δισεκατομμύρια ιστοσελίδες δημιουργούν terabytes νέων δεδομένων κάθε μέρα.



Εικόνα 1.1: Γράφημα Παγκόσμιας κατανάλωσης όγκου δεδομένων στο διαδίκτυο [πηγή www.statista.com]

Στο παραπάνω χρονοδιάγραμμα, από την statista.com, φαίνεται η παγκόσμια κατανάλωση όγκου δεδομένων στο διαδίκτυο. Το χρονοδιάγραμμα δείχνει μια πρόβλεψη για τα έτη 2019 2020 και 2021 του όγκου δεδομένων σε exabytes (1,000,000 terabytes) της παγκόσμιας κίνησης IP καταναλωτών. Το 2022, η παγκόσμια κίνηση IP καταναλωτών αναμένεται να φτάσει τα 333 exabytes ανά μήνα με σύνθετο ετήσιο ρυθμό ανάπτυξης 27%.

1.2 Ανάλυση Μεγάλων Δεδομένων

Σύμφωνα με μελέτη του MIT, οι επιχειρήσεις που χρησιμοποιούν τεχνικές ανάλυσης δεδομένων, είναι 5% έως 6% περισσότερο παραγωγικές από τους ανταγωνιστές τους. Η ανάλυση των μεγάλων δεδομένων παρέχει απαντήσεις σε ερωτήματα που οι αναλυτές/ερευνητές δεν ήξεραν ότι υπάρχουν. Πριν αρχίσει η διαδικασία της ανάλυσης, το σημαντικότερο σημείο είναι αυτό της εύρεσης του κατάλληλου συνόλου δεδομένων ανάμεσα στο χάος της διαθέσιμης πληροφορίας, ώστε να

υπάρχει δυνατότητα απόκτησης αξιόπιστης και αξιοποιήσιμης γνώσης.*[HCM Whitepaper: HR's Secret Weapon: The Power of Big Data]*

1.2.1 Επιχειρηματική Ευφυΐα (Business Intelligence BI)

Η Επιχειρηματική Ευφυΐα (Business Intelligence BI) ορίζεται ως *«ένα σύνολο από μεθόδους ανάλυσης, τεχνολογίες, ικανότητες και στρατηγικές, οι οποίες στόχο έχουν την επεξεργασία των διαθέσιμων δεδομένων και την εξαγωγή χρήσιμης πληροφορίας από αυτά, για την υποστήριξη της διαδικασίας λήψης επιχειρηματικών αποφάσεων»*.

Τα **συστήματα Επιχειρηματικής Ευφυΐας** δημιουργήθηκαν για την ικανοποίηση αυτών των σκοπών. Είναι εξειδικευμένα πληροφοριακά συστήματα τα οποία παρέχουν πληροφορίες με σκοπό την αποτελεσματική λήψη αποφάσεων. Τα δεδομένα συνδυάζονται με λογισμικό ικανό στην διεξαγωγή κατάλληλων αναλύσεων, με σκοπό την συνεχόμενη βελτίωση της ποιότητας των πληροφοριών. Στόχος αυτών των συστημάτων είναι, η ταχύτερη πρόσβαση στην πληροφορία, η ευκολότερη υποβολή ερωτημάτων στο σύστημα και σύνταξη αναφορών, η προχωρημένη ανάλυση δεδομένων, καθώς και η βελτίωση της ποιότητας των δεδομένων. Τα συστήματα αυτά, με την χρήση διαφόρων τεχνικών όπως η *OLAP (online analytical processing), η Στατιστική Ανάλυση, η Οπτικοποίηση (visualization) και η Εξόρυξη Δεδομένων (data mining)*, παρέχουν γνώση κρυμμένη μέσα σε αυτόν τον τεράστιο όγκο δεδομένων.

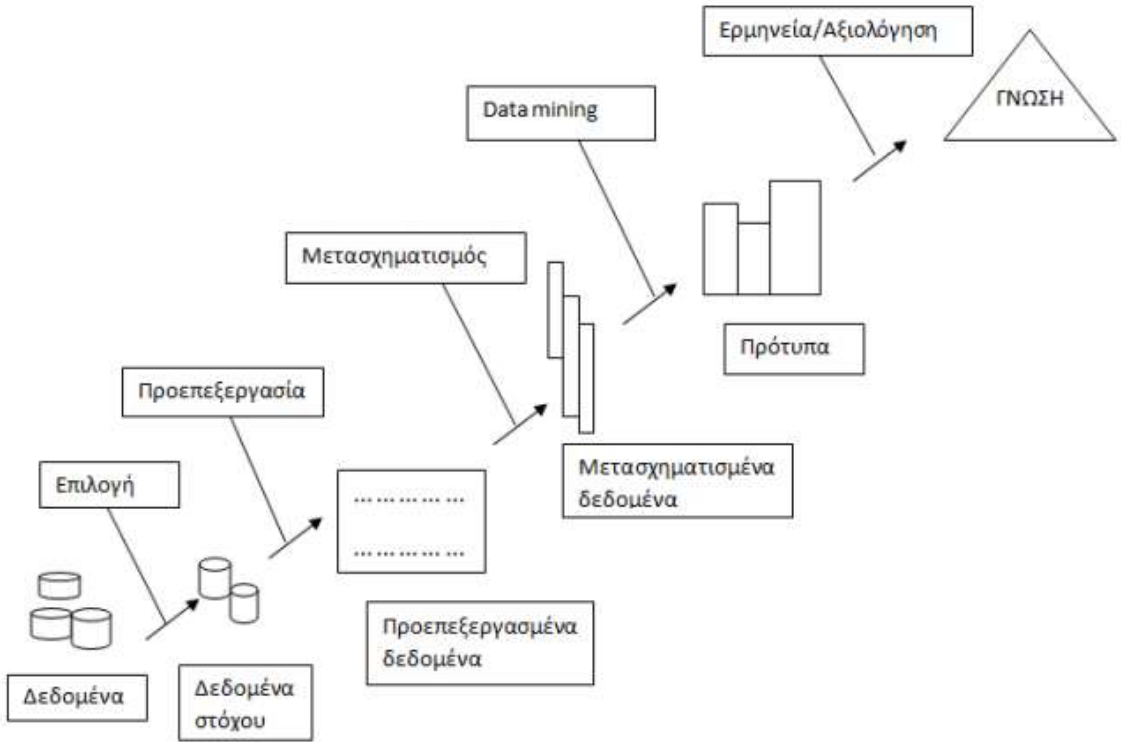
Τα συστήματα Επιχειρηματικής Ευφυΐας μπορούν να αναπαρασταθούν ως μια πυραμίδα τεσσάρων επιπέδων.

Εικόνα 1.2: Πυραμίδα Συστημάτων Επιχειρηματικής Ευφυΐας

Στην βάση της πυραμίδας βρίσκονται τα πρωτογενή δεδομένα, όπως αυτά λαμβάνονται από διάφορες πηγές (συστήματα ERP, Διαδίκτυο, κ.λπ.), και αποθηκεύονται σε Αποθήκες Δεδομένων (data warehouses) όπου γίνονται διεργασίες εξαγωγής, μετατροπής και φόρτωσης (Extract, Transform, Load (ETL)). Κατά την διερεύνηση των δεδομένων γίνονται κάποιες αρχικές εργασίες επεξεργασίας και μια πρώτη ανάλυση. Στο δεύτερο στάδιο, της Εξόρυξης Δεδομένων, γίνονται προχωρημένες αναλύσεις χρησιμοποιώντας διάφορες τεχνικές και εργαλεία όπως η Μηχανική Μάθησης (machine learning), η ανάλυση συστάδων, η κατηγοριοποίηση και η δημιουργία κανόνων συσχέτισης. Τρίτο στάδιο είναι το στάδιο της Βελτιστοποίησης κατά το οποίο ο αναλυτής επιλέγει από τα αποτελέσματα και τις μεθόδους που χρησιμοποιήθηκαν κατά την εξόρυξη, τα καταλληλότερα για το εν λόγω πρόβλημα που χρήζει λύσης. Τέλος, στην κορυφή της πυραμίδας βρίσκεται η Λήψη Αποφάσεων. Ο εκάστοτε ενδιαφερόμενος, σταθμίζοντας τα αποτελέσματα της ανάλυσης των προηγούμενων σταδίων, καταλήγει στην τελική απόφαση σύμφωνα με τις γνώσεις και τις ικανότητές του. *[Κύρκος, Ε. (2015). Επιχειρηματική Ευφυΐα και Εξόρυξη Δεδομένων]*

1.2.2 Ανακάλυψη Γνώσης από Βάσεις Δεδομένων

Η Ανακάλυψη Γνώσης από Βάσεις Δεδομένων (Knowledge Discovery in Databases (KDD)), είναι μια ευρύτερη διαδικασία που περιλαμβάνει τα στάδια της συλλογής δεδομένων, της προ επεξεργασίας (preprocessing) και καθορισμού (data cleaning) τους, του μετασχηματισμού τους (transformation), της εξόρυξης (data mining), της ερμηνείας (interpretation) και της αξιολόγησής τους (evaluation). Σκοπός του KDD είναι η ανακάλυψη νέας, χρήσιμης γνώσης μέσω αναγνώρισης μοτίβων στα δεδομένα.



Εικόνα 1.3: Ανακάλυψη Γνώσης από Βάσεις Δεδομένων [εικόνα από https://core.ac.uk/download/pdf/38467814.pdf]

Το βασικό πρόβλημα της κατανόησης των δεδομένων στην διαδικασία της KDD, είναι η χαρτογράφηση δεδομένων χαμηλού επιπέδου (ογκώδεις, δύσκολα στην κατανόηση και αφομοίωση τους) σε άλλες μορφές πιο συμπαγείς (πχ μια σύντομη αναφορά), πιο αφηρημένα (πχ μια περιγραφική προσέγγιση ή μοντέλο της

διαδικασίας που δημιούργησαν τα δεδομένα) ή πιο χρήσιμα (πχ ένα μοντέλο πρόβλεψης για την εκτίμηση μελλοντικών περιπτώσεων).

1.2.3 Μέθοδοι Ανάλυσης Στατιστικών Δεδομένων

Στην επιστήμη των δεδομένων (Data Science), που ασχολείται το παρόν κεφάλαιο, η εύρεση δομής και η πρόβλεψη αυτών των μεγάλων δεδομένων είναι τα σημαντικότερα βήματα για την ανάλυσή τους. Οι στατιστικές μέθοδοι είναι απαραίτητες γιατί είναι σε θέση να χειριστούν πολλά διαφορετικά αναλυτικά καθήκοντα. Οι βασικότερες μέθοδοι ανάλυσης στατιστικών δεδομένων είναι οι ακόλουθες:

1. *Έλεγχος υποθέσεων (Hypothesis testing)*. Ένα από τα σημαντικότερα στατιστικά εργαλεία ανάλυσης. Πολλές ερωτήσεις που προκύπτουν κατά την διαδικασία της ανάλυσης δεδομένων μπορούν να μεταφραστούν σε υποθέσεις. Με τους ελέγχους αυτούς, εξετάζονται κατά πόσο τα δεδομένα μπορεί να προβούν παραπλανητικά ή μη χρήσιμα. Η πολλαπλή χρήση των ίδιων δεδομένων σε διαφορετικά προβλήματα συχνά συμβαίνει με την ανάγκη διορθώσεων των επιπέδων σημασίας / σημαντικότητας (π.χ. στις φαρμακευτικές μελέτες).
2. *Ταξινόμηση (classification)*. Η ταξινόμηση των δεδομένων βοηθά στην καλύτερη κατανόηση, εντοπισμό και πρόβλεψη του χαοτικού όγκου αυτών των δεδομένων. Επίσης η μέθοδος αυτή χρειάζεται για τον εντοπισμό υπό πληθυσμών των δεδομένων – ομαδοποίηση δεδομένων.
3. *Παλινδρόμηση (regression)*. Η μέθοδος της παλινδρόμησης είναι το κυριότερο εργαλείο για την εύρεση σχέσεων μεταξύ μεταβλητών. Ανάλογα με την κατανομή που ακολουθεί το εκάστοτε δείγμα, μπορούν να εφαρμοστούν διαφορετικές προσεγγίσεις. Στην προϋπόθεση της κανονικότητας, η γραμμική ανάλυση παλινδρόμησης είναι η πιο κοινή, ενώ η περίπτωση της γενικευμένης γραμμικής παλινδρόμησης χρησιμοποιείται για άλλες κατανομές από την εκθετική οικογένεια.
4. *Ανάλυση χρόνο-σειρών (time series analysis)*. Δεδομένα με αλλαγή συμπεριφοράς στον χρόνο. Η ανάλυση τέτοιων δεδομένων γίνεται με την μέθοδο των χρόνο-

σειρών, αποσκοπώντας στην κατανόηση και στην πρόβλεψη της χρονικής τους δομής. Τυπικά παραδείγματα χρόνο-σειρών είναι οι επιστήμες συμπεριφοράς και οικονομίας, όπως επίσης και οι φυσικές επιστήμες και η μηχανική.

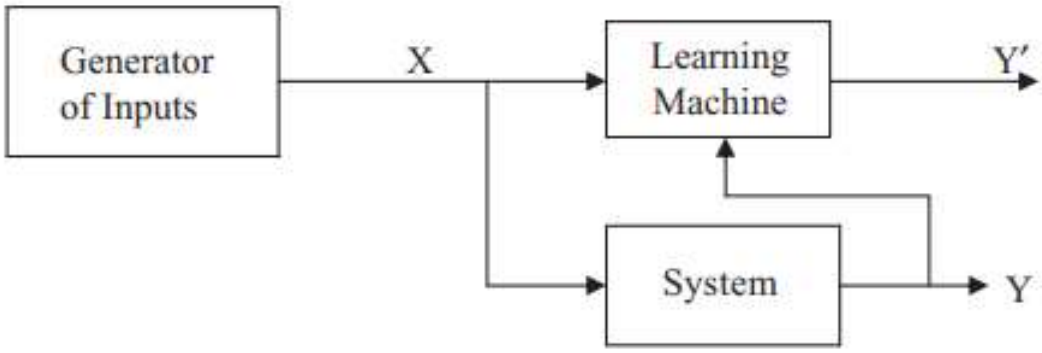
1.2.4 Μηχανή Εκμάθησης (Machine Learning)

Η μηχανή εκμάθησης, συνδυάζοντας την τεχνητή νοημοσύνη με τις αρχές της στατιστικής, έχει αποδειχθεί ένας από τους αποτελεσματικότερους τομείς της έρευνας, δημιουργώντας αλγόριθμους για την λύση συγκεκριμένων προβλημάτων που ορίζονται από τον ερευνητή. Αυτοί οι αλγόριθμοι ποικίλουν ανάλογα με το είδος του προβλήματος και των δεδομένων. Μια από τις πιο θεμελιώδεις μηχανές εκμάθησης είναι η επαγωγική, όπου από ένα γενικευμένο σύνολο δειγμάτων τυποποιείται σε μια συγκεκριμένη μορφή χρησιμοποιώντας διαφορετικές τεχνικές και μοντέλα.

Μπορούμε να ορίσουμε την επαγωγική μάθηση ως την διαδικασία εκτίμησης μιας άγνωστης εξάρτησης εισόδου (input) – εξόδου (output) χρησιμοποιώντας περιορισμένο αριθμό παρατηρήσεων ή μετρήσεων εισόδων και εξόδων του συστήματος. Για τα ζεύγη εισόδου – εξόδου χρησιμοποιείται ο όρος «δείγμα» και η διαδικασία επεξεργασίας περιλαμβάνει τρία συστατικά:

1. Μια γεννήτρια τυχαίων διανυσμάτων εισόδου X,
2. Ένα σύστημα που επιστρέφει μια έξοδο Y για ένα δεδομένο διάνυσμα εισόδου X, και
3. Μια μηχανή εκμάθησης – ένα μοντέλο, που εκτιμά μια άγνωστη (είσοδος X, έξοδος Y’) χαρτογράφηση του συστήματος από τα παρατηρούμενα δείγματα (είσοδος X, έξοδος Y).

Σχηματικά:



Εικόνα 1.3: Διαδικασία Μηχανής Εκμάθησης Σχηματικά

Αυτή είναι μια γενική διατύπωση που περιγράφει πολλές πρακτικές, επαγωγικής μάθησης, παλινδρόμησης, ταξινόμησης, ομαδοποίησης και εκτίμηση πυκνότητας. Το σύστημα παράγει ένα τυχαίο διάνυσμα Χ, το οποίο σχεδιάζεται ανεξάρτητα από οποιαδήποτε διανομή. Στη στατιστική ορολογία, αυτή η διαδικασία ονομάζεται *σύνθετη παρατήρηση*. Διαφέρει από την κλασσική σχεδιασμένη διαδικασία πειράματος, η οποία περιλαμβάνει την δημιουργία μιας τυχαίας δειγματοληψίας, βέλτιστη για μια συγκεκριμένη ανάλυση σύμφωνα με την πειραματική θεωρία σχεδιασμού. Η μηχανή εκμάθησης δεν ελέγχει ποιες τιμές εισόδου έχουν συμπεριληφθεί στο σύστημα, επομένως μιλάμε για μια προσέγγιση επαγωγικής μηχανής μάθησης.

1.2.5 Τεχνικές Ανάλυσης Μεγάλων Δεδομένων

Όπως έχει προαναφερθεί, η κατανόηση και ανάλυση των μεγάλων δεδομένων είναι αναγκαία για την αξιοποίηση και μετατροπή τους σε ένα χρήσιμο εργαλείο για την λήψη αποφάσεων. Οι οργανισμοί, με σκοπό την ανάλυση των δεδομένων, χρειάζονται αποτελεσματικές διαδικασίες για να μετατρέψουν μεγάλους όγκους δεδομένων σε χρήσιμη γνώση.

Η διαδικασία εξαγωγής τέτοιων χρήσιμων πληροφοριών από μεγάλα δεδομένα μπορεί να αναλυθεί σε δύο κατηγορίες. Διαχείριση δεδομένων και ανάλυση. Η διαχείριση δεδομένων περιλαμβάνει υποστηρικτικές τεχνολογίες για την αποθήκευση και προετοιμασία των δεδομένων για ανάλυση. Η δεύτερη μεγάλη

κατηγορία της ανάλυσης αφορά τεχνικές που χρησιμοποιούνται προκειμένου να αναλυθούν τα δεδομένα και για την άντληση πληροφοριών από αυτά – εξαγωγή γνώσεων. Ενδεικτικά, κάποιες από τις μεγάλες **αναλυτικές τεχνικές δεδομένων**, δομημένων και αδόμητων, παρουσιάζονται παρακάτω:

- **Text Analytics**. Η ανάλυση κειμένων γίνεται με τεχνικές οι οποίες εξάγουν πληροφορίες – δεδομένα από κείμενα. Πληροφορίες από μηνύματα ηλεκτρονικού ταχυδρομείου, ιστοσελίδες, ηλεκτρονικά φόρουμ, «feeds» μέσων κοινωνικής δικτύωσης, εταιρικά έγγραφα, κ.λπ. Οι αναλύσεις αυτές περιλαμβάνουν στατιστική ανάλυση, υπολογιστική γλωσσομάθεια και μηχανή μάθησης (machine learning). Τα δεδομένα που εξάγονται από τις αναλύσεις των κειμένων επιτρέπουν στις εταιρείες να λαμβάνουν πληροφορίες από μεγάλους όγκους κειμένων, που παράγονται από ανθρώπους, με σκοπό την λήψη αποφάσεων. Όπως για παράδειγμα οι αναλύσεις κειμένων που προέρχονται από οικονομικές ειδήσεις, μπορούν να χρησιμοποιηθούν στην πρόβλεψη χρηματιστηριακής αγοράς. Σύντομα παραδείγματα τέτοιων εργαλείων ανάλυσης κειμένου είναι τα εξής:
    - *Information Extraction (IE).* Οι τεχνικές εξαγωγής πληροφοριών εξάγουν δομημένα δεδομένα από αδόμητο κείμενο. Για παράδειγμα, σε μια φαρμακευτική εταιρεία, παρέχουν πληροφορίες όπως όνομα φαρμάκου, δοσολογία και συχνότητα ιατρικής συνταγής. Υπό-εργασίες στο IE είναι η Αναγνώριση Οντοτήτων (Entity Recognition) και η Εξόρυξη Συσχετισμού (Relation Extraction). Στην ER εξάγονται ονόματα από κείμενο και ταξινομούνται σε κατηγορίες, όπως άτομο, τοποθεσία, οργάνωση, ημερομηνία κ.λπ. Στην RE εξάγονται σημασιολογικές σχέσεις μεταξύ οντοτήτων (οργανισμών, ατόμων, φαρμάκων, γονιδίων κ.λπ.). Για παράδειγμα στο κείμενο «Steve Jobs συνιδρυτής της Apple Inc. Το 1976» το σύστημα RE εξάγει σχέσεις όπως FounderOf (Steve Jobs, Apple Inc.) ή FoundedIn (Apple Inc., 1976).
    - *Τεχνικές Συνοπτικής Τεκμηρίωσης.* Παράγονται αυτόματα σύντομες περιλήψεις κειμένων οι οποίες προσφέρουν τις βασικές πληροφορίες

των εγγράφων. Σε γενικές γραμμές η τεχνική αυτή περιλαμβάνει δύο προσεγγίσεις: την εξορυκτική και την αφηρημένη προσέγγιση. Στην εξορυκτική σύνοψη, η περίληψη που προκύπτει είναι ένα υποσύνολο του αρχικού εγγράφου – συνήθως προτάσεις. Η διατύπωση μιας πρότασης – σύνοψης περιλαμβάνει τον προσδιορισμό των χαρακτηριστικών μονάδων του κειμένου και την σύζευξη τους. Η σημασία αυτών των χαρακτηριστικών μονάδων αξιολογείται με ανάλυση της θέσης και της συχνότητάς τους στο κείμενο. Αντίθετα, στην αφηρημένη προσέγγιση γίνεται εξαγωγή σημασιολογικών πληροφοριών από το κείμενο. Οι μονάδες κειμένου που προκύπτουν από τις περιλήψεις δεν υπάρχουν απαραίτητα στο αρχικό έγγραφο.

    - *Τεχνικές Ανάλυσης Αισθήσεων (εξόρυξης γνώμης)*. Τα κείμενα τα οποία περιέχουν τις γνώμες των ανθρώπων για θέματα όπως προϊόντα, οργανώσεις, άτομα, επιχειρήσεις, πολιτικά θέματα κ.λπ., αναλύονται με κατανόηση ώστε οι επιχειρήσεις να καταγράφουν όλο και περισσότερα στοιχεία σχετικά με τα συναισθήματα και τις συνήθειες των πελατών τους. Αυτά τα δεδομένα οδηγούν στην αντίληψη των συναισθημάτων οδηγώντας το μάρκετινγκ, τις πολιτικές και κοινωνικές επιστήμες στην λήψη αποτελεσματικότερων αποφάσεων.

- **Audio analytics.** Οι αναλύσεις ήχου εξάγουν και αναλύουν πληροφορίες από μη δομημένα δεδομένα ήχου. Τα κέντρα τηλεφωνικής εξυπηρέτησης πελατών και υγειονομικής περίθαλψης αποτελούν τους κύριους τομείς εφαρμογής των αναλυτικών συστημάτων ήχου. Αυτή η τεχνική χρησιμοποιείται για την βελτίωση της πελατειακής εμπειρίας, την αξιολόγηση της απόδοσης των εργαζομένων στα τηλεφωνικά κέντρα και στην παρακολούθηση της συμμόρφωσης με διαφορετικές πολιτικές απορρήτου και ασφάλειας.
- **Video analytics.** Η ανάλυση των βίντεο, περιλαμβάνει μια πληθώρα τεχνικών για την ανάλυση και εξαγωγή σημαντικών πληροφοριών. Λόγω της αυξανόμενης ροής βίντεο των τελευταίων χρόνων, οι τεχνικές ανάλυσής τους είναι σε μικρότερο φάσμα από αυτές της ανάλυσης κειμένων και ήχου. Ωστόσο, έχουν αναπτυχθεί διάφορες τεχνικές εξόρυξης δεδομένων από

βίντεο σε πραγματικό χρόνο καθώς και είδη καταγεγραμμένων βίντεο. Η βασικότερη πρόκληση της ανάλυσης βίντεο είναι το μέγεθος των δεδομένων. Ένα δευτερόλεπτο ενός βίντεο υψηλής ευκρίνειας είναι ισοδύναμο με πάνω από 2.000 σελίδες κειμένου. Από στατιστικά του YouTube, κάθε λεπτό μεταφορτώνονται 100 ώρες βίντεο. Σημαντικές επίσης είναι οι αναλύσεις που γίνονται στο υλικό από κάμερες CCTV καταστημάτων λιανικής πώλησης. Τα δεδομένα που προκύπτουν μπορούν να εξαχθούν για επιχειρηματική ευφυΐα, διαχειρίζοντας το μάρκετινγκ των επιχειρήσεων. Για παράδειγμα, οι έξυπνοι αλγόριθμοι συλλέγουν πληροφορίες σχετικά με το προφίλ των πελατών. Τις αγοραστικές τους συνήθειες, την ηλικία, το φύλο, την εθνικότητα κ.λπ. Οι λιανοπωλητές μπορούν να μετρήσουν τον αριθμό των πελατών, τον χρόνο παραμονής τους στο κατάστημα, και τα πρότυπα κίνησης τους. Με την συσχέτιση αυτών των πληροφοριών μπορούν να διαμορφωθούν στρατηγικές τιμολόγησης, τοποθέτηση προϊόντων, σχεδιασμό προώθησης κ.λπ.

- **Social media analytics.** Οι αναλύσεις κοινωνικών μέσων αναφέρονται στην ανάλυση δομημένων και αδόμητων δεδομένων. Τα κοινωνικά μέσα αποτελούν ένα δίκτυο απεριόριστου φάσματος που επιτρέπει την αλληλεπίδραση των χρηστών με διάφορους τρόπους (κείμενο, ήχο, εικόνα και βίντεο). Τα κοινωνικά μέσα ταξινομούνται στους εξής τύπους: Κοινωνικά δίκτυα (Facebook, LinkedIn), Μικροσκόπια (Twitter, Tumblr), Κοινωνικές Ειδήσεις (Digg, Reddit), Κοινή Χρήση Μέσων (Instagram, YouTube), Ερωτήσεις Απαντήσεις (Yahoo Answers, Ask.com, TripAdvisor). Η έρευνα για την ανάλυση κοινωνικών μέσων δικτύωσης εκτείνεται σε διάφορους κλάδους όπως η ψυχολογία, η κοινωνιολογία, η ανθρωπολογία, τα μαθηματικά, η πληροφορική, η φυσική και η οικονομία. Το μάρκετινγκ αποτελεί την κύρια εφαρμογή των αναλυτικών μέσων κοινωνικής δικτύωσης τα τελευταία χρόνια, αφού η χρήση αυτών των μέσων από τους καταναλωτές αυξάνεται συνεχώς. Η Forrester Research Inc. αναφέρεται στα κοινωνικά μέσα ως το δεύτερο ταχύτερα αναπτυσσόμενο κανάλι μάρκετινγκ μεταξύ 2011 και 2016.

