

Οπτικοποίηση και ανάλυση δεδομένων δεικτών

Κολιάτος Δημήτριος

<Διπλωματική Εργασία>

Επιβλέπων: Κωσταντίνος Μπλέκας

Ιωάννινα, Ιούλιος, 2022



ΤΜΗΜΑ ΜΗΧ. Η/Υ & ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

UNIVERSITY OF IOANNINA

Εξεταστική επιτροπή:

- **Μπλέκας Κωνσταντίνος**, Καθηγητής Τμήμα Μηχανικών Η/Υ και Πληροφορικής, Πανεπιστήμιο Ιωαννίνων (Επιβλέπων)
- **Βλάχος Κωνσταντίνος**, Επίκουρος Καθηγητής Τμήμα Μηχανικών Η/Υ και Πληροφορικής, Πανεπιστήμιο Ιωαννίνων
- **Λύκας Αριστείδης**, Καθηγητής Τμήμα Μηχανικών Η/Υ και Πληροφορικής, Πανεπιστήμιο Ιωαννίνων

Αφιέρωση

Αφιερωμένο στην οικογένειά μου.

Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω τους γονείς μου για την πολύτιμη στήριξη και βοήθεια που μου παρείχαν σε κάθε επίπεδο όλα αυτά τα χρόνια.

Επιπλέον, θα ήθελα να ευχαριστήσω τον επιβλέπων καθηγητή της διπλωματικής αυτής εργασίας, κ. Κωνσταντίνο Μπλέκα, Καθηγητής στο Τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής της Πολυτεχνικής Σχολής του Πανεπιστημίου Ιωάννινων, που επέβλεπε το έργο και με καθοδήγησε μέχρι την ολοκλήρωσή του.

Και τέλος τους φίλους μου που με συμβούλευαν σε κάθε μου βήμα.

Περιεχόμενα

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ	7
ΠΕΡΙΛΗΨΗ.....	10
ABSTRACT	11
ΚΕΦΑΛΑΙΟ 1: ΕΙΣΑΓΩΓΗ	12
1.1 Αντικείμενο διπλωματικής εργασίας.....	12
1.2 Εργαλεία που χρησιμοποιήθηκαν	12
1.3 Οργάνωση του τόμου.....	13
ΚΕΦΑΛΑΙΟ 2: ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ	14
2.1 Κατηγορίες Μάθησης.....	15
2.1.1 Μάθηση με επίβλεψη.....	16
2.1.2 Μάθηση χωρίς επίβλεψη	17
2.1.3 Ενισχυτική μάθηση	17
2.2 Κατηγορίες Ομαδοποίησης.....	18
2.2.1 Μέθοδοι που βασίζονται στην πυκνότητα.....	18
2.2.2 Μέθοδοι με βάση την ιεραρχία.....	18
2.2.3 Μέθοδοι κατάτμησης	19
2.2.4 Μέθοδοι που βασίζονται σε πλέγμα.....	19
2.3 Ομαδοποίηση με τον αλγόριθμο K-means	19
2.3.1 Λειτουργία K-means	20
2.3.2 Αλγόριθμος K-means	21
2.3.3 Υπολογισμός βέλτιστης K τιμής	21
ΚΕΦΑΛΑΙΟ 3: ΥΛΟΠΟΙΗΣΗ ΑΛΓΟΡΙΘΜΟΥ ΟΜΑΔΟΠΟΙΗΣΗΣ	23
3.1 Αρχικοποίηση παραμέτρων	23
3.2 Μορφή δεδομένων προς ομαδοποίηση.....	24
3.3 Εκτέλεση K-means.....	25
3.4 Υλοποίηση Elbow	25
ΚΕΦΑΛΑΙΟ 4: ΔΕΔΟΜΕΝΑ ΚΑΙ ΟΠΤΙΚΟΠΟΙΗΣΗ	27
4.1 Οπτικοποίηση δεδομένων	28
4.2 Πηγή δεδομένων.....	30
4.3 Εργαλεία οπτικοποίηση δεδομένων.....	31
4.3.1 Γραμμικό διάγραμμα	31
4.3.2 Διάγραμμα ράβδων	33
4.3.3 Διάγραμμα περιοχής	34
4.3.4 Διάγραμμα διασποράς	35
4.3.5 Γεωγραφικό διάγραμμα	36
ΚΕΦΑΛΑΙΟ 5: BACKEND.....	37

5.1	Βάση δεδομένων.....	38
5.1.1	Περιβάλλον και εργαλεία	38
5.1.2	Σχεδιασμός βάσης δεδομένων	39
5.1.3	Δημιουργία βάσης δεδομένων.....	43
5.1.4	Προετοιμασία δεδομένων	46
5.1.5	Φόρτωση δεδομένων	48
5.1.6	Μη αυτοματοποιημένες παρεμβάσεις	49
5.1.7	Αντίγραφο ασφαλείας δεδομένων.....	50
5.2	Σύστημα και επικοινωνία βάσης δεδομένων με τον πελάτη	50
5.2.1	REST.....	51
5.2.2	Spring Boot.....	53
5.2.3	API	56
5.2.4	Λειτουργίες	56
ΚΕΦΑΛΑΙΟ 6: WEBSITE		58
6.1	Κατασκευή ιστοσελίδας.....	59
6.2	Οδηγός κατασκευής εφαρμογής σε React.....	60
6.3	Περιγραφή ιστοσελίδας.....	61
ΚΕΦΑΛΑΙΟ 7: ΑΠΕΙΚΟΝΙΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΠΡΟΣΟΜΟΙΩΣΕΩΝ		74
7.1	Αποτελέσματα προσομοίωσης χωρίς ομαδοποίηση	74
7.1.1	Οπτικοποίηση δεδομένων σε μορφή χάρτη	75
7.1.2	Οπτικοποίηση δεδομένων σε μορφή διαγράμματος.....	75
7.1.3	Οπτικοποίηση δεδομένων σε μορφή πίνακα.....	76
7.2	Αποτελέσματα προσομοίωσης με ομαδοποίηση.....	77
7.2.1	Απεικόνιση αποτελεσμάτων ομαδοποίηση ενός δείκτη.....	77
7.2.2	Απεικόνιση αποτελεσμάτων ομαδοποίηση περισσότερων από ένα δείκτες	78
ΚΕΦΑΛΑΙΟ 8: ΜΕΛΛΟΝΤΙΚΕΣ ΕΠΕΚΤΑΣΕΙΣ		81
8.1	Μεταφόρτωση δεδομένων	81
8.2	Εξαγωγή αρχείων δεδομένων	82
ΓΛΩΣΣΑΡΙ		83
ΒΙΒΛΙΟΓΡΑΦΙΑ		1

