

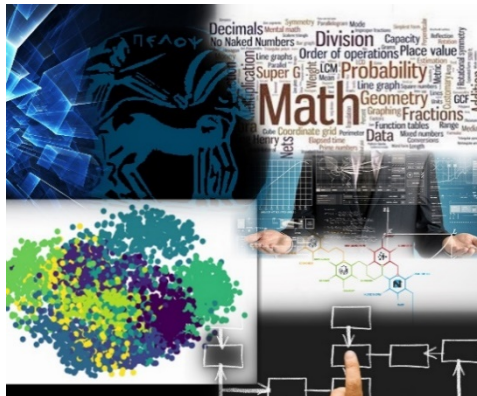


Πανεπιστήμιο Πελοποννήσου  
Σχολή Οικονομίας και Τεχνολογίας

Τμήμα Πληροφορικής και Τηλεπικοινωνιών  
Μ.Π.Σ. «Επιστήμη και Τεχνολογία Υπολογιστών»

## Διπλωματική εργασία

«Διερευνητική Ανάλυση Δεδομένων:  
Μέθοδοι και Μελέτες Περίπτωσης»



Στάθης Ανδρέας

A.M. 2022201602016

Επιβλέπων καθηγητής: κ. Βασιλάκης Κωνσταντίνος

Συνεπιβλέπων: καθηγητής: κ. Γουάλλες Μανώλης

Μάιος 2025

# Περιεχόμενα

<i>Περιεχόμενα</i>	<i>i</i>
<i>Ευρετήριο σχημάτων</i>	<i>ii</i>
<i>Περίληψη</i>	<i>1</i>
<b>1 Εισαγωγή</b>	<b>3</b>
<b>2 Διερευνητική ανάλυση δεδομένων</b>	<b>6</b>
<b>2.1 Εισαγωγή</b>	<b>6</b>
<b>2.2 Στόχοι της διερευνητικής ανάλυσης δεδομένων.</b>	<b>6</b>
<b>2.3 Διαδικασία διερευνητικής ανάλυσης δεδομένων</b>	<b>9</b>
2.3.1 Η συλλογή των δεδομένων σε επίπεδο καταγραφής ερευνητικών ερωτήσεων	10
2.3.2 Η χρήση στατιστικών τεχνικών παρατήρησης και γραφικών προσεγγίσεων	13
2.3.3 Διερευνητική ανάλυση, έλεγχος υποθέσεων και βελτιστοποιήσεις	13
<b>3 Το πλαίσιο αρχών FAIR για τα ανοικτά δεδομένα</b>	<b>15</b>
<b>4 Εργαλεία στατιστικής επεξεργασίας και οπτικοποίησης</b>	<b>17</b>
<b>4.1 Γλώσσα R</b>	<b>17</b>
<b>4.2 Grafana</b>	<b>17</b>
<b>4.3 Kibana</b>	<b>19</b>
<b>5 Προσχέδιο δεδομένων και διεργασιών (Blueprint and Data pre processing)</b>	<b>21</b>
<b>6 Μελέτη περίπτωσης - Διερευνητική ανάλυση δεδομένων (Case study - Exploratory data analysis)</b>	<b>24</b>
<b>6.1 Υλοποίηση.</b>	<b>24</b>
<b>6.2 Ερωτήματα διερευνητικής ανάλυσης με χρήσης της SQL</b>	<b>25</b>
<b>7 Συμπεράσματα</b>	<b>29</b>
<b>8 Βιβλιογραφία</b>	<b>30</b>

## Ευρετήριο σχημάτων

Εικόνα 1. Αναγνώριση προβλήματος στη διερευνητική ανάλυση	4
Εικόνα 2. Συμμετοχή ερευνητικών ερωτήσεων στη διαδικασία διερευνητικής ανάλυσης δεδομένων	11
Εικόνα 3. Γραφήματα για την πανδημία Covid-19 με χρήση του Grafana (πηγή: <a href="https://github.com/FortDigital/covid-19">https://github.com/FortDigital/covid-19</a> )	19
Εικόνα 4. Οπτικοποίηση των μισθών στη Γαλλία με χρήση του Kibana (πηγή: <a href="https://www.elastic.co/jp/blog/visualizing-france-salary-data-with-region-maps-in-kibana?blade=fbs">https://www.elastic.co/jp/blog/visualizing-france-salary-data-with-region-maps-in-kibana?blade=fbs</a> )	20
Εικόνα 5. Διάγραμμα οντοτήτων-συσχετίσεων για τα δεδομένα της εφαρμογής	21
Εικόνα 6. Ανίχνευση ακραίων τιμών με την τεχνική IQR (πηγή: <a href="https://www.cloudymml.com/blog/outlier-detection-and-treatment/">https://www.cloudymml.com/blog/outlier-detection-and-treatment/</a> )	23
Εικόνα 7. Μετρήσεις ανά ημέρα	26
Εικόνα 8. Περιγραφική στατιστική	26
Εικόνα 9. Σύγκριση συμβολαίων ταχύτητας με μετρήσεις.	27
Εικόνα 10. Πλατφόρμα διερευνητικής ανάλυσης δεδομένων.	28
Εικόνα 11. Αποτύπωση επιτυχής διαδικασίας διερευνητικής ανάλυσης δεδομένων της εργασίας.	29

**No table of figures entries found.**

## Περίληψη

Στη διερευνητική ανάλυση δεδομένων τα σύνολα δεδομένων αναλύονται για να παρουσιαστούν τα κύρια χαρακτηριστικά τους. Για την πιο εύληπτη και διαισθητική παρουσίαση των χαρακτηριστικών αυτών στους χρήστες, είναι ιδιαίτερα σύνηθες να χρησιμοποιείται η οπτικοποίηση των δεδομένων. Η σύγχρονη τεχνολογία, δημιουργεί αυξημένους όγκους δεδομένων που χρήζουν ανάλυσης, ενώ η ανάλυση αυτή πραγματοποιείται συχνά από διαφορετικές ομάδες χρηστών, με διαφορετικούς στόχους και διαφορετικές ανάγκες από την ανάλυση δεδομένων. Επιπρόσθετα, η διαθεσιμότητα των ανοικτών δεδομένων δημιουργεί πρόσθετες ευκαιρίες συσχέτισης δεδομένων για τη δημιουργία εμπλουτισμένων συνόλων δεδομένων, και κατόπιν εξαγωγής πληροφοριών και συμπερασμάτων από εμπλουτισμένα αυτά σύνολα.

Για τους λόγους αυτούς, είναι σημαντικό να υπάρχουν ευέλικτες και αποτελεσματικές μέθοδοι για τη συλλογή, αποθήκευση και διαχείριση των δεδομένων, την εξέταση των μεθόδων της στατιστικής επεξεργασίας και την οπτικοποίηση. Στο επίπεδο της αποθήκευσης και διαχείρισης των δεδομένων, η χρήση συστημάτων σχεσιακών βάσεων δεδομένων αποτελεί μία καλά δοκιμασμένη τεχνική η οποία εγγυάται καλή απόδοση, αποτελεσματικότητα και ευκολία στη διαχείριση. Για τον εντοπισμό και τη συλλογή δεδομένων προς εμπλουτισμό των αρχικών συλλογών δεδομένων μπορούν να αξιοποιηθούν τα αποθετήρια ανοικτών δεδομένων (open data repositories). Η σύγχρονη τάση μάλιστα, να διατίθενται όλο και περισσότερα και πιο πλούσια σύνολα δεδομένων ως ανοικτά δεδομένα, παράλληλα με την υιοθέτηση των αρχών FAIR (Findable, Accessible, Interoperable, Reusable – ανακαλύψιμα/ευρέσιμα, προσβάσιμα, διαλειτουργικά και επαναχρησιμοποιήσιμα) [6] από τα αποθετήρια ανοικτών δεδομένων, καθιστούν την εύρεση συνόλων ανοικτών δεδομένων ευχερέστερη και τη χρήση τους πιο προσοδοφόρα. Η διάθεση μάλιστα των ανοικτών δεδομένων σε μορφές που είναι ευχερές να εισαχθούν σε σχεσιακές βάσεις δεδομένων (όπως π.χ. CSV, XLS/XLX/JSON), κατ' εφαρμογή της αρχής της διαλειτουργικότητας, διευκολύνει τη δημιουργία των εμπλουτισμένων συνόλων δεδομένων.

Στον τομέα της στατιστικής επεξεργασίας των δεδομένων και της οπτικοποίησης, νέα σύγχρονα εργαλεία όπως η γλώσσα προγραμματισμού  $R^1$ , η οποία είναι

---

<sup>1</sup> <https://www.r-project.org/>

προσανατολισμένη στη στατιστική επεξεργασία των δεδομένων, καθώς και πιο υψηλού επιπέδου εργαλεία που ενοποιούν την πρόσβαση σε βάσεις δεδομένων με τη στατιστική επεξεργασία και την οπτικοποίηση, όπως π.χ. τα Grafana<sup>2</sup> και Kibana<sup>3</sup>, επιτρέπουν στους χρήστες να έχουν πρόσβαση σε προκατασκευασμένες οπτικοποιήσεις της πληροφορίας ή/και να αναπτύσσουν νέες οπτικοποιήσεις, προσαρμοσμένες στις δικές τους ανάγκες, πολλές φορές μάλιστα με λίγες ή καθόλου γνώσεις προγραμματισμού.

Στην παρούσα διπλωματική εργασία, αναλύονται αρχικά οι ανωτέρω συνιστώσες του πλαισίου της διερευνητικής ανάλυσης δεδομένων, όπως αυτό διαμορφώνεται με τις σύγχρονες εξελίξεις. Στη συνέχεια, πραγματοποιείται μία μελέτη περίπτωσης η οποία αντλείται από τον χώρο της αγοράς τηλεπικοινωνιών, όπου σύνολα ανοικτών δεδομένων αντλούνται, εισάγονται σε σύστημα διαχείρισης σχεσιακών βάσεων δεδομένων και τέλος αναπτύσσονται κατάλληλες στατιστικές επεξεργασίες και οπτικοποιήσεις, χρησιμοποιώντας το εργαλείο Grafana. Η συγκεκριμένη μελέτη περίπτωσης αφορά τα δεδομένα μέτρησης ταχύτητας διαδικτυακών συνδέσεων ΥΠΕΡΙΩΝ (<https://hyperiontest.gr/>), τα οποία συνδυάζονται με δεδομένα που αφορούν τις περιοχές της χώρας, και που αφορούν ιδίως την οικονομική δραστηριότητα.