  [VanBoskirk, Overby, & Takvorian, (2011) - Beyond the hype: Big data concepts, methods, and analytics]

- **Predictive analytics.** Οι προγνωστικές αναλύσεις περιλαμβάνουν τεχνικές και μέσα τα οποία βοηθούν στην πρόβλεψη μελλοντικών αποτελεσμάτων με βάση ιστορικά και τρέχοντα δεδομένα. Οι προβλέψεις αυτές για παράδειγμα, βοηθούν τις επιχειρήσεις, και όχι μόνο, να αποκτήσουν μια εικόνα των πωλήσεων τους τα επόμενα διαστήματα. Οι τεχνικές πρόγνωσης κατατάσσονται σε δύο ομάδες. Πρώτη ομάδα είναι αυτή των κινητών μέσων όρων, προσπαθούν να ανακαλύψουν τα ιστορικά μοτίβα της μεταβλητής των αποτελεσμάτων και να τα εξαγάγουν στο μέλλον. Στην δεύτερη ομάδα κατατάσσονται οι τεχνικές της γραμμικής παλινδρόμησης, στοχεύουν στην καταγραφή των αλληλεξαρτήσεων μεταξύ των ανεξάρτητων μεταβλητών και της εξαρτημένης μεταβλητή, αξιοποιώντας αυτές τις καταγραφές για να κάνουν προβλέψεις. Εκτός από αυτές τις ομάδες τεχνικών υπάρχει και μια Τρίτη που περιλαμβάνει τις τεχνικές μηχανικής μάθησης (π.χ. νευρωτικά δίκτυα). Μια άλλη ταξινόμηση βασίζεται στον τύπο των μεταβλητών: μεταβλητές συνεχούς έκβασης γραμμικής παλινδρόμησης (π.χ. τιμή πώλησης πετρελαίου) και διακριτές μεταβλητές αποτελεσμάτων - τυχαία δάση (π.χ. κατάσταση πιστοληπτικής ικανότητας). Οι προγνωστικές τεχνικές βασίζονται κυρίως στην στατιστική. Οι κλασσικές στατιστικές μέθοδοι δεν είναι πάντα αποτελεσματικές για τεράστιους όγκους δεδομένων, απαιτείται ανάπτυξη νέων στατιστικών μεθόδων για τους εξής λόγους: Πρώτον, στην στατιστική ανάλυση ένα μικρό δείγμα λαμβάνεται από τον πληθυσμό για την εξέταση της συμπεριφοράς του και τα αποτελέσματα συγκρίνονται με την πιθανότητα ύπαρξης του αποτελέσματος στον πληθυσμό με ένα συγκεκριμένο επίπεδο εμπιστοσύνης. Αντίθετα, τα μεγάλα δείγματα δεδομένων είναι τεράστια και επαρκή ώστε να αντιπροσωπεύουν την πλειοψηφία, αν όχι ολόκληρο τον πληθυσμό. Ως εκ τούτου η έννοια της στατιστικής σημασίας δεν είναι τόσο σημαντική για μεγάλα δεδομένα. Δεύτερον, πολλές συμβατικές μέθοδοι για μικρά δείγματα δεν κλιμακώνονται μέχρι τα μεγάλα δεδομένα, όσον αφορά την υπολογιστική αποτελεσματικότητα. Ο τρίτος παράγοντας που κάνει τις κλασσικές στατιστικές μεθόδους ελλιπείς, είναι τα χαρακτηριστικά των μεγάλων δεδομένων: η ανομοιογένεια, η συσσώρευση θορύβου και οι ψευδείς συσχετίσεις. (Challenges of Big Data Analysis Fan, Han, & Liu, 2014)

Πιο αναλυτικά:

- Ανομοιογένεια (Heterogeneity). Τα μεγάλα δεδομένα λαμβάνονται συχνά από διαφορετικές πηγές, επομένως αντιπροσωπεύουν πληροφορίες από διαφορετικούς υπό-πληθυσμούς, με αποτέλεσμα τα δεδομένα αυτά είναι εξαιρετικά ετερογενή.
- Συσσώρευση Θορύβου (Noise Accumulation). Η ταυτόχρονη εκτίμηση διαφόρων παραμέτρων στα προγνωστικά μοντέλα μεγάλων δεδομένων, είναι αναπόφευκτη. Με αποτέλεσμα να αγνοούνται μεγέθη μεταβλητών με σημαντική επεξηγηματική δύναμη για την εκτίμηση του μοντέλου – αυτό ονομάζεται συσσωρευμένο σφάλμα εκτίμησης (ή θόρυβος).
- Παράλογη Συσχέτιση (Spurious Correlation). Στα μεγάλα δεδομένα υπάρχει ο κίνδυνος της αποδοχής ψευδών και παραπλανητικών συσχετίσεων μη συσχετισμένων μεταβλητών λόγω του τεράστιου όγκου του συνόλου των δεδομένων. Το σφάλμα αυτό μπορεί να φανεί σε μια εκτίμηση παλινδρόμησης όταν ο συντελεστής συσχέτισης μεταξύ ανεξάρτητων τυχαίων μεταβλητών αυξάνεται όσο αυξάνεται το μέγεθος του συνόλου των δεδομένων.

Στην επόμενη ενότητα, παρουσιάζεται αναλυτικότερα η διαδικασία εξόρυξης μεγάλων δεδομένων σε πιο θεωρητικό υπόβαθρο.

1.3 Εξόρυξη Δεδομένων (Data Mining)

Όπως προαναφέρθηκε, αυτός ο τεράστιος όγκος δεδομένων είναι πολύ δύσκολο να διαχειριστεί από τις επιχειρήσεις και τους αναλυτές. Περίπου το 80% των δεδομένων παγκοσμίως δεν έχουν συγκεκριμένη δομή (ήμι-δομημένα, αδόμητα), συνεπώς η συνηθισμένες μέθοδοι που χρησιμοποιούνται για την ανάλυση των δομημένων δεομένων δεν είναι κατάλληλες. Αυτό έχει οδηγήσει στην ανάγκη ανάπτυξης νέων αλγορίθμων, με σκοπό την διαχείριση τέτοιων ιδιαίτερων δεδομένων που προέρχονται από διαφορετικές πηγές, πχ κείμενα, βίντεο, εικόνες. *[Hurwitz, J. S., Nugent, A., Halper, F., & Kaufman, M. (2013). Big Data For Dummies. John Wiley & Sons]*

Η ανάγκη κατανόησης και ερμηνείας αυτών των μεγάλων, περίπλοκων και πλούσιων σε πληροφορίες συνόλων δεδομένων, είναι κοινή για όλους τους τομείς της επιχείρησης, της επιστήμης και της μηχανικής. Στον επιχειρηματικό κόσμο, τα εταιρικά δεδομένα και τα δεδομένα των πελατών αναγνωρίζονται ως *στρατηγικό πλεονέκτημα*. Η δυνατότητα εξαγωγής χρήσιμων πληροφοριών που κρύβονται σε αυτά τα δεδομένα και η εκμετάλλευση αυτής της γνώσης, γίνεται όλο και πιο σημαντική στον σημερινό ανταγωνιστικό κόσμο. Η όλη διαδικασία εφαρμογής μεθοδολογίας βασισμένη σε υπολογιστές, συμπεριλαμβανομένων νέων τεχνικών, για την «μετάφραση» των δεδομένων και ανακάλυψη γνώσεων μέσα από αυτά, ονομάζεται *εξόρυξη δεδομένων (data mining).*

Η εξόρυξη δεδομένων είναι μια επαναληπτική διαδικασία, στην οποία η πρόοδος εξαρτάται από την ανακάλυψη αυτόματων είτε μη αυτόματων μεθόδων. Κατά την διαδικασία της εξόρυξης είναι αποτελεσματικότερο να μην υπάρχουν προκαθορισμένες αντιλήψεις για το τι θα αποτελούσε «ενδιαφέρον» αποτέλεσμα. Η εξόρυξη δεδομένων είναι η αναζήτηση νέων, πολύτιμων και μη ιδιωτικών πληροφοριών σε μεγάλους όγκους δεδομένων μέσω μιας συνεργατικής προσπάθειας ανθρώπων και υπολογιστών. Ο άνθρωπος περιγράφει το πρόβλημα και τον στόχο και ο υπολογιστής προσφέρει τις δυνατότητες αναζήτησης.

Στον επιχειρηματικό κόσμο, η εξόρυξη δεδομένων μπορεί να χρησιμοποιηθεί για την ανακάλυψη νέων τάσεων της αγοράς, σχεδιασμό επενδυτικών στρατηγικών και εντοπισμό μη λογικών δαπανών στην λειτουργία του οργανισμού. Μπορεί να βελτιώσει τις καμπάνιες μάρκετινγκ, έχοντας καταλληλότερη υποστήριξη και προσοχή στους πελάτες. Οι τεχνικές εξόρυξης δεδομένων μπορούν να εφαρμοστούν σε προβλήματα ανασχεδιασμού των επιχειρησιακών διαδικασιών, κατανοώντας καλύτερα την αλληλένδετη εξάρτηση κάθε τμήματος του οργανισμού.

1.3.1 Στόχος της Εξόρυξης Δεδομένων (Data Mining)

Οι δύο πρωταρχικοί στόχοι της εξόρυξης δεδομένων είναι η πρόβλεψη και η περιγραφή. Αναλυτικά:

- Η **πρόβλεψη** περιλαμβάνει την χρήση μεταβλητών στο σύνολο των δεδομένων για την πρόβλεψη άγνωστων ή μελλοντικών τιμών άλλων

μεταβλητών υπό εξέταση. Παράγει το μοντέλο του συστήματος το οποίο περιγράφεται από τα δεδομένα που δίνονται.

- Η **περιγραφή** επικεντρώνεται στην εύρεση μοτίβων που περιγράφουν τα δεδομένα που μπορούν να ερμηνευτούν από τους ανθρώπους. Παράγει νέες, μη εμπιστευτικές πληροφορίες με βάση το διαθέσιμο σύνολο δεδομένων.

Ο τελικός στόχος της εξόρυξης δεδομένων είναι να παράγει ένα μοντέλο εκφραζόμενο ως εκτελέσιμος κώδικας, ο οποίος μπορεί να χρησιμοποιηθεί για την ταξινόμηση, πρόβλεψη, εκτίμηση ή άλλες παρόμοιες εργασίες με σκοπό την κατανόηση του συστήματος που αναλύεται αποκαλύπτοντας μοτίβα και σχέσεις σε μεγάλα σύνολα δεδομένων.

Ο στόχος της πρόβλεψης και της περιγραφής επιτυγχάνεται χρησιμοποιώντας τεχνικές εξόρυξης δεδομένων, οι οποίες θα αναφερθούν αργότερα στο παρόν κεφάλαιο, για τις ακόλουθες δουλειές *(data mining tasks):*

1. *Ταξινόμηση (classification)* : λειτουργία πρόβλεψης που ταξινομεί ένα στοιχείο από τα δεδομένα μέσα σε ένα σύνολο είδη καθορισμένων / προσδιορισμένων δεδομένων.
2. *Παλινδρόμηση (regression)* : λειτουργία πρόβλεψης με την χαρτογράφηση ενός στοιχείου σε πραγματική μεταβλητή πρόβλεψης.
3. *Ομαδοποίηση (clustering)* : περιγραφική λειτουργία κατά την οποία γίνεται προσπάθεια για τον εντοπισμό ενός συνόλου κατηγοριών ή ομάδων για την περιγραφή των δεδομένων.
4. *Σύνοψη (summarization)* : περιγραφική λειτουργία που περιλαμβάνει μεθόδους για την εύρεση μιας ολοκληρωμένης / συμπαγής περιγραφής για ένα σύνολο ή υποσύνολο δεδομένων.
5. *Μοντελοποίηση εξάρτησης (dependency model)* : εύρεση ενός μοντέλου που περιγράφει σημαντικές εξαρτήσεις ανάμεσα στις μεταβλητές η στις τιμές των χαρακτηριστικών σε ένα σύνολο δεδομένων ή σε ένα μέρος ενός συνόλου δεδομένων.

6. *Ανίχνευση αλλαγών και απόκλισης (Change and Deviation Detection)* : εύρεση των σημαντικότερων αλλαγών στο σύνολο των δεδομένων.

1.3.2. Οι Ρίζες της Εξόρυξης Δεδομένων (Data Mining Roots)

Τα περισσότερα προβλήματα της εξόρυξης δεδομένων έχουν τις ρίζες τους στην κλασσική ανάλυση δεδομένων, στην στατιστική και στην μηχανική μάθησης (machine learning). Η στατιστική με την σειρά της έχει τις ρίζες της στα μαθηματικά. Η επιστήμη των μαθηματικών εκφράζεται με μια αυστηρότητα ως προς το θεωρητικό υπόβαθρο, δηλαδή είναι απαραίτητο να αποδειχθεί η λογική ενός στοιχείου προτού εφαρμοστεί στην πράξη. Αντίθετα, η μηχανική μάθησης (machine learning) έχει τις ρίζες της στην πρακτική του υπολογιστή. Αυτό οδηγεί σε πρακτικό προσανατολισμό, γίνονται συνεχώς δοκιμές για να αποδοθεί εν τέλη το άριστο χωρίς κάποια επίσημη απόδειξη.

Οι σύγχρονες στατιστικές μέθοδοι είναι σχεδόν εξ ολοκλήρου καθοδηγούμενες από την έννοια ενός μοντέλου ενώ η μηχανική μάθηση τείνει να δίνει έμφαση στους αλγόριθμους.

1.3.3 Σύστημα Ταυτοποίησης – Αναγνώριση Συστήματος

Οι βασικές αρχές μοντελοποίησης στην εξόρυξη δεδομένων έχουν επίσης ρίζες στην θεωρία ελέγχου, η οποία εφαρμόζεται κυρίως σε μηχανικά συστήματα και βιομηχανικές διαδικασίες. Το πρόβλημα προσδιορισμού ενός μαθηματικού μοντέλου για ένα άγνωστο σύστημα άγνωστων δεδομένων εισόδου – εξόδου, αναφέρεται ως *σύστημα ταυτοποίησης.* Οι σκοποί της αναγνώρισης των συστημάτων ταυτοποίησης είναι η πρόβλεψη της συμπεριφοράς των συστημάτων και η εξήγηση της αλληλεπίδρασης και των σχέσεων των μεταβλητών μέσα σε ένα σύστημα. Παρακάτω παρουσιάζονται αναλυτικά τα **βήματα αναγνώρισης συστήματος** ιεραρχικά.

1. *Προσδιορισμός δομής*: Σε αυτό το βήμα γίνεται ο προσδιορισμός μιας κατηγορίας μοντέλων για την διεξαγωγή του καταλληλότερου. Αυτή η κατηγορία μοντέλων συμβολίζεται από μια παραμετροποιημένη συνάρτηση **y = f (u, t)**, όπου **y** είναι η έξοδος (output) του μοντέλου, **u** είναι ένα διάνυσμα εισόδου (input) και **t** είναι ένα διάνυσμα

παραμέτρων. Ο προσδιορισμός της συνάρτησης f εξαρτάται από τον σχεδιασμό και την λειτουργία του εκάστοτε συστήματος και την εμπειρία του σχεδιαστή.

2. *Προσδιορισμός παραμέτρου*: Όταν η δομή του μοντέλου είναι γνωστή, στο δεύτερο βήμα εφαρμόζονται τεχνικές βελτιστοποίησης για τον προσδιορισμό του διανύσματος παραμέτρων t έτσι ώστε να προκύπτει μοντέλο **y * = f (u, t *)** το οποίο είναι το καταλληλότερο / αποτελεσματικότερο για την περιγραφή του συστήματος.

Η αναγνώριση συστήματος, η δομή και ο προσδιορισμός παραμέτρων, είναι μια διαδικασία που εκτελείται επαναλαμβανόμενα έως ότου βρεθεί το ικανοποιητικό μοντέλο. Αυτή η επαναλαμβανόμενη διαδικασία έχει ως εξής:

1. Προσδιορισμός και παραμετροποίηση μιας κατηγορίας μαθηματικών μοντέλων y * = f (u, t *) που αντιπροσωπεύει το σύστημα προς αναγνώριση (διαδικασία αναγνώρισης συστήματος).
2. Επιλογή των καταλληλότερων παραμέτρων για το διαθέσιμο σύνολο δεδομένων (δηλαδή τα «λάθη», η διαφορά y - y *, της παλινδρόμησης να είναι τα ελάχιστα).
3. Δοκιμές επικύρωσης (validation tests), κατά πόσο το αναγνωρισμένο μοντέλο ανταποκρίνεται σωστά αόρατο σύνολο δεδομένων (αναφέρεται ως σύνολο δοκιμών, επικύρωσης ή ελέγχου).
4. Τερματισμός διαδικασίας όταν τα αποτελέσματα επικύρωσης είναι τα θεμιτά.

1.3.4 Διαδικασία Εξόρυξης Δεδομένων (Data Mining Process)

Η διαδικασία της εξόρυξης δεδομένων δεν είναι απλά μια συλλογή μεμονωμένων εργαλείων και τυχαίων εφαρμογών στατιστικών και μηχανικών μεθόδων. Δεν είναι μια τυχαία επισκόπηση του χώρου των αναλυτικών τεχνικών, αλλά μια προσεκτικά σχεδιασμένη και μελετημένη διαδικασία λήψης αποφάσεων. Ο αναλυτής μελετά τα δεδομένα, τα εξετάζει χρησιμοποιώντας μια αναλυτική τεχνική, τροποποιώντας όπου είναι απαραίτητο τα εργαλεία φτάνοντας είτε σε καλύτερα είτε διαφορετικά αποτελέσματα. Η τροποποίηση μπορεί να γίνει πολλές φορές χρησιμοποιώντας κάθε

τεχνική για να διερευνήσει μια ελαφρώς διαφορετική οπτική των δεδομένων – μια διαφορετική ερώτηση από τα δεδομένα.

Η γενική πειραματική διαδικασία που έχει προσαρμοστεί στο πρόβλημα της εξόρυξης δεδομένων περιλαμβάνει τα εξής βήματα:

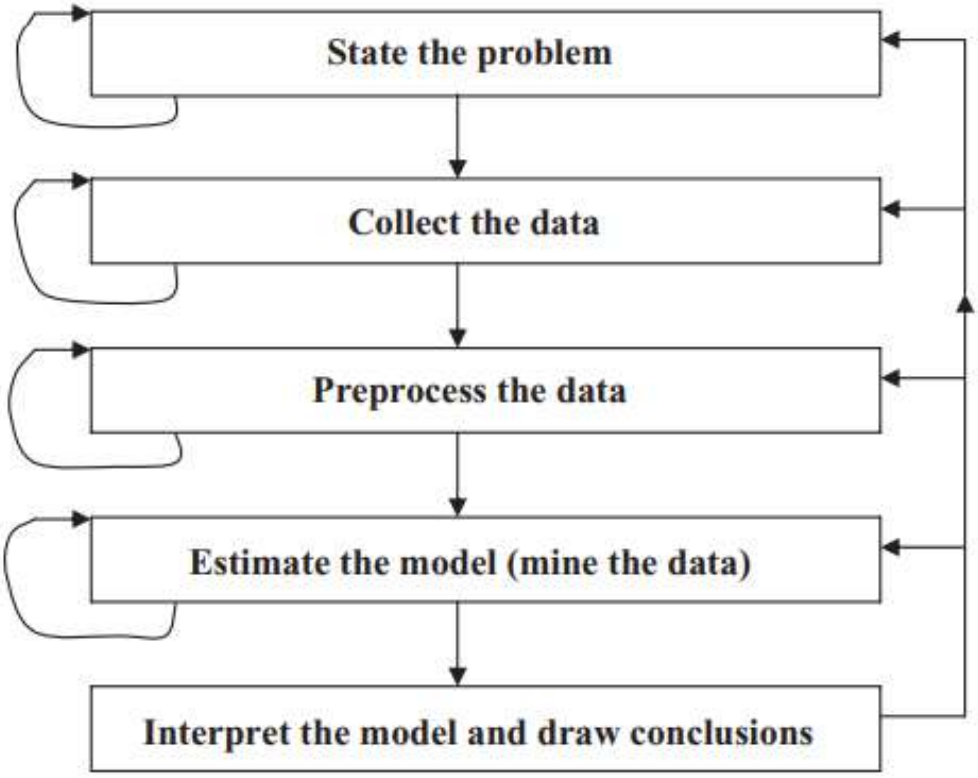
1. *Εντοπισμός του προβλήματος και διατύπωση της υπόθεσης*. Σε αυτό το βήμα καθορίζεται ένα σύνολο μεταβλητών για την άγνωστη εξάρτηση του μοντέλου, και εάν είναι δυνατόν, μια γενική εικόνα αυτής της εξάρτησης ως αρχική υπόθεση. Μπορεί να υπάρχουν πολλές υποθέσεις για ένα μόνο πρόβλημα σε αυτό το στάδιο. Το πρώτο βήμα απαιτεί μια συνδυαστική τεχνογνωσία ενός τομέα εφαρμογών (application domain) και ενός μοντέλου εξόρυξης δεδομένων (data mining model). Στην πράξη είναι απαραίτητη η συνεργασία των εμπειρογνωμόνων εξόρυξης δεδομένων και του ειδικού σχεδιαστή των εφαρμογών.
2. *Συλλογή δεδομένων*. Αυτό το βήμα αφορά τον τρόπο δημιουργίας και συλλογής των δεδομένων. Υπάρχουν δύο δυνατότητες. Στην πρώτη, η διαδικασία παραγωγής δεδομένων βρίσκεται υπό τον έλεγχο ενός ειδικού (μοντελοποιητής) – αυτή η προσέγγιση είναι γνωστή ως «σχεδιασμένο πείραμα» (designed experiment) . Η δεύτερη δυνατότητα είναι όταν ο μοντελοποιητής δεν μπορεί να επηρεάσει τα δεδομένα – γνωστή ως «προσέγγιση παρατήρησης» (observational approach), είναι μια τυχαία παραγωγή δεδομένων και η κατανομή δειγματοληψίας είναι άγνωστη μετά τη συλλογή των δεδομένων. Είναι σημαντικό εδώ να τονιστεί ότι μια εκ των προτέρων γνώση της θεωρητικής κατανομής των δεδομένων, μπορεί να είναι πολύ χρήσιμη για την μοντελοποίηση και αργότερα για την τελική ερμηνεία των αποτελεσμάτων. Τα δεδομένα που θα χρησιμοποιηθούν για την εκτίμηση ενός μοντέλου και αυτά που θα χρησιμοποιηθούν αργότερα για την δοκιμή αυτού του μοντέλου πρέπει να προέρχονται από την ίδια άγνωστη κατανομή δειγματοληψίας ειδάλλως το εκτιμώμενο μοντέλο δεν μπορεί να χρησιμοποιηθεί επιτυχώς.
3. *Προ επεξεργασία των δεδομένων*. Τα δεδομένα συνήθως συλλέγονται από τις υπάρχουσες βάσεις δεδομένων (databases), τις αποθήκες δεδομένων (data

warehouses) και τα data marts – αποθήκες δεδομένων του κάθε οργανισμού/επιχείρησης. Η προ επεξεργασία δεδομένων περιλαμβάνει δύο κοινές εργασίες:

a. ανίχνευση ακραίων τιμών (outliers) και αφαίρεσή τους. Οι ακραίες τιμές είναι ασυνήθιστες τιμές δεδομένων που δεν συμφωνούν με τις υπόλοιπες παρατηρήσεις. Αυτές οι τιμές συνήθως προκύπτουν από σφάλματα μέτρησης, κωδικοποίησης και εγγραφής, και μερικές φορές είναι φυσικές, μη φυσιολογικές τιμές. Η αντιμετώπιση τους γίνεται είτε αφαιρώντας τες είτε αναπτύσσοντας ισχυρές μεθόδους μοντελοποίησης μη ευαίσθητες στα ακραία σημεία.
b. Κλιμάκωση, κωδικοποίηση και επιλογή χαρακτηριστικών. Για παράδειγμα, ένα χαρακτηριστικό με εύρος [0, 1] και ένα άλλο με εύρος [-100, 1000] δεν θα έχουν το ίδιο βάρος στην εφαρμοσμένη τεχνική και θα επηρεάσουν διαφορετικά τα αποτελέσματα της εξόρυξης δεδομένων. Συνίσταται η κλιμάκωσή τους, φέρνοντας και τα δύο χαρακτηριστικά στο ίδιο βάρος για την περαιτέρω ανάλυση. Επίσης, συγκεκριμένοι μέθοδοι κωδικοποίησης εφαρμογών επιτυγχάνουν μείωση των διαστάσεων παρέχοντας μικρότερο αριθμό ενημερωτικών χαρακτηριστικών για τις ακόλουθες μοντελοποιήσεις δεδομένων.

4. *Εκτίμηση μοντέλου*. Η επιλογή και εφαρμογή της κατάλληλης τεχνικής εξόρυξης δεδομένων είναι το κύριο στοιχείο αυτής της φάσης. Η επιλογή του κατάλληλου μοντέλου δεν είναι απλή και χρειάζεται γνώση και αντίληψη των δεδομένων και του προβλήματος.
5. *Ερμηνεία μοντέλου και εξαγωγή συμπερασμάτων*. Σκοπός των μοντέλων εξόρυξης δεδομένων είναι να βοηθούν στην διαδικασία λήψης αποφάσεων. Για να είναι βοηθητικά πρέπει να είναι και κατανοητά από τον άνθρωπο, δηλαδή να είναι ερμηνεύσιμα και όχι περίπλοκα. Οι στόχοι της ακρίβειας του μοντέλου και της ακρίβειας της ερμηνείας του είναι κάπως αντιφατικά. Συνήθως τα απλά μοντέλα είναι πιο κατανοητά και ερμηνεύσιμα, αλλά και λιγότερο ακριβή. Οι σύγχρονες μέθοδοι εξόρυξης δεδομένων αναμένεται να αποφέρουν εξαιρετικά ακριβή αποτελέσματα χρησιμοποιώντας μοντέλα υψηλών διαστάσεων. Το πρόβλημα στην ερμηνεία τέτοιων μοντέλων

θεωρείται ξεχωριστή εργασία, με συγκεκριμένες τεχνικές για την εξακρίβωση των αποτελεσμάτων. Δεν είναι ευανάγνωστο και χρήσιμο για τον χρήστη να έχει να αντιμετωπίσει εκατοντάδες σελίδες αριθμητικών αποτελεσμάτων. Δεν μπορεί να τα κατανοήσει, να τα ερμηνεύσει και ως εκ τούτου να τα χρησιμοποιήσει για την λήψη αποφάσεων.



Εικόνα 1.4: Βήματα της διαδικασίας Εξόρυξης Γνώσης από Βάσεις Δεδομένων

Στο βήμα της προ επεξεργασίας των δεδομένων της διαδικασίας εξόρυξης, είναι αναγκαίο να προετοιμαστεί μια *ανάλυση ποιότητας δεδομένων.* Η ανάλυση επιδρά σημαντικά στην εικόνα του συστήματος και καθορίζει το αντίστοιχο μοντέλο που περιγράφεται. Με δεδομένα κακής ποιότητας, θα είναι σχεδόν αδύνατο για μια επιχείρηση να λάβει τις σωστές αποφάσεις. Υπάρχουν διάφοροι **δείκτες ποιότητας δεδομένων** που λαμβάνονται κατά την διάρκεια προ επεξεργασίας της διαδικασίας εξόρυξης δεδομένων:

1. Τα δεδομένα πρέπει να είναι ακριβή. Ο αναλυτής πρέπει να ελέγχει εάν το όνομα/η ετικέτα είναι γραμμένο σωστά, οι τιμές βρίσκονται σε συγκεκριμένο εύρος και είναι πλήρης, και ούτω κάθε εξής.

2. Τα δεδομένα πρέπει να αποθηκεύονται σύμφωνα με τον τύπο δεδομένων που απαρτίζονται. Ο αναλυτής πρέπει να διασφαλίσει ότι οι αριθμητικές τιμές δεν παρουσιάζονται σε μορφή χαρακτήρων, οι ακέραιες τιμές δεν είναι στην μορφή πραγματικών αριθμών και ούτω καθεξής.
3. Τα δεδομένα πρέπει να έχουν ακεραιότητα. Οι ενημερώσεις του συστήματος δεν πρέπει να χάνονται λόγω κακής συνεννόησης των χρηστών και οι λειτουργίες ασφάλειας και ανάκτησης θα πρέπει να εφαρμόζονται εάν δεν αποτελούν μέρος του συστήματος διαχείρισης βάσης δεδομένων.
4. Τα δεδομένα πρέπει να είναι συνεπή. Η φόρμα και το περιεχόμενο πρέπει να είναι τα ίδια μετά την ολοκλήρωση μεγάλων συνόλων δεδομένων από διαφορετικές πηγές.
5. Τα δεδομένα δεν πρέπει να είναι περιττά. Στην πράξη τα περιττά δεδομένα θα πρέπει να αποφεύγονται, να ελαχιστοποιούνται η να αιτιολογείται η τυχόν χρησιμότητά τους.
6. Τα δεδομένα πρέπει να είναι έγκαιρα. Ο χρονικός ορίζοντας των δεδομένων πρέπει να συμβαδίζει με τον χρόνο της ανάλυσης.
7. Τα δεδομένα πρέπει να είναι καλά κατανοητά.
8. Το σύνολο των δεδομένων πρέπει να είναι πλήρες. Η απουσία δεδομένων, τα οποία εμφανίζονται στην πραγματικότητα, πρέπει να ελαχιστοποιηθεί. Αυτές οι απώλειες θα μπορούσαν να μειώσουν την ποιότητα ενός μοντέλου.

Έχοντας τελειώσει την διαδικασία εξόρυξης των δεδομένων προχωράμε στην ανάλυσή τους την οποία θα δούμε πιο αναλυτικά στα επόμενα κεφάλαια.

ΚΕΦΑΛΑΙΟ 2

ΟΠΤΙΚΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

Έχοντας κάνει τις απαραίτητες αναλύσεις και τεχνικές εξόρυξης στα μεγάλα δεδομένα, ακολουθεί το σημείο όπου ο αναλυτής επεξεργάζεται τα αποτελέσματα με σκοπό τον εντοπισμό και την αναγνώριση δομών και ιδιοτήτων σε ένα σύνολο δεδομένων. Τα αποτελέσματα αυτά, τις περισσότερες φορές, τα παρέχει στα στελέχη των επιχειρήσεων για την λήψη αποφάσεων. Για την ευκολότερη ανάγνωση των δεδομένων και των αποτελεσμάτων, ο αναλυτής μετατρέπει τους αριθμούς σε εικόνες. Αυτή η τεχνική ανάλυσης ονομάζεται Οπτικοποίηση Δεδομένων και αποτελεί στάδιο της εξαγωγής γνώσης από τα δεδομένα (KDD).