Κατάλογος Σχημάτων

Σχήμα 2-1 Διάγραμμα Λειτουργίας k-means	20
Σχήμα 2-2 Τύπος υπολογισμού 3 Συμπλεγμάτων	21
Σχήμα 2-3 Διάγραμμα Συμπλεγμάτων-WCSS	22
Σχήμα 3-1 Εκτέλεση αλγορίθμου K-means.....	25
Σχήμα 3-2 Υλοποίηση αλγορίθμου Elbow	25
Σχήμα 4-1 Παράδειγμα γραμμικού διαγράμματος	32
Σχήμα 4-2 Παράδειγμα διαγράμματος ράβδων.....	33
Σχήμα 4-3 Παράδειγμα διαγράμματος περιοχής.....	34
Σχήμα 4-4 Παράδειγμα διαγράμματος διασποράς.....	35
Σχήμα 4-5 Παράδειγμα γεωγραφικού διαγράμματος	36
Σχήμα 5-1 Παράδειγμα Δομής Εφαρμογής	37
Σχήμα 5-2 Παράδειγμα μορφή Δεδομένα CSV με ένα Δείκτη	40
Σχήμα 5-3 Παράδειγμα μορφής Δεδομένων CSV με περισσότερους από ένα Δείκτη	40
Σχήμα 5-4 Παράδειγμα Δομής Πίνακα country της Βάσης Δεδομένων.....	41
Σχήμα 5-5 Παράδειγμα Δομής Πίνακα indicator της Βάσης Δεδομένων.....	41
Σχήμα 5-6 Παράδειγμα Δομής Πίνακα category της Βάσης Δεδομένων	42
Σχήμα 5-7 Παράδειγμα Δομής Πίνακα categorized_indicators της Βάσης Δεδομένων	42
Σχήμα 5-8 Παράδειγμα Δομής metric Πίνακα της Βάσης Δεδομένων.....	43
Σχήμα 5-9 DDL του Πίνακα country της Βάσης Δεδομένων	44
Σχήμα 5-10 DDL του Πίνακα indicator της Βάσης Δεδομένων	44
Σχήμα 5-11 DDL του Πίνακα category της Βάσης Δεδομένων	44
Σχήμα 5-12 DDL του Πίνακα categorized_indicators της Βάσης Δεδομένων	45
Σχήμα 5-13 DDL του Πίνακα metric της Βάσης Δεδομένων.....	45
Σχήμα 5-14 Σχήμα Βάσης Δεδομένων	45
Σχήμα 5-15 Διάγραμμα επεξεργασίας countries csv αρχείου	46

Σχήμα 5-16 Παράδειγμα έτοιμης βάσης χωρών	46
Σχήμα 5-17 Παράδειγμα τελικής μορφής countries αρχείου έτοιμο για φόρτωση ...	46
Σχήμα 5-18 Διάγραμμα επεξεργασίας indicators csv αρχείων	47
Σχήμα 5-19 Παράδειγμα τελικής μορφής indicators αρχείου έτοιμο για φόρτωση ..	47
Σχήμα 5-20 Παράδειγμα τελικής μορφής metrics αρχείου έτοιμο για φόρτωση	47
Σχήμα 5-21 Διάγραμμα UML για την προετοιμασία δεδομένων.....	48
Σχήμα 5-22 Παράδειγμα παραμετροποιήσεων για εισαγωγή αρχείου σε Πίνακα της Βάσης Δεδομένων.....	49
Σχήμα 5-23 αρχιτεκτονική Spring Boot.....	54
Σχήμα 5-24 Ροή εργασίας Spring Boot	55
Σχήμα 6-1 Αρχική μορφή σελίδας επιλογής δεικτών	62
Σχήμα 6-2 Απεικόνιση πλαισίου κατηγοριών δεικτών.....	62
Σχήμα 6-3 Απεικόνιση κατηγοριών δεικτών.....	63
Σχήμα 6-4 Επιλογή κατηγορίας και απεικόνιση κατάλληλων δεικτών	63
Σχήμα 6-5 Παράδειγμα επιλεγμένων δεικτών	64
Σχήμα 6-6 Εμφάνιση επιλογής Submit	65
Σχήμα 6-7 Αρχική μορφή σελίδας απεικόνισης δεδομένων	65
Σχήμα 6-8 Επιλογές περιοχών	66
Σχήμα 6-9 Παράδειγμα επιλογών δεικτών.....	66
Σχήμα 6-10 Παράδειγμα απεικόνισης δεδομένων σε μορφή χάρτη	67
Σχήμα 6-11 Μπάρα απεικόνισης χρονιών.....	67
Σχήμα 6-12 Επιλογές τύπων διαγραμμάτων	68
Σχήμα 6-13 Επιλογή χωρών προς απεικόνιση.....	68
Σχήμα 6-14 Παράδειγμα απεικόνισης δεδομένων σε μορφή διαγράμματος	69
Σχήμα 6-15 Παράδειγμα απεικόνισης δεδομένων σε μορφή πίνακα	69
Σχήμα 6-16 Παράδειγμα απεικόνισης αποτελεσμάτων ομαδοποίησης σε μορφή χάρτη.....	70
Σχήμα 6-17 Επιλογές αριθμών ομάδων	71
Σχήμα 6-18 Τύπος υπολογισμού τυπικής απόκλισης.....	71
Σχήμα 6-19 Παράδειγμα αποτελεσμάτων για τρία κέντρα και τρεις επιλεγμένους δείκτες.....	72

Σχήμα 6-20 Παράδειγμα αποτελεσμάτων ομαδοποίησης σε μορφή χάρτη	72
Σχήμα 6-21 Παράδειγμα επιλογής χρονικής περιόδου για τον αλγόριθμο ομαδοποίησης	72
Σχήμα 7-1 Παράδειγμα επιλογής ενός δείκτη	74
Σχήμα 7-2 Αποτελέσματα προσομοίωσης ενός δείκτη σε μορφή χάρτη	75
Σχήμα 7-3 Επιλεγμένες χώρες	75
Σχήμα 7-4 Οπτικοποίηση δεδομένων ενός δείκτη σε μορφή διαγραμμάτων	76
Σχήμα 7-5 Οπτικοποίηση δεδομένων ενός δείκτη σε μορφή πινάκων	76
Σχήμα 7-6 Απεικόνιση αποτελεσμάτων ομαδοποίησης με τέσσερις ομάδες	77
Σχήμα 7-7 Απεικόνιση αποτελεσμάτων ομαδοποίησης με επιλογή auto για ομάδες	78
Σχήμα 7-8 Παράδειγμα επιλογής περισσότερων από έναν δείκτη	78
Σχήμα 7-9 Απεικόνιση αποτελεσμάτων ομαδοποίησης με τέσσερις ομάδες και περισσότερους από έναν δείκτες	79
Σχήμα 7-10 Απεικόνιση αποτελεσμάτων ομαδοποίησης με έξι ομάδες και περισσότερους από έναν δείκτες	79

Περίληψη

Δημήτριος Κολιάτος, φοιτητής στο Τμήμα Μηχανικών Η/Υ και Πληροφορικής,
Πολυτεχνική Σχολή, Πανεπιστήμιο Ιωαννίνων, Ιούλιος 2022.

Οπτικοποίηση και ανάλυση δεδομένων δεικτών.