**Λέξεις κλειδιά:** Διερευνητική ανάλυση δεδομένων, Συχνές ερωτήσεις, Διαμόρφωση στόχων/ενεργειών, Μελέτη περίπτωσης.

---

<sup>2</sup> <https://grafana.com/>

<sup>3</sup> <https://www.elastic.co/kibana>

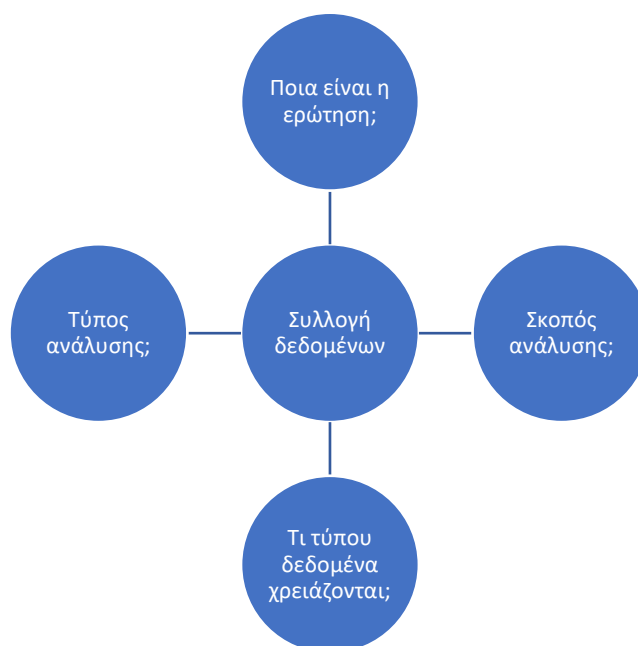
# 1 Εισαγωγή

Η διερευνητική ανάλυση δεδομένων είναι μία προσέγγιση όπου τα σύνολα δεδομένων αναλύονται για να παρουσιαστούν τα κύρια χαρακτηριστικά τους. Για την πιο εύληπτη και διαισθητική παρουσίαση των χαρακτηριστικών αυτών στους χρήστες, είναι ιδιαίτερα σύνηθες να χρησιμοποιείται η οπτικοποίηση των δεδομένων και των χαρακτηριστικών τους. Υπάρχει ένας άτυπος κανόνας που αναφέρεται στην διαφορά μεταξύ χρήσης οπτικοποιήσεων και χρήσης στατιστικών αλγορίθμων: όταν το πρόβλημα αφορά την εύρεση πληροφορίας μέσω παρατήρησης των δεδομένων, η οπτικοποίηση έχει πρωτεύοντα ρόλο και χρησιμοποιείται καθώς η οπτική αναπαράσταση μπορεί να υποβοηθήσει τις ανθρώπινες ικανότητες συλλογιστικής και συμπερασμού προς την κατεύθυνση του εντοπισμού ιδιαίτερων χαρακτηριστικών των δεδομένων ή/και της διευκόλυνσης της κατανόησης αυτών. Αντίθετα, όταν το πρόγραμμα απαιτεί τον υπολογισμό εκ των προτέρων γνωστών στατιστικών μεγεθών, π.χ. των μέσων όρων και ελαχίστων/μεγίστων τιμών, τότε η χρήση στατιστικών αλγορίθμων είναι κατάλληλη για την εύρεση της απάντησης μέσω των ικανοτήτων του αριστερού-ημισφαιρίου του εγκεφάλου και την αναλυτική φύση που έχει. Στην εργασία αυτή η χρήση συστήματος οπτικοποίησης στοχεύει στην παροχή υποστήριξης προς τους χρήστες για εύρεση μοτίβων(patterns), σχέσεων σε μεμονωμένα σύνολα δεδομένων ή συνδυασμούς συνόλων δεδομένων από πηγές (π.χ. ανοιχτών δεδομένων) και εξέταση αυτών. Στην όλη διαδικασία της διερευνητικής ανάλυσης, είναι δυνατόν να υπάρξουν διαφορετικές προσεγγίσεις από διαφορετικούς χρήστες ως προς την απάντηση μιας συγκεκριμένης ερώτησης, όπως π.χ. διαφορετικές πράξεις κανονικοποίησης, συσχέτισης, μείωσης του πλήθους των διαστάσεων των δεδομένων κ.ο.κ.

Με βάση τα ανωτέρω, στο πλαίσιο της διερευνητικής ανάλυσης, γύρω από τη *συλλογή δεδομένων* η οποία υπόκειται κάθε φορά σε ανάλυση περιστρέφεται ένα πλήθος ερωτημάτων-παραμέτρων της ανάλυσης, τα οποία περιλαμβάνουν:

- Ποια είναι η ερευνητική ερώτηση;
- Ποιος είναι ο σκοπός της ανάλυσης;
- Τι τύπου δεδομένα χρειάζονται;
- Ποιος τύπος ανάλυσης θα εφαρμοστεί;

Τα ερωτήματα αυτά συνδιαμορφώνουν και το ίδιο το σύνολο δεδομένων, σε μία εξελικτική πορεία γνώσης του αντικειμένου ανάλυσης. Η διαδικασία αυτή απεικονίζεται στην Εικόνα 1.



*Εικόνα 1. Αναγνώριση προβλήματος στη διερευνητική ανάλυση*

Μια τέτοια διαδικασία χαρακτηρίζεται από αρχές επιχειρηματικής ευφυΐας: στο πλαίσιο της παρούσας εργασίας, ο σκοπός της ανάλυσης είναι η διερεύνηση της τρέχουσας κατάστασης της αγοράς των τηλεπικοινωνιών ώστε να αναδειχθούν ευκαιρίες για νέα ή βελτιωμένα προϊόντα και υπηρεσίες. Οι χρήστες μπορούν να αναλύσουν τα διαθέσιμα δεδομένα, παρατηρώντας στατιστικούς δείκτες, εξελίξεις και τάσεις, εξετάζοντας την επίδραση διαφορετικών παραγόντων μέσω συσχετίσεων δεδομένων ή προβάλλοντας οπτικοποιήσεις, προκειμένου να διευκολυνθούν στην εξαγωγή συμπερασμάτων. Στα κεφάλαια που ακολουθούν, θα παρουσιαστούν ο σκοπός και ο ρόλος της διερευνητικής ανάλυσης υπό το πλαίσιο του προβλήματος που καλείται το σύστημα να υποστηρίξει.

Για την υποστήριξη της προαναφερθείσας διερευνητικής ανάλυσης, θα υλοποιηθεί ένα κατάλληλο σύστημα που θα βασίζεται στην αρχιτεκτονική εξυπηρέτη-εξυπηρετούμενου (client-server), ενώ η υλοποίηση του συστήματος θα ενσωματώνει και κατάλληλα εργαλεία για την υλοποίηση των απαιτούμενων λειτουργιών.

Οι στατιστικές τεχνικές που χρησιμοποιούνται στη διερευνητική ανάλυση (Exploratory analysis) ξεκίνησαν από τον Αμερικανό στατιστικόλόγο Tukey (1977) να

αντιμετωπίζονται συχνότερα με γραφικά οπτικοποιημένα μέσα και να βασίζονται στο τρόπο παρατήρησης των δεδομένων. Πολλοί αναλυτές δεδομένων χρησιμοποιούν γραφικές αναπαραστάσεις και στατιστικές τεχνικές για να αντιμετωπίσουν το πρόβλημα που σχετίζεται με τα δεδομένα. Στο σημείο αυτό διευκρινίζετε ότι η *Διερευνητική Ανάλυση των Δεδομένων (Exploratory Data Analysis)* αποτελεί μια επιστημονική μέθοδο που προσεγγίζετε σε φιλοσοφικό επίπεδο μέσα από τις εργασίες των Burrill- Biehler [3] και του Good [7] ως μια διανοητική δραστηριότητα, που συχνά στην επιστήμη των υπολογιστών αποκαλείται ως γνώση υψηλού επιπέδου περιέχοντας γνωστικές περιοχές πολλών διαφορετικών επιστημών, όπως για παράδειγμα η στατιστική, τα μαθηματικά και η ψυχολογία.

Η παρούσα διπλωματική εργασία είναι διαρθρωμένη ως ακολούθως: αρχικά, στο δεύτερο κεφάλαιο πραγματοποιείται μια επισκόπηση της *εφαρμογής διερευνητικής ανάλυσης (EDA)*, παρουσιάζοντας διαφορετικές περιπτώσεις τεχνικών που μπορούν να εφαρμοστούν για την κάλυψη διαφορετικών αναγκών/στόχων. Στο τρίτο κεφάλαιο, εξετάζονται τα *ανοικτά δεδομένα*, μία τάση αλλά και τεχνολογία η οποία είναι δυνατόν να επαυξήσει το αρχικό σύνολο δεδομένων με πρόσθετες διαστάσεις, και συνακόλουθα να οδηγήσει μέσω της διερευνητικής ανάλυσης στην ανάδειξη συσχετίσεων. Στο τέταρτο κεφάλαιο παρουσιάζονται *εργαλεία στατιστικής ανάλυσης και οπτικοποίησης* που μπορούν να χρησιμοποιηθούν για την υλοποίηση ενός συστήματος υποστήριξης διερευνητικής ανάλυσης, ενώ στο πέμπτο κεφάλαιο παρουσιάζεται ως μελέτη περίπτωσης η *κατασκευή και χρήση ενός συστήματος για την ανάλυση δεδομένων τηλεπικοινωνιακών συνδέσεων*. Τέλος, στο έκτο κεφάλαιο, παρατίθεται *σύνοψη* της εργασίας και σκιαγραφούνται πιθανές επεκτάσεις.

## 2 Διερευνητική ανάλυση δεδομένων

### 2.1 Εισαγωγή

Η διερευνητική ανάλυση δεδομένων είναι μία θεμελιώδης διαδικασία που παράγει με τη χρήση στατιστικών τεχνικών και γραφικών αναπαραστάσεων το σκοπό να αποκτηθεί μια βαθύτερη γνώση από τα δεδομένα.[1] Πρόκειται για μια επιστημονική μέθοδο ανάλυσης δεδομένων που χρησιμοποιείται κυρίως σε επιστημονικά προβλήματα ως μια διαδικασία μάθησης των δεδομένων, με στόχο την ανακάλυψη, τη λήψη αποφάσεων και την εξέταση ερευνώντας τα δεδομένα κατά τη λειτουργία τους. Η αποτύπωση μέσω της οπτικοποίησης και η βελτίωση της διαδικασίας ανάλυσης για την εκάστοτε περίπτωση αναδεικνύει την εφαρμογή της σε διαφορετικούς επιστημονικούς χώρους και χρησιμοποιείται από διαφορετικές ομάδες ειδικών, σε ποικίλες γνωστικές περιοχές. Η παρουσίαση της θεωρίας της διερευνητικής ανάλυσης στην εργασία αυτή βασίζεται στο κεφάλαιο του James J. Filliben του εγχειριδίου από το Εθνικό Ινστιτούτο Προτύπων και Τεχνολογιών (NIST) [2]. Στη συνέχεια του κεφαλαίου γίνεται αναφορά στους στόχους της διερευνητικής ανάλυσης, την ανάλυση της διαδικασίας σε προβλήματα συστηματικής<sup>4</sup> περιγραφής και τα διαφορετικά σημεία από κλασικές αναλύσεις δεδομένων σύμφωνα με τη προσέγγιση από το εγχειρίδιο και σύγχρονες τάσεις απόδοσης αναλύσεων(analytics).

### 2.2 Στόχοι της διερευνητικής ανάλυσης δεδομένων.

Τα δεδομένα ενός συστήματος, οργανισμού ή μιας έρευνας αποτυπώνονται σε ένα πλήθος ποσοτικών ή/και ποιοτικών διαστάσεων. Κατόπιν τα δεδομένα αναλύονται με τρόπο τέτοιο ώστε να μπορεί να είναι εφικτή η οπτικοποίηση τους σε ένα πλήθος περιπτώσεων και να μπορούν να υποστηριχθεί η κατάρτιση σχεδίων για βελτίωση διαδικασιών ή/και επίλυση προβλημάτων της επιχείρησης/του οργανισμού. Για να γίνει αυτό, διαμορφώνεται και χρησιμοποιείται ένα σύνολο γραφικών και στατιστικών μεθόδων που αποτελούν τα εργαλεία της διερευνητικής ανάλυσης των δεδομένων της επιχείρησης/του οργανισμού και έχουν ως στόχο (α) να επιτρέψουν στον αναλυτή να κατανοήσει τα δεδομένα και τους τομείς και τις δραστηριότητες του οργανισμού τους οποίους αυτά αντιπροσωπεύουν και να εξάγει κατάλληλα συμπεράσματα και (β) προκειμένου για την υποστήριξη του στόχου [α], να παρέχουν με ακρίβεια όλα τα

---

<sup>4</sup> Ο όρος συστηματική χρησιμοποιείται για να περιγράψει το τρόπο που η διαδικασία λαμβάνει χώρα.