2.1 Οπτικοποίηση Δεδομένων

Η διαδικασία της οπτικοποίησης περιλαμβάνει τεχνικές που χρησιμοποιούνται προκειμένου να αναπαραστήσουμε τα δεδομένα με χρήση γραφικών μέσων. Σκοπός αυτών των τεχνικών είναι μια πιο «ζωντανή» προσέγγιση του υπό εξέταση συνόλου δεδομένων. Με την βοήθεια γραφημάτων, μπορούν να αναπαρασταθούν ιδιότητες των δεδομένων, συγκρίσεις τιμών, γεωγραφική διασπορά συμβάντων, σχέσεις συνάφειας, ανοδικές και καθοδικές τάσεις, επιμερισμός συνόλων σε υποσύνολα κ.λπ. Ο ανθρώπινος εγκέφαλος επεξεργάζεται καλύτερα πληροφορίες που προέρχονται από εικόνες, παρά όταν προέρχονται από συμπαγή κείμενα και ατελείωτους αριθμούς. Πέρα από την χρησιμότητα της γραφικής αναπαράστασης, η οπτικοποίηση των δεδομένων καθιστά τα αποτελέσματα πιο καλαίσθητα και ευχάριστα για ανάγνωση. Αυτές οι ιδιότητες έχουν καταστήσει την οπτικοποίηση ένα χρήσιμο εργαλείο της ανάλυσης δεδομένων και την εξαγωγή συμπερασμάτων. *[Κύρκος, Ε. (2015). Επιχειρηματική Ευφυΐα και Εξόρυξη Δεδομένων]*

Με την εξέλιξη των υπολογιστικών συστημάτων και τα δεδομένα να γίνονται διαρκώς μεγαλύτερα, μπορεί να γίνει με ασφάλεια η πρόβλεψη ότι η χρήση τεχνικών οπτικοποίησης θα συνεχίσει να μεγαλώνει και να αποκτά διαρκώς αυξανόμενη αξία.

Ο βασικός σκοπός της χρήσης των τεχνικών οπτικοποίησης, είναι η παρουσίαση πολλαπλών και περίπλοκων δομών των δεδομένων, με κατανοητό και απλό τρόπο. Με την χρήση γραφημάτων επιτυγχάνεται η κατανόηση συνθηκών που σε άλλες περιπτώσεις θα ήταν δύσκολο να εντοπιστούν, όπως ο εντοπισμός ακραίων τιμών (outliers). Επίσης είναι δυνατό να προβλεφθούν μελλοντικές τιμές και να απεικονιστούν πιθανές τάσεις μεταβλητών. *[Miller, J. D. (2017). Big Data Visualization. Packt Publishing Ltd.]*

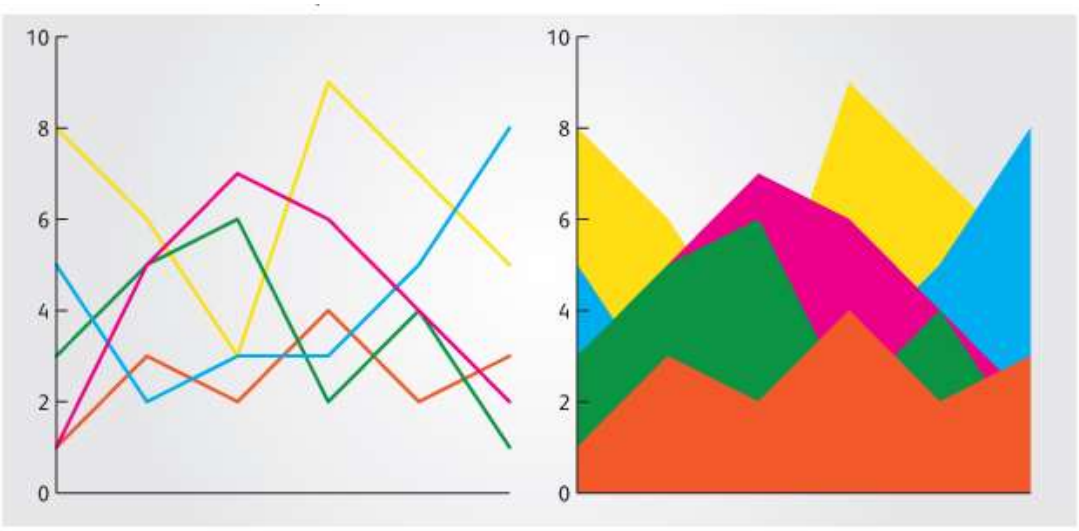
2.2 Τεχνικές Οπτικοποίησης

Υπάρχει μεγάλο εύρος τεχνικών για την οπτικοποίηση των δεδομένων. Χρησιμοποιούνται διαφορετικά γραφήματα/διαγράμματα ανάλογα με την πληροφορία που ο εκάστοτε αναλυτής θέλει να τονίσει. Κάποια από τα βασικότερα παρουσιάζονται παρακάτω.

2.2.1 Γραφήματα χρονικών τάσεων (Time-Series Graphs)

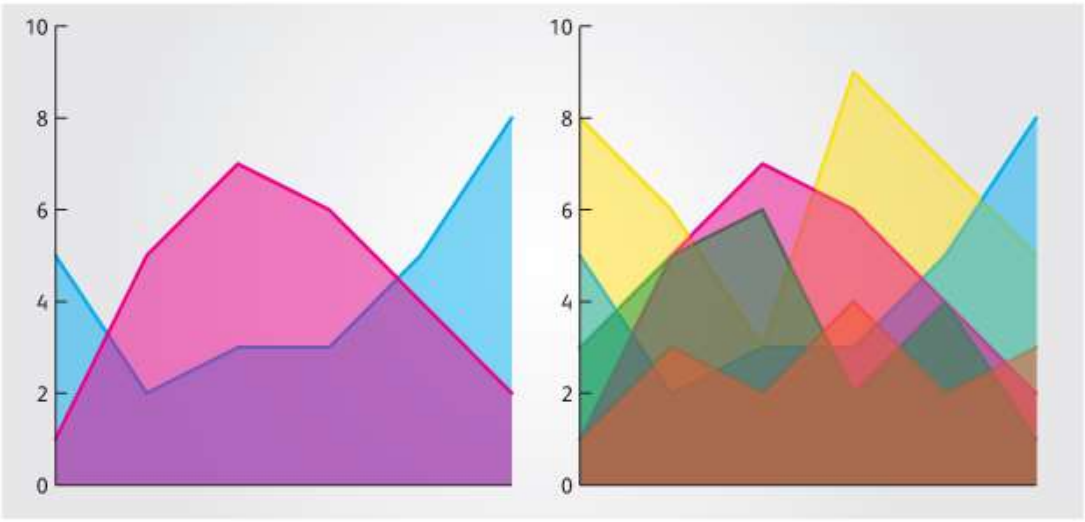
Στην περίπτωση μεταβλητών κινούμενων στον χρόνο, π.χ. τα έσοδα και τα έξοδα μιας επιχείρησης, η απεικόνιση αυτών των αλλαγών/τάσεων σε ένα γράφημα είναι εξαιρετικά ενδιαφέρον και βοηθητικό στα στελέχη της εκάστοτε επιχείρησης. Στην αναπαράσταση των χρονοσειρών, μπορούν να χρησιμοποιηθούν δύο είδη γραφημάτων: *γράφημα γραμμών (line chart)* και *γράφημα περιοχής (area chart).*

Τα γραφήματα γραμμών και περιοχής σχετίζονται πολύ στενά. Χρησιμοποιούνται και τα δύο για δεδομένα χρονοσειρών, δείχνουν μια συνέχεια σε ένα σύνολο δεδομένων και τονίζουν τις τάσεις των τιμών των μεταβλητών και όχι μεμονωμένες τιμές. Ωστόσο, δεν μπορούν πάντα να χρησιμοποιηθούν εναλλακτικά, καθώς έχουν κάποιες μικρές, αλλά σημαντικές, διαφορές.



Εικόνα 2.1: Γραφήματα Χρονικών Τάσεων – line chart, area chart.

Όπως φαίνεται στην παραπάνω εικόνα, στο line chart αριστερά διακρίνονται όλες οι κορυφές όλων των γραμμών, ενώ αντίθετα στο area chart δεξιά έχουν χαθεί πολλές πληροφορίες στην σύμπτυξη των χρωμάτων. Ένας τρόπος για την επίλυση αυτού του προβλήματος είναι η διαφάνεια των περιοχών, αλλά και πάλι αν οι περιοχές είναι παραπάνω από τρεις η εικόνα θα είναι χαοτική.



Εικόνα 2.2: Γραφήματα Περιοχών – line chart, area chart με διαφάνεια.

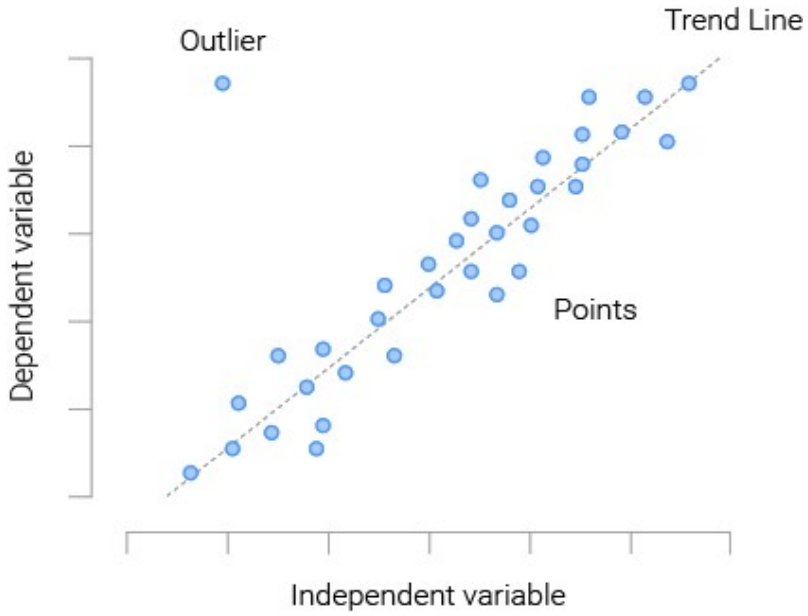
Ο λόγος που χρησιμοποιούνται τα διαγράμματα περιοχών είναι για την επισήμανση φυσικών μετρήσεων και όχι άυλων ποσοστών. Παράδειγμα οι τιμές του πληθυσμού. Ένα γράφημα γραμμών θα ήταν αποτελεσματικό για την εμφάνιση της καθαρής

μεταβολής του πληθυσμού με την πάροδο του χρόνου, ενώ ένα γράφημα περιοχής δείχνει την τον συνολικό πληθυσμό σε αυτή την πάροδο χρόνου.

[Πηγή Εικόνων: https://visual.ly/blog/line-vs-area-charts/#:~:text=A%20line%20chart%20would%20be,rather%20than%20an%20intangible%20rate.]

2.2.2 Διάγραμμα Σχέσης Μεταβλητών – διασποράς (scatter plot)

Στην στατιστική ανάλυση, μια από τις χρησιμότερες πληροφορίες που αντλούνται από τα δεδομένα είναι η σχέση μεταξύ των υπό εξέταση μεταβλητών, ώστε να υπάρχει δυνατότητα επεξήγησης τους και εύρεση εκτιμώμενων τιμών. Ένα από τα βασικά γραφήματα που χρησιμοποιούνται για την απεικόνιση αυτής της σχέσης είναι το *διάγραμμα συσχετίσεων (scatter plot).* Επιπλέον, είναι ευκολότερος ο εντοπισμός ακραίων τιμών (outliers) όπως φαίνεται στην παρακάτω εικόνα.



Εικόνα 2.3: Διάγραμμα Συσχετίσεων (scatter plot)

Κάθε κουκίδα (point) αντιστοιχεί σε μια τιμή της ανεξάρτητης μεταβλητής X (independent variable) του οριζόντιου άξονα, και σε μια τιμή της εξαρτημένης μεταβλητής Y (dependent variable) του κάθετου άξονα. Η ανεξάρτητη μεταβλητή X χρησιμοποιείται για την ερμηνεία της μεταβλητότητας της εξαρτημένης μεταβλητής Y. Η γραμμή τάσης (trend line) σχηματίζεται βάση της θέσεις των σημείων στο

γράφημα, ώστε να σχηματιστεί μια γενική εικόνα αυτής της σχέσης των δύο μεταβλητών. Οι μεταβλητές Χ και Υ μπορεί να έχουν, θετική ή αρνητική συσχέτιση ή να μην συσχετίζονται με κανένα τρόπο.

Εικόνα 2.4: Διαγράμματα Διασποράς Διαφορετικών Συσχετίσεων

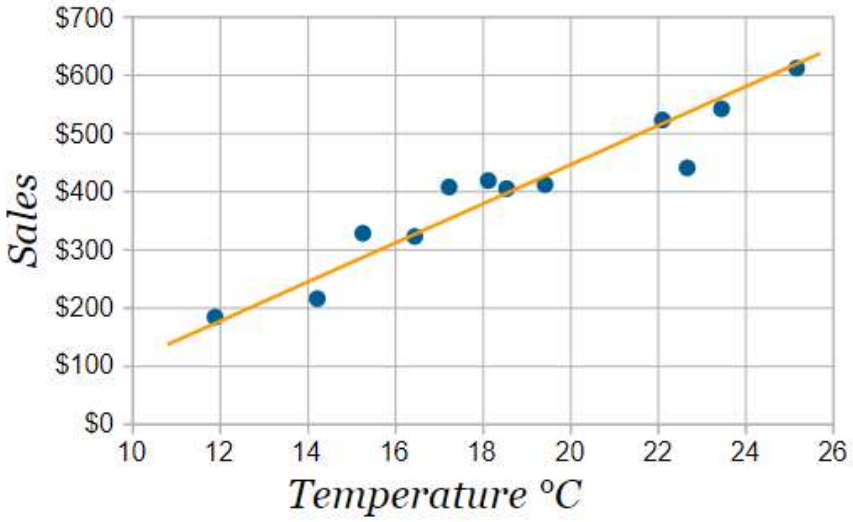
Ένα απλό παράδειγμα για την κατανόηση των εννοιών και του διαγράμματος διασποράς:

Έστω ότι το εργοστάσιο Α παραγωγής παγωτού θέλει να προβλέψει τις πωλήσεις του ανάλογα με την ημερήσια θερμοκρασία. Κατέγραψε τα δεδομένα των τελευταίων 12 ημερών στον παρακάτω πίνακα:

Ice Cream Sales vs Temperature	
Temperature °C	Ice Cream Sales
14,2°	$215
16,4°	$325
11,9°	$185
15,2°	$332
18,5°	$406
22,1°	$522
19,4°	$412
25,1°	$614
23,4°	$544
18,1°	$421
22,6°	$445
17,2°	$408

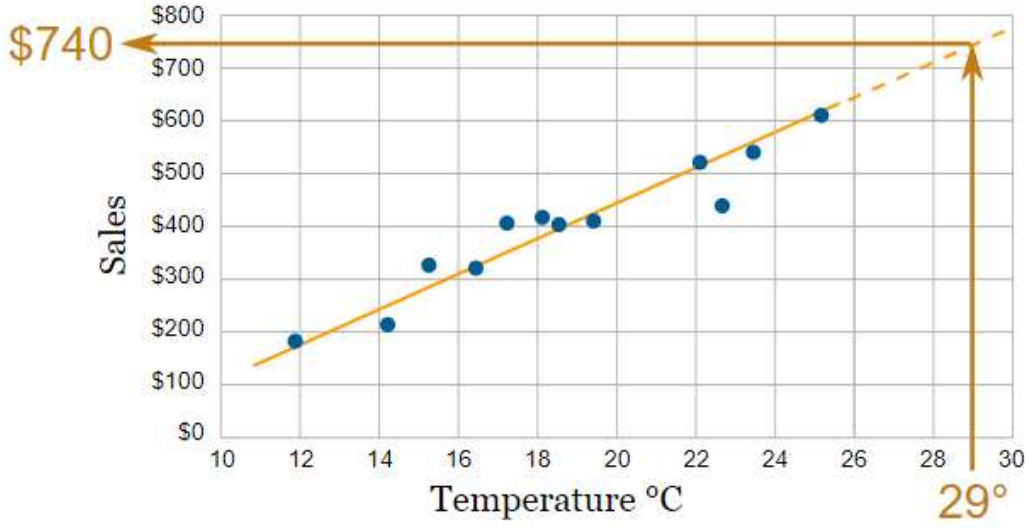
Εικόνα 2.5: Πωλήσεις Παγωτού Εργοστασίου Α

Και τα ίδια δεδομένα σε μορφή scatter plot:



Εικόνα 2.6: Διάγραμμα Scatter Plot – Πωλήσεις-Θερμοκρασία

Όπως φαίνεται από το διάγραμμα, υπάρχει μια θετική σχέση ανάμεσα στις πωλήσεις παγωτού και την ημερήσια θερμοκρασία. Όσο αυξάνεται η θερμοκρασία, αυξάνονται και οι πωλήσεις. Έχοντας μια γραμμική σχέση ανάμεσα στις μεταβλητές, μπορούμε να προβλέψουμε τις πωλήσεις για μια θερμοκρασία που δεν έχει καταγραφεί για ένα συγκεκριμένο ύψος πωλήσεων. Επίσης, είναι εφικτό να κάνουμε προβλέψεις για θερμοκρασίες υψηλότερες από 22,6 που είναι η μεγαλύτερη παρατήρηση που έχουμε. Παράδειγμα:



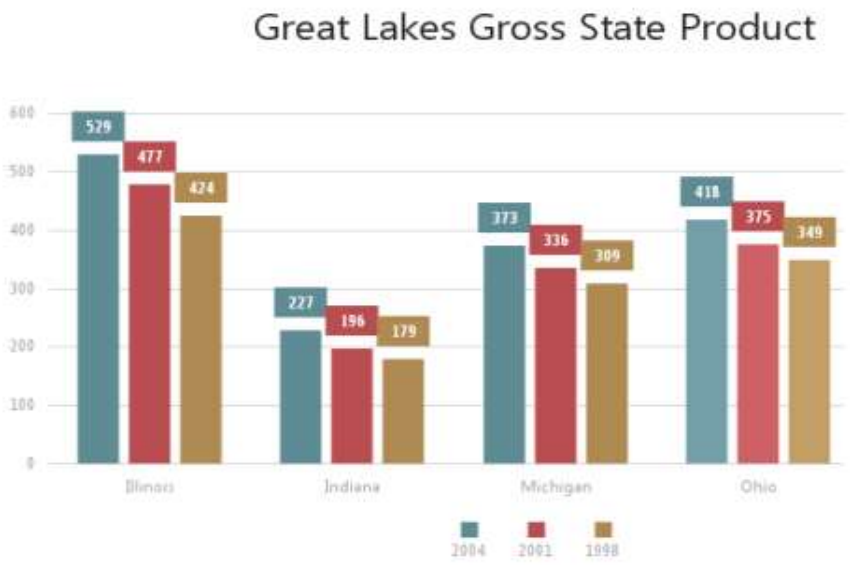
Εικόνα 2.7: Διάγραμμα Scatter Plot – Πωλήσεις-Θερμοκρασία - ΠΡΟΒΛΕΨΗ

Η εκτίμηση των πωλήσεων για θερμοκρασία 29, μεγαλύτερη από κάθε άλλη καταγεγραμμένη μέτρηση, ονομάζεται *γραμμική παρέκταση*.

2.2.3 Διαγράμματα Σύγκρισης Τιμών (Bar Chart and Histogram)

α) Ραβδόγραμμα (Bar Chart) ή Διάγραμμα Pareto.

Είναι ένα από τα πιο συνηθισμένα διαγράμματα στην στατιστική ανάλυση, από το οποίο αντλούμε πληροφορίες για τις τιμές των δεδομένων συγκρίνοντας τα ύψη ή τα μήκη των ράβδων. Τα χαρακτηριστικά που πρέπει να έχει ένα ραβδόγραμμα είναι: ο τίτλος που εξηγεί το περιεχόμενο του διαγράμματος, τα ονόματα των αξόνων, η αριθμητική κλίμακα και οι αποστάσεις ανάμεσα στις ράβδους πρέπει να είναι ίσες. Ακολουθεί ένα παράδειγμα με δεδομένα το ακαθάριστο εγχώριο προϊόν (ΑΕΠ) για τέσσερις πολιτείες τις Αμερικής που βρίσκονται κοντά στις Μεγάλες Λίμνες.

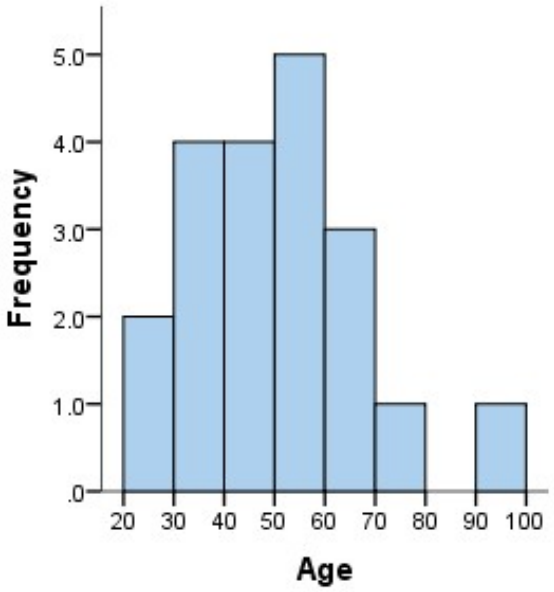


Εικόνα 2.8: Ραβδόγραμμα (Bar Chart) – Διάγραμμα Pareto

β) Ιστόγραμμα (Histogram) ή Διάγραμμα Συχνοτήτων

Το ιστόγραμμα είναι η αναπαράσταση μια μεταβλητής σε ένα σύνολο δεδομένων και εμφανίζει την συχνότητα ή τον αριθμό εμφάνισης μιας συγκεκριμένης τιμής της μεταβλητής σε έναν συγκεκριμένο χρόνο. Η δημιουργία ενός ιστογράμματος απαιτεί:

την συγκέντρωση των δεδομένων, την συγκέντρωσή τους σε κλάσεις ταξινόμησης, την δημιουργία ενός πίνακα συχνοτήτων και την σχεδίαση των ραβδογραμμάτων. http://dione.lib.unipi.gr/xmlui/bitstream/handle/unipi/167/DT2003-0107.pdf?sequence=1&isAllowed=y

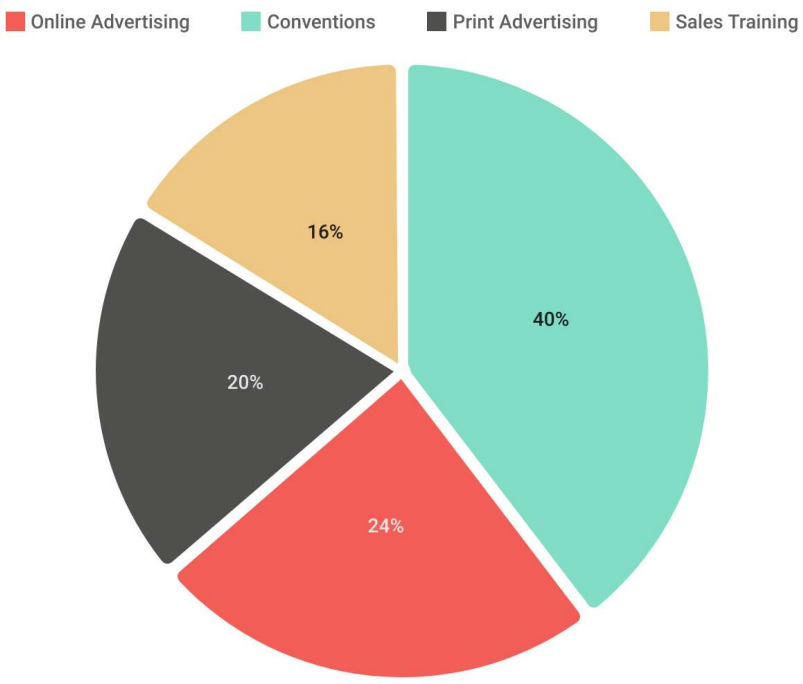


Εικόνα 2.9: Ιστόγραμμα (Histogram) ή Διάγραμμα Συχνοτήτων

2.2.4 Διαγράμματα Μέρους Συνόλου (Pie Chart and Tree Map)

α) Διάγραμμα Πίτας (Pie Chart)

Από τους πιο δημοφιλής τύπους διαγραμμάτων, συναντάται στην καθημερινότητα σε εσημερίδες, αναφορές επιχειρήσεων κ.λπ. Παρόλο που χρησιμοποιείται ευρέως, η χρήση του βρίσκεται αντιμέτοπη με την ισχύ μιας βασικής προϋπόθεσης: ότι τα κομμάτια του κύκλου πρέπει να συνοψίζονται σε ένα κατανοητό σύνολο. Έχοντας επιλέξει ένα σύνολο δεδομένων από έναν συγκεκριμένο πληθυσμό, μπορούμε να χωρίσουμε αυτόν τον πληθυσμό βάση ενός χαρακτηριστικού που έχουμε επιλέξει για την ανάλυση. Κάθε μέρος της πίτας αναπαριστά το ποσοστό του πληθυσμού που εμπεριέχει ένα συγκεκριμένο στοιχείο. Παράδειγμα, μια επιχείρηση θέλει να εξετάσει τον διαμερισμό του συνολικού διαθέσιμου χρηματικού ποσού για το μάρκετινγκ. Διαγραμματικά:



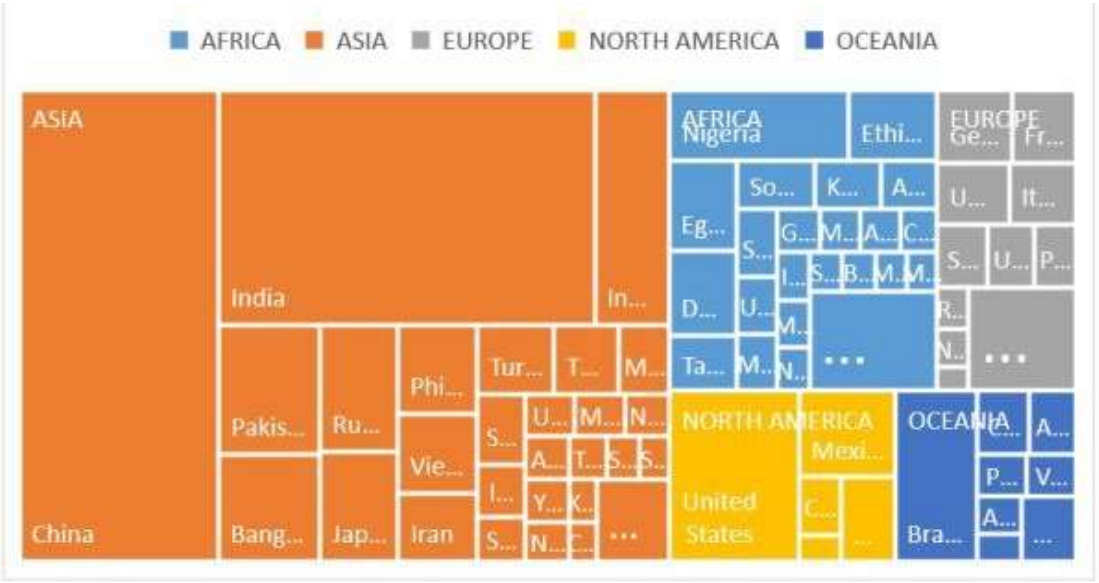
Εικόνα 2.10: Διάγραμμα Πίτας (Pie Chart)

β) Δενδροχάρτης (Tree Map)

Για την αναπαράσταση δεδομένων με ιεραρχική δομή σε μια κατανοητή εικόνα, οι Johnson and Shneiderman (1991) πρότειναν τη μέθοδο των Δενδροχαρτών (Tree maps). Η κατασκευή δενδροχαρτών επιτυγχάνεται με την χρήση διαγραμμάτων δένδρων. Ένα διάγραμμα δένδρου αποτελείται από κόμβους συνδεδεμένους με τέτοιο τρόπο ώστε να αποτυπώνουν την ιεραρχική δομή. Ο λόγος μετασχηματισμού τους σε δενδροχάρτες είναι ο όγκος και η πολυπλοκότητα τους, καθώς και στο ότι τα διαγράμματα δένδρων παρέχουν πληροφορίες σχετικά με την δομή και όχι με το περιεχόμενο των δεδομένων. Αντίθετα, στους δενδροχάρτες παρέχονται πληροφορίες για το περιεχόμενο και την δομή των δεδομένων, επιμερίζοντας τα σε τμήματα ενός ορθογωνίου παραλληλόγραμμου, όσοι είναι οι κόμβοι του πρώτου επιπέδου. Το κάθε τμήμα επιμερίζεται σε τόσα υποτμήματα, όσοι είναι οι κόμβοι που ανήκουν σε αυτό έως ότου εξαντληθούν όλα τα επίπεδα και οι κόμβοι. Ο επιμερισμός γίνεται αρχικά κάθετα και η αλλαγή της κατεύθυνσης επιμερισμού – σε οριζόντια, γίνεται για κάθε μετάβαση σε χαμηλότερο επίπεδο. Το μέγεθος κάθε κόμβου εξαρτάται από το «βάρος» των δεδομένων. Αν για παράδειγμα, το δένδρο αναπαριστά την δομή των φακέλων σε έναν υπολογιστή, το βάρος θα μπορούσε να

είναι το συνολικό μέγεθος των αρχείων που περιέχει ο κάθε φάκελος (περιεχόμενο δεδομένων). Άλλες πληροφορίες μπορούν να εξαχθούν από τους δενδροχάρτες με την χρήση διαφορετικών χρωμάτων, υφών, πλαισίων κ.λπ. https://repository.kallipos.gr/bitstream/11419/1232/2/Kef._5.pdf

Παράδειγμα, στο παρακάτω δενδροχάρτη παρουσιάζονται οι χώρες ανά ήπειρο, με το μέγεθος του κάθε τετραγώνου να αναπαριστά το ανάλογο μέγεθος κάθε χώρας.