Επιβλέπων: Κωνσταντίνο Μπλέκα, Καθηγητής στο Τμήμα Μηχανικών Ηλεκτρονικών
Υπολογιστών και Πληροφορικής της Πολυτεχνικής Σχολής του Πανεπιστημίου
Ιωάννινων.

Η παρούσα διπλωματική εργασία ασχολείται με την οπτικοποίηση και ανάλυση δεδομένων δεικτών. Η οπτικοποίηση δεδομένων πραγματοποιείται με διάφορες τεχνικές οπτικοποίησης. Οι τεχνικές αυτές μπορεί να περιέχουν διάφορους τύπους διαγραμμάτων όπως γραμμικά διαγράμματα ή διαγράμματα ράβδων αλλά και άλλες τεχνικές όπως πίνακες ή χάρτες. Μία από τις τεχνικές που χρησιμοποιούνται για την ανάλυση δεδομένων είναι η ομαδοποίηση. Η ομαδοποίηση δεδομένων πραγματοποιείται μέσω αλγορίθμου μηχανικής μάθησης. Ο αλγόριθμος ομαδοποίησης που χρησιμοποιήθηκε είναι ο k-means. Το μέσο με το οποίο γίνεται η αλληλεπίδραση με τον χρήστη είναι μια ιστοσελίδα. Τα δεδομένα των δεικτών αντλήθηκαν από αρχεία csv τα οποία τροποποιήθηκαν και αποθηκεύτηκαν σε μία βάση δεδομένων. Τα δεδομένα αυτά με την βοήθεια ενός διακομιστή επιστρέφονται στην ιστοσελίδα σε κατάλληλη μορφή. Είναι γνωστό ότι με τόσα πολλά δεδομένα που αντλούνται την σημερινή εποχή δεν είναι δυνατή η εξαγωγή συμπερασμάτων χωρίς την οπτικοποίηση τους. Η συγκεκριμένη ιστοσελίδα μπορεί να προσφέρει σε ερευνητές ή ακόμα και επιχειρηματίες την οπτικοποίηση και ομαδοποίηση δεδομένων με σκοπό την διευκόλυνσή τους στην εξαγωγή συμπερασμάτων για τους επιλεγμένους δείκτες.

Λέξεις Κλειδιά: Μηχανική μάθηση, k-means, δεδομένα, οπτικοποίηση, διαγράμματα, βάση δεδομένων, διακομιστής, ιστοσελίδα

Abstract

Dimitrios Koliatos, student of Department of Computer Science and Engineering,
School of Engineering, University of Ioannina, Greece, July 2022.

Visualization and analysis of indicators data.

Advisor: Mplekas Konstantinos, Professor

The present thesis deals with visualization and analysis of indicators data. Data visualization is carried out with various visualization techniques. These techniques may contain various types of charts such as line charts or bar charts but also other techniques such as tables or maps. Data clustering is carried out through a machine learning algorithm. The algorithm that is being used is the k-means clustering. A website provides accessibility for the user. Also, the data of indicators has been extracted from csv files which were modified and saved in a database. This data return in an appropriate format from a server to the website. A common phenomenon nowadays is that handling and extracting big amounts of data is difficult without optimization. The current website can offer data visualization and clustering to researchers or businessmen, to draw conclusions about the selected indicators.

Key words: Machine learning, k-means, data, visualization, diagrams, database, server, website

Κεφάλαιο 1: Εισαγωγή

1.1 Αντικείμενο διπλωματικής εργασίας

1.2 Εργαλεία που χρησιμοποιήθηκαν

1.3 Οργάνωση του τόμου

1.1 Αντικείμενο διπλωματικής εργασίας

Η παρούσα διπλωματική εργασία επικεντρώνεται στην κατασκευή ιστοσελίδας για την οπτικοποίηση και ανάλυση δεδομένων δεικτών. Ο χρήστης μπορεί να επιλέξει έναν ή περισσότερους δείκτες για την εξαγωγή συμπερασμάτων. Η ιστοσελίδα προβάλλει με διάφορες τεχνικές οπτικοποίησης τα δεδομένα ανάλογα με το αν έχουν υποστεί ομαδοποίηση ή όχι. Σε περίπτωση που δεν έχουν υποστεί ομαδοποίηση οι δείκτες προβάλλονται ο καθένας ξεχωριστά με διάφορες τεχνικές οπτικοποίησης όπως διαγράμματα, χάρτη αλλά και πίνακα. Αντίθετα αν έχουν υποστεί ομαδοποίηση μπορούν να προβληθούν ταυτόχρονα περισσότεροι από έναν δείκτες μόνο με την μορφή χάρτη. Αυτή η ιστοσελίδα θα είναι χρήσιμη για οποιαδήποτε μορφή επιστήμονα ή επαγγελματία επιθυμεί να προβάλλει δεδομένα δεικτών αλλά ακόμα και να ομαδοποιήσει περισσότερους από ένα δείκτες για την εξαγωγή συμπερασμάτων.

1.2 Εργαλεία που χρησιμοποιήθηκαν

Η ομαδοποίηση έχει υλοποιηθεί χρησιμοποιώντας μηχανική μάθηση. Πιο συγκεκριμένα χρησιμοποιήθηκε ο αλγόριθμος k-means μέσω της γλώσσας προγραμματισμού Python και την βιβλιοθήκης Sklearn. Ο λόγος που επιλέχτηκε ο

συγκεκριμένος αλγόριθμος είναι λόγω την εύκολης υλοποίησης του, την εύκολη προσαρμογή σε νέα παραδείγματα, την κλιμάκωση σε μεγάλα σύνολα δεδομένων αλλά και την (2021). Στην συνέχεια για την κατασκευή της ιστοσελίδας χρησιμοποιήθηκε η React όπου είναι μια βιβλιοθήκη της JavaScript για την κατασκευή ιστοσελίδων. Επιπλέον για την αποθήκευση και την επιλογή δεδομένων στην ιστοσελίδα, κατασκευαστικά ένας διακομιστής και μια βάση δεδομένων. Για την κατασκευή αυτών των δύο χρησιμοποιήθηκε η spring boot και η MySQL αντίστοιχα. Για την απεικόνιση δεδομένων χρησιμοποιήθηκε μια βιβλιοθήκη της Google όπου διαθέτει τεχνικές οπτικοποίησης με διαγράμματα και χάρτη. Όλα όσα προαναφέρθηκαν θα αναλυθούν στα επόμενα κεφάλαια.