στοιχεία που θα χρειαζόταν ο αναλυτής, τα οποία περιλαμβάνουν τουλάχιστον τα 8 σημεία που περιγράφει στο εγχειρίδιο ο Filliben [2] (1.1.4 σελ. 19), ήτοι:

- Ένα μοντέλο που έχει προσαρμοστεί κατάλληλα,
- Μία λίστα με ακραίες τιμές,
- Την πεποίθηση ότι τα συμπεράσματα που εξάγονται είναι επαρκώς θεμελιωμένα,
- Ακριβείς εκτιμήσεις των παραμέτρων του μοντέλου,
- Ακριβή εικόνα σχετικά με τον βαθμό αβεβαιότητας εκτιμήσεων των παραμέτρων,
- Έναν κατάλογο με τους σημαντικούς παράγοντες, ταξινομημένους με βάση τον βαθμό σημαντικότητας,
- Συμπεράσματα σχετικά με το εάν οι μεμονωμένοι παράγοντες είναι στατιστικά σημαντικοί,
- Ένα σύνολο βέλτιστων ρυθμίσεων.

Για τα 8 αυτά σημεία γίνεται σαφές ότι η απόκτηση γνώσης μέσω της περιγραφής των δεδομένων και της επεξήγηση των δομών τους είναι δύο διαφορετικού τύπου προβλήματα, τα οποία όμως τέμνονται και είναι σκόπιμο να αντιμετωπιστούν με κοινά σημεία και παραδοχές. Η περιγραφή των δεδομένων, γίνεται συνυπολογίζοντας εννοιολογικά και πραγματολογικά στοιχεία, εφαρμόζοντας σχετικές στατιστικές μεθόδους, όπως η ανάπτυξη ενός νοητικού μοντέλου σύμφωνα με την εργασία του John T. Behrens [5] ή εφαρμόζοντας μια προσέγγιση λύσης προβλημάτων όπως αυτή που αποτυπώνεται στην εργασία του Jukka-Matti Turtiainen [4]. Για το σκέλος της ερμηνείας των δεδομένων στο πλαίσιο του διαμοιρασμού/κοινής χρήσης τους, τυγχάνουν εφαρμογής οι αρχές του μοντέλου FAIR [5], το οποίο θα παρουσιαστεί στο επόμενο κεφάλαιο.

Ο βασικός στόχος της διερευνητικής ανάλυσης είναι να μεγιστοποιήσει την επίγνωση του χρήστη για τα δεδομένα καθώς και τη δομή τους. Η μελέτη της δομής των δεδομένων, με χρήση κατάλληλων μορφών παρουσίασής των, είναι ένα πολύ βασικό στοιχείο στο πλαίσιο της απόκτησης επίγνωσης, αλλά δεν είναι από μόνη της επαρκής: Σύμφωνα με τον Filliben [2], η απόκτηση επίγνωσης περιλαμβάνει εξέταση, μελέτη και κατανόηση του τι περιέχουν και τι δεν περιέχουν τα δεδομένα, εξαγωγή μοτίβων κ.ο.κ., στοιχεία που μπορούν να παραχθούν μόνο με χρήση γνωσιακών λειτουργιών,

όπως η σχηματική αναγνώριση μοτίβων και η συγκριτική συλλογιστική πάνω σε γραφικές αναπαραστάσεις των δεδομένων.

Σύμφωνα με τον Good [7], οι στόχοι της διερευνητικής ανάλυσης μπορούν να συνοψισθούν σε 5 σημεία ως ακολούθως:

1. **Παρουσίαση των δεδομένων.** Η αναπαράσταση των δεδομένων με τέτοιο τρόπο που να διευκολύνει την κατανόησή τους και την εξαγωγή συμπερασμάτων. Θα πρέπει να υποστηρίζονται τόσο οι γνωσιακές λειτουργίες του αναλυτή που αφορούν οπτική κατανόηση (ερμηνεία σχημάτων, χωρική ανάλυση κ.τ.λ.) αλλά και την κατανόηση αριθμών, κατηγοριών και σχέσεων μεταξύ τους.
2. **Αναγνώριση προτύπων/μοτίβων και τάσεων.** Τα δεδομένα θα πρέπει να παρουσιάζονται με τρόπο ώστε να αναδεικνύονται μοτίβων ή/και τάσεις και να είναι ευχερές να εντοπισθούν. Η παροχή υποστήριξης για λογική θεμελίωση των μοτίβων είναι επίσης σημαντική.
3. **Διατύπωση υποθέσεων.** Οι υποθέσεις αποτελούν υψηλού επιπέδου διατυπώσεις σχετικά με το τι είναι αληθές και τι όχι, οι οποίες εξηγούν με εύλογο τρόπο τα μοτίβα και τις τάσεις που παρατηρούνται στα δεδομένα. Η διατύπωση των υποθέσεων βασίζεται στην ερμηνεία των γραφικών αναπαραστάσεων και των οπτικοποιήσεων, οι οποίες συνεισφέρουν στην κατανόηση των σχέσεων μεταξύ των δεδομένων.
4. **Αναζήτηση υποθέσεων με μεγαλύτερη επεξηγηματική ικανότητα.** Οι αρχικά διατυπωμένες και επιβεβαιωμένες εκλεπτόνονται ή συμπληρώνονται με πρόσθετες υποθέσεις, όπως για παράδειγμα η ερμηνεία των αποκλίσεων από τα μοτίβα και τις τάσεις που αναδείχθηκαν και καταγράφηκαν. Η προοδευτική διαμόρφωση υποθέσεων είναι ουσιαστικό στοιχείο της διερευνητικής ανάλυσης δεδομένων.
5. **Αναζήτηση της κατάλληλης ισορροπίας μεταξύ της ποιότητας των λύσεων και του κόστους διαμόρφωσής τους.** Αυτό είναι ιδιαίτερα σημαντικό διότι η διερευνητική ανάλυση παρέχει πολύ μεγάλο πλήθος εργαλείων και μεθόδων για τη μελέτη και ανάλυση των δεδομένων και η χρήση όλων των διαθέσιμων επιλογών και των συνδυασμών τους θα εκτόξευε το κόστος της διαμόρφωσης των λύσεων [11]. Για την επίτευξη του στόχου αυτού θα πρέπει να ληφθούν υπ' όψιν τόσο οι γενικότερες περιγραφές των δεδομένων όσο και τα ίδια τα

δεδομένα, στοιχεία τα οποία θα καθοδηγήσουν τις αποφάσεις για την επιλογή των καταλληλότερων μεθόδων και εργαλείων.

Για την επίτευξη των στόχων αυτών, είναι καθοριστικό να ικανοποιούνται τα 8 σημεία που αναφέρονται στην εργασία του Filliben [2] και παρατέθηκαν ανωτέρω.

Συνολικά, το συμπέρασμα που προκύπτει είναι ότι, η χρήση γραφικών αναπαραστάσεων είναι αναντικατάστατο στοιχείο της διαδικασίας, καθώς μπορεί να υποστηρίξει την επίγνωση, τροφοδοτώντας με τα κατάλληλα ερεθίσματα τις γνωσιακές ικανότητες του χρήστη, επιτρέποντάς του να συλλογισθεί πάνω στα δεδομένα και να καταλήξει σε αποφάσεις κατόπιν ενδεδειγμένης ανάλυσης των δεδομένων και συνακόλουθης συναγωγής σχετικών κρίσεων και αποφάσεων πάνω στα δεδομένα τα οποία εξετάζει. Για τον λόγο αυτό, ακολουθώντας τις εξελίξεις και τις δυνατότητες που δίνει η τεχνολογία, η χρήση εξειδικευμένου λογισμικού με αυτοματοποιημένες ή μη λειτουργίες σκιαγραφεί κάποιες λειτουργίες όπως: γραφικές αναπαραστάσεις διαφορετικών στατιστικών τεχνικών, δυνατότητα συλλογής δεδομένων και διαχείριση του σκοπού της διερευνητικής ανάλυσης. Με τις λειτουργίες αυτές καλύπτεται ένα ευρύ σύνολο από διερευνητικές ερωτήσεις, ώστε να εξετάζονται διάφορες πτυχές του προβλήματος που αναλύεται, να μπορούν να εκφράζονται με εύληπτο τρόπο τα στατιστικά αποτελέσματα και να υποστηρίζεται η λήψη αποφάσεων.

### **2.3 Διαδικασία διερευνητικής ανάλυσης δεδομένων**

Προκειμένου για την καθοδήγηση των χρηστών στην εκτέλεση της διερευνητικής ανάλυσης, δημιουργείται ένα πλαίσιο εργασίας (framework) ώστε να αποτυπωθεί μία σαφής και αποτελεσματική ροή εργασιών.

Ακολουθώντας την προσέγγιση της καθοδήγησης από ζητήματα, ο Thomas Davenport και Jinho Kim [8] παρουσίασαν ένα μοντέλο τριών επιπέδων αποτελούμενο από (i) τη διαμόρφωση του προβλήματος, (ii) τη λύση του προβλήματος και (iii) τη διατύπωση των αποτελεσμάτων. Κάθε ένα επίπεδο αποτελεί μια οντότητα στο πλαίσιο της οποίας πραγματοποιείται επεξεργασία εισόδων και διαμόρφωση εξόδων, με τις εξόδους να συμβάλλουν στο επόμενο επίπεδο, συμβάλλοντας στην επίτευξη των στόχων και τη κάλυψη των αναγκών. Πιο αναλυτικά, τα τρία επίπεδα που προσδιορίζουν οι Davenport και Kim [8] έχουν ως ακολούθως:

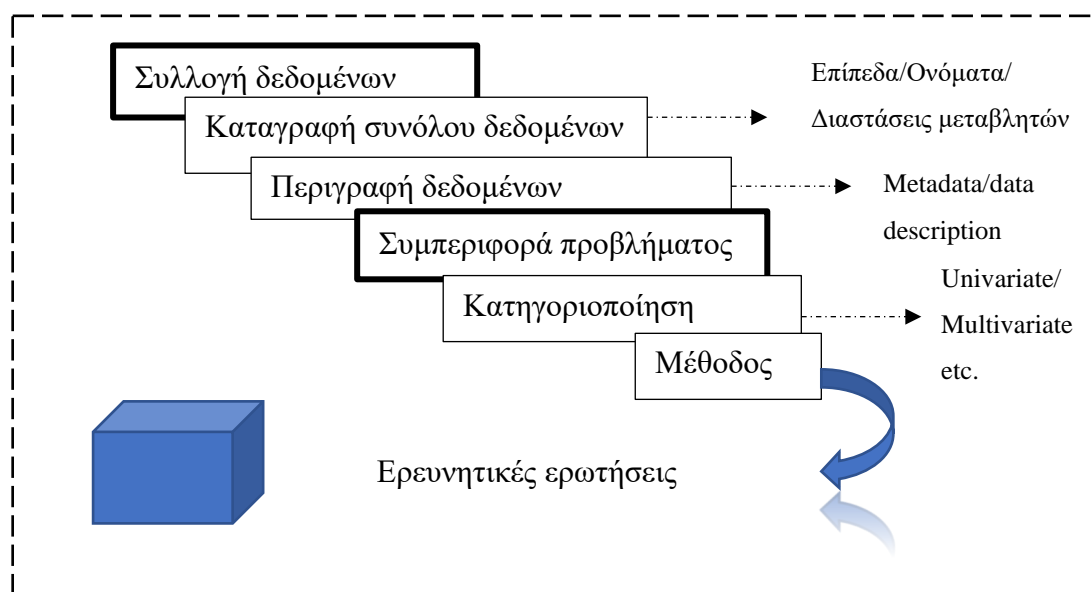
- *Προσδιορισμός προβλήματος.* Απόκτηση γνώσης των δεδομένων υπό την καθοδήγηση κύριων ερευνητικών ερωτήσεων που αφορούν όρους προσδιορισμού της ανάλυσης,
- *Λύση προβλήματος.* Διαμόρφωση του μοντέλου ανάλυσης με χρήση στατιστικών τεχνικών και γραφικών αναπαραστάσεων ,
- *Διατύπωση αποτελεσμάτων.* Εξαγωγή συμπερασμάτων από την χρήση του μοντέλου ανάλυσης.