Εικόνα 2.11: Δενδροχάρτης (Tree Map)

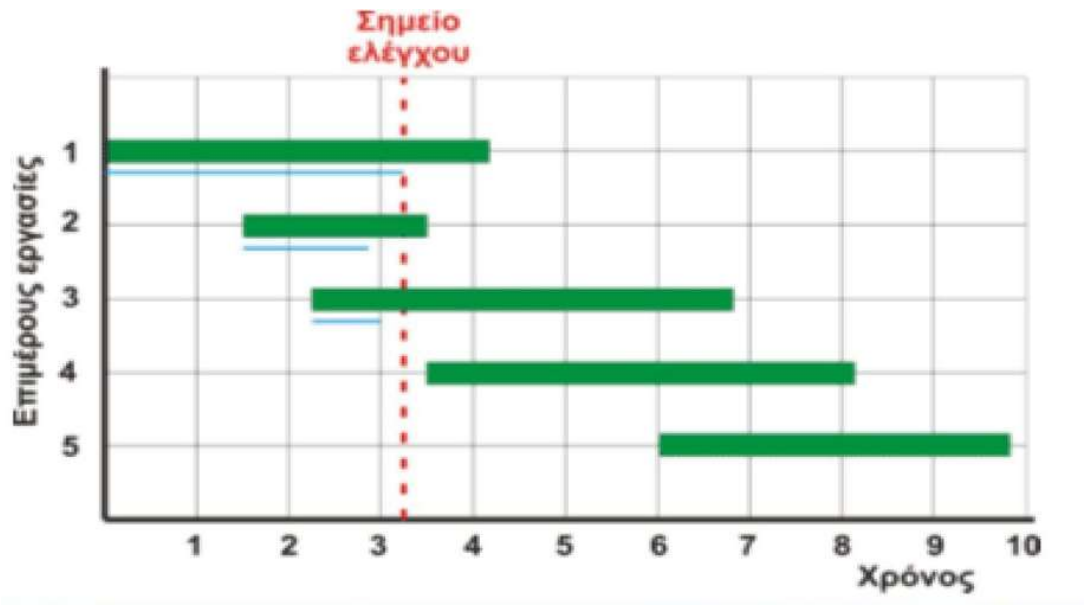
2.2.5 Διάγραμμα Διαχείρισης Έργου (Gantt)

Η κατασκευή ενός διαγράμματος Gantt χρησιμοποιείται για τον σχεδιασμό και τον συντονισμό των διαφόρων εργασιών σε ένα έργο. Είναι ένα οριζόντιο ραβδόγραμμα που απεικονίζει την σχέση των εργασιών αυτών, μέσα στον χρόνο. Ο οριζόντιος άξονας χωρίζεται σε ίσα χρονικά διαστήματα (π.χ. μέρες, μήνες) και ορίζει την συνολική διάρκεια του έργου. Στον κάθετο άξονα τοποθετούνται οι επι μέρους δραστηριότητες του έργου, στην κάθε μία αντιστοιχεί μια χρονική διάρκεια.

Η βασικότερη χρήση των διαγραμμάτων Gantt είναι η παρακολούθηση της προόδου ενός έργου. Η κατασκευή του απαιτεί τα εξής: απαρίθμηση όλων των δραστηριοτήτων του έργου με την αντίστοιχη χρονική τους διάρκεια, διάταξη των δραστηριοτήτων στον κάθετο άξονα – συνήθως με σειρά προτεραιότητας ως προς την ημερομηνία έναρξης τους, σχεδιασμός του άξονα του χρόνου στις κατάλληλες

χρονικές μονάδες και ο σχεδιασμός των δραστηριοτήτων ως ράβδοι σε οριζόντια διάταξη με μήκος ανάλογο με την χρονική διάρκεια που απαιτούν για την ολοκλήρωσή τους. Τέλος, για κάθε ράβδο σχεδιάζεται μια δεύτερη ράβδος που απεικονίζει την πρόοδο υλοποίησης κάθε δραστηριότητας. Την χρονική στιγμή που θέλουμε να ελέγξουμε την πορεία του έργου, μπορούμε να τραβήξουμε μια κάθετη γραμμή ως προς τον άξονα του χρόνου, για να δούμε την πρόοδο του έργου σε σχέση με τον αρχικό προγραμματισμό.

https://repository.kallipos.gr/bitstream/11419/747/1/02_chapter_9.pdf

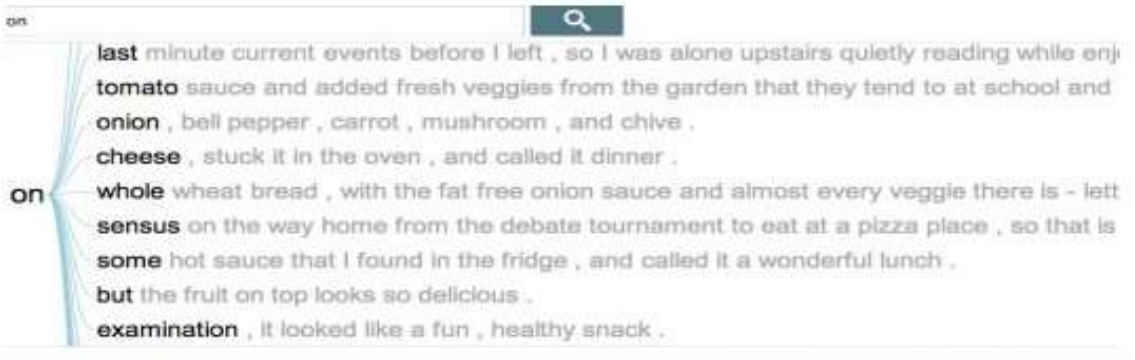


Εικόνα 2.12: Διάγραμμα Διαχείρισης Έργου (Gantt)https://slideplayer.gr/slide/12562996/

2.2.6 Ανάλυση Κειμένου (Word Tree and Word/Tag Cloud)

α) Δένδρο Κειμένου (word tree)

Είναι ένα εργαλείο εντοπισμού λέξεων η φράσεων μέσα σε ένα κείμενο. Ο χρήστης αναζητά την λέξη ή φράση που επιθυμεί και το πρόγραμμα του επιστρέφει όλες τις προτάσεις στις οποίες η λέξη/φράση εμφανίζεται. Τα αποτελέσματα απεικονίζονται με την μορφή διακλάδωσης ενός δέντρου με σκοπό την εμφάνιση επαναλαμβανόμενων λέξεων/φράσεων. Αν στην αναζήτηση δεν οριστεί κάποια λέξη ή φράση, αυτόματα το πρόγραμμα θα πάρει ως είσοδο την λέξη που εμφανίζεται συχνότερα στο κείμενο.



Εικόνα 2.12: Δένδρο Κειμένου (word tree)

β) Σύννεφα Λέξεων/Ετικετών (word/tag clouds)

Για την περιγραφή του περιεχομένου ενός κειμένου ή ιστότοπου χρησιμοποιούνται οι οπτικές αναπαραστάσεις που ονομάζονται σύννεφα λέξεων ή ετικετών. Το αποτέλεσμα αυτής της διαδικασίας είναι η οπτικοποίηση των λέξεων σε διαφορετικό μέγεθος, ανάλογο με την συχνότητα εμφάνισής τους στο υπό εξέταση κείμενο. Ένα παράδειγμα ενός σύννεφου λέξεων για τα δεδομένα (data) θα ήταν κάπως έτσι:

Εικόνα 2.13: Σύννεφα Λέξεων/Ετικετών (word/tag clouds)

Η οπτικοποίηση δεδομένων χρησιμοποιείται ευρέως από τις επιχειρήσεις για την παρουσίαση αποτελεσμάτων στις διάφορες εργασίες που αναθέτονται στα τμήματα για τον λόγο του ότι μπορούν εύκολα να καταλάβουν τα αποτελέσματα όλοι οι ενδιαφερόμενοι ανεξαρτήτου τμήματος και χωρίς να χρειάζεται να γνωρίζουν περεταίρω τις διαδικασίες της ανάλυσης. Μια εικόνα είναι πάντα πιο ευχάριστη στην όψη και πιο εύκολα κατανοητή από το κοινό.

ΚΕΦΑΛΑΙΟ 3

ΕΠΙΧΕΙΡΗΜΑΤΙΚΗ ΑΝΑΛΥΤΙΚΗ ΚΑΙ ΑΝΑΛΥΤΙΚΗ ΜΑΡΚΕΤΙΝΓΚ

Η σημερινή εποχή μπορεί να χαρακτηριστεί και ως εποχή της ψηφιακής τεχνολογίας καθιστώντας δυνατό, όπως ειπώθηκε και στα προηγούμενα κεφάλαια, το να συλλέγονται τεράστιες ποσότητες δεδομένων.

Τα δεδομένα έχουν φτάσει σε τέτοιον όγκο που πολλές επιχειρήσεις διαμαρτύρονται ήδη για το ότι πλέον, η ανάλυσή τους είναι πολύ δύσκολη έως αδύνατη με συμβατικές μεθόδους. Ωστόσο, υπάρχουν και εταιρείες που βλέπουν αυτήν την αυξητική τάση του όγκου των δεδομένων, ως πηγή ανταγωνιστικού πλεονεκτήματος. Στην πραγματικότητα, ένα από τα πιο διαδεδομένα θέματα που απασχολούν τον επιχειρηματικό κόσμο σήμερα είναι αυτό της Επιχειρηματικής Ανάλυσης (Business Analytics).

3.1 Επιχειρηματική Αναλυτική

Η Επιχειρηματική Ανάλυση είναι μια σχετικά νέα επιστήμη. Κάποιοι λένε ότι δεν είναι πραγματικά επιστήμη αλλά περισσότερο ένα σύνολο εργασιών και δραστηριοτήτων. Κάποιοι άλλοι λένε ότι είναι επιστήμη και πιο συγκεκριμένα ένα σύνολο επιχειρησιακών τεχνικών το οποίο δυστυχώς, απουσίαζε από τον πολύπλοκο κόσμο των επιχειρήσεων για πολύ καιρό.

Το International Institute of Business Analysis καθορίζει το Business Analytics με τους εξής ορισμούς:

• Με τον όρο Business Analytics αναφερόμαστε σε ένα σύνολο δράσεων που περιλαμβάνουν την εφαρμογή αλλαγών μέσα σε ένα οργανωτικό πλαίσιο, προσδιορίζοντας τις ανάγκες και προτείνοντας αξιόλογες και αποτελεσματικές λύσεις στους ενδιαφερόμενους (stakeholders) με σκοπό κάποιο κέρδος.

• Με τον όρο Business Analytics (BA) αναφερόμαστε στις δεξιότητες, τεχνολογίες και πρακτικές που είναι αναγκαίες για την εύρεση κρυμμένων πληροφοριών και γνώσεων μέσα από συνεχή διερεύνηση και ερμηνεία δεδομένων που αφορούν παρελθοντικές επιδόσεις της επιχείρησης. Οι γνώσεις και οι πληροφορίες αυτές με διάφορες στατιστικές μεθόδους και σε συνδυασμό με την εμπειρία και ευστροφία αυτού που καλείται να τις ερμηνεύσει, οδηγούν στην ανάπτυξη νέων ιδεών και γενικότερα σε αρτιότερο επιχειρηματικό σχεδιασμό.

Η επιστήμη του Business Analytics μπορεί να διεξαχθεί στο πλαίσιο δραστηριοτήτων που σκοπό έχουν να επιφέρουν αλλαγές στην επιχείρηση. Αυτές οι επιδιωκόμενες αλλαγές μπορεί να είναι στρατηγικής σημασίας, μπορεί να είναι αλλαγές τακτικής, ή απλώς αλλαγές στον τρόπο λειτουργίας της επιχείρησης.

Το Business Analytics μπορεί να λάβει χώρα κατά την διάρκεια κάποιου project ή γενικότερα μέσα στην επιχείρηση, βοηθώντας συνεχώς στην εξέλιξη και την συνεχή βελτίωσή της. Οι τεχνικές του BA μπορούν να χρησιμοποιηθούν για την καλύτερη κατανόηση της τρέχουσας κατάστασης της επιχείρησης, της μελλοντικής κατάστασης που αυτή επιδιώκει να επιτύχει και για τον εντοπισμό των δραστηριοτήτων που απαιτούνται για να μεταφερθεί από την τρέχουσα στην επιδιωκόμενη μελλοντική κατάσταση.

3.2 Αναλυτική Μάρκετινγκ

Η διαδικασία ανάλυσης μάρκετινγκ (marketing analytics) δεν αφορά μόνο στην διατήρηση αποφυγής προβλημάτων της επιχείρησης ή του οργανισμού, αλλά την βοηθά στο να επιτύχει την απόδοσή της στο υψηλότερο δυνατό επίπεδο.

Για τις περισσότερες εταιρείες, η επιστήμη της αναλυτικής βρίσκεται στο επίκεντρο του τρόπου διαχείρισης της επιχείρησης. Το χρηματοοικονομικό τμήμα ασχολείται άμεσα με τις μετρήσεις και τα αναλυτικά στοιχεία όπως: έσοδα, κέρδος, απόδοση επένδυσης (Rate Of Interest (ROI)), απόδοση ιδίων κεφαλαίων (Rate Of Equity (ROE)) και πολλά άλλα. Το κατασκευαστικό τμήμα παρακολουθεί μετρήσεις όπως έξοδα και ελαττώματα. Το τμήμα ανθρώπινου δυναμικού θα μετρήσει την απόδοση των εργαζομένων. Συνεπώς, κάθε τμήμα μιας επιχείρησης χρησιμοποιεί αναλυτικά

στοιχεία για τον έλεγχο της απόδοσης και επίδοσης με βάση τους προκαθορισμένους στόχους.

Ακόμα και σήμερα, παρά την ραγδαία τεχνολογική εξέλιξη και τους διαθέσιμους πόρους, μεγάλο ποσοστό των επιχειρήσεων δεν κατέχουν και δεν χρησιμοποιούν αναλυτικά στοιχεία

Σύμφωνα με τον Jerry Rackley, η αναλυτική μάρκετινγκ ορίζεται ως η διαδικασία προσδιορισμού μετρήσεων, που είναι έγκυρες ενδείξεις της απόδοσης του μάρκετινγκ στην επίτευξη των στόχων του, παρακολουθώντας αυτές τις μετρήσεις με την πάροδο του χρόνου και την χρήση των αποτελεσμάτων με σκοπό την βελτίωση του τρόπου λειτουργίας του μάρκετινγκ.

Τα βασικά στοιχεία αυτού του ορισμού εξετάζονται παρακάτω πιο αναλυτικά:

- Έγκυροι δείκτες (valid indicators): υπάρχουν πολλά μετρήσιμα στοιχεία, εργασίες και αποτελέσματα, που μπορούν να χρησιμοποιηθούν στην ανάλυση του μάρκετινγκ. Κάποια από αυτά είναι αληθινοί δείκτες απόδοσης. Ωστόσο, η διαδικασία ανάλυσης πρέπει να καθορίσει ποιες μετρήσεις έχουν νόημα και αντιπροσωπεύουν καλύτερα την αξία που δημιουργεί το μάρκετινγκ στον οργανισμό.
- Επιδίωξη στόχων: η διαδικασία ανάλυσης είναι ιδανικά κατασκευασμένη ώστε να μετρά την πρόοδο προς ένα σύνολο στόχων. Πρώτα καθορίζονται αυτοί οι στόχοι και στη συνέχεια προσδιορίζονται οι σχετικές μετρήσεις απόδοσης.
- Μετρήσεις παρακολούθησης με την πάροδο του χρόνου: η διαδικασία ανάλυσης δεν αφορά την λήψη ενός τυχαίου, στιγμιότυπου μιας εικόνας μέτρησης, αλλά την παρακολούθηση μετρήσεων με την πάροδο του χρόνου παρακολουθώντας τις τάσεις και την κατεύθυνση της απόδοσης.
- Βελτίωση του τρόπου λειτουργίας του μάρκετινγκ: υπάρχουν πολλοί λόγοι που μια επιχείρηση εφαρμόζει μια διαδικασία ανάλυσης, όπως αιτιολογία χρήσης πόρων, αλλά τελικά ο πιο σημαντικός λόγος είναι να βελτιώσει την απόδοσή της.

Οι όροι αναλυτική και μετρήσεις έχουν παρόμοιες έννοιες. Η αναλυτική είναι τόσο η διαδικασία όσο και η συλλογή των αποτελεσμάτων και πληροφορίων απόδοσης. Οι μετρήσεις είναι η «ατομική μονάδα» των της αναλυτικής. Η αναλυτική μάρκετινγκ αποτελείται από την δημιουργία μιας σειράς μετρήσεων σε συγκεκριμένους τομείς.

Με βάση τον ορισμό της αναλυτικής, είναι σαφές ότι η διαδικασία της αναλυτικής δεν είναι απλά ένα σύνολο αριθμών, αλλά αποτελείται από τα παρακάτω κύρια συστατικά.

1. Άνθρωποι. Η αναλυτική μάρκετινγκ δημιουργείται, εκτελείται και διαχειρίζεται από ανθρώπους ειδικούς στον τομέα. Στις περισσότερες εταιρείες μάρκετινγκ, ο συντονιστής της διαδικασίας αναλυτικής είναι ο υπεύθυνος του τμήματος μάρκετινγκ (chief marketing officer (CMO)) ή ο διευθυντής μάρκετινγκ.
2. Βήματα. Η διαδικασία αναλυτικής μάρκετινγκ αποτελείται από μια ακολουθία βημάτων, τα οποία πρέπει να εκτελούνται με την σωστή σειρά ώστε να υπάρχει συνοχή και τα αποτελέσματα να είναι βάσιμα.
3. Εργαλεία και Τεχνολογία. Ενώ η διαδικασία αναλυτικής μάρκετινγκ δεν είναι απαραίτητα περίπλοκη, τα εργαλεία και η τεχνολογία προσφέρουν μεγαλύτερη αξία στις επιχειρήσεις μάρκετινγκ γρηγορότερα από ότι χωρίς αυτά.
4. Είσοδος και Έξοδος (input and output): τα δεδομένα τροφοδοτούν την διαδικασία αναλυτικής μάρκετινγκ με πληροφορίες και αποφάσεις ως «έξοδο» (output) της διαδικασίας.

Όπως πολλές διαδικασίες, η αναλυτική μάρκετινγκ θα πρέπει να έχει μια οριοθετημένη αρχή αλλά χωρίς κάποιο τέλος, είναι μια διαδικασία που μόλις ξεκινήσει πρέπει να συνεχιστεί επ’ αόριστο.

3.3 Επιπτώσεις της Αναλυτικής Μάρκετινγκ

Η αντιμετώπιση της διαδικασίας αναλυτικής μάρκετινγκ απλά ως ένα εργαλείο διαχείρισης της απόδοσης, χάνεται η ουσία αυτής της επιστήμης. Μια σωστή εφαρμογή των αναλυτικών στοιχείων μπορεί να βοηθήσει το μάρκετινγκ να λειτουργεί πιο αποδοτικά και πιο αποτελεσματικά. Ωστόσο, η αναλυτική διαδικασία

θα πρέπει να παρέχει τα δεδομένα τα οποία «εξοικονομούν» στο μάρκετινγκ χρόνο και χρήμα, κάνοντας το καλύτερο δυνατό χρησιμοποιώντας τους λιγότερους δυνατούς πόρους ("do more with less"). Η καλύτερη χρήση των αναλυτικών στοιχείων βοηθά στην λειτουργία του μάρκετινγκ ως ένα «κέντρο εσόδων». Μέλημα της κάθε επιχείρησης είναι η βιωσιμότητα και φυσικά το κέρδος. Αυτό επιτυγχάνεται με την σωστή διαχείριση και χρήση των αναλυτικών στοιχείων από το μάρκετινγκ.

Μια γενική άποψη βασίζεται στο γεγονός ότι η κύρια χρήση των δεδομένων ανάλυσης είναι να υποστηρίζει την διαδικασία λήψης αποφάσεων. Επίσης μια σημαντική χρήση αυτών των δεδομένων είναι στην παροχή αναφορών στο διοικητικό συμβούλιο και στους μετόχους της εκάστοτε εταιρίας. Η λήψη σωστών αποφάσεων από το διοικητικά στελέχη, και η σωστή εκτέλεση των εργασιών που απαιτούνται για την εκπλήρωση των στόχων, οδηγούν στο κέρδος. Επομένως, η αναλυτική μάρκετινγκ είναι ένα «εργαλείο δημιουργίας κέρδους». Πως όμως τα αναλυτικά στοιχεία φτάνουν στο σημείου της δημιουργίας κέρδους; Όλα εξαρτώνται από τον πελάτη και την ζήτηση για το προϊόν/υπηρεσία που εξετάζεται από την αναλυτική. Πότε θα γίνει η αγορά, γιατί θέλει να αγοράσει το συγκεκριμένο προϊόν, πως θα φτάσει στο σημείο πώλησης και τελικά φτάνουμε στο σημείο απόφασης για την αγορά. Σκοπός του μάρκετινγκ είναι η κατανόηση και η προετοιμασία για αυτόν τον «κύκλο αγοράς».

Η κατανόηση αυτού του κύκλου αγοράς γίνεται με διάφορους τρόπους. Κατά τα πρώτα χρόνια της εκτέλεσης των διαδικασιών μάρκετινγκ, τα εργαλεία και τα μέσα ήταν περιορισμένα, οι όροι μάρκετινγκ και διαφήμιση ήταν συνώνυμοι και εκτελούνταν με τον κλασσικό έντυπο τρόπο. Ωστόσο, ακόμα και τότε υπήρχε η ανάγκη για δεδομένα ανάλυσης. Φτάνοντας στον 21^{ο} αιώνα, οι marketers διαθέτουν πληθώρα καναλιών, παραδοσιακά και ψηφιακά, για τον σχεδιασμό στρατηγικής μάρκετινγκ και τον εντοπισμό εν δυνάμει πελατών. Η πολυπλοκότητα των δεδομένων σήμερα- ήχος, εικόνα, βίντεο, κείμενα- μαζί με την δημιουργικότητα των ανθρώπων στα μέσα κοινωνικής δικτύωσης και όχι μόνο, το μάρκετινγκ δεν μπορεί να επιτύχει «ευφυείς» αποφάσεις χωρίς την χρήση αναλυτικής.

Η προσπάθεια των marketers να «ελέγξουν» την αγορά γίνεται μέσω των εκστρατειών «campaigns», που έχουν ως στόχο την προσέλκυση των δυνητικών πελατών, αναφερόμενων ως "leads": άτομα και επιχειρήσεις οι οποίοι έχουν δείξει

κάποιο ενδιαφέρον πάνω στην εκστρατεία αφήνοντας στοιχεία επικοινωνίας για μελλοντική συνεργασία. Για τις περισσότερες επιχειρήσεις οι leads είναι τα «καύσιμα» στην «μηχανή» πωλήσεων. Στόχος του μάρκετινγκ είναι η συνεχής αύξηση και διατήρηση των leads. Στη συνέχεια, οι καμπάνιες διαμορφώνονται βάση των χαρακτηριστικών που έχουν συλλεχθεί για τους leads, δηλαδή των αναλυτικών στοιχείων, για την στόχευση διαφορετικών ομάδων leads εξατομικεύοντας τις καμπάνιες για αυτές τις ομάδες. Τα αναλυτικά στοιχεία βοηθούν επίσης τους marketers να αποφασίσουν ποιες καμπάνιες θα κρατήσουν, ποιες θα αποβάλουν, και ποιες χρίζουν βελτίωσης. Η διαδικασία της αναλυτικής, έχει αντίκτυπο και στους παρακάτω τομείς:

- Αναγνώριση της μάρκας (Brand Recognition). Κατανόηση της ιδεολογίας και του κοινού στο οποίο αναφέρεται το brand της εταιρείας και τι συναισθήματα προκαλεί στους πελάτες.
- Περιεχόμενο (content). Αναγνώριση ποιο περιεχόμενο του μάρκετινγκ καταναλώνεται, κοινοποιείται ευρέως και παράγει την καλύτερη μετατροπή.
- Βελτιστοποίηση καναλιών (channel optimization). Σύγκριση της απόδοσης διαφόρων καναλιών μάρκετινγκ, όπως e-mail η “pay-per-click”, για την βελτίωση της απόδοσής τους ή επενδύοντας σε αυτό που την υψηλότερη.
- Κατανόηση των πελατών (customer understanding). Κατανόηση των αναγκών και προτιμήσεων των πελατών. Περισσότερα πάνω στο συγκεκριμένο θέμα αναλύονται σε παρακάτω ενότητα αυτού του κεφαλαίου.
- Προγνωστική νοημοσύνη (predictive intelligence). Όσο το δυνατόν μεγαλύτερη ακρίβεια στην πρόβλεψη μεταβλητών στην αρχή του κύκλου αγοράς, όπως το πότε θα αγοράσουν οι πελάτες, που θα το αγοράσουν κ.λπ.

Τα παραπάνω αντιπροσωπεύουν ορισμένα από τα μέρη όπου τα αναλυτικά στοιχεία μπορούν να έχουν σημαντικό αντίκτυπο, και όχι μόνο στην λειτουργία μάρκετινγκ, αλλά σε ολόκληρο το σύνολο των διεργασιών και διαδικασιών οργάνωσης μιας επιχείρησης. Ωστόσο, η διαδικασία της αναλυτικής μάρκετινγκ μπορεί να έχει θετικές επιπτώσεις μόνο εάν βασίζεται στις σωστές μετρήσεις, οι οποίες παρακολουθούνται με επιμέλεια και αμερόληπτα, και επηρεάζουν την λήψη αποφάσεων σε συνδυασμό με μια ολοκληρωμένη εικόνα της λειτουργίας της επιχείρησης.

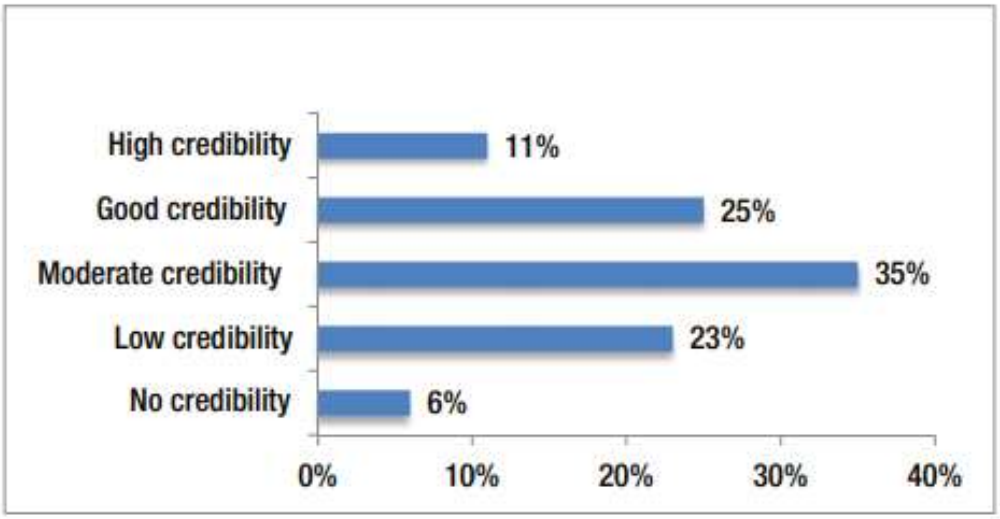
3.3.1 Αξιοπιστία Δεδομένων Ανάλυσης

Όπως έχει αναφερθεί και στο πρώτο κεφάλαιο, τα δεδομένα σε τόσο μεγάλο όγκο δεν μπορούν να είναι πάντα αξιόπιστα ώστε οι marketers να τα εμπιστεύονται τυφλά για την λήψη σημαντικών αποφάσεων. Η αμφισβήτηση των δεδομένων μπορεί να επέλθει από το εσωτερικό και το εξωτερικό περιβάλλον της επιχείρησης. Η αμφιβολία πηγάζει από το γεγονός ότι η διαδικασία είναι κατά κάποιο τρόπο ύποπτη ή τα δεδομένα εισαγωγής έχουν προβλήματα ποιότητας. Άλλες φορές, οι αναλυτές δεν θέλουν να βασίζονται ή να πιστέψουν στο τι αποκαλύπτουν αυτά τα δεδομένα.