1.3 Οργάνωση του τόμου

Η εργασία διαθέτει οχτώ κεφάλαια. Στο δεύτερο κεφάλαιο, γίνεται αναφορά στην μηχανική μάθηση, στην ομαδοποίηση και στον αλγόριθμο ομαδοποίησης που επιλέξαμε συγκεκριμένα. Στο τρίτο κεφάλαιο περιγράφεται ο ακριβής τρόπος με τον οποίο υλοποιήθηκε ο αλγόριθμος ομαδοποίησης k-means. Στην συνέχεια στο τέταρτο κεφάλαιο αναφέρεται οι λόγοι για τους οποίους είναι χρήσιμη η οπτικοποίηση δεδομένων στις μέρες μας, το μέσο από το οποίο αντλήθηκαν τα δεδομένα και διάφορες τεχνικές και εργαλεία για την οπτικοποίηση τους. Επιπλέον υπάρχει το πέμπτο κεφάλαιο που περιγράφει αναλυτικά την κατασκευή της βάσης δεδομένων, στην οποία αποθηκεύτηκαν τα δεδομένα δεικτών καθώς και του διακομιστή ο οποίος τα επιστρέφει στην ιστοσελίδα σε κατάλληλη μορφή. Ακόμα το έκτο κεφάλαιο αφορά τον τρόπο που κατασκευάστηκε η ιστοσελίδα μαζί με τις λειτουργίες της και τον τρόπο που πρέπει να γίνει ο χειρισμός της. Επιπρόσθετος το έβδομο κεφάλαιο περιέχει μια μεγάλη γκάμα τυχαίων παραδειγμάτων με ομαδοποίηση δεδομένων ή χωρίς. Το τελευταίο κεφάλαιο, όγδοο κεφάλαιο, περιγράφει μερικές από τις μελλοντικές επεκτάσεις που μπορούν να εφαρμοστούν πάνω στην υπάρχουσα ιστοσελίδα.

Κεφάλαιο 2: Μηχανική Μάθηση

2.1 Κατηγορίες Μάθησης

2.2 Κατηγορίες Ομαδοποίησης

2.3 Ομαδοποίηση με τον αλγόριθμο K-means

Η μηχανική μάθηση αποκτά μεγαλύτερη σημασία λόγω του αυξανόμενου όγκου και της ποικιλίας δεδομένων, της πρόσβασης και της οικονομικής προσιτότητας της υπολογιστικής ισχύος καθώς και της διαθεσιμότητας διαδικτύου υψηλής ταχύτητας. Αυτοί οι παράγοντες ψηφιακού μετασχηματισμού καθιστούν δυνατή την ταχεία και αυτόματη ανάπτυξη μοντέλων που μπορούν να αναλύσουν γρήγορα και με ακρίβεια εξαιρετικά μεγάλα και πολύπλοκα σύνολα δεδομένων.

Υπάρχουν πολλές περιπτώσεις χρήσης στις οποίες μπορεί να εφαρμοστεί η μηχανική μάθηση. Κάποιες από αυτές είναι η μείωση του κόστους, τον μετριάσμο κινδύνων και τη βελτίωση της συνολικής ποιότητας ζωής. Επίσης χρησιμοποιείται για την ανίχνευση παραβιάσεων της κυβερνοασφάλειας και την ενεργοποίηση των αυτοοδηγούμενων αυτοκινήτων. Όσο μεγαλώνει η πρόσβαση σε δεδομένα και υπολογιστική ισχύ, η μηχανική μάθηση επεκτείνεται όλο και περισσότερο και σύντομα θα ενσωματωθεί σε ακόμα περισσότερες πτυχές της ανθρώπινης ζωής.

Μηχανική μάθηση είναι ένα πεδίο έρευνας αφιερωμένο στην κατανόηση και τη δημιουργία μεθόδων που μαθαίνουν, δηλαδή μεθόδους που αξιοποιούν δεδομένα για τη βελτίωση της απόδοσης τους σε κάποιο σύνολο εργασιών. Θεωρείται ως μέρος της τεχνητής νοημοσύνης. Οι αλγόριθμοι μηχανικής μάθησης δημιουργούν ένα μοντέλο που βασίζεται σε δείγματα δεδομένων, γνωστά ως

δεδομένα εκπαίδευσης, προκειμένου να κάνουν προβλέψεις ή να λαμβάνουν αποφάσεις χωρίς να έχουν προγραμματιστεί ρητά για αυτό. Οι αλγόριθμοι μηχανικής μάθησης χρησιμοποιούνται σε μια μεγάλη ποικιλία εφαρμογών, όπως στην ιατρική, το φιλτράρισμα email, την αναγνώριση ομιλίας αλλά και την υπολογιστή όραση, όπου είναι δύσκολο ή ανέφικτο να αναπτυχθούν συμβατικοί αλγόριθμοι για την εκτέλεση των απαραίτητων εργασιών.

Ένα υποσύνολο της μηχανικής μάθησης σχετίζεται στενά με τις υπολογιστικές στατιστικές, οι οποίες επικεντρώνονται στην πραγματοποίηση προβλέψεων χρησιμοποιώντας υπολογιστές. Δεν είναι όλη η μηχανική μάθηση στατιστική μάθηση. Η μελέτη της μαθηματικής βελτιστοποίησης παρέχει μεθόδους, θεωρίες και τομείς εφαρμογής στο πεδίο της μηχανικής μάθησης. Η εξόρυξη δεδομένων είναι ένα σχετικό πεδίο μελέτης, που εστιάζει στην διερευνητική ανάλυση δεδομένων μέσω της μάθησης χωρίς επίβλεψη. Ορισμένες υλοποιήσεις μηχανικής μάθησης χρησιμοποιούν δεδομένα και νευρωνικά δίκτυα με τρόπο που μιμείται τη λειτουργία ενός βιολογικού εγκεφάλου. Στην εφαρμογή της σε επιχειρηματικά προβλήματα, η μηχανική μάθηση αναφέρεται επίσης ως προγνωστική ανάλυση.

2.1 Κατηγορίες Μάθησης

Υπάρχουν διάφοροι τρόποι εκπαίδευσης αλγορίθμων μηχανικής μάθησης. Ο καθένας με τα δικά του πλεονεκτήματα και μειονεκτήματα. Για να γίνουν κατανοητά τα πλεονεκτήματα και τα μειονεκτήματα κάθε τύπου μηχανικής μάθησης, πρέπει πρώτα να γίνει αντιληπτό τι είδους δεδομένα προσλαμβάνουν. Υπάρχουν δύο είδη δεδομένων, με ετικέτα ή χωρίς.

Τα δεδομένα με ετικέτα έχουν τις παραμέτρους εισόδου και εξόδου σε ένα εντελώς αναγνώσιμο από μηχανή μοτίβο, αλλά απαιτούν μεγάλη ανθρώπινη εργασία για την επισήμανσή τους. Τα δεδομένα χωρίς ετικέτα έχουν μόνο μία ή καμία από τις παραμέτρους σε μορφή αναγνώσιμη από μηχανή. Αυτό αναιρεί την ανάγκη για ανθρώπινη εργασία, αλλά απαιτεί πιο σύνθετες λύσεις.

Υπάρχουν επίσης ορισμένοι τύποι αλγορίθμων μηχανικής μάθησης που χρησιμοποιούνται σε πολύ συγκεκριμένες περιπτώσεις. Εμείς θα ασχοληθούμε με τις τρεις κύριες μέθοδοι που χρησιμοποιούνται σήμερα.