### **2.3.1 Η συλλογή των δεδομένων σε επίπεδο καταγραφής ερευνητικών ερωτήσεων**

Ο προσδιορισμός του προβλήματος βασίζεται στη καταγραφή και συμπεριφορά του προβλήματος, η καταγραφή του προβλήματος αφορά την αναγνώριση του βάσης της περιγραφής των δεδομένων. Ποιο συγκεκριμένα η αποτύπωση του καλούμενου προβλήματος προς καταγραφή συχνά διασπάτε σε μικρά βήματα ώστε το τμήμα που εξετάζετε να συγκεκριμενοποιεί την διατύπωση της ιδέας καθώς και την ανάλυση που εξελίσσετε. Η διερευνητική ανάλυση μπορεί να χρησιμοποιηθεί ώστε να επιβεβαιώσει ότι τα σύνολα δεδομένων που έχουν συλλεχθεί ανταποκρίνονται σε μια περιγραφή μεταδεδομένων, ή να δημιουργήσει μια τέτοια περιγραφή εφόσον δεν υπάρχει. Κάθε μια από τις περιπτώσεις μπορεί να θεωρηθεί ότι χτίζει μια συμπεριφορά για τα δεδομένα αναφέρει στο βιβλίο του ο Antonio Badia [9] (3.5.1 σελ. 167).

Η συμπεριφορά και τη κατηγοριοποίηση των προβλημάτων όπως αναφέρει ο Filliben στη διερευνητική και την κλασική ανάλυση χαρακτηρίζονται ως: Univariate/control, Comparative/screening, Optimization/regression, Time series/multivariate, κατηγορίες που καλύπτει και η περιγραφική ανάλυση, η χρήση της οποίας ικανοποιεί εν μέρη την οπτική επίγνωση. Για παράδειγμα η μέθοδος των πινάκων συνάφειας (contingency tables) βοηθά την ικανότητα του χρήστη ώστε να κατανοήσει τις μετρήσεις του συγκεκριμένου προβλήματος καθώς οι πίνακες είναι αφαιρέσεις και το αντικείμενο ποιο συγκεκριμένο, δεν αποτελεί μια ικανή αποτύπωση όπου ο χρήστης μπορεί να έχει πολλές διαφορετικές αποδόσεις σε κλίμακες πολλών μεταβλητών καταστάσεων όπως στην χρήση διαγράμματος μωσαϊκού τύπου, και δια δραστικού περιεχομένου. Η κατηγορία των πολλών μεταβλητών (multivariate) είναι το συνηθέστερο παράδειγμα όπου η διερευνητική ανάλυση έχει την ικανότητα να αποδώσει πολλαπλές διαφορετικές οπτικές σε σχέση με την περιγραφική ανάλυση μέσα από διαγράμματα συχνά διαδραστικού τύπου που αναπαριστούν χρονοσειρές δεδομένων (time series) .

Για να αξιολογηθούν τα δεδομένα σε επίπεδο ερευνητικών ερωτήσεων υπάρχουν κάποιες ερωτήσεις που αναφέρονται σε κάθε σχεδόν τύπο ανάλυσης δεδομένων σύμφωνα με το Filliben (ενότητα 1.3.2). Το πλεονέκτημα των ερευνητικών ερωτήσεων βασίζεται ότι μπορούν να εφαρμοστούν σε ποιοτικές και ποσοτικές μελέτες δεδομένου ότι αναγνωρίζετε η συμπεριφορά του καλούμενου προβλήματος μέσω της καταγραφής και της περιγραφής των διαθέσιμων συνόλων δεδομένων.



Εικόνα 2. Συμμετοχή ερευνητικών ερωτήσεων στη διαδικασία διερευνητικής ανάλυσης δεδομένων

Οι ερευνητικές ερωτήσεις καθορίζουν τον στόχο της ανάλυσης και χαρακτηρίζονται ως το βήμα το οποίο διατυπώνει υποθέσεις για ύπαρξη συγκεκριμένων σχέσεις μεταξύ των μεταβλητών που εξετάζονται. Στη διερευνητική ανάλυση, καθώς προχωράμε ανάμεσα στα στάδια της συλλογής των δεδομένων, της μοντελοποίησης και της περιγραφής τους και της απόκτησης περισσότερης γνώσης τόσο για τα δεδομένα όσο και για τον χώρο τον οποίο αφορούν, στη διάρκεια αυτής της πορείας εξελίσσονται και συμπληρώνονται και οι ερευνητικές ερωτήσεις. Για τη διατύπωση των ερευνητικών ερωτήσεων θα πρέπει να εφαρμόζονται οι ακόλουθοι μη τεχνικοί κανόνες:

- Να είναι σύντομες,
- Να είναι σαφείς και ξεκάθαρες,
- Να σχετίζονται με την έρευνα,
- Η απάντησή τους να είναι σημαντική για την περιοχή του προβλήματος,
- Να μπορούν να απαντηθούν,

- Παρά την ανεξαρτησία τους, θα πρέπει να σχηματίζουν μια συνεκτική ερευνητική κατεύθυνση.

Για να γίνει κατανοητό τι είδους ερωτήσεις εφαρμόζονται στη διερευνητική ανάλυση ο Filliben (1.3.2) παρουσιάζει τις πιο κοινές:

1. Ποια είναι η τυπική τιμή για κάποιο μέγεθος;
2. Ποιο είναι το μέγεθος της αβεβαιότητας για μία προσδιορισθείσα τυπική τιμή;
3. Ποια είναι μια καλή στατιστική κατανομή που ταιριάζει σε μεγάλο βαθμό σε ένα σύνολο αριθμών;
4. Ποια είναι η κατανομή των αριθμών σε ποσοστημόρια (percentiles);
5. Είχε επίπτωση μία αλλαγή μηχανικής στις μετρήσεις (π.χ. αν έχουν ληφθεί μετρήσεις προ και κατόπιν της αλλαγής μιας διαδικασίας);
6. Έχει κάποιος συγκεκριμένος παράγοντας επίδραση (π.χ. ο καιρός στην κατανάλωση γάλακτος);
7. Ποιοι είναι οι πιο σημαντικοί παράγοντες;
8. Είναι ισοδύναμες μετρήσεις που προέρχονται από διαφορετικές πηγές;
9. Ποια συνάρτηση συσχετίζει με βέλτιστο τρόπο μια εξαρτώμενη μεταβλητή με ένα σύνολο ανεξάρτητων μεταβλητών;
10. Ποιες είναι οι καλύτερες ρυθμίσεις για τις τιμές των παραμέτρων και υπερ-παραμέτρων που σχετίζονται με τους παράγοντες;
11. Μπορεί να διαχωρισθεί το σήμα από το θόρυβο σε δεδομένα που εξαρτώνται από το χρόνο;
12. Μπορεί να εξαχθεί οποιαδήποτε δομή από πολυμεταβλητά δεδομένα (multivariate data);
13. Έχουν τα δεδομένα ακραίες τιμές;

Για τη διατύπωση των ερευνητικών ερωτήσεων ανάλυσης, είναι σκόπιμο να χρησιμοποιείται ένα κατάλληλο λεξιλόγιο και ένας συγκεκριμένος τρόπος έκφρασης των ερωτήσεων. Στην εργασία [10] προτείνεται ένας συστηματικός τρόπος ανάπτυξης του στατιστικού τρόπου σκέψης για ερευνητικές ερωτήσεις, ο οποίος περιλαμβάνει τη χρήση κατευθυντήριων ρημάτων όπως επηρεάζει, καθορίζει, σχετίζεται, κ.λπ. ενώ προτείνεται να αποφεύγονται εκφράσεις που έχουν διερευνητική έννοια, όπως ανακάλυψη, αναζήτηση, εξερεύνηση, περιγραφή, αναφορά, κ.λπ.

### **2.3.2 Η χρήση στατιστικών τεχνικών παρατήρησης και γραφικών προσεγγίσεων**

Η προσέγγιση του σώματος των δεδομένων μπορεί να θεωρηθεί ως η ανάλυση αριθμητικών τιμών που σχετίζονται με ένα θέμα προς εξέταση, δηλαδή το περιεχόμενο τους να εστιάζει στη κατανόηση των δεδομένων με όρους:

- παραδοχών που αφορούν τη διαδικασία του ελέγχου,
- επιλογής μοντέλου,
- επικύρωσης μοντέλου,
- επιλογής εκτιμητών [12],
- αναγνώρισης των σχέσεων,
- προσδιορισμού βαθμού επίπτωσης των διάφορων παραγόντων,
- ανίχνευση ακραίων τιμών

Για κάθε ένα από τους παραπάνω όρους χρησιμοποιούνται στατιστικές μέθοδοι με την αναπαράσταση/οπτικοποίηση μέσω γραφικών, ώστε να αναδεικνύουν την υποκειμενική εξήγηση των δεδομένων [2]. Η ποσοτική ανάλυση δεν επαρκεί, άπαυτής για να καλύψει τους στόχους της διερευνητικής ανάλυσης, υπό την έννοια ότι οι αριθμητικές τιμές να μην ποσοτικοποιούνται μαθηματικά, όμως δεν περιγράφουν στο σύνολό τους τους όρους που περιγράφονται παραπάνω. Το αντικείμενο της διερευνητικής ανάλυσης είναι η καταγραφή μεταξύ των συγκλίσεων των μετρήσεων έως να καταστεί σαφές ότι από τα δεδομένα μπορεί να κατασκευαστεί βελτιστοποιημένο μοντέλο κατάλληλο για την διατύπωση έγκυρων και επικυρωμένων επιστημονικών και μηχανικών συμπερασμάτων με τη χρήση των γραφικών αναπαραστάσεων. Γραφικές τεχνικές για την αναπαράσταση των δεδομένων που εξετάζονται αναφέρονται στο εγχειρίδιο στατιστικής του οργανισμού NIST στην ενότητα 1.3.3 [2].

### **2.3.3 Διερευνητική ανάλυση, έλεγχος υποθέσεων και βελτιστοποιήσεις**

Μολονότι η διερευνητική ανάλυση των δεδομένων είναι ένα σημαντικό συστατικό της όλης διαδικασίας, είναι σημαντικό να διαχωριστεί σαφώς από τον έλεγχο των υποθέσεων. Οι αποφάσεις σχετικά με τα μοντέλα που πρέπει να ελεγχθούν πρέπει να γίνονται εκ των προτέρων, με βάση την πλήρη (κατά το δυνατόν) κατανόηση του ερευνητή αναφορικά με το σύστημα/το πεδίο το οποίο αφορούν τα δεδομένα. Σε περίπτωση που η κατανόηση δεν έχει αναπτυχθεί ακόμη σε επαρκή βαθμό, η διερεύνηση των δεδομένων μπορεί να αξιοποιηθεί στο πλαίσιο της δοκιμαστικής

διαμόρφωσης υποθέσεων που πρόκειται να επαληθευθούν, αλλά και πάλι τα στάδια της διατύπωσης υποθέσεων και του ελέγχου των υποθέσεων αυτών είναι διαφορετικά. Επί παραδείγματι, ο εντοπισμός μοτίβων κατά τη διερευνητική ανάλυση μπορεί να μας δώσει τη βάση και το κίνητρο για την ύπαρξη των μοτίβων αυτών, ωστόσο θα πρέπει να αποφευχθεί η άμεση εξαγωγή συμπερασμάτων για το αν και σε ποια έκταση ισχύουν τα μοτίβα: για την επαλήθευση της ύπαρξης των μοτίβων στο σύνολο των δεδομένων ή σε υποσύνολο αυτών, θα πρέπει να ακολουθηθεί ο κύκλος της διατύπωσης και επαλήθευσης υποθέσεων με χρήση στατιστικής ανάλυσης και ελέγχων ανεξαρτησίας. Είναι σημαντικό όταν γίνεται μία δήλωση για τα δεδομένα και το μοντέλο που αυτά ακολουθούν, να διατυπώνεται με σαφήνεια εάν η δήλωση βασίζεται σε διαίσθηση που προέκυψε από τη διερευνητική ανάλυση ή σε αποτελέσματα στατιστικών ελέγχων και να δίνεται το πλήρες πλαίσιο που μας οδήγησε στη διατύπωση αυτή.