Πολλά από τα δεδομένα που χρησιμοποιούνται στην αναλυτική μάρκετινγκ είναι συχνά αποθηκευμένα σε συστήματα διαχείρισης πελατειακών σχέσεων CRM (customer relationship management) και συστήματα αυτοματισμού μάρκετινγκ. Αυτά τα δεδομένα περνάνε μέσα από αυτά τα συστήματα μέσω πολλών διαδρομών, αντιμετωπίζοντας προβλήματα στην διατήρηση της ακρίβειας και τις ποιότητας τους ώσπου να φτάσουν στο σημείο εξόδου.

Ιδανικά, μια επιχείρηση κατέχει συστήματα και μηχανισμούς για την διασφάλιση της ποιότητας των δεδομένων και τον «καθαρισμό» τους, διότι όσα περισσότερα «βρώμικα» δεδομένα περνάνε στο σύστημα για ανάλυση τόσο μεγαλύτερη η πιθανότητα να τεθεί σε κίνδυνο η διαδικασία της αναλυτικής μάρκετινγκ, οι καμπάνιες, οι άνθρωποι, ή άλλα συστήματα που τα χρησιμοποιούν. Οι περισσότερες επιχειρήσεις παρ’ όλα αυτά δεν έχουν τέτοια συστήματα «καθαρισμού» των δεδομένων, κάτι που οδηγεί σε ζητήματα έλλειψης αξιοπιστίας της αναλυτικής διαδικασίας.

Η αξιοπιστία των δεδομένων επηρεάζει άμεσα το κέρδος της επιχείρησης. Όσο παραπλανητικά είναι τα αναλυτικά στοιχεία τόσο λανθασμένες θα είναι και οι προβλέψεις για τον κύκλο αγοράς, οι καμπάνιες δεν θα έχουν τα επιθυμητά αποτελέσματα αφού η διαμόρφωσή τους θα έχει βασιστεί σε λανθασμένα στοιχεία για τους leads, με αποτέλεσμα την απώλεια leads και φυσικά, πωλήσεων.



Εικόνα 3.1: Διάγραμμα Ποσοστού Εμπιστοσύνης στην Διαδικασία Αναλυτικής Μαρκετινγκ

Η παραπάνω εικόνα δείχνει πόση εμπιστοσύνη υπάρχει από τις επιχειρήσεις στην διαδικασία αναλυτικής μάρκετινγκ. Δεδομένου του χαρακτήρα της διαδικασίας ανάλυσης και πως αυτή θα έπρεπε να λειτουργεί, ένας κρίσιμος παράγοντας της επιτυχίας της είναι η εμπιστοσύνη. Τα αποτελέσματα τις διαδικασίας μπορούν να επηρεάσουν την λήψη αποφάσεων μόνο εάν οι μέτοχοι και το διοικητικό συμβούλιο έχει εμπιστοσύνη στην διαδικασία και στα ευρήματα αυτής. Η αντιληπτή αξιοπιστία οποιωνδήποτε δεδομένων είναι ίσως ο πιο κρίσιμος παράγοντας για την επιτυχή χρήση αυτών των δεδομένων ως εργαλείο μέτρησης. Σύμφωνα με έρευνες, οι μικρές επιχειρήσεις αποδίδουν μεγαλύτερη αξιοπιστία στα δεδομένα ανάλυσης μάρκετινγκ από ότι οι μεγάλες (το 44% των μικρών επιχειρήσεων αξιολογούν την αξιοπιστία των δεομένων τους ως «καλή» ή «πολύ καλή», ενώ μόνο το 28% των μεγάλων επιχειρήσεων έχει την ίδια άποψη.) file:///C:/Users/User/Downloads/marketing_analytics_benchmark_report.pdf

3.3.2 Λήψη Αποφάσεων

Βάση όσων έχουν ειπωθεί, η διαδικασία αναλυτικής μάρκετινγκ πρέπει να παρέχει δεδομένα με σκοπό την διευκόλυνση της αποτελεσματικότερης λήψης αποφάσεων. Αυτό δεν σημαίνει ότι η απόφαση που θα παρθεί είναι απόλυτη και εξαρτάται αποκλειστικά από τα αποτελέσματα της ανάλυσης, αλλά είναι ένας συνδυασμός

πολλών παραγόντων από όλα τα τμήματα της επιχείρησης όπως έχει αναφερθεί και στην ενότητα 3.1 .

Για την καλύτερη κατανόηση του τρόπου με τον οποίο η διαδικασία της αναλυτικής μπορεί να παρέχει πληροφορίες και στη συνέχεια ελέγχει τα αποτελέσματα των αποφάσεων, ας εξετάσουμε μια υποθετική διαδικασία λήψης αποφάσεων από το τμήμα μάρκετινγκ.

Υποθετικό παράδειγμα λήψης αποφάσεων

Κατά την διάρκεια μιας συνάντησης της ομάδας μάρκετινγκ μιας επιχείρησης, ο επικεφαλής οικονομικός διευθυντής (Chief Financial Officer, CFO) αναφέρει ότι αναφορές του τελευταίου τριμήνου δείχνουν μια πτώση των εσόδων κατά 500.000 €. Ο γενικός διευθυντής (CEO) στρέφεται στον υπεύθυνο του τμήματος μάρκετινγκ (CMO) αναθέτοντάς του την εύρεση τρόπου για την κάλυψη αυτού του ελλείμματος. Μετά την συνάντηση με τους επικεφαλής, ο CMO μαζί με το τμήμα πωλήσεων, πρέπει να αποφασίσουν από που θα καλυφθεί αυτό το «κενό», δηλαδή από που θα προκύψουν έσοδα: από νέους πελάτες, από είδη υπάρχοντες ή από έναν συνδυασμό των δύο. Οι μετρήσεις που θα χρειαστούν για μια σοφή απόφαση είναι διαθέσιμες στα στελέχη, αφού η επιχείρηση εφαρμόζει συστήματα διαχείρισης πελατειακών σχέσεων CRM και συστήματα αυτοματοποίησης μάρκετινγκ. Οι μετρήσεις αυτές δείχνουν ότι,

- Ο μέσος κύκλος πωλήσεων για να δημιουργηθεί μια νέα προοπτική πωλήσεων (qualified prospect) είναι 6 μήνες.
- Το τμήμα πωλήσεων μετατρέπει το 30% των στοχευμένων προοπτικών πωλήσεων (qualified prospects) σε πωλήσεις (conversion rate),
- Έστω μια μέση πώληση 5.000 €, για να κλείσει το έλλειμμα των εσόδων (500.000 €) θα χρειαστούν 100 νέες πωλήσεις.

Δεδομένου του ποσοστού μετατροπής 30%, το τμήμα μάρκετινγκ πρέπει να εντοπίσει 333 νέες στοχευμένες προοπτικές πωλήσεων (qualified prospects) για τον στόχο των 500.000 ευρώ εσόδων (0,3 x 100 = 30 νέες πωλήσεις → για 100 νέες πωλήσεις χρειάζονται 100/0,3 = 333 qualifies prospects).

Το πρόβλημα που αντιμετωπίζει το τμήμα μάρκετινγκ είναι ότι δεν υπάρχει ο χρόνος στόχευσης 333 νέων δυνητικών πελατών (qualified leads) και η απόκτηση τουλάχιστον 100 από αυτών μέσα στο διάστημα του εξάμηνου κύκλου πωλήσεων. Παρ' όλα αυτά, συνεχίζονται οι προσπάθειες των δύο τμημάτων, μάρκετινγκ και πωλήσεων, για την επανεξέταση των ευκαιριών από τις εκστρατείες πωλήσεων (sales pipeline) που είναι σε ισχύ, με την βοήθεια των διαθέσιμων δεδομένων από το σύστημα CRM. Στην ανάλυση αυτών των δεδομένων αφαιρούνται οι leads που έχουν είδη ενσωματωθεί στο εξεταζόμενο διάστημα 6 μηνών. Τα αποτελέσματα τις ανάλυσης δείχνουν την ύπαρξη 285 leads στο sales pipeline, αλλά για διάφορους λόγους δεν έχουν πάρει την απαραίτητη προσοχή από το τμήμα πωλήσεων ή το μάρκετινγκ. Αυτοί οι 285 qualified leads αντιπροσωπεύουν την καλύτερη ευκαιρία απόκτησης των 500.000€ μέχρι το τέλος του εξαμήνου. Δεδομένου όμως του ποσοστού μετατροπής 30%, από τους 285 qualified leads η επιχείρηση θα κερδίσει 85 πωλήσεις ή 425.000€ (285 x 0,3 = 85 ή 85 x 5.000€ = 425.000€) και απομένει το ποσό των 75.000€ για την κάλυψη του συνολικού ελλείμματος (500.000€ - 425.000€ = 75.000€). Οι υπεύθυνοι των δύο τμημάτων αποφάσισαν να δράσουν με διαφορετική τακτική με σκοπό να αυξήσουν το ποσοστό μετατροπής, αφού ακολουθώντας τις ίδιες προσεγγίσεις θα οδηγηθούν στα ίδια αποτελέσματα.

Η ομάδα του τμήματος μάρκετινγκ εξετάζει τα εγχειρήματα που μπορούν να ακολουθήσουν. Οι βασικές εκστρατείες για την στόχευση των leads είναι τα emails που περιέχουν συνδέσμους προς τις σελίδες προορισμού που προσφέρουν κάποιο είδος ειδικού περιεχομένου. Η ομάδα μάρκετινγκ εξειδικεύεται στην δημιουργία ελκυστικού περιεχομένου, το οποίο δεν θα περάσει απαρατήρητο από τους leads. Οι σελίδες αυτές κατασκευάζονται σε σκοπό να μπορέσουν να συλλέξουν πληροφορίες για τους leads, διατηρώντας μια φόρμα που απαιτεί από τους επισκέπτες να καταγράψουν τα στοιχεία επικοινωνίας τους.

Είναι απαραίτητο να εξεταστούν τα αποτελέσματα προηγούμενων εκστρατειών emails με την χρήση των αναλυτικών στοιχείων από το CRM και των αυτόματων συστημάτων μάρκετινγκ, που παρέχουν πληροφορίες όπως: πόσοι, ποιοι και πόσες φορές άνοιξαν το email και για τα «κλικ» που πραγματοποιήθηκαν στον σύνδεσμο αντίστοιχα. Τα αποτελέσματα έδειξαν ότι τα ποσοστά των «κλικ» είναι χαμηλά σε

σχέση με αυτά του κλάδου, που σημαίνει ότι η ανταπόκριση στο προωθητικό email είναι ισχυρή αλλά λίγοι είναι οι ενδιαφερόμενη για την σελίδα προορισμού του email. Τα αναλυτικά στοιχεία επομένως δείχνουν που είναι το πρόβλημα.

Μετά τον εντοπισμό του προβλήματος στο αρχικό στάδιο της εκστρατείας, εξετάζονται τα αναλυτικά στοιχεία στο διαδίκτυο που αφορούν τις σελίδες προορισμού. Η ομάδα μάρκετινγκ εντόπισε πρόβλημα στην ανταπόκριση των leads που έχουν κάνει «κλικ» στην σελίδα προορισμού. Πολλοί λίγοι από αυτούς επιλέγουν να «πατήσουν» το «κουμπί παρότρυνσης», λεγόμενο ως “call-to-action button”, δηλαδή «Εγγραφή» ή «Κατέβασε το Άρθρο».

Μέχρι στιγμής, η επιχείρηση δεν είχε αντιμετωπίσει προβλήματα ελλείμματος εσόδων, οι μέθοδοι στόχευσης εξυπηρετούσαν τους στόχους του μάρκετινγκ και δεν είχαν χρειαστεί διαρθρωτικές αλλαγές στον τρόπο προσέγγισης πιθανών πελατών, με αποτέλεσμα τα αναλυτικά στοιχεία δεν είχαν την απαραίτητη προσοχή και εξέταση.

Το ερώτημα είναι ποια εναλλακτική προσέγγιση πρέπει να ακολουθήσουν τα τμήματα μάρκετινγκ και πωλήσεων με σκοπό την αύξηση του ποσοστού μετατροπής 30%. Το τμήμα μάρκετινγκ στοχεύει στην παραγωγή διαδραστικών βίντεο για την ενσωμάτωσή τους στις σελίδες προορισμού. Αυτά τα βίντεο έχουν διάρκεια το πολύ δύο λεπτών και στο τέλος της προβολής παρουσιάζονται στον θεατή κάποια call-to-action buttons. Τα αναλυτικά δεδομένα έδειξαν ότι ενσωματώνοντας διαδραστικά βίντεο στις σελίδες προορισμού τα «κλικ» στα call-to-action buttons, «Επικοινωνήστε μαζί μου», ήταν διπλάσια από πριν. Το τμήμα μάρκετινγκ αποφάσισε να δημιουργηθούν καμπάνιες με διαδραστικά βίντεο.

Η επόμενη σημαντική απόφαση για την νέα καμπάνια είναι στα χαρακτηριστικά αυτών των βίντεο (περιεχόμενο, διάρκεια κλπ). Έρευνες έχουν δείξει ότι τα πιο επιτυχημένα και αποτελεσματικά μάρκετινγκ βίντεο είναι αυτά που διαρκούν έως 90 δευτερόλεπτα. Για την εύρεση του καλύτερα στοχευμένου περιεχομένου, η ομάδα μάρκετινγκ εξετάζει και πάλι τα αναλυτικά δεδομένα παλαιότερων καμπανιών για να καταλήξουν σε αυτό που είχε την μεγαλύτερη απήχηση στους qualified prospects και εν τέλη ενσωματώθηκαν στον κύκλο πωλήσεων. Τα αποτελέσματα αυτών των αναλύσεων έδειξαν ότι το μεγαλύτερο ποσοστό μετατροπής είχε επιτευχθεί όταν το

περιεχόμενο της καμπάνιας διέθετε μια αληθινή ιστορία ενός πελάτη της επιχείρησης που ανέλυε τα προνόμια που απολάμβανε.

Βασιζόμενοι σε αυτά τα δεδομένα, η ομάδα μάρκετινγκ αποφάσισε να παράγει ένα διαδραστικό βίντεο 90 δευτερολέπτων το οποίο θα «αφηγείται» μια μελέτη περίπτωσης ενός πελάτη. Έχοντας τα δεδομένα για την καλύτερη στιγμή αποστολής του email, ημέρα και ώρα, η ομάδα μάρκετινγκ ξεκινάει την καμπάνια των emails με τις σελίδες προορισμού, οι οποίες πλέον εμπεριέχουν τα διαδραστικά βίντεο, για τους 285 leads. Τα αυτοματοποιημένα συστήματα μάρκετινγκ και το CRM είναι συνδεδεμένα και αντλούν στοιχεία για την διάρκεια παρακολούθησης των βίντεο και το ποσοστό που τα κοινοποιεί σε άλλους χρήστες. Επιπλέον, όταν ένας lead παρακολουθήσει ολόκληρο το βίντεο οι ομάδα των πωλήσεων ενημερώνεται με email έτσι ώστε να δράσουν ανάλογα σε πραγματικό χρόνο.

Η αποτελεσματικότητα της καμπάνιας φαίνεται από τις πρώτες μέρες δράσης της, έτσι ώστε αν κάτι δεν λειτουργεί σωστά η ομάδα μάρκετινγκ να προβεί σε εναλλακτικό σχέδιο. Στο υποθετικό παράδειγμα η νέα καμπάνια φαίνεται να λειτουργεί αποδοτικά, τα ποσοστά παρακολούθησης των βίντεο είναι υψηλά, αλλά οι μετρήσεις των εσόδων είναι ακόμη άγνωστες. Ο επικεφαλής του τμήματος μάρκετινγκ συναντά τον διευθυντή οικονομικών για ενημέρωση των εσόδων. Πολλοί ενδιαφερόμενοι που παρακολούθησαν το βίντεο με την αληθινή ιστορία ενός πελάτη, επικοινώνησαν με το τμήμα πωλήσεων για να δηλώσουν τις δικές τους εμπειρίες ή και προβλήματα. Σε διάρκεια λίγων εβδομάδων, οι προβλέψεις του τριμήνου είχαν ανοδική πορεία.

Όλοι συμφώνησαν με την άποψη ότι σε κάθε περίπτωση νέων ιδεών που στοχεύουν στην απόκτηση νέων πωλήσεων, ο διευθυντής οικονομικών δεν θα αρνηθεί μια επένδυση κεφαλαίων για την στήριξη αυτών των ιδεών, όπως ήταν στο παράδειγμα η παραγωγή διαδραστικών μάρκετινγκ βίντεο.

Όπως έδειξε το υποθετικό σενάριο, η διαδικασία της αναλυτικής μάρκετινγκ επηρεάζει σε μεγάλο βαθμό την λήψη αποφάσεων σε ένα σύγχρονο οργανισμό μάρκετινγκ με πολλούς τρόπους. Στο συγκεκριμένο παράδειγμα, τα αναλυτικά

στοιχεία χρησιμοποιήθηκαν σε αποφάσεις που συνδεόντουσαν απευθείας με τα έσοδα.

3.4 Διαδικασία της Αναλυτικής Μάρκετινγκ

Οι επιχειρήσεις, για να είναι βιώσιμες, πρέπει πάντα να ακολουθούν συγκεκριμένες διαδικασίες για την εκτέλεση των διαφόρων λειτουργιών τους. Διαδικασία πωλήσεων, εξυπηρέτησης πελατών, λογιστικής, παραγωγής, εκπαίδευσης προσωπικού και άλλες. Ακόμα και στην προσωπική μας ζωή, ακολουθούμε κάποιες συγκεκριμένες διαδικασίες για τις καθημερινές μας συνήθειες, για παράδειγμα η διαδικασία της εκτέλεσης ενός φαγητού – συνταγή, προμήθεια των υλικών, προετοιμασία επι μέρους τροφίμων και εκτέλεση συνταγής. Από τις πιο μικρές μέχρι τις πιο μεγάλες εργασίες, οι διαδικασίες είναι ο τρόπος εκτέλεσής τους.

Οι διαδικασίες βοηθούν στην εκτέλεση των εργασιών αποτελεσματικά και με τους σωστούς τρόπους (*efficiently and effectively*). Η διαδικασία που θα αποδειχθεί η πιο αποτελεσματική, εφαρμόζεται για την εκτέλεση της εργασίας και γίνονται συνεχείς βελτιώσεις στα επιμέρους στοιχεία που την απαρτίζουν. Στο τέλος κάθε διαδικασίας εκτελείται ο έλεγχος, κατά τον οποίο εντοπίζονται πιθανά αδύναμα σημεία για βελτίωση. *[Ν. Γεωργόπουλος – Σημειώσεις Μαθήματος «Στρατηγικό Μανατζμεντ», Πανεπιστήμιο Πειραιώς 2019]*

Το μάρκετινγκ είναι μία μάκρο διαδικασία (macro process) ενός συστήματος, που απαρτίζεται από πολλές υπό-διαδικασίες (subprocesses). Σε αυτές τις διαδικασίες δεν λείπει ποτέ και η δημιουργικότητα, η οποία είναι ένα απαραίτητο στοιχείο που δεν μπορεί να λείπει από την ομάδα μάρκετινγκ.

Αν υπήρχε ένα κτήριο σαν μέτρο αποτελεσματικότητας, οι διαδικασίες αναγκάζουν την διοίκηση να λειτουργεί στην κορυφή αυτού του κτηρίου. Πως όμως μετριέται η αποδοτικότητα και αποτελεσματικότητα των διαδικασιών; Σε αυτό το σημείο είναι απαραίτητες οι μετρήσεις - αναλυτική μάρκετινγκ. Όλες οι ενέργειες του μάρκετινγκ - email μάρκετινγκ, μάρκετινγκ περιεχομένου, αναγνώριση leads, μάρκετινγκ εκδηλώσεων, διαφήμιση pay-per-click – υπάρχουν οι ιδανικές μετρήσεις για κάθε μία από τις διαδικασίες μάρκετινγκ προκειμένου να γίνουν οι απαραίτητες αναλύσεις για

την αναγνώριση των σημαντικών μεταβλητών για την περεταίρω ανάλυσή τους και ενδεχομένως για βελτίωση αδυναμιών.

Η διαδικασία της αναλυτικής μάρκετινγκ είναι ένας συνεχής κύκλος μετρήσεων, αναλύσεων και βελτιώσεων. Τα βήματα αυτής της επαναλαμβανόμενης διαδικασίας αναλύονται παρακάτω.

3.4.1 Βήματα Διαδικασίας Αναλυτικής Μάρκετινγκ

ΒΗΜΑ 1 - Εντοπισμός Μετρήσεων.

Σε κάθε προσπάθεια δράσης μάρκετινγκ, οι μετρήσεις έχουν κυρίαρχο ρόλο αφού είναι ο δρόμος για να αποφανθεί η επιτυχία ή η αποτυχία. Φυσικά, όπως προαναφέρθηκε, κύριο συστατικό του μάρκετινγκ είναι η δημιουργικότητα και η φαντασία. Ωστόσο, η ομάδα μάρκετινγκ χρειάζεται αποτελέσματα, τα οποία μπορούν να φανούν μόνο με αριθμούς που θα μετρήσουν αν η δημιουργικότητα είχε το θεμιτό αποτέλεσμα ή η εκάστοτε καμπάνια χρειάζεται βελτιώσεις. Με την σειρά της η επιχείρηση αξιολογεί τις ενέργειες του τμήματος μάρκετινγκ βάση των κερδών της περιόδου που «έτρεχε» η καμπάνια. Τα κέρδη του οργανισμού είναι αποτέλεσμα των πωλήσεων, άρα μία από τις μετρήσεις που θα μπορούσαν να χρησιμοποιηθούν για το μάρκετινγκ είναι οι πωλήσεις, παρόλο που το μάρκετινγκ σχετίζεται με αυτές έμμεσα και όχι άμεσα.

Άλλες μετρήσεις, άμεσα συσχετισμένες με το μάρκετινγκ, που θα μπορούσε το τμήμα μάρκετινγκ να παρέχει στην διοίκηση είναι τα likes στα μέσα κοινωνικής δικτύωσης, τις κοινοποιήσεις, τα posts, τα tweets και άλλα. Το email μάρκετινγκ παρέχει μετρήσεις για το πόσοι άνοιξαν το email, πόσες φορές το διάβασαν, αν το κοινοποίησαν, αν έκαναν εγγραφή/αποσύνδεση και άλλα. Από τις ιστοσελίδες μπορούν να αντλήσουν δεδομένα για την επισκεψιμότητα, τις προτιμήσεις των επισκεπτών, τις πιο συχνές αναζητήσεις, σχετικές ιστοσελίδες και άλλα. Στο «μοντέρνο» μάρκετινγκ οι μετρήσεις που μπορούν να χρησιμοποιηθούν, αυξάνονται συνεχώς καθώς αυξάνονται και οι δυνατότητες των τεχνολογικών μέσων που χρησιμοποιούνται στις διαδικασίες του μάρκετινγκ.

Μια πρόκληση που αντιμετωπίζει το τμήμα μάρκετινγκ είναι η επιλογή αυτών των μετρήσεων. Ποιες μετρήσεις αξίζουν περεταίρω ανάλυση και ποιες είναι αυτές που δεν αποσκοπούν σε χρήσιμα στοιχεία. Ορισμένες μετρήσεις μπορεί να προβούν και στον αποπροσανατολισμό των διαδικασιών. Επίσης, μία άλλη δυσκολία που αντιμετωπίζουν οι marketers είναι στην επιλογή των μετρήσεων που θα παρουσιάσουν στον CEO. Για παράδειγμα, η πιο κοινή μέτρηση του μάρκετινγκ είναι αυτή των αντιδράσεων και των συναισθημάτων, δηλαδή στο πως αντιδρούν οι άνθρωποι – πελάτες ή πιθανοί πελάτες – στην καμπάνια. Παρουσιάζοντας αυτές τις μετρήσεις στον CEO, πιθανών να μην αντιληφθεί την σημασία τους, δεν γνωρίζουν όλα τα στελέχη τι επιρροή μπορεί να έχει μια θετική ή μια αρνητική αντίδραση ενός υποψήφιου πελάτη σε μια καμπάνια μάρκετινγκ. Οι CEO θέλουν να δουν απευθείας μια μέτρηση εκφρασμένη σε κέρδος ή ζημία. Οι marketers επομένως θα πρέπει να μεταφράσουν τα δεδομένα τους σε όρους κατανοητούς από την διοίκηση. Έχοντας ιστορικά στοιχεία πωλήσεων και αντιδράσεων, μπορούν να κάνουν πρόβλεψη των κερδών χρησιμοποιώντας τις τωρινές μετρήσεις αντιδράσεων. Το πρόβλημα σε αυτή την περίπτωση είναι ο βαθμός απόκλισης της εκτίμησης του τμήματος μάρκετινγκ με κίνδυνο να χάσουν την αξιοπιστία τους.

Είναι γνωστό ότι το μάρκετινγκ «χτίζει» την δύναμη της μάρκας. Μια πρόκληση που αντιμετωπίζουν οι επιχειρήσεις είναι η αναγνώριση δεδομένων που μπορούν να μετρήσουν την επωνυμία ως περιουσιακό στοιχείο (brand equity). Πολλές μάρκες χρεώνουν υψηλή τιμή λόγο της δύναμης που έχει το brand τους. Δυνατές μάρκες στην αγορά έχουν πολλοί περισσότερους «ακόλουθους». Ένας τρόπος μέτρησης αυτού του περιουσιακού στοιχείου είναι η αναγνώριση του μεριδίου αγοράς που απολαμβάνει η επιχείρηση σε συνδυασμό με την δύναμη της ανταγωνιστικότητας που απωθεί κινδύνους από άλλες μάρκες.

Οι περισσότερες μετρήσεις του μάρκετινγκ δεν σχετίζονται άμεσα με τα κέρδη και είναι δύσκολο να μετρηθούν αλλά είναι απαραίτητες για την αποτελεσματική λειτουργία του τμήματος μάρκετινγκ. μια κατεύθυνση για το πως θα μετρηθούν οι ενέργειες του μάρκετινγκ είναι η εξής:

επικέντρωση στους στόχους → μέτρηση αποτελεσματικότητας → μέτρηση αποδοτικότητας

Για την κατανόηση της παραπάνω κατεύθυνσης, ορίζουμε έναν πολύ κοινό στόχο μιας επιχείρησης, την αύξηση των κερδών. Το τμήμα μάρκετινγκ αναπτύσσει μια στρατηγική για την συμμετοχή του στην αύξηση των κερδών δημιουργώντας δύο στόχους:

1. Αύξηση της διείσδυσης αγοράς με τα ήδη υπάρχοντα προϊόντα στην υπάρχουσα αγορά,
2. Βελτίωση κινήσεων στην διατήρησης πελατών.

Αυτοί οι δύο στόχοι είναι εύκολα μετρήσιμοι για το τμήμα μάρκετινγκ και μπορούν να κατευθύνονται κάθε στιγμή στην σωστή κατεύθυνση διορθώνοντας στην πορεία τυχόν λάθη. Υπάρχει μια συμβιωτική σχέση ανάμεσα στην στρατηγική μάρκετινγκ και των στόχων της και στην αναλυτική μάρκετινγκ. Η στρατηγική πρέπει να προηγείται της αναλυτικής διαδικασίας, και οι στόχοι αυτής της στρατηγικής πρέπει να είναι η βάση για την αναγνώριση των μετρήσεων για την ανάλυση μάρκετινγκ.

ΒΗΜΑ 2 - Ανάλυση των μετρήσεων.

Το δεύτερο βήμα της διαδικασίας είναι η ανάλυση των μετρήσεων που εντοπίστηκαν στον βήμα 1. Βασιζόμενοι σε αυτές τις μετρήσεις το τμήμα μάρκετινγκ, με συγκεκριμένες τεχνικές ανάλυσης, αναλύει τα δεδομένα με σκοπό την ανίχνευση ενεργών πληροφοριών. Αυτές οι πληροφορίες μεταφέρουν γνώση σχετικά με την πρόοδο του μάρκετινγκ και την επίτευξη των στόχων του, ωστόσο οι πληροφορίες αυτές δεν είναι πάντα προφανείς. Έτσι με την ανάλυση και την σωστή ερμηνεία τους, το τμήμα μάρκετινγκ και ο CMO καταφέρνουν να «μεταφράσουν» τις ασαφής πληροφορίες σε χρήσιμη γνώση.

Στην σύγχρονη εποχή όπου το μάρκετινγκ βασίζεται κυρίως σε ψηφιακά συστήματα, οι μετρήσεις που προσδιορίζονται για την ανάλυση προέρχονται από τους ψηφιακούς μηχανισμούς του διαδικτύου. Επομένως για να γίνει σωστά η ανάλυση πρέπει οι μηχανισμοί αυτοί και τα συστήματα να είναι πλήρως αξιόπιστα για την αποτελεσματική εξαγωγή χρήσιμων πληροφοριών. Οι ιστότοποι είναι από τα πλέον σημαντικότερα κανάλια ψηφιακού μάρκετινγκ. Οι καταναλωτές προκειμένου να λάβουν τις αγοραστικές τους αποφάσεις, συμβουλεύονται τους ιστότοπους τον επιχειρήσεων, από τους οποίους οι οργανισμοί μάρκετινγκ λαμβάνουν τα δεδομένα

για την ανάλυσή τους. Οι ιστότοποι αυτοί πρέπει να έχουν ενεργοποιημένο το σύστημα αναλυτικής (analytic system) έτσι ώστε οι έμποροι να κατανοήσουν την απόδοση και να βελτιστοποιήσουν την ιστοσελίδα της επιχείρησής τους. Το Google Analytics είναι η πλέον δημοφιλέστερη, δωρεάν λύση που οι περισσότερες επιχειρήσεις χρησμιποιούν για αυτό το σκοπό.