2.1.1 Μάθηση με επίβλεψη

Η μάθηση με επίβλεψη είναι ένας από τους πιο βασικούς τύπους μηχανικής μάθησης. Σε αυτόν τον τύπο, ο αλγόριθμος μηχανικής μάθησης εκπαιδεύεται σε δεδομένα με ετικέτα. Παρόλο που τα δεδομένα πρέπει να επισημαίνονται με ακρίβεια για να λειτουργήσει είναι εξαιρετικά ισχυρή όταν χρησιμοποιείται στις σωστές συνθήκες.

Στην μάθηση με επίβλεψη, δίνεται στον αλγόριθμο μηχανικής μάθησης ένα μικρό σύνολο δεδομένων εκπαίδευσης για να εργαστεί. Αυτό το σύνολο δεδομένων εκπαίδευσης είναι ένα μικρότερο μέρος του μεγαλύτερου συνόλου δεδομένων και χρησιμεύει για να δώσει στον αλγόριθμο μια βασική ιδέα για το πρόβλημα, τη λύση και τα σημεία δεδομένων που πρέπει να αντιμετωπιστούν. Το σύνολο δεδομένων εκπαίδευσης είναι επίσης πολύ παρόμοιο με το τελικό σύνολο δεδομένων ως προς τα χαρακτηριστικά του και παρέχει στον αλγόριθμο τις επισημασμένες παραμέτρους που απαιτούνται για το πρόβλημα.

Στη συνέχεια, ο αλγόριθμος βρίσκει σχέσεις μεταξύ των παραμέτρων που δίνονται, καθιερώνοντας ουσιαστικά μια σχέση αιτίου και αποτελέσματος μεταξύ των μεταβλητών στο σύνολο δεδομένων. Στο τέλος της εκπαίδευσης, ο αλγόριθμος έχει μια εκτίμηση για το πώς λειτουργούν τα δεδομένα και τη σχέση μεταξύ της εισόδου και της εξόδου.

Αυτή η λύση στη συνέχεια αναπτύσσεται για χρήση με το τελικό σύνολο δεδομένων, από το οποίο μαθαίνει με τον ίδιο τρόπο όπως το σύνολο δεδομένων εκπαίδευσης. Αυτό σημαίνει ότι οι αλγόριθμοι μηχανικής μάθησης με επίβλεψη θα συνεχίσουν να βελτιώνονται ακόμη και μετά την ανάπτυξη, ανακαλύπτοντας νέα μοτίβα και σχέσεις καθώς εκπαιδεύεται σε νέα δεδομένα.

2.1.2 Μάθηση χωρίς επίβλεψη

Η μάθηση χωρίς επίβλεψη έχει το πλεονέκτημα της δυνατότητας εργασίας με δεδομένα χωρίς ετικέτα. Αυτό σημαίνει ότι δεν απαιτείται ανθρώπινη εργασία για να γίνει το σύνολο δεδομένων αναγνώσιμο από μηχανή, επιτρέποντας την επεξεργασία πολύ μεγαλύτερων συνόλων δεδομένων από το πρόγραμμα.

Στην μάθηση με επίβλεψη, οι ετικέτες επιτρέπουν στον αλγόριθμο να βρει την ακριβή φύση της σχέσης μεταξύ οποιωνδήποτε δύο σημείων δεδομένων. Ωστόσο, η μάθηση χωρίς επίβλεψη δεν έχει ετικέτες για να λειτουργήσει, με αποτέλεσμα τη δημιουργία κρυφών δομών. Οι σχέσεις μεταξύ των σημείων δεδομένων γίνονται αντιληπτές από τον αλγόριθμο με αφηρημένο τρόπο, χωρίς να απαιτείται παρέμβαση από τον άνθρωπο.

Η δημιουργία αυτών των κρυφών δομών είναι αυτό που κάνει τους αλγόριθμους μάθησης χωρίς επίβλεψη ευέλικτους. Αντί για μια καθορισμένη δήλωση προβλήματος, οι αλγόριθμοι μάθησης χωρίς επίβλεψη μπορούν να προσαρμοστούν στα δεδομένα αλλάζοντας δυναμικά κρυφές δομές. Αυτό προσφέρει περισσότερη ανάπτυξη από ό,τι οι αλγόριθμοι εκμάθησης με επίβλεψη.

2.1.3 Ενισχυτική μάθηση

Η ενισχυτική μάθηση εμπνέεται άμεσα από το πώς μαθαίνουν τα ανθρώπινα όντα από δεδομένα στη ζωή τους. Διαθέτει έναν αλγόριθμο που βελτιώνεται μόνος του και μαθαίνει από νέες καταστάσεις χρησιμοποιώντας μια μέθοδο δοκιμής και λάθους. Τα ευνοϊκά αποτελέσματα ενθαρρύνονται ή ενισχύονται και τα μη ευνοϊκά αποτελέσματα αποθαρρύνονται ή τιμωρούνται.

Με βάση την έννοια της προετοιμασίας, η ενισχυτική μάθηση λειτουργεί βάζοντας τον αλγόριθμο σε περιβάλλον εργασίας με διερμηνέα και σύστημα ανταμοιβής. Σε κάθε επανάληψη του αλγορίθμου, το αποτέλεσμα εξόδου δίνεται στον διερμηνέα, ο οποίος αποφασίζει εάν το αποτέλεσμα είναι ευνοϊκό ή όχι.

Σε περίπτωση που το πρόγραμμα βρει τη σωστή λύση, ο διερμηνέας ενισχύει τη λύση παρέχοντας μια ανταμοιβή στον αλγόριθμο. Εάν το αποτέλεσμα δεν είναι

ευνοϊκό, ο αλγόριθμος αναγκάζεται να επαναλάβει μέχρι να βρει ένα καλύτερο αποτέλεσμα. Στις περισσότερες περιπτώσεις, το σύστημα ανταμοιβής συνδέεται άμεσα με την αποτελεσματικότητα του αποτελέσματος.

Σε τυπικές περιπτώσεις χρήσης ενισχυτικής μάθησης, όπως η εύρεση της συντομότερης διαδρομής μεταξύ δύο σημείων σε έναν χάρτη, η λύση δεν είναι απόλυτη τιμή. Αντίθετα, παίρνει μια βαθμολογία αποτελεσματικότητας, εκφρασμένη σε ποσοστιαία τιμή. Όσο υψηλότερη είναι αυτή η ποσοστιαία τιμή, τόσο μεγαλύτερη ανταμοιβή δίνεται στον αλγόριθμο. Έτσι, το πρόγραμμα εκπαιδεύεται ώστε να δίνει την καλύτερη δυνατή λύση για την καλύτερη δυνατή ανταμοιβή.

2.2 Κατηγορίες Ομαδοποίησης

Η ομαδοποίηση είναι το έργο της διαίρεσης του πληθυσμού ή των σημείων δεδομένων σε έναν αριθμό ομάδων, έτσι ώστε τα σημεία δεδομένων στις ίδιες ομάδες να είναι πιο παρόμοια με άλλα σημεία δεδομένων της ίδια ομάδα και ανόμοια με τα σημεία δεδομένων άλλων ομάδων. Είναι βασικά μια συλλογή αντικειμένων με βάση την ομοιότητα και την ανομοιότητα μεταξύ τους. Ανήκει στην κατηγορία της μάθησης χωρίς επίβλεψη.