Ένα συναφές πλαίσιο εργασίας παρέχεται από την εργασία [13], όπου προτείνεται ένα πρωτόκολλο για τη διερευνητική ανάλυση δεδομένων, το οποίο δύναται να ανιχνεύσει κοινά σφάλματα στη διαδικασία που ακολουθείται. Το πρωτόκολλο περιλαμβάνει τα ακόλουθα βήματα:

1. Εύρεση έκτοπων τιμών (outliers)
2. Έλεγχος ομογένειας της διακύμανσης
3. Έλεγχος για ύπαρξη κανονικής κατανομής
4. Ύπαρξη μεγάλου πλήθους μηδενικών τιμών
5. Ύπαρξη συγγραμμικότητας μεταξύ των μεταβλητών ελέγχου (covariates)
6. Εξέταση των συσχετίσεων και αλληλεπιδράσεων μεταξύ των μεταβλητών
7. Έλεγχος αν πληρούνται οι προϋποθέσεις ανεξαρτησίας στις μετρήσεις των μεταβλητών

### 3 Το πλαίσιο αρχών FAIR για τα ανοικτά δεδομένα

Η αυξανόμενη διάθεση και χρήση ανοιχτών δεδομένων στο διαδίκτυο έχει δημιουργήσει ανάγκες ως προς τη δυνατότητα εύρεσης και αξιοποίησής των δεδομένων. Στο πλαίσιο αυτό, η πρωτοβουλία GOFAIR [6] έχει διατυπώσει ένα σύνολο αρχών που θα πρέπει να διέπουν τη διάθεση των ανοικτών δεδομένων έτσι ώστε να μεγιστοποιείται η αξία και η χρησιμότητά τους για την κοινότητα. Οι κύριοι άξονες του πλαισίου αυτού αφορούν τις διαστάσεις της Ευρεσιμότητας, Προσβασιμότητα, Διαλειτουργικότητας, Επαναχρησιμοποίησης (Findability, Accessibility, Interoperability, Reusability - FAIR).

Οι αρχές του πλαισίου GOFAIR επιτρέπουν την εύρεση, ανάκτηση, ερμηνεία και χρήση των δεδομένων, , επιτρέποντας έτσι την αξιοποίηση των δεδομένων για ανάλυση, εξαγωγή συμπερασμάτων και ανάπτυξη εφαρμογών προστιθέμενης αξίας, είτε με μεμονωμένα σύνολα δεδομένων είτε συνδυάζοντας σύνολα δεδομένων σύμφωνα με το υπόδειγμα του Linked Open Data (LOD) [15]. Οι διαστάσεις FAIR αναλύονται στις επόμενες παραγράφους:

- *Ευρεσιμότητα (Findability)*: Η πρώτη ανάγκη αναφέρεται στη φάση εύρεσης των δεδομένων, όπου ιδιαίτερα σημαντικό είναι να καταγραφούν οι περιγραφές τους μέσω μετα-δεδομένων (Metadata).
- *Προσβασιμότητα (Accessibility)*: μετά την εύρεση των δεδομένων, είναι σημαντικό να μπορούν να ανακτηθούν και να υποβληθούν σε επεξεργασία. Έτσι τα δεδομένα θα πρέπει να είναι ελεύθερα διαθέσιμα για ανάκτηση μέσω τυποποιημένων δικτυακών πρωτοκόλλων (π.χ. HTTP), να μην έχουν περιορισμούς πρόσβασης (συμπεριλαμβάνοντας τυχόν εξουσιοδοτήσεις), να διατίθενται σε ευρέως διαδεδομένους μορφότυπους (formats), και να μην εισάγονται περιορισμοί από την άδεια χρήσης.
- *Διαλειτουργικότητα (Interoperability)*: τα ανοικτά δεδομένα θα πρέπει να μπορούν να χρησιμοποιηθούν από ένα ευρύ σύνολο γλωσσών προγραμματισμού, εφαρμογών επεξεργασίας και ανάλυσης και συστημάτων, ενώ θα πρέπει να μπορούν να επεξεργαστούν συνδυαστικά με άλλα σύνολα δεδομένων με τρόπο που να επιτρέπει το ταίριασμα όμοιων οντοτήτων που βρίσκονται στα δύο σύνολα και τη συνδυαστική ερμηνεία και ανάλυση των περιεχομένων των συνόλων δεδομένων. Για τον σκοπό αυτό είναι απαραίτητη

η χρήση ευρέως διαδεδομένων και αποδεκτών μορφοτύπων (formats), καθώς και τυποποιημένων προσδιοριστών για οντότητες (π.χ. ονόματα χωρών, κωδικοί περιοχών κ.λπ.) που να επιτρέπουν την σύνδεση μεταξύ οντοτήτων σε διαφορετικά σύνολα δεδομένων.

- *Επαναχρησιμοποιησιμότητα (Reusability)*: Τα ανοικτά δεδομένα πρέπει να μπορούν να επαναχρησιμοποιηθούν σε μεγάλο εύρος περιβαλλόντων. Κατά συνέπεια θα πρέπει να περιγράφονται με ευρύ σύνολο χαρακτηριστικών (attributes) τα οποία να είναι σχετικά με τα δεδομένα και ακριβή, να διατίθενται με άδεια χρήσης που να είναι σαφής, προσβάσιμη και να παρέχει ευρύ σύνολο δικαιωμάτων, να περιγράφεται πλήρως η προέλευσή τους και η διαδικασία παραγωγής τους και να συμμορφώνονται με τα πρότυπα που ισχύουν στο σχετικό πεδίο και που χρησιμοποιούνται από την οικεία κοινότητα.

Οι αρχές του πλαισίου αναφέρονται σε τρεις διακριτούς τύπους οντοτήτων: Στα δεδομένα, στις περιγραφές τους και την υποδομή τους. Η σημασιολογία των ανοιχτών συνόλων δεδομένων αναφέρεται στην εργασία [14] κεφάλαιο 1 ως η διαδικασία για Linked Open Data που επιτρέπει την απόκτηση περισσότερης γνώσης για τα δεδομένα για το τομέα που αναφέρονται χρησιμοποιώντας ανοιχτά πρωτόκολλα. Πολλές από τις διαδικασίες παρέχονται αυτοματοποιημένα από διάφορα τεχνολογικά εργαλεία και συστήματα, άλλες αφορούν τη διαχείριση, την οργάνωση και την ανάγκη για περαιτέρω επεξεργασία. Στην παρούσα εργασία η διερευνητική ανάλυση των δεδομένων δεν προϋποθέτει ότι τα ανοικτά δεδομένα ακολουθούν το πλαίσιο αρχών FAIR: αν τα ανοικτά δεδομένα ακολουθούν το πλαίσιο αρχών FAIR, τότε μπορούν να ανακτηθούν και να συνδυαστούν σύνολα δεδομένων με αυτοματοποιημένο τρόπο μέσω των τεχνολογικών εργαλείων, διευκολύνοντας έτσι τη διερευνητική ανάλυση. Αν τα δεδομένα δεν ακολουθούν το πλαίσιο αρχών FAIR, τότε και πάλι είναι δυνατή η αξιοποίησή τους στο πλαίσιο της διερευνητικής ανάλυσης, ωστόσο ένα σύνολο εργασιών προεπεξεργασίας, ερμηνείας και ομογενοποίησης θα πρέπει να πραγματοποιηθούν με ανθρώπινη παρέμβαση.

## 4 Εργαλεία στατιστικής επεξεργασίας και οπτικοποίησης

### 4.1 Γλώσσα R

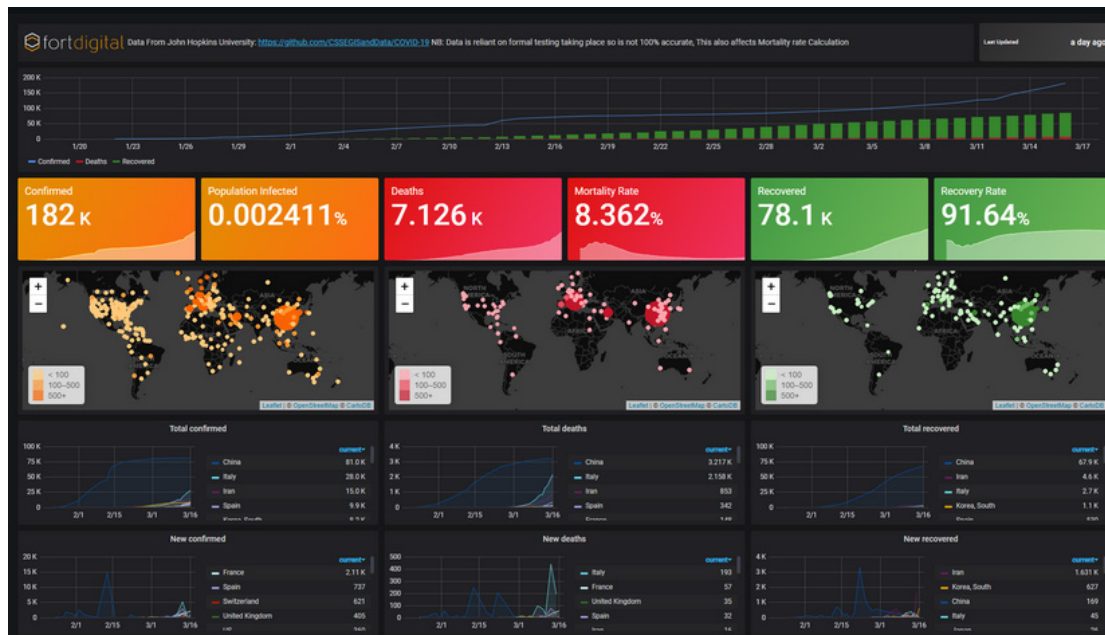
Η γλώσσα προγραμματισμού R, η οποία είναι προσανατολισμένη στη στατιστική επεξεργασία δεδομένων, επιτρέπει την επεξεργασία των δεδομένων και την εξαγωγή συμπερασμάτων μέσα από ένα περιβάλλον το οποίο παρέχει τη δυνατότητα συγγραφής κώδικα εμπλουτισμένου με χρήση στατιστικών μεθόδων και τεχνικών οπτικοποίησης. Είναι ένα έργο ανοιχτού κώδικα και το περιβάλλον εργασίας τυπικά περιλαμβάνει το R Studio, ένα πρόγραμμα με δυνατότητες οργάνωσης και συγγραφής κώδικα με τη δομή της γλώσσας R. Απαιτεί καλή γνώση των εννοιών του προγραμματισμού, ενώ επιπροσθέτως μπορεί να συνδυαστεί με άλλες γλώσσες προγραμματισμού όπως η SQL προκειμένου για τη σύνδεση με Βάσεις Δεδομένων και την ανάκτηση και τη διαχείριση των δεδομένων που αυτές περιέχουν. Υπάρχουν προγραμματιστικά περιβάλλοντα όπως το Visual Studio της Microsoft, IntelliJ της JetBrains, το Eclipse κα. τα οποία μπορούν και ενσωματώνουν τη γλώσσα R με την κατάλληλη δόμηση σε έργα που υποστηρίζει το εκάστοτε περιβάλλον και επίσης μπορεί να συνδυαστεί με άλλες γλώσσες. Η γλώσσα R έχει διαθέσιμο ένα ευρύ σύνολο βιβλιοθηκών, από το οποίο μπορεί να επιλεγεί το κατάλληλο υποσύνολο, έτσι ώστε σε κάθε περίπτωση να είναι διαθέσιμα τα απαραίτητα εργαλεία για τη υποστήριξη της εργασίας με τρόπο που ταιριάζει στη στατιστική επεξεργασία των δεδομένων που επιθυμούμε να εφαρμοστεί.