Τρία βασικά συστήματα που χρησιμοποιούνται από τους οργανισμούς μάρκετινγκ για την καταγραφή των δεδομένων ανάλυσης είναι τα εξής: αναλυτική ιστού(web analytics), διαχείριση πελατειακών σχέσεων (CRM) και αυτοματοποίηση μάρκετινγκ (marketing automation). Φυσικά υπάρχουν και άλλα συστήματα που παρέχουν δεδομένα στο μάρκετινγκ, όπως δεδομένα συναλλαγών αγοράς, αλλά τα προαναφερόμενα είναι τα πιο βασικά. Οι marketers συλλέγουν τα δεδομένα για την ανάλυσή τους και δημιουργούν ταμπλό (dashboards) για την καλύτερη ερμηνεία και οπτικοποίηση των μετρήσεων.

Κατά την διαδικασία της ανάλυσης, το κυριότερο σημείο που πρέπει να βρεθούν οι marketers είναι η κατανόηση της τρέχουσας κατάστασης σε σύγκριση με την ιδανική βρίσκοντας τα σημεία απόκλισης των δύο αυτών καταστάσεων. Οι ιστορικές μετρήσεις βοηθούν στην σύγκριση αυτή και στο κατά πόσο αποτελεσματικές ήταν οι δράσεις του μάρκετινγκ. Για παράδειγμα σε ένα ερωτηματολόγιο ικανοποίησης των πελατών, με βαθμολογία 1 έως 5 (1 καθόλου ικανοποιημένος, 5 απόλυτα ικανοποιημένος) αν ο μέσος όρος είναι 2,5 οι marketers δεν μπορούν να καταλήξουν σε κάποιο συμπέρασμα της εκάστοτε καμπάνιας αν δεν έχουν τα αποτελέσματα της προηγούμενης περιόδου πριν την καμπάνια για να συγκρίνουν την ικανοποίηση των πελατών.

Η ανάλυση μάρκετινγκ είναι η βάση για την αξιολόγηση της απόδοσης του μάρκετινγκ. Τα δεδομένα που χρησιμοποιούνται στην ανάλυση είναι πολύ σημαντικό να είναι ακριβή και πλήρης έτσι ώστε η αξιολόγηση των δράσεων μάρκετινγκ να είναι βάσιμη.

ΒΗΜΑ 3 – Διορθωτικές Δράσεις.

Στο βήμα 2 – ανάλυση των μετρήσεων του μάρκετινγκ, η διαδικασία καταλήγει σε αποτελέσματα που βοηθούν το μάρκετινγκ να πάρει αποφάσεις, όπως έχει

προαναφερθεί, χωρίς όμως να δίνει λύσεις όταν τα αποτελέσματα αυτά δεν είναι τα επιθυμητά. Σε αυτό το σημείο της διαδικασίας αναλυτικής μάρκετινγκ, η ομάδα μάρκετινγκ πρέπει να εντοπίσει πιθανές δράσεις και αλλαγές ώστε να βελτιωθούν τα αποτελέσματα και να οδηγηθούν όσο το δυνατόν πλησιέστερα των επιθυμητών. Ορισμένες φορές οι πράξεις βελτίωσης είναι προφανείς, όπως για παράδειγμα σε μια καμπάνια emails αν έχουν καταχωρηθεί λανθασμένες διευθύνσεις παραληπτών, τα emails θα είναι «μπλοκαρισμένα», σε αυτό το πρόβλημα η λύση είναι απλή και προφανής, η ομάδα που διαχειρίζεται την καμπάνια θα πρέπει να διορθώσει τις διευθύνσεις email. Υπάρχουν όμως περιπτώσεις που οι λύσεις και οι δράσεις των προβλημάτων δεν είναι καθόλου σαφής.

Στις περιπτώσεις αυτές, είναι αναγκαίο να γίνονται κάποια τεστ από την ομάδα μάρκετινγκ όσον αφορά το περιεχόμενο της καμπάνιας. Τα χρώματα που έχουν χρησιμοποιήσει στο email, το στήσιμο του κειμένου, πόσο σαφές είναι το μήνυμα που θέλουν να περάσουν, αν ο σύνδεσμος είναι αρκετά ευδιάκριτος και μπορεί κάποιος γρήγορα και απλά να κάνει το «κλικ», και άλλα.

Από όλα τα βήματα της διαδικασίας αναλυτικής μάρκετινγκ, το βήμα των διορθωτικών δράσεων είναι το λιγότερο σχεδιασμένο – δρομολογημένο. Το εύρος των διορθωτικών αλλαγών έχει όρια με άξονα την δημιουργικότητα και την έμπνευση της ομάδας μάρκετινγκ. Οι marketers δεν απορρίπτουν ποτέ μια ιδέα και πάντα είναι ανοιχτοί να ακούσουν τις δημιουργικές σκέψεις όλων των συνεργατών της ομάδας τους. *[Jerry Rackley (2015): Marketing Analytics Roadmap Methods, Metrics, and Tools. Apress]*

3.5 Πίνακας Ελέγχου Μάρκετινγκ (Marketing Dashboards)

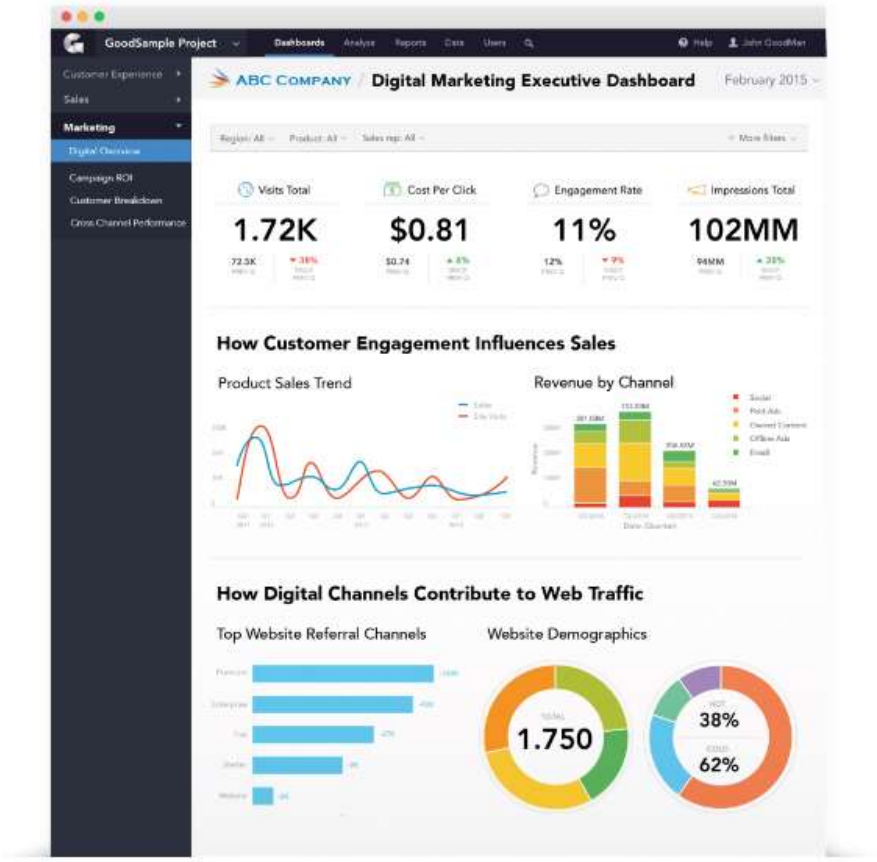
Σε όρους μάρκετινγκ οι πίνακες ελέγχου (dashboards) είναι εργαλεία διαχείρισης της επίδοσης που παρέχουν μια οπτική περιγραφή, περίληψη των *βασικών δεικτών απόδοσης* (*Key Performance Indicators* (*KPIs*)). Αυτά τα dashboars υπάρχουν σε διάφορες μορφές, από την απλούστερη που είναι αυτή των γραφημάτων που βοηθούν στην επεξήγηση μετρήσεων σε μια συνάντηση στελεχών, έως τους πίνακες που ενημερώνονται σε πραγματικό χρόνο. Όπως και στην οπτικοποίηση των δεδομένων που αναφέρεται στο κεφάλαιο 1, τα dashboards αποδίδουν τον μεγάλο

όγκο δεδομένων, που παράγονται από μια διαδικασία ανάλυσης μάρκετινγκ, σε ένα νόημα κατανοητό σε μια πρακτική εικονογραφημένη σύνοψη, όπως είναι τα διαγράμματα οι χάρτες θερμότητας οι γραφικές παραστάσεις και τα χρονοδιαγράμματα, παρέχοντας πληροφορίες στα άτομα του τμήματος μάρκετινγκ αλλά και στα στελέχη της επιχείρησης και σε άτομα μη έχοντας άμεση συμμετοχή στις ενέργειες του μάρκετινγκ.

Κάθε επιχείρηση επιλέγει το λογισμικό μάρκετινγκ που ταιριάζει καλύτερα στις ανάγκες της, επομένως υπάρχουν αρκετά dashboards που χρησιμοποιούνται ευρέως στον κόσμο των επιχειρήσεων. Κάποια παραδείγματα τέτοιων λογισμικών φαίνονται παρακάτω.

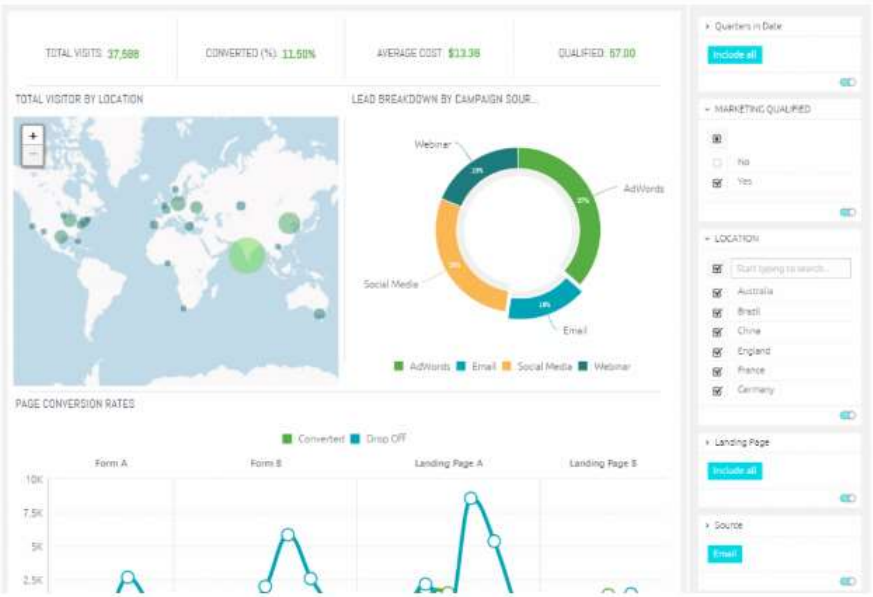
Εικόνα 3.2: Πίνακας Ελέγχου (Dashboard) Λογισμικού «Cyfe»

GoodData



Εικόνα 3.3: Πίνακας Ελέγχου (Dashboard) Λογισμικού «GoodData»

Sisense



Εικόνα 3.4: Πίνακας Ελέγχου (Dashboard) Λογισμικού «Sisense»

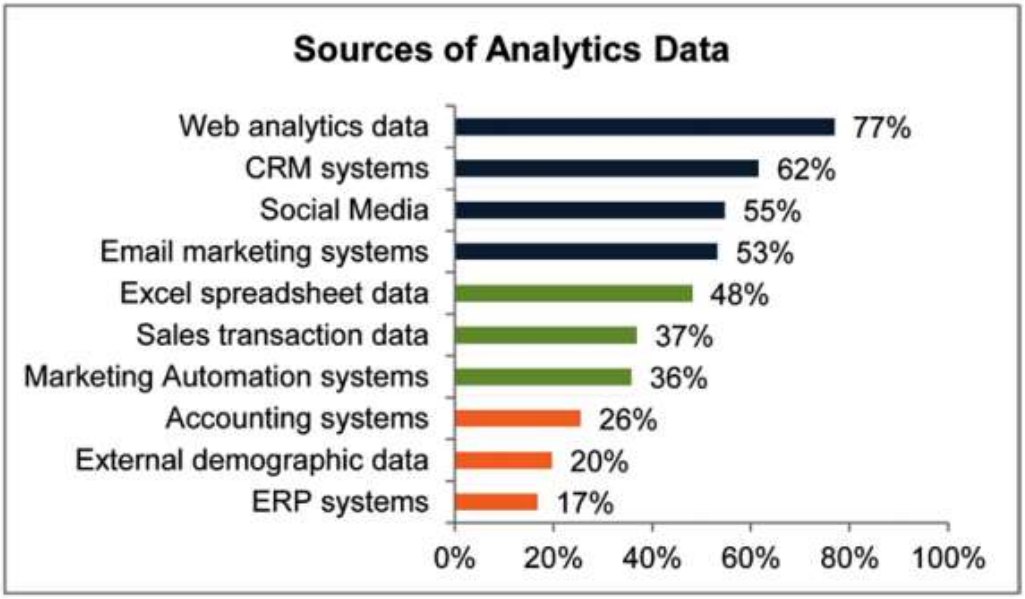
Tableau

Εικόνα 3.5: Πίνακας Ελέγχου (Dashboard) Λογισμικού «Tableau»

[Πηγή Εικόνων:https://technologyadvice.com/blog/marketing/6-best-marketing-dashboards-visualizing-performance/]

Όπως έχει προαναφερθεί, η μεγαλύτερη δυσκολία που αντιμετωπίζουν οι αναλυτές όταν εξετάζουν μεγάλα δεδομένα είναι το ποσοστό της αξιοπιστίας αυτών των δεδομένων. Έτσι και το τμήμα μάρκετινγκ αντιμετωπίζει το ίδιο πρόβλημα με τα δεδομένα που θα χρησιμοποιηθούν για την κατασκευή των dashboards.

Τα δεδομένα που τροφοδοτούν τα dashboards προέρχονται από διάφορες πηγές πραγματικού χρόνου καταγραφής δεδομένων (real-time data), όπως για παράδειγμα από τις ιστοσελίδες που επισκέπτονται οι χρήστες καθημερινά, τα μέσα κοινωνικής δικτύωσης, τα προγράμματα CRM και άλλα.



Εικόνα 3.6: Πηγές πραγματικού χρόνου καταγραφής δεδομένων

["Sales & Marketing Analytics: Gauging and Optimizing Performance, Demand Metric, December 2013, p. 13; http://www.demandmetric.com/content/sales-marketinganalytics-benchmark-report.]

3.6 Διαχείριση Πελατειακών Σχέσεων μέσω της Αναλυτικής Μάρκετινγκ (CRM)

Το μάρκετινγκ, σαν πολλές ενέργειες μαζί, αφορά κυρίως την κατανόηση της συμπεριφοράς των καταναλωτών όσων αφορά τις αγοραστικές τους συνήθειες. Οι

καταναλωτές αγοράζουν ένα προϊόν την στιγμή που το χρειάζονται σε μια τιμή που εκτιμούν οι ίδιοι ότι είναι διατεθειμένοι να διαθέσουν για το συγκεκριμένο προϊόν. Οι επιχειρήσεις ψάχνουν πιστούς πελάτες και ο βασικός τους σκοπός θα πρέπει να είναι η ικανοποίηση των αναγκών των πελατών τους και η ανάπτυξη των πελατειακών τους σχέσεων. Η αγορά είναι ένας χώρος αλληλεπίδρασης πελατών και επιχειρήσεων και το μάρκετινγκ είναι η διαδικασία που κινεί αγοραστές και πωλητές τον έναν προς τον άλλον.

Με το πέρασμα του χρόνου και καθώς η αγορά γίνεται όλο και περισσότερο ανταγωνιστική, το μάρκετινγκ εξελίσσεται σε ένα σύνολο ενεργειών και λειτουργιών με κύριο σκοπό την επιβίωση στην πλέον ανταγωνιστικότερη αγορά. Οι επιχειρήσεις προσπαθούν να πολεμήσουν τους ανταγωνιστές τους ξεχνώντας την σημαντικότερη ενέργεια του μάρκετινγκ, την εστίαση στην συμπεριφορά των καταναλωτών.

Η διαδικασία της αναλυτικής μάρκετινγκ, με τις στατιστικές μεθόδους και την οικονομετρική προσέγγιση, μπορεί να δημιουργήσει την καταλληλότερη στρατηγική για την εκάστοτε επιχείρηση. Με τις σωστές αναλυτικές τεχνικές , τα κατάλληλα μαθηματικά μοντέλα και τα αξιόπιστα δεδομένα, το τμήμα μάρκετινγκ μπορεί να προσδιορίσει, μέσα σε ένα ικανοποιητικό διάστημα εμπιστοσύνης και με την μικρότερη δυνατή απόκλιση, την αγοραστική συμπεριφορά των καταναλωτών με σκοπό την διαμόρφωση της πλέον αποτελεσματικότερης στρατηγικής μάρκετινγκ.

Παράδειγμα προσφοράς στα αλκοολούχα ποτά.

Ο SEO μεγάλης αλυσίδας σούπερ μάρκετ ανέθεσε στο τμήμα μάρκετινγκ να βρει μια στρατηγική προσφορών στο τμήμα των ποτών. Οι marketers γνωρίζοντας τις δυνατότητες που μπορεί να τους παρέχουν τα αναλυτικά στοιχεία της google αποφάσισαν να χρησιμοποιήσουν τα δεδομένα από τις τοποθεσίες των καταναλωτών με σκοπό την ανάλυση και τμηματοποίηση της αγοράς. Τα δεδομένα πολιτικής και απορρήτου στα smartphones, μετά την αποδοχή τους από τον χρήστη, επιτρέπουν στην google να αποθηκεύει δεδομένα για την τρέχουσα τοποθεσία του χρήστη. Έχοντας πρόσβαση στην τοποθεσία του συγκεκριμένου καταστήματος της αλυσίδας, υπάρχει δυνατότητα συλλογής δεδομένων για την επισκεψιμότητα των καταναλωτών σε αυτό το κατάστημα. Ο χρήστης για να μπορέσει να έχει πρόσβαση

στις εφαρμογές της google θα πρέπει να είναι συνδεδεμένος στον λογαριασμό του στην google έχοντας περάσει τα στοιχεία του, όπως το όνομά του, η ηλικία του, ο τόπος κατοικίας του κ.α. Οι αναλυτές συλλέγουν τα δεδομένα της επισκεψιμότητας στο κατάστημα για όλες τις μέρες τις εβδομάδας και στη συνέχεια ομαδοποιούν τους καταναλωτές μες βάση την ηλικία και το φύλο τους.

Τα αποτελέσματα της ανάλυσης επιτρέπουν στο τμήμα μάρκετινγκ να δημιουργήσουν προσφορές και προωθητικές ενέργειες για το συγκεκριμένο υποκατάστημα της αλυσίδας σούπερ μάρκετ. Μετά την τμηματοποίηση των καταναλωτών με βάση την ηλικία και το φύλο τους, το τμήμα μάρκετινγκ πήρε την ομάδα άνδρες ηλικίας 25 – 35 ετών και παρατήρησε από τα αποτελέσματα της ανάλυσης ότι η επισκεψιμότητα της συγκεκριμένης ομάδας είναι συγκεντρωμένη τις μέρες Πέμπτη Παρασκευή και Σάββατο τις ώρες 18:00 – 21:00. Το τμήμα μάρκετινγκ πρότεινε στον διευθυντή του καταστήματος προωθητικές ενέργειες για αυτές τις μέρες και ώρες της εβδομάδας, να βάλουν προσφορές στις μπύρες και στα πατατάκια ως συμπληρωματικό αγαθό της μπύρας.

3.7 Η Μαθηματική Πλευρά της Αναλυτικής Μάρκετινγκ

Για να καταστεί πιο εύκολη η κατανόηση της τεχνικής πλευράς της αναλυτικής μάρκετινγκ πρέπει πρώτα να γνωρίζουμε άρτια τις επιστήμες της Επιχειρησιακής Έρευνας (Operational Research). Έχοντας κάνει γνωστά τα παραπάνω, μπορούμε να πούμε ότι η αναλυτική μάρκετινγκ, με την βάση της πάντα στην επιχειρηματική ανάλυση, επιστρατεύει όλες τις τεχνικές της επιχειρησιακής έρευνας προσπαθώντας να τις εφαρμόσει σε τεράστιο όγκο δεδομένων. Είναι δηλαδή μια νέα πτυχή σε κάτι που οι άνθρωποι κάνουν εδώ και πολλά χρόνια.

Τα μαθηματικά μοντέλα που χρησιμοποιούνται στα πλαίσια της αναλυτικής μάρκετινγκ είναι πολυάριθμα και διαφορετικά για κάθε είδος προβλήματος. Κάποια από τα σημαντικότερα μοντέλα με συχνή χρήση είναι:

- Κλασσική γραμμική παλινδρόμηση και
- Πολλαπλή γραμμική παλινδρόμηση (για πρόβλεψη με συνεχή μεταβλητή απόκρισης)
- Λογιστική παλινδρόμηση (για πρόβλεψη με διακριτή μεταβλητή απόκρισης)

- Μη παραμετρική παλινδρόμηση (π.χ. τοπική πολυωνυμική παλινδρόμηση)
- Εκτίμηση μη-γραμμικών μοντέλων
- Δένδρα αποφάσεων (για αποφάσεις και προβλέψεις)
- Προσομοίωση
- Ανάλυση επιβίωσης (ανάλυση δεδομένων που αφορούν στον χρόνο που μεσολαβεί μέχρι κάποιο συγκεκριμένο συμβάν)
- Clustering (για ομαδοποίηση δεδομένων τα οποία έχουν κοινά χαρακτηριστικά)
- Ανάλυση χρονοσειρών
- Αλγόριθμοι βελτιστοποίηση

και πολλά άλλα.

Υπάρχουν πολλές γλώσσες προγραμματισμού αλλά και πακέτα τα οποία διευκολύνουν την εργασία στα πλαίσια της αναλυτικής μάρκετινγκ. Στην παρούσα διπλωματική εργασία θα χρησιμοποιήσουμε ένα από τα πλέον πιο γνωστά εργαλεία του τομέα, την γλώσσα προγραμματισμού R, η παρουσίαση της οποίας θα γίνει σε επόμενο κεφάλαιο όπως και μελέτη περίπτωσης ανάλυσης πραγματικών δεδομένων με χρήση κώδικα στην R.

ΚΕΦΑΛΑΙΟ 4

ΓΛΩΣΣΑ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΥ R

Στα πλαίσια της επιχειρηματικής ανάλυσης και της αναλυτικής μάρκετινγκ, ένα από τα σημαντικότερα και αποτελεσματικότερα εργαλεία επίλυσης προβλημάτων είναι η γλώσσα προγραμματισμού R. Είναι το πλέον ιδανικότερο εργαλείο για στατιστικούς υπολογισμούς και γραφήματα.

Η R δημιουργήθηκε από τους Ross Ihaka και Robert Gentleman στο πανεπιστήμιο University of Auckland της Νέας Ζηλανδίας, και αυτή τη στιγμή αναπτύσσεται από την ομάδα R Development Core Team, της οποίας ο Chambers, (δημιουργός της γλώσσας προγραμματισμού S), είναι μέλος. Δημιουργήθηκε το 1992, με την αρχική της έκδοση να κυκλοφορεί το 1995 και την σταθερή έκδοση beta να κυκλοφορεί το 2000. Βλέπουμε λοιπόν ότι η R σαν γλώσσα είναι αρκετά νέα σε σχέση με άλλες.

Η R παρέχει μία ποικιλία στατιστικών τεχνικών όπως είναι ο γραμμικός και ο μη γραμμικός προγραμματισμός, η ανάλυση χρονοσειρών, το clustering κ.ά. Επίσης εμπεριέχει πολλές τεχνικές για δημιουργία πολλών τύπων γραφημάτων ενώ οι βιβλιοθήκες της επεκτείνονται συνεχώς.

Στην σημερινή εποχή, καθώς μηχανές και συσκευές παράγουν όλο και περισσότερα δεδομένα, η δημοτικότητα αυτής της γλώσσας προγραμματισμού αναμένεται να αυξάνεται με επιταχυνόμενους ρυθμούς. Παρόλα αυτά η R εκτός από πλεονεκτήματα έχει και μειονεκτήματα τα οποία ένας προγραμματιστής ή χρήστης θα πρέπει να γνωρίζει πολύ καλά.

4.1 Πλεονεκτήματα και Μειονεκτήματα της R

4.1.1 Κύρια Πλεονεκτήματα της R

1. Η R αποτελεί γλώσσα προγραμματισμού

Τα εργαλεία που χαρακτηρίζονται ως «package» (πακέτο, πχ SPSS) δίνουν ένα περιορισμένο αριθμό από ξεχωριστές λειτουργίες που μπορούν να πραγματοποιήσουν, ενώ μια γλώσσα προγραμματισμού επιτρέπει στον χρήστη να

καθορίσει την εκτέλεση εργασιών και γενικά να αντιμετωπίσει το πρόβλημα χωρίς περιορισμούς με όποιον τρόπο επιλέξει.

2. Αναρίθμητες επεκτάσεις της R και η σημασία του ελεύθερου κώδικα.

Ένα από τα μεγάλα πλεονεκτήματα της R είναι το μεγάλο περιβάλλον από επιπρόσθετες επεκτάσεις που βοηθούν τον χρήστη να έχει μια πιο ολοκληρωμένη εμπειρία νέων τεχνικών. Έτσι είναι εύκολο να καταλάβουμε ότι η R είναι μια επεκτάσιμη γλώσσα και προσφέρει μεγάλη λειτουργικότητα στους προγραμματιστές έτσι ώστε να μπορούν εύκολα και γρήγορα να δημιουργήσουν τα δικά τους εργαλεία και να αναλύουν δεδομένα. Επίσης αποτελεί ελεύθερο λογισμικό (open source), όχι μόνο διανέμεται δωρεάν στον ίντερνετ αλλά και ο κύριος κώδικάς της R είναι ανοιχτός με αποτέλεσμα να μπορεί να επεκτείνεται από οποιονδήποτε χωρίς να χρειάζεται κάποια άδεια.

3. Εύκολος τρόπος σκέψης.

Ένα από τα επιτεύγματα της R είναι ότι αντικατοπτρίζει τον τρόπο σκέψης των ανθρώπων. Ένα άλλο χαρακτηριστικό είναι ότι δουλεύει με διανύσματα (στήλες) που σημαίνει ότι τα δεδομένα αντιμετωπίζονται ολικά και όχι ως μια συλλογή μεμονωμένων αριθμών, πράγμα πιο κοντινό στην λογική των ανθρώπων.

4. Οι επιχειρήσεις χρησιμοποιούν R.

Η R χρησιμοποιείται εκτεταμένα από πολλές από τις μεγαλύτερες εταιρείες στον κόσμο (Google, Facebook, Microsoft). Πέρα από τους κολοσσούς της τεχνολογίας, η R χρησιμοποιείται σε μεγάλη κλίμακα από ένα ευρύ φάσμα εταιρειών, Bank of America, Ford, TechCrunch, Uber, Trulia.

5. Οπτικοποίηση δεδομένων.

Το πακέτο plot και η επέκτασή ggplot2, είναι κάποια από τα καλύτερα εργαλεία οπτικοποίησης δεδομένων που μπορεί κάποιος να χρησιμοποιήσει στην R. Είναι εύκολα στην χρήση και το αποτέλεσμα της απεικόνισης των δεδομένων είναι κατανοητό όχι μόνο στους χρήστες αλλά και στους άμμεσα ενδιαφερόμενους άλλων τμημάτων της εταιρείας.

6. Η R δεν απευθύνεται μόνο σε προχωρημένους προγραμματιστές.

Είναι σχεδιασμένη για να αποτελέσει ένα χρήσιμο εργαλείο για ανθρώπους που έχουν να αντιμετωπίσουν προβλήματα που σχετίζονται με ανάλυση δεδομένων χωρίς να έχουν κάποιο υπόβαθρο στις επιστήμες προγραμματισμού.

4.1.2 Κύρια Μειονεκτήματα της R

1. Το μειονέκτημα της R σε σχέση με την διαχείριση της μνήμης.

Οι βασικές αρχές της R προέρχονται από γλώσσες προγραμματισμού που υπήρχαν από την δεκαετία του 1960 συνεπώς εμπεριέχει κατά κάποιο τρόπο παλαιό κώδικα λόγω του τρόπου που αρχικά σχεδιάστηκε. Αυτό έχει ως αποτέλεσμα κάποιες φορές να δημιουργούνται προβλήματα όταν υπάρχει τεράστιος όγκος δεδομένων.

2. Το πρόβλημα της R σε σχέση με την ασφάλεια.