2.2.1 Μέθοδοι που βασίζονται στην πυκνότητα

Αυτές οι μέθοδοι θεωρούν τα συμπλέγματα ως την πυκνή περιοχή που έχει κάποιες ομοιότητες και διαφορές από την περιοχή του κενού χώρου. Αυτές οι μέθοδοι έχουν καλή ακρίβεια και τη δυνατότητα συγχώνευσης δύο συμπλεγμάτων. Παραδείγματα αλγορίθμων ομαδοποίησης που βασίζονται στην πυκνότητα είναι DBSCAN (Density-Based Spatial Clustering of Applications with Noise), OPTICS (Ordering Points to Identify Clustering Structure).

2.2.2 Μέθοδοι με βάση την ιεραρχία

Τα συμπλέγματα που σχηματίζονται σε αυτήν τη μέθοδο σχηματίζουν μια δομή δέντρου που βασίζεται στην ιεραρχία. Νέα συμπλέγματα σχηματίζονται χρησιμοποιώντας την προηγούμενη δομή. Χωρίζονται σε δύο κατηγορίες. Τις

συγκεντρωτικές (προσέγγιση από κάτω προς τα πάνω) και τις διαιρετικές (προσέγγιση από πάνω προς τα κάτω). Παραδείγματα αλγορίθμων ομαδοποίησης που βασίζονται στην ιεραρχία είναι CURE (Clustering Using Representatives), BIRCH (Balanced Iterative Reducing Clustering and using Hierarchies).

2.2.3 Μέθοδοι κατάτμησης

Αυτές οι μέθοδοι χωρίζουν τα αντικείμενα σε k συμπλέγματα και κάθε διαμέρισμα σχηματίζει ένα σύμπλεγμα. Αυτή η μέθοδος χρησιμοποιείται για τη βελτιστοποίηση μιας συνάρτησης ομοιότητας αντικειμενικού κριτηρίου. Μια κύρια παράμετρος μπορεί να είναι η απόσταση. Παραδείγματα αυτής της μεθόδου είναι K-means, CLARANS (Clustering Large Applications based upon Randomized Search).

2.2.4 Μέθοδοι που βασίζονται σε πλέγμα

Η προσέγγιση ομαδοποίησης που βασίζεται σε πλέγμα χρησιμοποιεί μια δομή δεδομένων πλέγματος πολλαπλής ανάλυσης. Κβαντίζει τον χώρο του αντικειμένου σε έναν πεπερασμένο αριθμό κελιών που σχηματίζουν μια δομή πλέγματος στο οποίο εκτελούνται όλες οι λειτουργίες για ομαδοποίηση. Παραδείγματα αλγορίθμων ομαδοποίησης που βασίζονται σε πλέγμα είναι, STING (Statistical Information Grid) και σύμπλεγμα κυμάτων και CLIQUE (Clustering In Quest).

2.3 Ομαδοποίηση με τον αλγόριθμο K-means

Η ομαδοποίηση K-Means είναι ένας αλγόριθμος μάθησης χωρίς επίβλεψη, ο οποίος ομαδοποιεί το σύνολο δεδομένων χωρίς ετικέτα σε διαφορετικά συμπλέγματα. Συνήθως, οι αλγόριθμοι χωρίς επίβλεψη εξάγουν συμπεράσματα από σύνολα δεδομένων χρησιμοποιώντας μόνο διανύσματα εισόδου χωρίς να αναφέρονται σε γνωστά ή με ετικέτα αποτελέσματα. Το K ορίζει τον αριθμό των προκαθορισμένων συμπλεγμάτων που πρέπει να δημιουργηθούν στη διαδικασία.

Μας επιτρέπει να ομαδοποιήσουμε τα δεδομένα σε διαφορετικές ομάδες και έναν βολικό τρόπο για να ανακαλύψουμε τις κατηγορίες ομάδων στο μη επισημασμένο σύνολο δεδομένων από μόνος του χωρίς την ανάγκη εκπαίδευσης. Είναι ένας αλγόριθμος που βασίζεται στο κέντρο, όπου κάθε σύμπλεγμα σχετίζεται

με αυτό. Ο κύριος στόχος αυτού του αλγορίθμου είναι να ελαχιστοποιήσει το άθροισμα των αποστάσεων μεταξύ του σημείου δεδομένων και των αντίστοιχων συμπλεγμάτων τους.

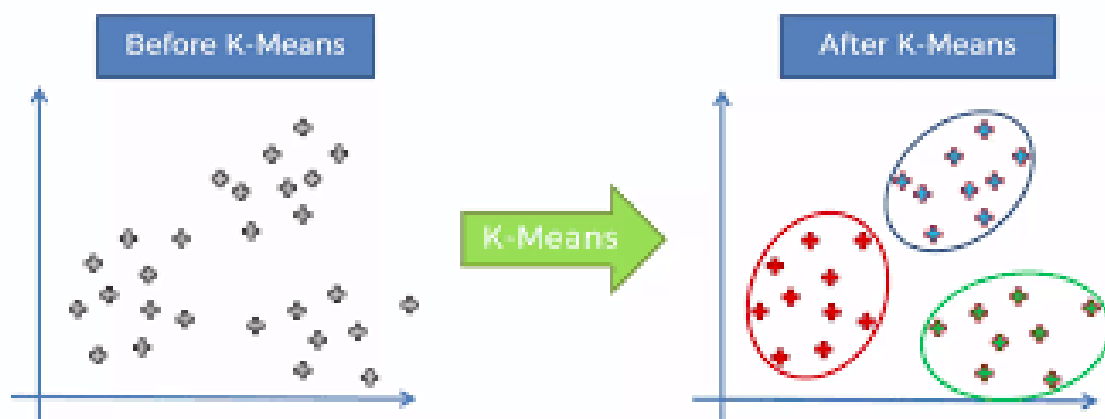
2.3.1 Λειτουργία K-means

Ο αλγόριθμος παίρνει το σύνολο δεδομένων χωρίς ετικέτα ως είσοδο, διαιρεί το σύνολο δεδομένων σε k -αριθμό συμπλεγμάτων και επαναλαμβάνει τη διαδικασία μέχρι να μην βρει τα καλύτερα συμπλέγματα. Η τιμή του k θα πρέπει να είναι προκαθορισμένη σε αυτόν τον αλγόριθμο.

Ο αλγόριθμος ομαδοποίησης k -means εκτελεί κυρίως δύο εργασίες:

- Καθορίζει την καλύτερη τιμή για K κεντρικά σημεία ή κεντροειδή με μια επαναληπτική διαδικασία.
- Αντιστοιχίζει κάθε σημείο δεδομένων στο πλησιέστερο k -κέντρο του. Αυτά τα σημεία δεδομένων που βρίσκονται κοντά στο συγκεκριμένο κέντρο k , δημιουργούν ένα σύμπλεγμα.