### 4.2 Grafana

Το Grafana είναι ένα λογισμικό ανοιχτού κώδικα που έχει δομηθεί με βάση την αρχιτεκτονική πελάτη-εξυπηρετητή (client-server) και παρέχει λειτουργικότητα σχετική με την οπτικοποίηση δεδομένων και την ανάλυση αυτών. Η τεχνολογία του επιτρέπει τη σύνδεση με Βάσεις Δεδομένων, την άντληση/ανάκτηση και οπτικοποίηση των δεδομένων, καθώς και την ανάλυση αυτών. Η υλοποίηση των λειτουργιών αυτών πραγματοποιείται μέσω προγραμματισμού με τη γλώσσα SQL ή/και με την παραμετροποίηση προτύπων JSON της γλώσσας JavaScript. Οι δυνατότητες που προσφέρει καλύπτουν ένα κενό που υπήρχε στην οπτικοποίηση των δεδομένων που αποθηκεύονται σε βάσεις δεδομένων και διατίθεται ως ένα τεχνολογικό εργαλείο αξιοποίησης των δεδομένων μέσω της προ-επεξεργασίας (preprocessing) και της

ανάλυσης των επιλεγμένων δεδομένων (data analysis). Υπάρχουν και οι εκδόσεις που απαιτούν συνδρομή, οι οποίες ενσωματώνουν επιπλέον δυνατότητες και αφορούν κυρίως επιχειρήσεις στον χώρο του διαδικτύου των πραγμάτων (IoT) με τα κύρια χαρακτηριστικά της εμπλουτισμένης πλατφόρμας να αφορούν αναλύσεις χρονοσειρών, με βασική εφαρμογή τις μετρήσεις που λαμβάνονται από κάποιο ελεγκτή ή κάποια συσκευή IoT. Στην παρούσα εργασία η χρήση του εργαλείου Grafana γίνεται με στόχο την οπτικοποίηση δεδομένων από τη Βάση Δεδομένων MySQL και την αναπαράσταση σε δυναμικούς πίνακες γραφημάτων (dashboard) ώστε να γίνει η διερευνητική ανάλυση των προτυποποιημένων μεταβλητών και των επιλεγμένων συνόλων δεδομένων από τους τελικούς χρήστες, ανάλογα με την εκάστοτε στατιστική μέθοδο. Η επιλογή οπτικοποιήσεων και σχετικών παραμέτρων στο πλαίσιο της διερευνητικής ανάλυσης μπορεί να πραγματοποιείται μέσω ενός drop down menu και εν γένει ο σκοπός του χρήστη είναι η αποτελεσματική εξαγωγή μοτίβων από τα δεδομένα. Η χρήση προγραμματισμού από τους τελικούς χρήστες δεν είναι απαραίτητη, παρ' όλα αυτά παρέχεται η δυνατότητα παραμετροποίησης του κώδικα ώστε να είναι δυνατή η αναπροσαρμογή των συνόλων δεδομένων ή των παραμέτρων της οπτικοποίησης από τους χρήστες.

Στο παρακάτω παράδειγμα διακρίνονται 3 δυναμικοί πίνακες που έχουν δημιουργηθεί μέσω του Grafana, οι οποίοι περιλαμβάνουν δεδομένα που αφορούν τη πορεία ιού Covid. Συγκεκριμένα καταγράφονται το πλήθος και το ποσοστό των κρουσμάτων, το πλήθος και το ποσοστό των θανάτων και το πλήθος και το ποσοστό των ιαθέντων περιστατικών ανά χώρα. Τα δεδομένα που χρησιμοποιούνται για το γράφημα ενημερώνονται με δυναμικό τρόπο μέσω API από φορείς που λειτουργούν σε κάθε χώρα. Η εταιρία που ανέπτυξε το λογισμικό παρέχει πλήρες αναφορά στο τρόπο που συλλέγονται και υπόκεινται σε επεξεργασία τα δεδομένα μέσω της ιστοσελίδας <https://github.com/FortDigital/covid-19>.

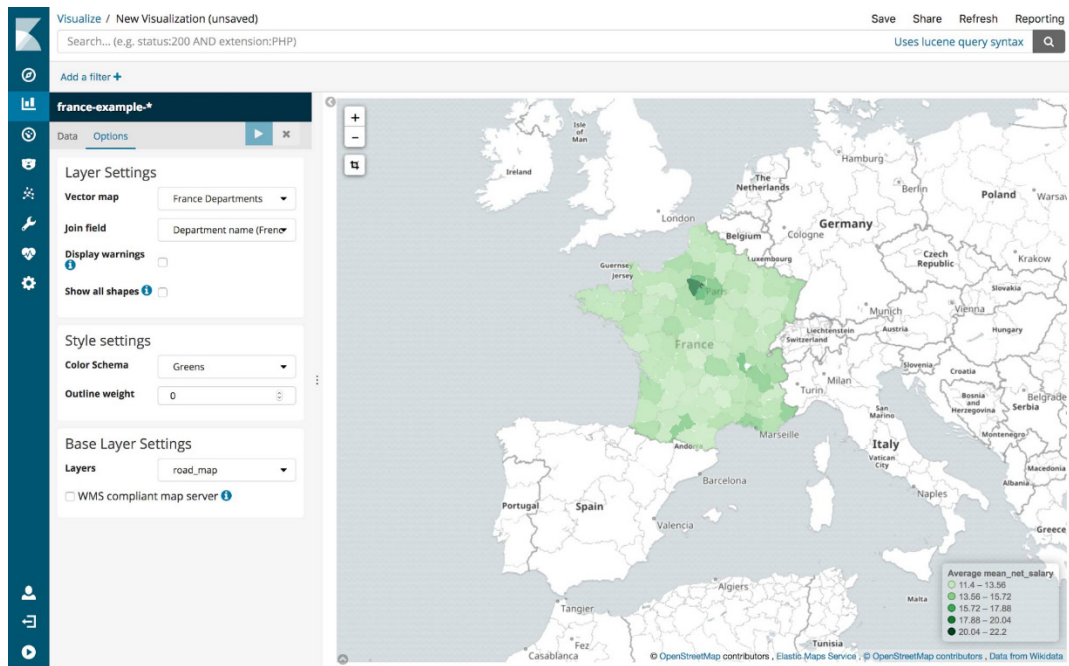


Εικόνα 3. Γραφήματα για την πανδημία Covid-19 με χρήση του Grafana (πηγή: <https://github.com/FortDigital/covid-19>)

### 4.3 Kibana

Το Kibana είναι ένα τεχνολογικό εργαλείο που είναι κυρίως προσανατολισμένο στην αναπαράσταση/οπτικοποίηση δεδομένων που αποθηκεύονται στη βάση δεδομένων Elasticsearch. Το Elasticsearch είναι μία μηχανή αναζήτησης για κάθε κατηγορία δεδομένων. Διατίθεται επί πληρωμή με πλήρες σύνολο λειτουργιών, αλλά υπάρχουν και δωρεάν διατιθέμενες εκδόσεις στις οποίες κάποιες λειτουργίες δεν είναι διαθέσιμες. Οι κύριες λειτουργίες του αφορούν την αναπαράσταση των δεδομένων σε γραφήματα με σκοπό την ανάλυση και την ερμηνεία αυτών.

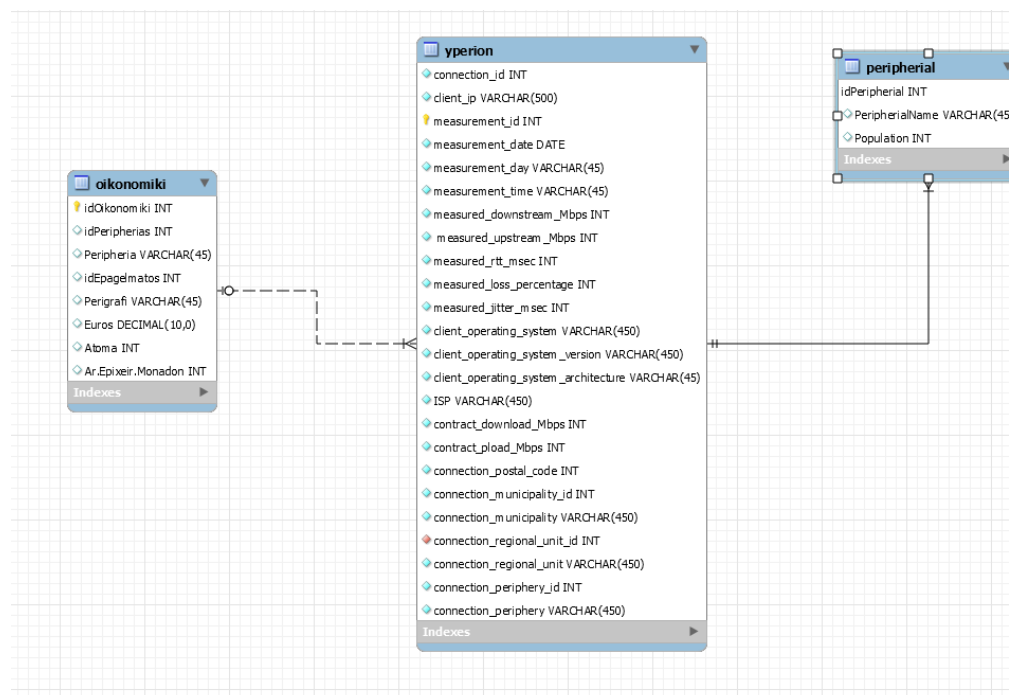
Ένα παράδειγμα οπτικοποίησης των μισθών στη Γαλλία με χρήση του Kibana φαίνεται στην Εικόνα 4.



Εικόνα 4. Οπτικοποίηση των μισθών στη Γαλλία με χρήση του Kibana (πηγή: <https://www.elastic.co/jp/blog/visualizing-france-salary-data-with-region-maps-in-kibana?blade=fbs>)

## 5 Προσχέδιο δεδομένων και διεργασιών (Blueprint and Data pre processing)

Στη μελέτη περίπτωσης της παρούσας εργασίας εξετάζεται το σύνολο των δεδομένων που παρέχεται από την εφαρμογή της ΕΕΤΤ «Υπερίων» για το πρώτο εξάμηνο του έτους 2021, λαμβάνοντας υπ' όψιν και την ενσωμάτωση της απογραφής πληθυσμού ανά περιφέρεια της ΕΛΣΤΑΤ που πραγματοποιήθηκε το έτος 2021. Τα δεδομένα που αφορούν την οικονομική δραστηριότητα ανά περιφέρεια του μητρώου επιχειρήσεων της ΕΛΣΤΑΤ το έτος 2016 ενσωματώθηκαν για την ανάδειξη του κλάδου που θα επιχειρηθεί να ανακαλύψει νέα προϊόντα και υπηρεσίες που βασίζονται στο διαδίκτυο και θα ερευνηθεί αν η κλίμακα της ταχύτητας των συνδέσεων εξυπηρετεί την ανάδειξη αυτή ή όχι. Τα σύνολα δεδομένων συλλέχθηκαν σε μορφή CSV και εισήχθησαν σε RDBMS (Relational Data Base Management System) με βάση το παρακάτω σχήμα ERD (Entity Relation Diagram).