Η R δεν εμπεριέχει κάποιο ενσωματωμένο κώδικα ασφαλείας με αποτέλεσμα οι λειτουργίες της να μην μπορούν να ανταπεξέλθουν στις απαιτήσεις των «web-like» ή «internet-like» εφαρμογών. Το θέμα της ασφάλειας ωστόσο έχει μειωθέι αρκετά με χρήση των «virtual x=containers» στην πλατφόρμα της Amazon Web Services cloud.

Τα παραπάνω είναι πιθανόν οι μεγαλύτερες προκλήσεις που αντιμετωπίζει η R σήμερα.

4.2 Παρουσίαση Αλγορίθμων – Γενικό Πρότυπο Γραμμικής Παλινδρόμησης

Τέτοιου είδους μοντέλου προβλέψεων (πολλαπλής παλινδρόμησης) όπως αυτού που θα παρουσιαστεί παρακάτω, έχει ευρεία επιστημονική αποδοχή διότι θεωρείται ισχυρό και ευέλικτο στατιστικό εργαλείο με πλήθος εφαρμογών σε διαφορετικά ερευνητικά πεδία (Draper & Smith, 1989, Pedhazur, 1997, Weisburg, 1985), όπως για παράδειγμα:

• Διοίκηση επιχειρήσεων και έρευνα αγοράς - εκτίμηση του βαθμού επίδοσης του προσωπικού μίας εταιρίας, διαχείριση του αριθμού και έκτασης των παραπόνων πελατών, προβλέψεις κ.ά.

• Προβλήματα οδικής συγκοινωνίας και logistics - καθορισμός του είδους μεταφορικού μέσου στο χρόνο εκπλήρωσης μίας μετακίνησης.

• Επιστημονική πρόβλεψη της συμπεριφοράς κάποιας εξαρτημένης μεταβλητής με την επίδραση κάποιων άλλων μεταβλητών (π.χ. το πως επιδρούν οι διαφημιστικές δαπάνες στις πωλήσεις των επιχειρήσεων). Θεωρητικά, όλα τα στατιστικά πακέτα περιέχουν μηχανισμούς εύκολους στον χειρισμό για την εκτίμηση των μοντέλων παλινδρόμησης. Δίνουν εκτιμήσεις των συντελεστών της παλινδρόμησης, των τυπικών σφαλμάτων, των στατιστικών στοιχείων που αφορούν την προσαρμοστικότητα του μοντέλου αλλά και προβλέψεις για νέες περιπτώσεις. Πάντως τις περισσότερες φορές στην R τα πράγματα είναι πιο εύκολα καθώς η όλη διαδικασία που περιγράφουμε μπορεί να γίνει με σχετικά μικρό κώδικα, με την προϋπόθεση όμως ότι τα δεδομένα μας στο dataframe είναι στην σωστή μορφή και με την προϋπόθεση ότι δόθηκε η απαραίτητη προσοχή για την αποφυγή απρόβλεπτων τιμών στο δείγμα (outliers), που μπορούν να επηρεάσουν το αποτέλεσμα. Δεύτερον, με την R έχουμε καλύτερο έλεγχο καθώς τα δεδομένα μας είναι σε dataframe το οποίο μπορούμε να το επεξεργαστούμε έτσι ώστε να το φτάσουμε στην επιθυμητή μορφή.

Υπάρχουν διάφορες μέθοδοι παλινδρόμησης, όπως είναι η κλασσική πολλαπλή παλινδρόμηση, (όπου η μεταβλητή απόκρισης είναι συνεχής), που θα παρουσιάσουμε παρακάτω, η λογιστική παλινδρόμηση (logistic regression) και η πολλαπλή λογιστική παλινδρόμηση. Στις δύο τελευταίες μεθόδους παλινδρόμησης, η μεταβλητή απόκρισης είναι διακριτή και δεν θα παρουσιαστούν στα πλαίσια αυτής της διπλωματικής εργασίας.

<u>Παρουσίαση Προτύπου Γραμμικής Παλινδρόμησης:</u>

Στο πρότυπο γραμμικής παλινδρόμησης πολλαπλών μεταβλητών, η απόκριση Y είναι μία συνεχής μεταβλητή μέτρησης, όπως για παράδειγμα είναι οι πωλήσεις ή το κέρδος. Το μοντέλο γραμμικής παλινδρόμησης με k ανεξάρτητες μεταβλητές, έχει την εξής μορφή:

$$Y = f\,(X1, X2, \ldots, Xp\,) + \varepsilon = a + b1X1 + b2X2 + \cdots + bkXk + \varepsilon$$

Για να διερευνηθεί η σχέση μεταξύ της Y και των $X1$,$X2$, … , Xk λαμβάνεται δείγμα μεγέθους n και για κάθε παρατήρηση του δείγματος καταγράφονται οι τιμές των συγκεκριμένων μεταβλητών. Π.χ. για κάθε i-παρατήρηση του δείγματος καταγράφονται οι τιμές (yi , $xi1$, $xi2$, … , xi,k) για i = 1, 2, … , n. Έτσι λοιπόν έχουμε $yi = fi\,(X1, X2, \ldots, Xk\,) + \varepsilon i = a + b1xi1 + b2xi2 + \cdots + bkxik + \varepsilon i$ όπου τα «σφάλματα» $\varepsilon 1$, $\varepsilon 2$, … , εn θεωρούνται ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν κανονική κατανομή, ενώ οι μεταβλητές παλινδρόμησης $X1$, $X2$, … , Xk δεν θεωρούνται τυχαίες.

4.3 Εκτίμηση Συντελεστών & Οπτικοποίηση Αποτελεσμάτων

Σκοπός της εκτίμησης των παραμέτρων της γραμμικής παλινδρόμησης είναι η διαδικασία πρόβλεψης, πχ πωλήσεων, σε μια εταιρία για το επόμενο χρονικό διάστημα. Έχοντα τα αποτελέσματα της παλινδρόμησης και τους εκτιμητές, μπορούν να προβούν σε περεταίρω ανάλυση των επιμέρους στοιχείων που τους ενδιαφέρουν, βάση του μοντέλου που έχει κατασκευαστεί.

Στην παρούσα εργασία δεν θα ασχοληθούμε εκτεταμένα με τις εκτιμήσεις των συντελεστών, όσο με την οπτικοποίηση των δεδομένων. Στο επόμενο κεφάλαιο θα δούμε αναλυτικά μια μελέτη περίπτωσης και πως, με την βοήθεια της R, γίνεται η οπτικοποίηση των αποτελεσμάτων για την εύκολη κατανόηση τους.

Τα γραφήματα που δημιουργούνται είναι κατανοητά από όλους τους ενδιαφερόμενους μιας επιχείρησης χωρίς να πρέπει να γνωρίζουν σε βάθος στατιστική ανάλυση. Σε ένα συμβούλιο η παρουσίαση γίνεται αχάριστη και εύκολη έχοντας τέτοια γραφήματα εξηγώντας στους CEO πως θα μπορούσαν να πάρουν μια απόφαση έχοντας συμβουλές από τα συγκεκριμένα γραφήματα.

ΚΕΦΑΛΑΙΟ 5

ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗΣ (case study) – ΤΑΙΝΙΕΣ ΤΟΥ HOLLYWOOD

5.1 Μελέτη περίπτωσης μεγάλων δεδομένων από ταινίες του Hollywood

[Πηγή δεδομένων: https://data.world/data-society/imdb-5000-movie-dataset

https://www.superdatascience.com/pages/rcourse]

Στην παρούσα μελέτη παρουσιάζεται μια ερευνητική προσέγγιση που χρησιμοποιεί εξόρυξη δεδομένων για την πρόβλεψη διάφορων μετρήσεων ταινιών των τελευταίων χρόνων από το imdb.

Η βιομηχανία της παραγωγής ταινιών βρίσκει μια τέτοια μελέτη πολύ χρήσιμη για την λήψη αποφάσεων που σχετίζονται με την προώθηση ταινιών, την επιλογή δημοφιλέστερης κατηγορίας στο κοινό, την πρόβλεψη του προϋπολογισμού για την παραγωγή μιας συγκεκριμένης ταινίας βάση των ιστορικών δεδομένων που υπάρχουν στην βάση δεδομένων που θα αναλυθεί, και άλλες πολλές σημαντικές πληροφορίες που θα μπορέσουν να αντλήσουν από μια τέτοιου είδους ανάλυση.

5.2 Δεδομένα & Εξόρυξη

Για την διεξαγωγή της έρευνας συμπεριλήφθηκαν δεδομένα 608 ταινιών 70 χρόνων (1939 – 2015) . Οι μεταβλητές που θα χρησιμοποιηθούν είναι οι εξής:

Δεδομένα	Τύπος Δεδομένων
Ημέρα κυκλοφορίας (day)	(κατηγ. μεταβλητή: «Δευτέρα», «Τρίτη»,… «Κυριακή»)
Σκηνοθέτης (director)	(κατηγ. μεταβλητή: «Brad Bird», «Scott Waugh»,..)
Είδος (genre)	(κατηγ. μεταβλητή: «Κωμωδία», «Δράμα», «Κοινωνική»,..)
Τίτλος (title)	(κατηγ. μεταβλητή: «Tomorrowland», «Need for Speed»,..)
Studio	(κατηγ. μεταβλητή: «Buena Vista Studios», « Lionsgate»,..)
Προϋπολογισμός σε εκατομ. $ (budget)	(αριθμ. μεταβλητή)
Έσοδα σε εκατομ. $ (gross)	(αριθμ. μεταβλητή)
Κρητική στο imdb	(αριθμ. μεταβλητή)
Κέρδος σε εκατομ. $ (profit)	(αριθμ. μεταβλητή)
Διάρκεια σε λεπτά (runtime)	(αριθμ. μεταβλητή)

Ο κυριότερος σκοπός μιας εταιρίας παραγωγής, προκειμένου να παραμένει στην αγορά, είναι να στοχεύει σε όσο το δυνατό υψηλότερες πωλήσεις σε συνδυασμό με την αποτελεσματικότερη λειτουργία των δραστηριοτήτων της με το χαμηλότερο κόστος.

Ως εξαρτημένη μεταβλητή στο μοντέλο θα χρησιμοποιήσουμε κυρίων τα κέρδη έχοντας όλες τις υπόλοιπες μεταβλητές ως επεξηγηματικές του μοντέλου. Έχοντας υπόψιν την γραμμική σχέση κέρδη = έσοδα – προϋπολογισμός, θα χειριστούμε κατάλληλα το μοντέλο για την αποφυγή πολυσυγγραμμικότητας.

5.3 Εισαγωγή δεδομένων στην R

Εισάγουμε τα δεδομένα στο Rstudio κάνοντας τους απαραίτητους χειρισμούς έτσι ώστε να είναι κατάλληλα για την ανάλυση.

```
movies<-read.csv(file.choose())
head(movies)
movies<-data.frame(movies$Day.of.Week,movies$Director,movies$Genre,movies$Movie.Title,movies$Studio,
                   movies$Budget...mill.,movies$Gross...mill.,movies$IMDb.Rating,movies$Profit...mi
                   movies$Profit.,movies$Runtime..min.)

colnames(movies)<-c("day", "director", "genre","title", "studio", "budget.mill","gross.mill",
                    "IMDBrating","profit.mill","profit","runtime")
movies$gross.mill<-movies$budget.mill*(movies$profit/100+1)
movies$profit.mill<-movies$gross.mill-movies$budget.mill
movies$profit<-NULL

str(movies)
```

```
> str(movies)
'data.frame':   608 obs. of  10 variables:
 $ day         : chr  "Friday" "Friday" "Friday" "Friday" ...
 $ director    : chr  "Brad Bird" "Scott Waugh" "Patrick Hughes" "Phil Lord, Chris Miller" ...
 $ genre       : chr  "action" "action" "action" "comedy" ...
 $ title       : chr  "Tomorrowland" "Need for Speed" "The Expendables 3" "21 Jump Street" ...
 $ studio      : chr  "Buena Vista Studios" "Buena Vista Studios" "Lionsgate" "Sony" ...
 $ budget.mill: num  170 66 100 42 150 80 50 85 70 5 ...
 $ gross.mill : num  202 203 206 202 205 ...
 $ IMDBrating : num  6.7 6.6 6.1 7.2 8 5.8 6 6.8 6.3 5.9 ...
 $ profit.mill: num  32.1 137.3 106.2 159.6 55.3 ...
 $ runtime     : int  130 132 126 109 131 134 125 115 92 84 ...
```

Μετατρέπουμε τις κατηγορικές μεταβλητές σε “factors” για να μπορούμε να τις χρησιμοποιήσουμε στην ανάλυση των γραφημάτων. Στην R οι μεταβλητές «factors» έχουν «levels» δηλαδή τις ομαδοποιεί ανάλογα με το περιεχόμενό τους. Για παράδειγμα η μεταβλητή «Μέρα Κυκλοφορίας» έχει μετατραπεί σε factor με 7 levels.

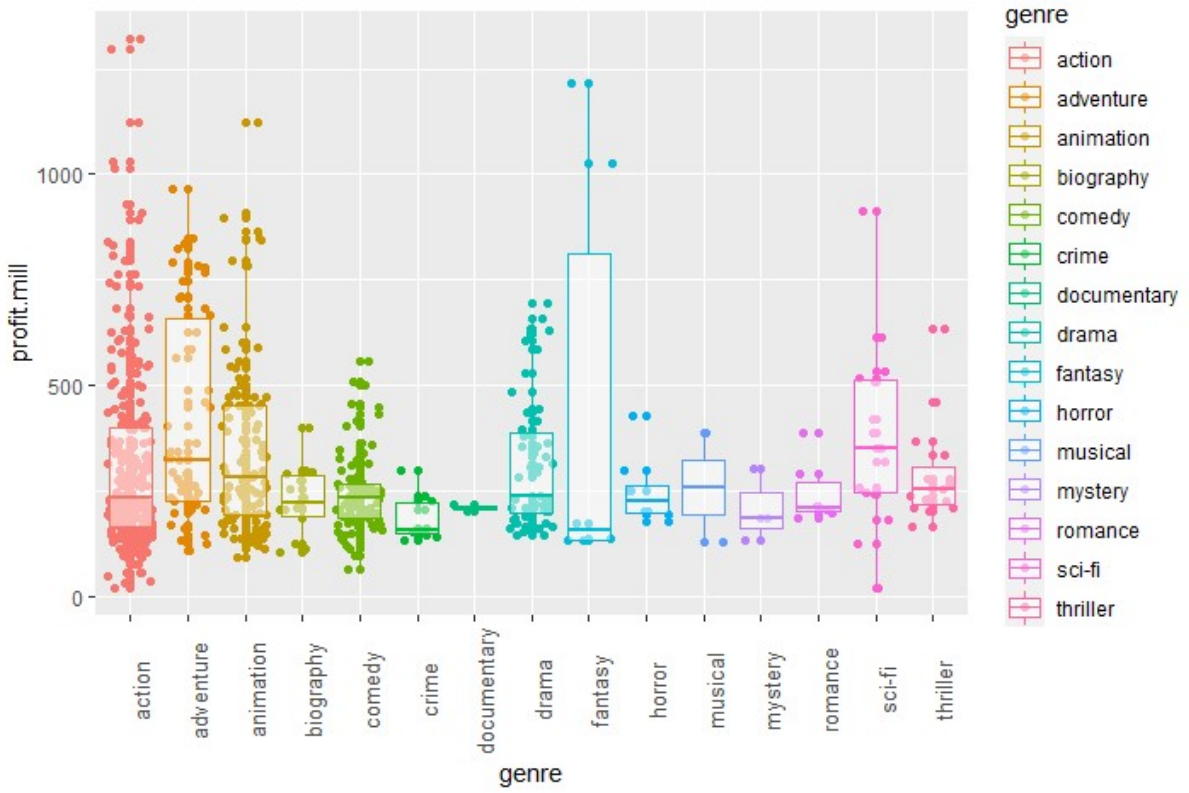
```
movies$day<-factor(movies$day)
movies$director<-factor(movies$director)
movies$genre<-factor(movies$genre)
movies$studio<-factor(movies$studio)



str(movies)
```

```
data.frame':   608 obs. of  10 variables:
$ day         : Factor w/ 6 levels "Friday","Saturday",..: 1 1 1 1 1 1 4 1 1 1 ...
$ director    : Factor w/ 337 levels "Aaron Blaise, Robert A. Walker",..: 31 297 233 256 287 76 276 71
108 126 ...
$ genre       : Factor w/ 15 levels "action","adventure",..: 1 1 1 5 1 1 2 1 1 10 ...
$ title       : chr  "Tomorrowland" "Need for Speed" "The Expendables 3" "21 Jump Street" ...
$ studio      : Factor w/ 36 levels "Art House Studios",..: 2 2 11 25 25 25 2 31 31 20 ...
$ budget.mill: num  170 66 100 42 150 80 50 85 70 5 ...
$ gross.mill : num  202 203 206 202 205 ...
$ IMDBrating : num  6.7 6.6 6.1 7.2 8 5.8 6 6.8 6.3 5.9 ...
$ profit.mill: num  32.1 137.3 106.2 159.6 55.3 ...
$ runtime     : int  130 132 126 109 131 134 125 115 92 84 ...
```

5.4 Οπτικοποίηση Αποτελεσμάτων

Για την ανάλυση των δεδομένων των ταινιών, ζητήθηκε από τον αναλυτή να παρουσιάσει με γραφήματα τα στοιχεία των ταινιών και να προτείνει στο διοικητικό συμβούλιο σε ποια είδη ταινιών να επενδύσουν και με ποια studio να συνεργαστούν.

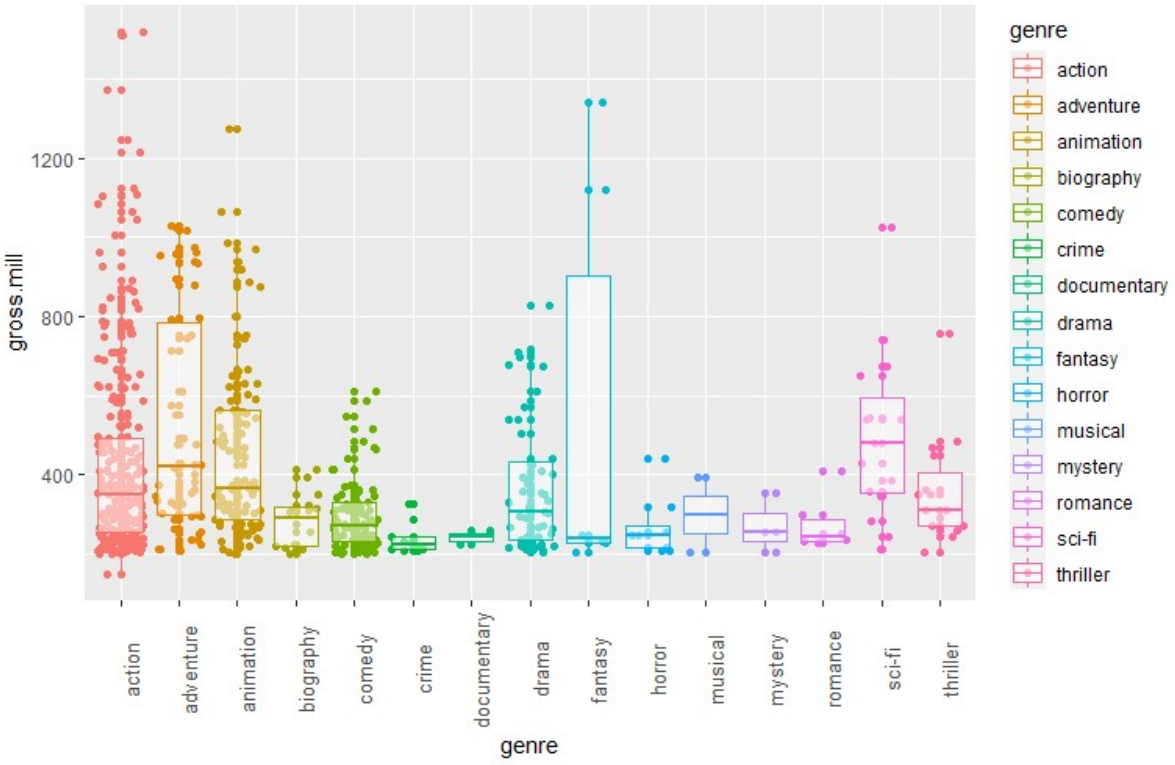
Γράφημα 1: Έσοδα από πωλήσεις εισιτηρίων ανάλογα το είδος της ταινίας.



Εξετάζοντας το παραπάνω boxplot μπορούμε να πούμε ως συμπέρασμα ότι στην κατηγορία «action» υπήρξαν οι περισσότερες δημιουργίες ταινιών. Ωστόσο το μεγαλύτερο κατά μέσο όρο ποσό εσόδων από την πώληση των εισιτηρίων ήταν για την κατηγορία ταινιών «sci-fi» και «adventure» με αρκετά μικρότερο πλήθος τέτοιων ταινιών. Η κατηγορία όμως «action» έχει το μεγαλύτερο εύρος εσόδων από οποιαδήποτε άλλη.

Ένας άλλος τρόπος απεικόνισης είναι να χρησιμοποιήσουμε στον άξονα y αντί τα έσοδα, τις πωλήσεις των εισιτηρίων ανεξαρτήτως προϋπολογισμού, δηλαδή τα κέρδη.

Γράφημα 2: Πωλήσεις εισιτηρίων ανάλογα το είδος της ταινίας.



Στην σύγκριση μεταξύ των δυο γραφημάτων μπορούμε να πούμε ότι δεν υπάρχει κάποια ουσιαστική διαφορά. Δηλαδή ο προϋπολογισμός των ταινιών είναι ανάλογος των πωλήσεων των εισιτηρίων για αυτό το λόγο το καθαρό έσοδο δεν διαφέρει ουσιαστικά από το κέρδος.

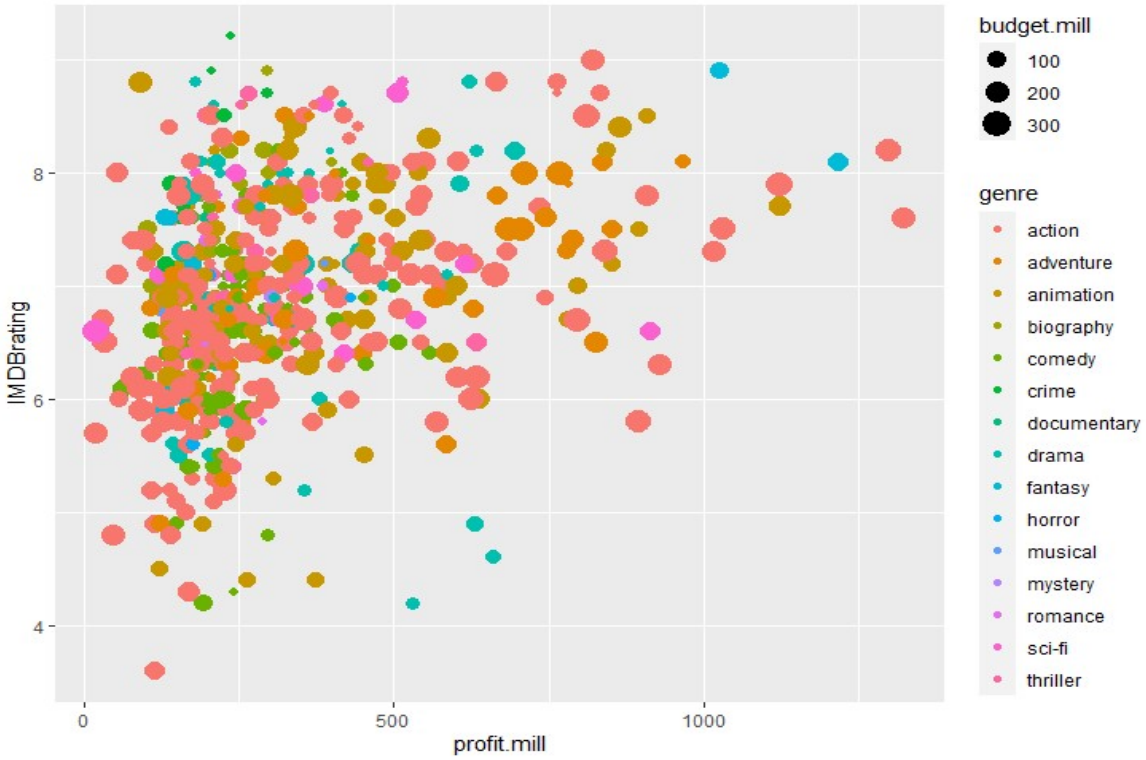
Σε επόμενα γραφήματα θα αναλύσουμε περεταίρω και με άλλες μεταβλητές, την συμπεριφορά αυτών των ταινιών στα έσοδα μιας εταιρίας παραγωγής.

Ο κώδικας που χρησιμοποιήθηκε για την διεξαγωγή του γραφήματος 1 (και 2 αντίστοιχα με την αλλαγή του άξονα Υ):

```
s<-ggplot(data=movies,aes(x=genre,y=profit.mill,colour=genre
))
s+geom_point()+geom_jitter()+geom_boxplot(alpha=0.5)+ theme(axis.text.x = element_text(
  size=10, angle=90))
```

Γράφημα 3: Έσοδα σχετικά με τις κριτικές στο IMDB

Στο παρακάτω γράφημα θα δείξουμε την σχέση ανάμεσα στις κριτικές των ταινιών στο IMDB και στα έσοδα από τις πωλήσεις.



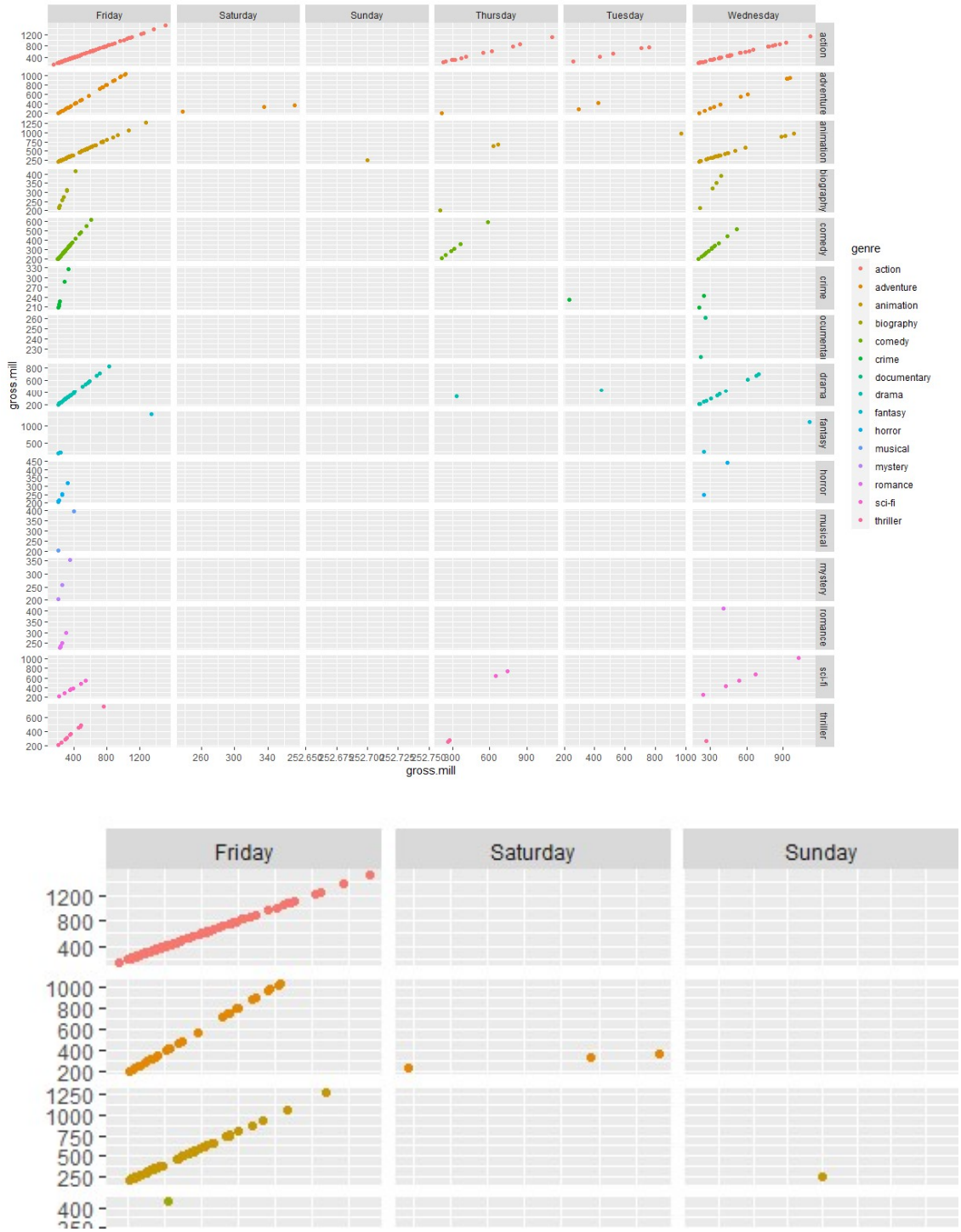
Επιπλέον το γράφημα δίνει πληροφορίες και για τον προϋπολογισμό των ταινιών, ο οποίος είναι ανάλογος με το μέγεθος των σημείων στο γράφημα σε 3 κλίμακες [0-100, 101-200, 201-300] σε εκατομμύρια δολάρια.