Το παρακάτω διάγραμμα εξηγεί τη λειτουργία του αλγορίθμου ομαδοποίησης K -means:



Σχήμα 2-1 Διάγραμμα Λειτουργίας k -means

2.3.2 Αλγόριθμος K-means

Η λειτουργία του αλγόριθμου K-Means εξηγείται στα παρακάτω βήματα:

- Βήμα 1. Επιλογή του αριθμού των K συμπλεγμάτων.
- Βήμα 2. Επιλογή τυχαίων σημείων για τα K. (Μπορεί να είναι άλλο από το σύνολο δεδομένων εισόδου).
- Βήμα 3. Αντιστοίχιση κάθε σημείο δεδομένων στο πλησιέστερο κέντρο, το οποίο θα σχηματίσει τα προκαθορισμένα συμπλέγματα K.
- Βήμα 4. Υπολογισμός της διακύμανσης και τοποθέτηση ενός νέο κέντρου για κάθε σύμπλεγμα.
- Βήμα 5. Επανάληψη των τριών βημάτων, που σημαίνει εκ νέου αντιστοίχιση κάθε σημείου δεδομένων στο νέο πλησιέστερο κέντρο κάθε συμπλέγματος.
- Βήμα 6. Εάν συμβεί κάποια εκ νέου αντιστοίχιση, τότε γίνεται μετάβαση στο Βήμα 4, διαφορετικά τέλος.
- Βήμα 7. Το μοντέλο έχει ολοκληρωθεί.

2.3.3 Υπολογισμός βέλτιστης K τιμής

Η επιλογή του βέλτιστου αριθμού συμπλεγμάτων είναι περίπλοκη υπόθεση. Υπάρχουν μερικοί διαφορετικοί τρόποι για να βρεθεί ο βέλτιστος αριθμός συμπλεγμάτων, αλλά εδώ αναφέρεται η καταλληλότερη μέθοδο εύρεσης του αριθμού των συμπλεγμάτων ή την τιμή του K. Η μέθοδος αυτή ονομάζεται Elbow.

Η μέθοδος Elbow είναι ένας από τους πιο δημοφιλείς τρόπους εύρεσης του βέλτιστου αριθμού συμπλεγμάτων. Αυτή η μέθοδος χρησιμοποιεί την έννοια της τιμής WCSS. Το WCSS σημαίνει Within Cluster Sum of Squares, το οποίο ορίζει τις συνολικές παραλλαγές μέσα σε ένα σύμπλεγμα. Ο τύπος για τον υπολογισμό της τιμής του WCSS (για 3 συμπλέγματα) δίνεται παρακάτω:

$$WCSS = \sum_{P_i \text{ in Cluster1}} \text{distance}(P_i C_1)^2 + \sum_{P_i \text{ in Cluster2}} \text{distance}(P_i C_2)^2 + \sum_{P_i \text{ in Cluster3}} \text{distance}(P_i C_3)^2$$

Σχήμα 2-2 Τύπος υπολογισμού 3 Συμπλεγμάτων

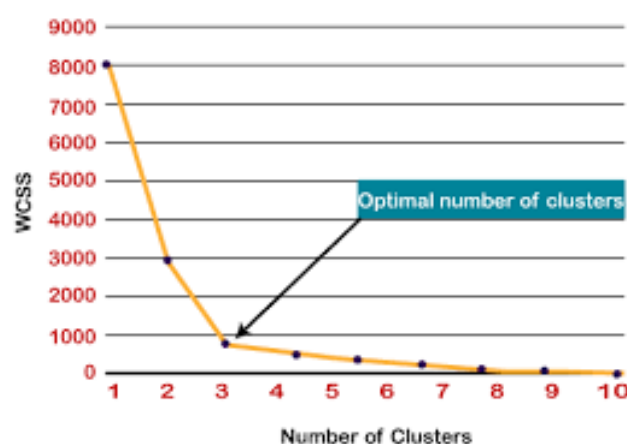
Στον παραπάνω τύπο του WCSS, $\sum P_i$ στην απόσταση $Cluster1(P_i C1)^2$: Είναι το άθροισμα του τετραγώνου των αποστάσεων μεταξύ κάθε σημείου δεδομένων και του κέντρου του μέσα σε ένα σύμπλεγμα. Το ίδιο ισχύει και για τους άλλους 2 όρους.

Για να υπολογιστεί η απόσταση μεταξύ των σημείων δεδομένων και του κέντρου, μπορεί να χρησιμοποιηθεί οποιαδήποτε μέθοδο, όπως η Ευκλείδεια απόσταση ή η απόσταση του Μανχάταν.

Για να βρείτε τη βέλτιστη τιμή των συμπλεγμάτων, η μέθοδος Elbow ακολουθεί τα παρακάτω βήματα:

- Βήμα 1. Εκτέλεση της ομαδοποίησης K-means σε ένα δεδομένο σύνολο δεδομένων για διαφορετικές τιμές K (εύρος από 1-10).
- Βήμα 2. Για κάθε τιμή του K, υπολογίζεται η τιμή WCSS.
- Βήμα 3. Σχεδίαση μια καμπύλη μεταξύ των υπολογισμένων τιμών WCSS και του αριθμού των K συμπλεγμάτων.
- Βήμα 4. Το αιχμηρό σημείο κάμψης είναι το σημείο που θεωρείται ως η καλύτερη τιμή του K.

Δεδομένου ότι το γράφημα δείχνει την απότομη κάμψη, η οποία μοιάζει με αγκώνα, είναι γνωστή ως μέθοδος αγκώνα (Elbow).



Σχήμα 2-3 Διάγραμμα Συμπλεγμάτων-WCSS

Κεφάλαιο 3: Υλοποίηση Αλγορίθμου Ομαδοποίησης

3.1 Αρχικοποίηση παραμέτρων

3.2 Μορφή δεδομένων προς ομαδοποίηση

3.3 Εκτέλεση K-means

3.4 Υλοποίηση Elbow

Η υλοποίηση του αλγορίθμου ομαδοποίησης K-means σε γλώσσα προγραμματισμού Python έγινε με την βοήθεια της Scikit-learn. Η Scikit-learn (Sklearn) είναι η πιο χρήσιμη και ισχυρή βιβλιοθήκη για μηχανική μάθηση στην Python. Παρέχει μια επιλογή αποτελεσματικών εργαλείων για μηχανική μάθηση και στατιστική μοντελοποίηση, συμπεριλαμβανομένης της ταξινόμησης, της παλινδρόμησης, της ομαδοποίησης και της μείωσης διαστάσεων μέσω μιας διεπαφής συνέπειας στην Python. Αυτή η βιβλιοθήκη, η οποία είναι σε μεγάλο βαθμό γραμμένη σε Python, βασίζεται στα NumPy, SciPy και Matplotlib.

3.1 Αρχικοποίηση παραμέτρων

Για την εκτέλεση του αλγορίθμου ομαδοποίησης θα καλεστεί η μέθοδος k-means από την βιβλιοθήκη Sklearn. Η μέθοδος αυτή χρειάζεται τις κατάλληλες παραμέτρους. Πολλές από τις παραμέτρους της δεν χρειάζονται αρχικοποίηση γιατί έχουν προκαθορισμένες τιμές.