Εικόνα 5. Διάγραμμα οντοτήτων-συσχετίσεων για τα δεδομένα της εφαρμογής

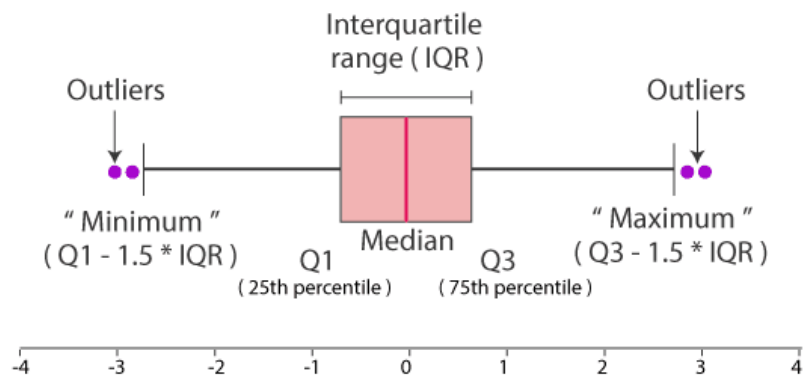
Συνολικά χρησιμοποιήθηκαν:

- 55.000 εγγραφές οργανωμένες σε τρεις πίνακες στη Σχεσιακή Βάση Δεδομένων της MySql.

- 16 μεταβλητές χρησιμοποιήθηκαν για να περιγράψουν την ταχύτητα ροής καθόδου των συνδέσεων ανά πάροχο σύνδεσης ανά περιφέρεια.
- Τα χαρακτηριστικά των συνόλων των δεδομένων, που αποτελούν τους παράγοντες της ανάλυσης, αντιπροσωπεύουν τη γεωγραφική κατανομή για τον υπολογισμό της μέσης ταχύτητας ροής καθόδου, λαμβάνοντας υπόψη τη μέση τιμή των μετρήσεων των χρηστών ανά πάροχο τηλεπικοινωνιών.
- Στη συνέχεια τα σύνολα δεδομένων ελέγχονται για τυχόν ανωμαλίες.
- Εάν τα σύνολα δεδομένων έχουν κανονική κατανομή εφαρμόζεται η τεχνική του διατεταρτημοριακού διαστήματος (Interquartile Range, IQR) για τον εντοπισμό ακραίων τιμών (outliers). Ο εντοπισμός των ακραίων τιμών με χρήση της τεχνικής του διατεταρτημοριακού διαστήματος πραγματοποιείται ως ακολούθως:
  - Διατάσσονται οι τιμές του συνόλου δεδομένων και υπολογίζεται ο διάμεσος.
  - Ο διάμεσος χωρίζει το σύνολο των δεδομένων σε δύο υποσύνολα, (α) το υποσύνολο L που προηγείται του διαμέσου στη διάταξη και το υποσύνολο H που έπεται του διαμέσου στη διάταξη.
  - Για το υποσύνολο L, υπολογίζεται ο διάμεσός του που συμβολίζεται με Q1.
  - Για το υποσύνολο H, υπολογίζεται ο διάμεσός του που συμβολίζεται με Q3.
  - Υπολογίζεται το διατεταρτημοριακό διάστημα  $IQR = Q3 - Q1$ .
  - Μία τιμή  $x$  του αρχικού συνόλου δεδομένων χαρακτηρίζεται ως *ακραία* εάν ισχύει:

$$x < Q1 - 1.5 * IQR \vee Q3 + 1.5 * IQR < x$$

Η ανίχνευση ακραίων τιμών με χρήση της τεχνικής IQR απεικονίζεται στην Εικόνα 6.



Εικόνα 6. Ανίχνευση ακραίων τιμών με την τεχνική IQR (πηγή: <https://www.cloudymml.com/blog/outlier-detection-and-treatment/>)

## 6 Μελέτη περίπτωσης - Διερευνητική ανάλυση δεδομένων (Case study - Exploratory data analysis)

### 6.1 Υλοποίηση.

Σε συνέχεια της προεπεξεργασίας των δεδομένων που αναπτύχθηκε στο προηγούμενο κεφάλαιο, υλοποιήθηκε σύστημα οπτικοποίησης για την εφαρμογή των βημάτων διερευνητικής ανάλυσης με το λογισμικό Grafana. Η ανάπτυξη σχεσιακής Βάσης δεδομένων στο σύστημα της MySQL, η σύνδεση της Βάσης Δεδομένων με το λογισμικό Grafana δημιουργεί μια πλατφόρμα που βασίζεται στην αρχιτεκτονική client-server. Η χρήση της προγραμματιστικής γλώσσας SQL αφορά την ανάλυση των δεδομένων σε επίπεδο διερεύνησης όσον αφορά:

1. την εύρεση των μοτίβων στα δεδομένα που εξετάζονται,
2. την ενίσχυση της αποτελεσματικότητας των μοντέλων, καλύπτοντας ένα μεγάλο εύρος.

Η κεντροποιημένη προσέγγιση στη διερευνητική ανάλυση βασίστηκε στο βιβλίο του Antonio Badia [9]. Η ανάπτυξη ενός συστήματος και των εφαρμογών του, τοποθετώντας τα δεδομένα σε κεντρικό ρόλο στη ροή των διεργασιών, διευκολύνεται από την τοποθέτηση και οργάνωση των δεδομένων σε μία βάση δεδομένων. Στην παρούσα εργασία, χρησιμοποιήθηκε μία βάση δεδομένων σχεσιακού τύπου. Η αλληλεπίδραση με τη βάση δεδομένων πραγματοποιήθηκε με χρήση της γλώσσας SQL.

Η προγραμματιστική γλώσσα SQL είναι μία ιδιαίτερα ισχυρή γλώσσα διαχείρισης δεδομένων, μέσω της οποίας με δηλωτικό τρόπο ορίζεται ο τρόπος επεξεργασίας των δεδομένων και η δομή και η μορφή των αποτελεσμάτων. Στο πλαίσιο της παρούσας εργασίας, τα ερωτήματα SQL συνδυάζουν δεδομένα από διαφορετικές πηγές, ενώ ερωτήματα SQL έχουν επίσης χρησιμοποιηθεί και για τον καθαρισμό και την προεπεξεργασία των δεδομένων. Εν γένει η SQL μπορεί να χρησιμοποιηθεί σε διαφορετικού τύπου αριθμητικές και γραφικές μεθόδους όπως:

- *Αριθμητικές μέθοδοι*
  1. Μέτρα σχετικής θέσης
    - Τεταρτημόρια – (Binning with quartiles)
    - Προβληματικές τιμές – (Dummy variables)

## 2. Μέτρα μεταβλητότητας

Δειγματική τυπική απόκλιση (Sample standard deviation)

Ενδοτεταρτημοριακό εύρος IQR (Interquartile Range)

- *Γραφικές μέθοδοι*

1. Ιστογράμματα

2. Θηκογράμματα και απομακρυσμένα σημεία

Το επόμενο παράδειγμα εξετάζει αριθμητικά την κατανομή πληθυσμού σε σχέση με την ταχύτητα ροής καθόδου και τους παρόχους, περιορισμένα στους 10 πρώτους χρήστες.

```
Select ISP, population, peripheralName
from yperion inner join peripheral on
yperion.connection_periphery_id=peripheral.idPeripheral
order by measured_downstream_Mbps desc
limit 10
```

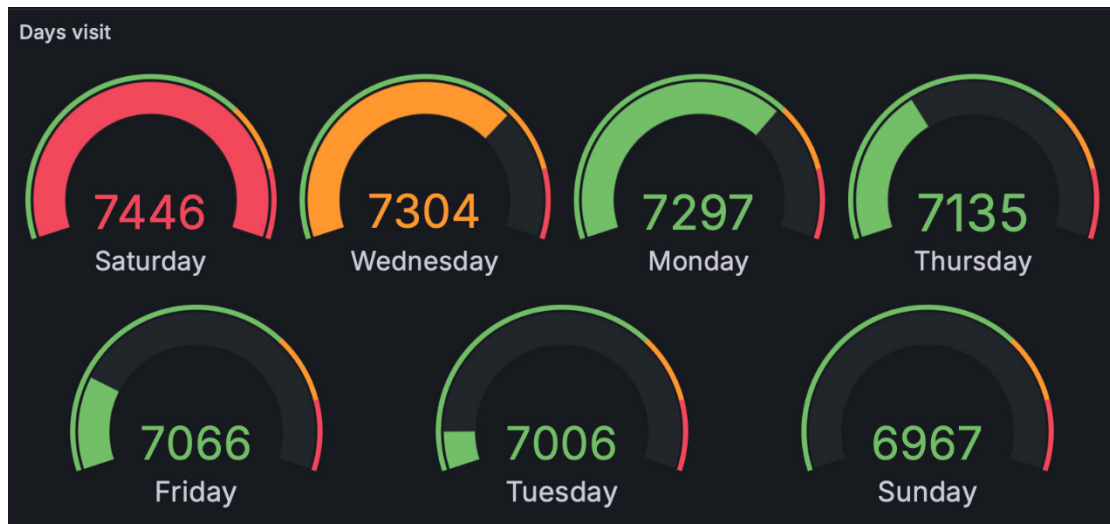
Σε συνέχεια η εξέταση αφορά την επέκταση της εύρεσης κατάλληλης λύσης με μετασχηματισμούς και γραφικές παραστάσεις μέχρις ότου ερευνηθούν οι σχέσεις και τα μοτίβα τα οποία ίσως υπάρχουν. Η οπτικοποίηση με γραφικές και αριθμητικές παραστάσεις είναι το κύριο αντικείμενο της εργασίας συχνά με επαναληπτική σειρά (iterative process) για την αντιμετώπιση ενός ερωτήματος όπου συνδυάζει στατιστικές μεθόδους οι οποίες είτε έχουν αναλυθεί στη διερευνητική ανάλυση είτε πρέπει να καλυφθούν ξανά.

## 6.2 Ερωτήματα διερευνητικής ανάλυσης με χρήσης της SQL

Στη συνέχεια καταγράφονται ερωτήματα διερευνητικής ανάλυσης που εισήχθησαν έναντι της βάσης δεδομένων προκειμένου για τη δημιουργία οπτικοποιήσεων και την εξαγωγή συμπερασμάτων.

1. Ποιες οι ημέρες που επισκέφθηκαν οι χρήστες την εφαρμογή του ΥΠΕΡΙΩΝ.