Μπορούμε να πούμε ότι φαίνεται μια σχετικά θετική σχέση στις δυο μεταβλητές, αν μια ταινία βαθμολογηθεί υψηλά στο IMDB θα έχει και σχετικά υψηλές πωλήσεις. Χωρίς όμως αυτό να είναι αναγκαίο, διακρίνοντας και υψηλές κριτικές στο γράφημα χωρίς να έχουν και υψηλές πωλήσεις αντίστοιχα.

Ο κώδικας που χρησιμοποιήθηκε για την διεξαγωγή του γραφήματος:

```
s<-ggplot(data=movies, aes(x=profit.mill,y=IMDBrating,colour=genre, size=budget.mill
                           ))
s+geom_point()
```

Γράφημα 4: Πωλήσεις εισιτηρίων – Μέρες προβολής.

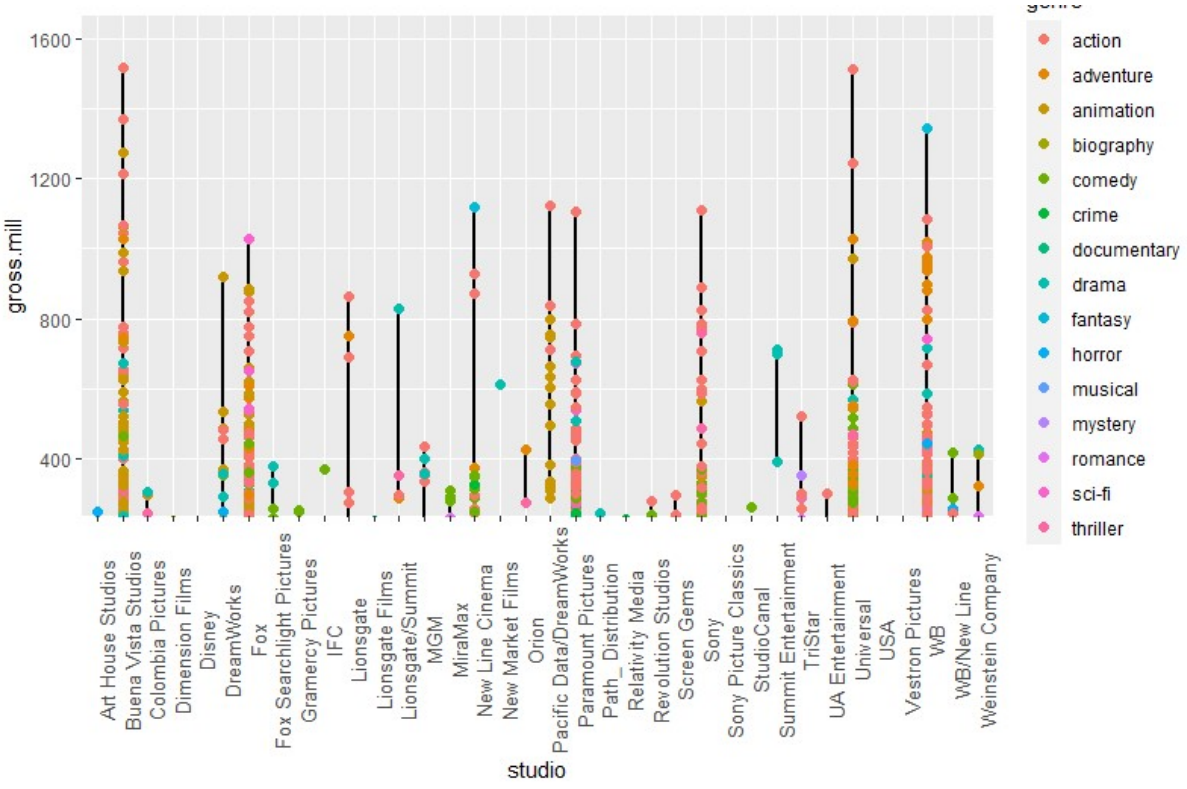


Παραπάνω δίνονται πληροφορίες σχετικά με τις πωλήσεις των εισιτηρίων ανά ημέρα προβολής των διαφόρων ειδών ταινιών. Για παράδειγμα αν η εταιρία παραγωγής αποφασίσει να δημιουργήσει μια ταινία δράσης η καλύτερη ημέρα πρώτης προβολής θα είναι η Παρασκευή.

Ο κώδικας που χρησιμοποιήθηκε για την διεξαγωγή του γραφήματος:

```
#------studio
s<-ggplot(data=movies, aes(x=studio,y=gross.mill, size=budget))

s+geom_line(size=1)+geom_point(aes(colour=genre),size=2)+
  coord_cartesian(ylim=c(300,1600))+
  theme(axis.text.x = element_text(
    size=10, angle=90))
```

Γράφημα 5: Επιλογή studio για συνεργασία.

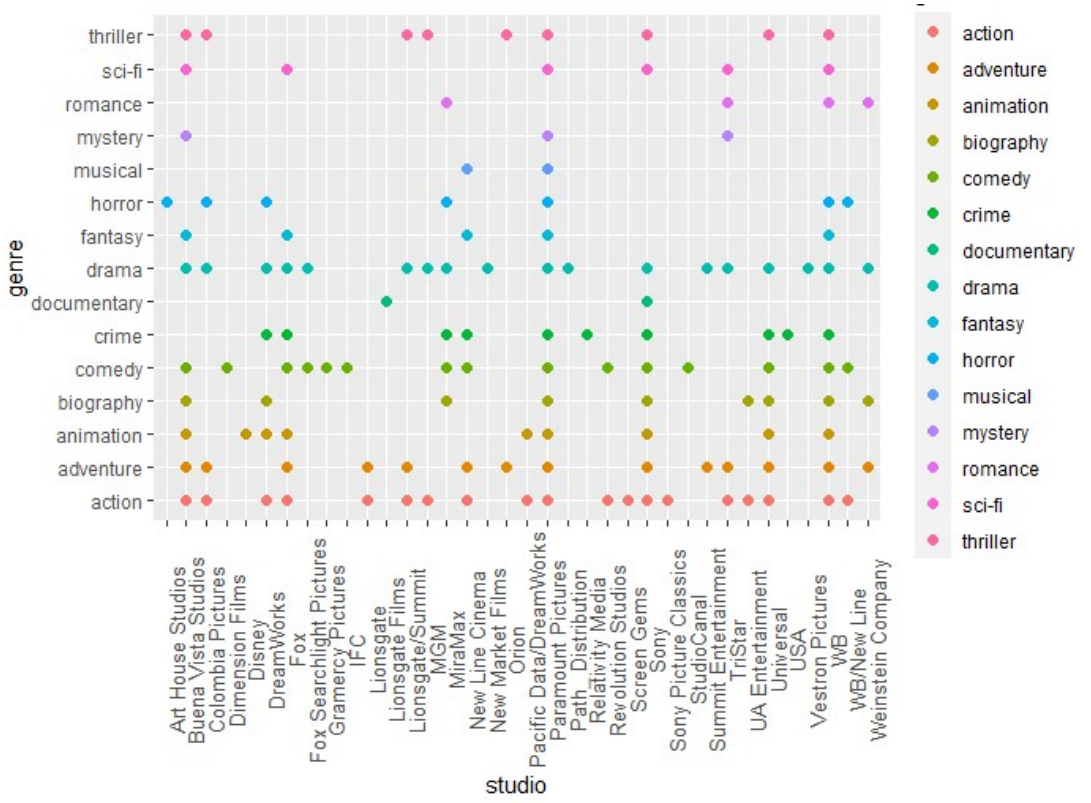


Στο συγκεκριμένο γράφημα παρουσιάζεται η αναλογία πωλήσεων – studio που αναλαμβάνουν την εκάστοτε ταινία. Διακρίνουμε τις εξής πληροφορίες:

1. Το πιο «κερδοφόρο» studio.
2. Για κάθε studio, τι είδος ταινιών αναλαμβάνει κυρίως.
3. Έχοντας επιλέξει το είδος που θα δημιουργηθεί, μπορούμε να βρούμε ποιο studio μας συμφέρει περισσότερο ώστε να την αναλάβει.

Για την καλύτερη κατανόηση των παραπάνω κατασκευάστηκε το βοηθητικό γράφημα για την επιλογή studio βάσει του είδους.

Βοηθητικό Γράφημα: Είδη ταινιών που υποστηρίζει κάθε studio.



Παράδειγμα: Η εταιρία παραγωγής αποφάσισε να κατασκευάσει μια ταινία «mystery». Από το βοηθητικό γράφημα βλέπουμε ότι τέτοιου είδους ταινίες αναλαμβάνουν τα studio «Buena Vista Studios», «Paramount Pictures» και «TriSTar». Από το γράφημα 5 γίνεται η διαλογή του πιο κερδοφόρου από αυτά τα τρία studio, δηλαδή το «Buena Vista Studios».

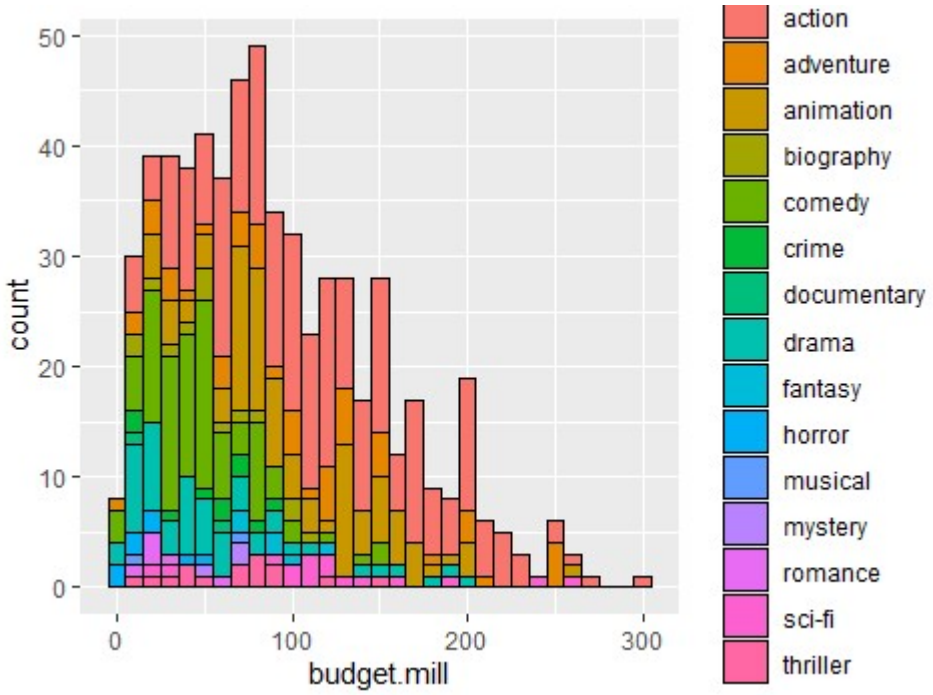
Ο κώδικας που χρησιμοποιήθηκε για την διεξαγωγή του γραφήματος 5:

```
s<-ggplot(data=movies, aes(x=studio,y=gross.mill))

s+geom_line(size=1)+geom_point(aes(colour=genre),size=2)+
  coord_cartesian(ylim=c(300,1600))+
  theme(axis.text.x = element_text(
    size=10, angle=90))
```

Ο κώδικας που χρησιμοποιήθηκε για την διεξαγωγή του βοηθητικού γραφήματος:

```
s<-ggplot(data=movies, aes(x=studio,y=genre, colour=genre))

s+geom_point(aes(colour=genre),size=2)+
  theme(axis.text.x = element_text(
    size=10, angle=90))
```

Γράφημα 6: Ιστόγραμμα Προϋπολογισμού

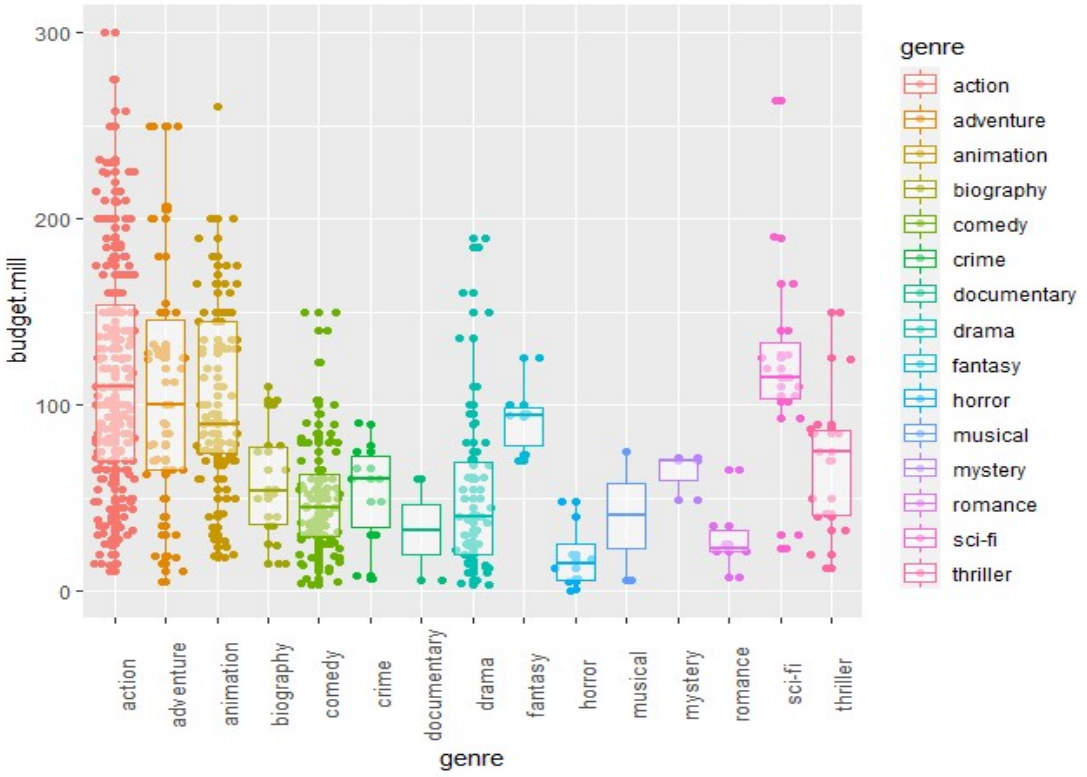


Στο τρίτο γράφημα παραπάνω φαίνεται το ιστόγραμμα του μεγέθους του προϋπολογισμού μαζί με το είδος των ταινιών. Το μεγαλύτερο ποσοστό των εξόδων κυμαίνεται μέχρι και 100 εκατομμύρια δολάρια. Επίσης από τα χρώματα διακρίνουμε το «budget» για κάθε είδος ταινίας. Αναμφίβολα βλέπουμε ότι οι περισσότερες ταινίες στα συγκεκριμένα δεδομένα είναι δράσης.

Ο κώδικας που χρησιμοποιήθηκε για την διεξαγωγή του γραφήματος 6:

```
library(ggplot2)
s<-ggplot(data=movies, aes(x=budget.mill))
s+geom_histogram(binwidth = 10, aes(fill=genre), colour="Black")
```

Γράφημα 7: Προϋπολογισμός ανάλογα με το είδος.

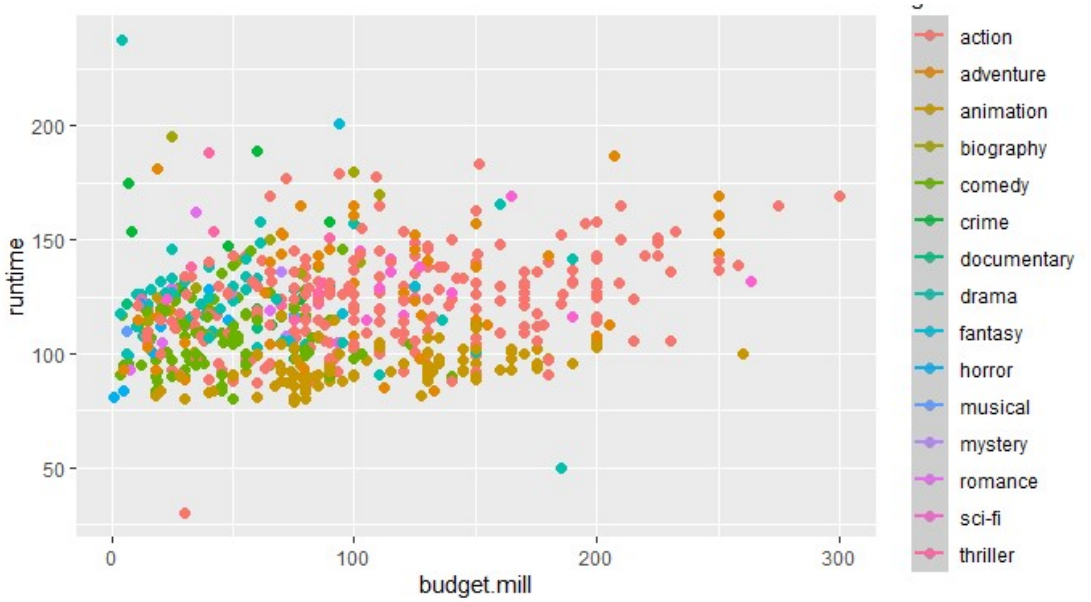


Τα υψηλότερα έξοδα κατά μέσο όρο για την διεξαγωγή ταινίας χρειάστηκαν για το είδος «sci-fi» και μετά ακολουθούν οι ταινίες «action», «adventure» και «fantasy». Ωστόσο το μεγαλύτερο εύρος εξόδων είναι στις ταινίες «action».

Ο κώδικας που χρησιμοποιήθηκε για την διεξαγωγή του γραφήματος 7:

```
s<-ggplot(data=movies,aes(x=genre,y=budget.mill,colour=genre
))
s+geom_point()+geom_jitter()+geom_boxplot(alpha=0.5)+ theme(axis.text.x = element_text(
  size=10, angle=90))
```

Γράφημα 8: Έξοδα βάση της συνολικής διάρκειας.



Παρατηρείται ότι υπάρχει θετική σχέση ανάμεσα στα έξοδα και στην διάρκεια μιας ταινίας. Ωστόσο ο βαθμός συσχέτισης δεν είναι τόσο μεγάλος, αν υπήρχε μια ευθεία γραμμή ανάμεσα στα σημεία του γραφήματος, αυτή θα είχε μικρή κλίση, ταινίες με την ίδια διάρκεια μπορεί να έχουν διαφορετικό προϋπολογισμό εξόδων. Βέβαια μπορούμε να διακρίνουμε ότι όσο αυξάνεται η διάρκεια μιας ταινίας αυξάνονται και τα έξοδά της.

Ο κώδικας που χρησιμοποιήθηκε για την διεξαγωγή του γραφήματος 8:

```
s<-ggplot(data=movies, aes(x=budget.mill, y=runtime, colour=genre))
s+geom_point()
```

5.5 Πρόταση δημιουργίας ταινίας.

Παρουσιάζοντας στο διοικητικό συμβούλιο της εταιρίας παραγωγής «Χ» τα αποτελέσματα της ανάλυσης των δεδομένων των ταινιών του Hollywood για τα τελευταία 76 χρόνια, ο CEO αποφάσισε τα εξής:

- Βάσει των πωλήσεων, η ταινία θα είναι είδους «sci-fi»,
- η πρώτη προβολή της θα είναι ημέρα Τετάρτη,
- το studio που θα αναλάβει την δημιουργία της ταινίας θα είναι το «Buena Vista Studios»,

- ο προϋπολογισμός για την υλοποίηση θα κυμανθεί στα 110 – 120 εκατομμύρια δολάρια και τέλος
- με δεδομένο τον προϋπολογισμό αυτό, η ταινία θα έχει διάρκεια περίπου 130 λεπτά.

Συμπεράσματα

Στην προκειμένη διπλωματική εργασία δημιουργήθηκε ο προβληματισμός για το πως οι σύγχρονες επιχειρήσεις μπορούν και διαχειρίζονται τον τεράστιο όγκο δεδομένων προς όφελός τους.

Ο πιο κοινός ορισμός των μεγάλων δεδομένων (Big Data) αναφέρεται στις καταστάσεις όπου η ποσότητα των δεδομένων είναι συντριπτική και δεν μπορεί να αντιμετωπιστεί με τις παραδοσιακές τεχνολογικές βάσεις δεδομένων και υπολογισμών. Οι άνθρωποι βλέπουν ουσιαστικά το τελευταίο στάδιο των δεδομένων, έχοντας δομή, μορφή και μπορώντας να βγάλουν ένα συμπέρασμα από την ανάγνωσή τους, ενώ στην πραγματικότητα απαρτίζονται από άπειρα αδόμητα, ημιδομημένα και χωρίς συγκεκριμένη μορφή δεδομένα.

Με την βοήθεια τεχνικών εξόρυξης ,συστημάτων και ανάλυσης αυτών των δεδομένων, ακολουθεί η ανακάλυψη γνώσης από τις βάσεις δεδομένων. Τα συστήματα αυτά, με την χρήση διαφόρων τεχνικών, παρέχουν γνώση κρυμμένη μέσα σε αυτόν τον τεράστιο όγκο δεδομένων με σκοπό την λήψη αποφάσεων. Σύμφωνα με μελέτες, οι επιχειρήσεις που χρησιμοποιούν τεχνικές ανάλυσης δεδομένων, είναι 5% έως 6% περισσότερο παραγωγικές από τους ανταγωνιστές τους. Η μηχανή εκμάθησης είναι ένα τέτοιο εργαλείο της διαδικασίας εξόρυξης και προ επεξεργασίας των δεδομένων, συνδυάζοντας την τεχνητή νοημοσύνη με τις αρχές της στατιστικής, έχει αποδειχθεί ένας από τους αποτελεσματικότερους τομείς της έρευνας, δημιουργώντας αλγόριθμους για την λύση συγκεκριμένων προβλημάτων που ορίζονται από τον ερευνητή.

Στον επιχειρηματικό κόσμο, τα εταιρικά δεδομένα και τα δεδομένα των πελατών αναγνωρίζονται ως στρατηγικό πλεονέκτημα. Ο ορισμός data mining (εξόρυξη δεδομένων) στηρίζεται στην διαδικασία εφαρμογής μεθοδολογίας βασισμένη σε υπολογιστές, συμπεριλαμβανομένων νέων τεχνικών, για την «μετάφραση» των δεδομένων και ανακάλυψη γνώσεων μέσα από αυτά. Σκοπός της εξόρυξης είναι η ανακάλυψη νέων τάσεων της αγοράς, ο σχεδιασμός επενδυτικών στρατηγικών και ο εντοπισμός μη λογικών δαπανών, η βελτίωση καμπάνιας μάρκετινγκ και τέλος ο

ανασχεδιασμός των επιχειρησιακών διαδικασιών κατανοώντας καλύτερα την αλληλένδετη εξάρτηση των τμημάτων.

Η οπτικοποίηση της ανάλυσης δεδομένων παρουσιάζει ενδιαφέρον για το λόγο ότι κάθε ενδιαφερόμενο μέλος μιας επιχείρησης μπορεί κοιτώντας ένα γράφημα να βγάλει συμπεράσματα για την λήψη αποφάσεων χωρίς να πρέπει να γνωρίζει σε βάθος το αντικείμενο της στατιστικής ανάλυσης και χωρίς να χρειάζεται να εισάγει δεδομένα σε κάποιο σύστημα για κάποιο αριθμητικό αποτέλεσμα. Επίσης μέσω της οπτικοποίησης είναι εφικτό να παρουσιαστεί ομαδοποίηση στοιχείων για μια πλήρη εικόνα πολλών παραγόντων μαζί.

Συνδυάζοντας την Επιχειρηματική Ευφυΐα, την αναλυτική δεδομένων με την στατιστική και την οπτικοποίηση, επιτυγχάνεται η λήψη αποτελεσματικών αποφάσεων με στόχο την καλύτερη λειτουργία του Μάρκετινγκ και όλων των λειτουργιών μιας επιχείρησης. Στα πλαίσια της επιχειρηματικής Ανάλυσης βασικό στοιχείο είναι και η αξιολόγηση των δεδομένων. Αν τα δεδομένα δεν είναι αξιόπιστα, τα αποτελέσματα θα είναι λανθασμένα και το κόστος των λάθος αποφάσεων θα είναι μοιραίο για την πορεία της επιχείρησης.

Βασικό μέλημα ενός οργανισμού που θέλει να παραμείνει βιώσιμος είναι η δημιουργία κερδών. «Δημιουργία κέρδους: Όλα εξαρτώνται από τον πελάτη. Πότε θα γίνει η αγορά, γιατί θέλει να αγοράσει το συγκεκριμένο προϊόν, πως θα φτάσει στο σημείο πώλησης και τελικά φτάνουμε στο σημείο απόφασης για την αγορά. Σκοπός του μάρκετινγκ είναι η κατανόηση και η προετοιμασία για αυτόν τον «κύκλο αγοράς».»

Στην έρευνα που διεξήχθη στο τέλος της διπλωματικής εργασίας, ασχοληθήκαμε με τον κλάδο της δημιουργίας ταινιών. Με την βοήθεια της γλώσσας προγραμματισμού R. Το πλεονέκτημα της χρήσης αυτής της γλώσσας είναι ότι αφήνει στον χρήστη την ελευθερία να δημιουργήσει δικά του γραφήματα χωρίς περιορισμούς, έχει πολύ μεγάλο χώρο αποθήκευσης δεδομένων για ταυτόχρονη χρήση και η λογική του κώδικα που χρησιμοποιεί είναι σε γενικά πλαίσια βατός. Παράλληλα όμως η διαχείριση τόσου μεγάλου όγκου δεδομένων από πηγές του διαδικτύου καθιστά την ανάλυση αρκετά δύσκολη. Ο χρήστης πρέπει να γνωρίζει άψογα στατιστική ανάλυση

και θεωρία των μεγάλων δεδομένων και να ξέρει πως να συμπεριφερθεί σε αυτά. Επίσης πρέπει να γνωρίζει τον τρόπο με τον οποίο «κινούνται» τα δεδομένα έτσι ώστε να βρει τους σωστούς στατιστικούς κώδικες που θα χρησιμοποιήσει.

Ενδιαφέρον θα παρουσίαζε και η εισαγωγή πολυπλοκότερων δεδομένων για την διεξαγωγή περισσότερων γραφημάτων και αναλύσεων σε διαφορετικό αντικείμενο επιχείρησης.

Βιβλιογραφία

Βιβλία:

1. Dominik Ryzko (2020) . Modern BD architecture. John Wiley & Sons, Inc, New Jersey.
2. Κύρκος, Ε. (2015). Επιχειρηματική Ευφυΐα και Εξόρυξη Δεδομένων
3. Hurwitz, J. S., Nugent, A., Halper, F., & Kaufman, M. (2013). Big Data For Dummies. John Wiley & Sons
4. Miller, J. D. (2017). Big Data Visualization. Packt Publishing Ltd.
5. Jerry Rackley (2015). Marketing Analytics Roadmap Methods, Metrics, and Tools. Apress
6. Ν. Γεωργόπουλος (2019) – Σημειώσεις Μαθήματος «Στρατηγικό Μανατζμεντ». Πανεπιστήμιο Πειραιώς
7. Mehmed Kantardzic (2011). DATA MINING Concepts Models, Methods, and Algorithms. A JOHN WILEY & SONS, INC. Canada.
8. Omer Artun (2018). Predictive Marketing Easy Ways Every Marketer Can Use Customer Analytics and Big Data. John Wiley & Sons. Canada
9. Gary Cokins (2014). STRATEGIC BUSINESS MANAGEMENT From planning to performance. AICPA New York
10. Simon Walkowiak (2016). Big Data Analytics with R. Packt Publishing. Birmingham

Άρθρα:

1. HCM Whitepaper: HR's Secret Weapon: The Power of Big Data
2. VanBoskirk, Overby, & Takvorian, (2011) - Beyond the hype: Big data concepts, methods, and analytics
3. Fan, Han, & Liu, (2014).Challenges of Big Data Analysis.
4. Zhenning Xu, Gary L. Frankwick, Edward Ramirez (2015). Effects of big data analytics and traditional marketing analytics on new product success: A knowledge fusion perspective.
5. William J. Hauser Marketing analytics (2017): the evolution of marketing research in the twenty-first century. The University of Akron, Akron, Ohio, USA
6. Sales & Marketing Analytics: Gauging and Optimizing Performance, Demand Metric, December 2013

Ηλεκτρονικές Πηγές:

1. https://www.ibm.com/gr-en
2. https://statista.com
3. https://repository.kallipos.gr/bitstream/11419/1232/2/Kef._5.pdf

4. http://dione.lib.unipi.gr/xmlui/bitstream/handle/unipi/167/DT2003-0107.pdf?sequence=1&isAllowed=y
5. file:///C:/Users/User/Downloads/marketing_analytics_benchmark_report.pdf
6. https://technologyadvice.com/blog/marketing/6-best-marketing-dashboards-visualizing-performance/
7. http://www.demandmetric.com/content/sales-marketinganalytics-benchmark-report