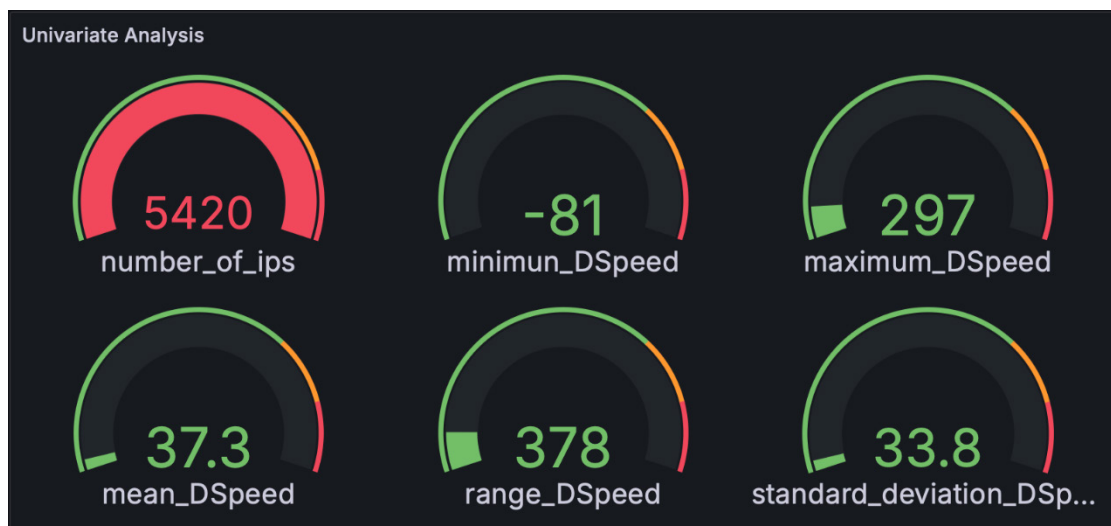
```
select measurement_day as day_visit,
count(*) as number_visits
from yperion
group by day_visit
order by number_visits desc;
```



Εικόνα 7. Μετρήσεις ανά ημέρα

2. Μονοπαραμετρική ανάλυση (univariate analysis), εξέταση περιγραφικής στατιστικής για τη μεταβλητή ροής καθόδου (measured\_downstream\_Mbps).

```
select count (distinct client_ip) as number_of_ips,
       min(measured_downstream_Mbps) as minimum_DSspeed,
       max(measured_downstream_Mbps) as maximum_DSspeed,
       avg (measured_downstream_Mbps) as mean_DSspeed,
       max(measured_downstream_Mbps) -
       min(measured_downstream_Mbps) as range_DSspeed,
```



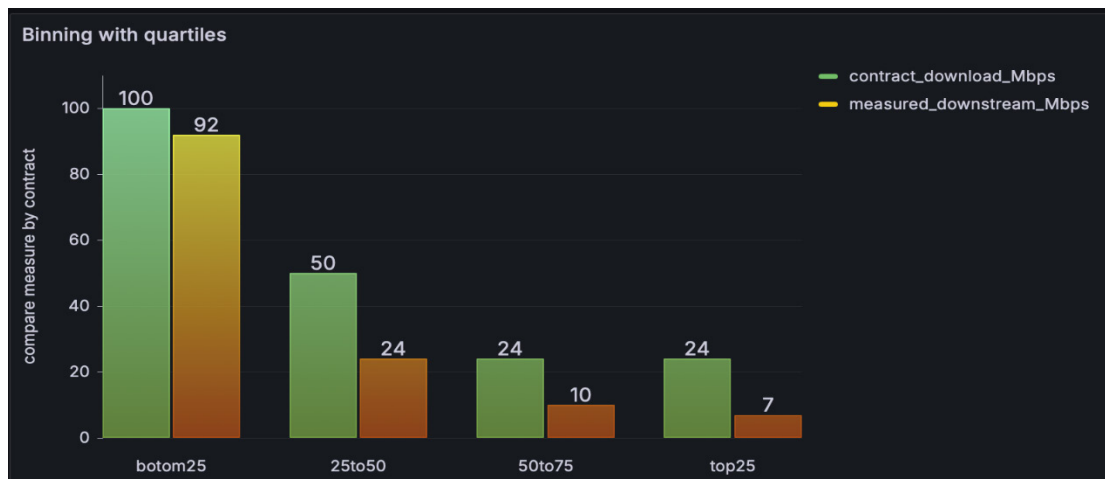
Εικόνα 8. Περιγραφική στατιστική

Από τις μετρήσεις της περιγραφικής στατιστικής που διατυπώνονται παραπάνω, υπάρχουν ακραίες/α τιμές/ή στις ελάχιστες ταχύτητες ροής καθόδου, καθώς η τιμή -81 δηλώνει πως είτε υπάρχει σφάλμα στη μέτρηση, είτε κάποια άλλη αιτία έχει εισάγει παραθορά στα δεδομένα στη χρήση αυτής

τιμής της μεταβλητής. Επίσης η διακύμανση, το εύρος και η διάμεσος εξετάζονται για τις τιμές της μεταβλητής ροής καθόδου.

### 3. Διερεύνηση των συμβολαίων ταχύτητας και των μετρήσεων που έχουν γίνει.

```
WITH OrderedData as
(SELECT contract_download_Mbps, measured_downstream_Mbps
FROM grafana.yperion,
(SELECT count(*)/4 as number from grafana.yperion) as T
order by measured_downstream_Mbps)
(select 'botom25' as quartile, contract_download_Mbps,
measured_downstream_Mbps
from OrderedData
LIMIT 1)
UNION
(SELECT '25to50' as quartile, contract_download_Mbps,
measured_downstream_Mbps
FROM OrderedData
LIMIT 1 OFFSET 30)
UNION
(select '50to75' as quantile, contract_download_Mbps,
measured_downstream_Mbps
FROM OrderedData
LIMIT 1 OFFSET 100)
UNION
(select 'top25' as quartile,
contract_download_Mbps,measured_downstream_Mbps
from OrderedData
LIMIT 1 OFFSET 1000;
```



Εικόνα 9. Σύγκριση συμβολαίων ταχύτητας με μετρήσεις.

Τα αποτελέσματα των ερωτήσεων με τη χρήση της SQL.

- Τα συμβόλαια με ταχύτητα 100Mbps έχουν τη μικρότερη απόκλιση σε σχέση με την ταχύτητα καθόδου.
- Τη μεγαλύτερη απόκλιση την έχουν τα συμβόλαια με ταχύτητα 50Mbps.

- Τα συμβόλαια με 24Mbps έχουν τη μεγαλύτερη απήχηση και έχουν χωριστεί σε δυο τεταρτημόρια για την καλύτερη αποτύπωση της ανάλυσης. Οι διαφορές είναι πανομοιότυπες με αποκλίσεις μεσαίου βαθμού.

Σε συνέχεια της εξέτασης των δεδομένων πραγματοποιείται έλεγχος σε αριθμητικές μεθόδους, διάμεσος για εύρεση τοποθεσίας στο πληθυσμό και δυναμικές μετρήσεις μέσω αλληλεπίδρασης της ολοκληρωμένης πλατφόρμας που παρουσιάζετε σε πραγματικό χρόνο.



Εικόνα 10. Πλατφόρμα διερευνητικής ανάλυσης δεδομένων.

## 7 Συμπεράσματα

Στα συστήματα διερευνητικής ανάλυσης δεδομένων οι εργασίες χαρακτηρίζονται από ροές δεδομένων οι οποίες δημιουργούν καταστάσεις όπως φόρτωση δεδομένων (data loading), καθαρισμός δεδομένων (data cleaning), προ επεξεργασία (pre-processing). Ο μετασχηματισμός δεδομένων (data transformation) πραγματοποιήθηκε χρησιμοποιώντας διαφορετικά επίπεδα χρήσης της προγραμματιστικής γλώσσας SQL καθώς η συμβολή της σε διεργασίες ήταν απαραίτητη για την επιτυχή αποτύπωση της ανάλυσης των δεδομένων.



Εικόνα 11. Αποτύπωση επιτυχής διαδικασίας διερευνητικής ανάλυσης δεδομένων της εργασίας.

Στο παραπάνω διάγραμμα Ven κάθε σημείο αλληλεπίδρασης των δεδομένων αντικατοπτρίζεται από την εκάστοτε κατάσταση ροής των δεδομένων. Η επιτυχής ή μη αποτύπωση της ανάλυσης σύμφωνα με τη συμβολή της προγραμματιστικής γλώσσας SQL παρατηρείται στο διάγραμμα ως το κρίσιμο σημείο όπου επικαλύπτονται οι τρεις κύκλοι.

Κάθε ερευνητική ερώτηση βασίστηκε στην επιδίωξη να κατανοηθούν τα δεδομένα, να εξάχθουν κατάλληλα συμπεράσματα και να παρέχουν με ακρίβεια όλα τα στοιχεία τα οποία περιγράφει στο εγχειρίδιο του ο James J. Filliben [2].

## 8 Βιβλιογραφία

- [1] A. Ghosh, M. Nashaat, J. Miller, S. Quader, and C. Marston, “A comprehensive review of tools for exploratory analysis of tabular industrial datasets,” *Vis. Informatics*, vol. 2, no. 4, pp. 235–253, 2018.
- [2] James J. Filliben, “NIST/SEMATECH e-Handbook of Statistical Methods”.
- [3] G. Burrill and R. Biehler, “Fundamental Statistical Ideas in the School Curriculum and in Training Teachers,” in *Teaching Statistics in School Mathematics-Challenges for Teaching and Teacher Education: A Joint ICMI/IASE Study: The 18th ICMI Study*, C. Batanero, G. Burrill, and C. Reading, Eds. Dordrecht: Springer Netherlands, 2011, pp. 57–69.
- [4] J. T. Behrens, “Principles and Procedures of Exploratory Data Analysis,” *Psychol. Methods*, vol. 2, no. 2, pp. 131–160, 1997.
- [5] Jukka-Matti Turtiainen, “MENTAL MODEL FOR EXPLORATORY DATA ANALYSIS APPLICATIONS FOR STRUCTURED PROBLEM-SOLVING.”
- [6] GO FARE, “FAIR Principles - GO FAIR.” 2018. <https://www.go-fair.org/fair-principles/> [Πρόσβαση 24/11/2019].
- [7] I. J. Good, “The Philosophy of Exploratory Data Analysis,” *Philos. Sci.*, vol. 50, no. 2, pp. 283–295, 1983.
- [8] T. H. Davenport and J. Kim, *Keeping Up with the Quants*, 2013th ed., vol. 1. Harvard Business School Publishing Corporation, 2013.
- [9] Antonio Badia, *SQL for Data*, 2020/12/11. Springer Nature.
- [10] Η Στατιστική στο νέο Π.Σ. των Μαθηματικών: Πιθανές Δυσκολίες Εφαρμογής, Αριστούλα Κοντογιάννη και Ευγενία Κολέζα. (Σελ. 198-213). [http://www.mathlab.upatras.gr/wp-content/uploads/2012/09/FINAL\\_Conference\\_Proceedings\\_Oct\\_16\\_-\\_10\\_2012.pdf](http://www.mathlab.upatras.gr/wp-content/uploads/2012/09/FINAL_Conference_Proceedings_Oct_16_-_10_2012.pdf)
- [11] I. J. Good. Twenty-seven principles of rationality. In V. P. Godambe and D. A. Sprott, editors, *Foundations of Statistical Inference*, pages 108–141. Holt, Rinehart, Winston, Toronto, 1971.
- [12] [https://www.researchgate.net/publication/283787595\\_Estimator\\_selection](https://www.researchgate.net/publication/283787595_Estimator_selection)
- [13] Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, 1(1), 3–14. <https://doi.org/10.1111/j.2041-210x.2009.00001.x>

[14] Poblet Marta and Casanovas, P. and R.-D. V. (2019). Introduction to Linked Data. In *Linked Democracy: Foundations, Tools, and Applications* (pp. 1–25). Springer International Publishing. [https://doi.org/10.1007/978-3-030-13363-4\\_1](https://doi.org/10.1007/978-3-030-13363-4_1)

[15] Florian Bauer, Martin Kaltenböck. (2012). *Linked Open Data: The Essentials. A Quick Start Guide for Decision Makers*. ISBN: 978-3-902796-05-9