Μια σημαντική παράμετρος από αυτές είναι η `init`. Η παράμετρος αυτή ασχολείται με την αρχικοποίηση των κέντρων (cluster centers). Αυτή η παράμετρος έχει τρεις τεχνικές. Η πρώτη τεχνική ονομάζεται `Forgy`. Είναι μία από τις ταχύτερες μεθόδους αρχικοποίησης για το `k-Means`. Εάν επιλέξουμε να έχουμε `k` συμπλέγματα, η μέθοδος `Forgy` επιλέγει οποιαδήποτε `k` σημεία από τα δεδομένα τυχαία ως αρχικά σημεία. Αυτή η μέθοδος έχει νόημα επειδή τα συμπλέγματα που ανιχνεύονται μέσω του `k-Means` είναι πιο πιθανό να βρίσκονται κοντά στις καταστάσεις που υπάρχουν στα δεδομένα. Επιλέγοντας τυχαία σημεία από δεδομένα, καθιστάτε πιο πιθανό να ληφθεί ένα σημείο που βρίσκεται κοντά στις λειτουργίες. Η επόμενη τεχνική ονομάζεται μέθοδος τυχαίας κατανομής (random partition method). Σε αυτή τη μέθοδο, εκχωρείται τυχαία κάθε σημείο στα δεδομένα σε ένα τυχαίο αναγνωριστικό συμπλέγματος. Στην περίπτωση που χρησιμοποιήθηκε επιλέξαμε την τεχνική `kmeans++`. Επιλέγει αρχικά κέντρα συμπλέγματος για ομαδοποίηση `k-mean` με έξυπνο τρόπο για να επιταχύνει τη σύγκλιση. Αυτή είναι μια τυπική μέθοδος η οποία γενικά λειτουργεί καλύτερα από τη μέθοδο του `Forgy` και τη μέθοδο τυχαίας κατανομής για την προετοιμασία του `k-Means`.

Ακόμα υπάρχει η παράμετρο `n_init` όπου της δίνεται η τιμή εκατό. Αυτή η παράμετρος μας δηλώνει τον αριθμό που θα εκτελεστεί ο αλγόριθμος `k-means` με διαφορετικά κέντρα. Τα τελικά αποτελέσματα θα είναι η καλύτερη έξοδος `n_init` διαδοχικών εκτελέσεων όσον αφορά την αδράνεια.

Τέλος πρέπει να δοθεί τιμή για τον αριθμό των ομάδων που θα χωρίσει ο αλγόριθμος τα δεδομένα.

3.2 Μορφή δεδομένων προς ομαδοποίηση

Αφού αρχικοποιηθούν οι παράμετροι στην μέθοδο `K-means` πρέπει να δοθούν στην μέθοδο τα δεδομένα που θα ομαδοποιήσει. Αυτό γίνεται μέσω του `fit`. Τα δεδομένα μας πρέπει να έχουν την κατάλληλη μορφή πριν την καταχώριση τους.

Στην περίπτωση της υλοποίησης αυτής τα δεδομένα θα αποθηκεύονται σε ένα πίνακα ο οποίος θα έχει τόσες διαστάσεις όσοι και οι επιλεγμένοι δείκτες. Κάθε

διάσταση θα περιέχει τα δεδομένα από έναν δείκτη. Πέρα από αυτό θα πρέπει να πραγματοποιηθεί μια κανονικοποίηση των δεδομένων ανάμεσα στο 0 και το 1 για την κάθε διάσταση ξεχωριστά. Αυτό μας βοηθάει στο να αντιμετωπιστούν οι δείκτες στην ίδια κλίμακα τιμών και όχι σε διαφορετική.

3.3 Εκτέλεση K-means

Σύμφωνα με την αρχικοποίηση παραμέτρων και την τροφοδότηση των δεδομένων ο αλγόριθμος K-means εκτελείται σύμφωνα με το παρακάτω παράδειγμα κώδικα.

```
def kMeans(n_clusters, size):  
    kMeans = KMeans(n_clusters, init='k-means++', n_init = 100)  
    kMeans.fit(getDataFromFile(size))
```

Σχήμα 3-1 Εκτέλεση αλγορίθμου K-means

Τα αποτελέσματα από την εκτέλεση του αλγορίθμου επιστέφονται από την κλήση των `kMeans.cluster_centers` και `kMeans.labels_`. Η πρώτη μέθοδος μας επιστρέφει ένα πίνακα με τα κέντρα που επέλεξε ο αλγόριθμος. Ο αριθμός των κέντρων είναι ίσος με τον αριθμό των κέντρων που δόθηκαν από την παράμετρο και τις διαστάσεις των δεδομένων. Η δεύτερη μέθοδος μας επιστρέφει έναν πίνακα με τον αριθμό των ομάδων που αντιστοιχίστηκε το κάθε στιγμιότυπο.

3.4 Υλοποίηση Elbow

Στην υλοποίηση αυτή υπάρχει η δυνατότητα να μην δοθεί ο αριθμός των ομάδων (`n_clusters`) που θα χωριστούν τα δεδομένα. Σε αυτή την περίπτωση ενεργοποιείται ο αλγόριθμος Elbow που περιγράψαμε στο κεφάλαιο 2.3.3.

```
def elbow(size, n_clusters):  
    if(n_clusters == 0):  
        WCSS = []  
        K = range(1,10)  
        for k in K:  
            kMeans = KMeans(k, init='k-means++', n_init = 100)  
            kMeans.fit(getDataFromFile(size))  
            WCSS.append(kMeans.inertia_)  
        return findBestNumberForClusters(WCSS)  
    return n_clusters
```

Σχήμα 3-2 Υλοποίηση αλγορίθμου Elbow

Ο αλγόριθμος αυτός ουσιαστικά εκτελεί τον K-means αλγόριθμο για αριθμό ομάδων από ένα μέχρι δέκα. Σε κάθε επανάληψη υπολογίζεται η αδράνεια. Η αδράνεια μετρά πόσο καλά ομαδοποιήθηκε ένα σύνολο δεδομένων από το K-Means. Ο τρόπος υπολογισμού της αναλύθηκε σε προηγούμενο κεφάλαιο. Η επιλογή του κατάλληλου αριθμού για τις ομάδες επιλέγεται με βάση τον συνδυασμό χαμηλής αδράνειας και μικρό αριθμό συστάδων.

Κεφάλαιο 4: Δεδομένα και Οπτικοποίηση

4.1 Οπτικοποίηση Δεδομένων

4.2 Πηγή Δεδομένων

4.3 Εργαλεία οπτικοποίηση δεδομένων

Από την εφεύρεση των υπολογιστών, οι άνθρωποι χρησιμοποίησαν τον όρο δεδομένα για να αναφερθούν σε πληροφορίες υπολογιστών, και αυτές οι πληροφορίες είτε μεταδίδονταν είτε αποθηκεύονταν. Αλλά αυτός δεν είναι ο μόνος ορισμός δεδομένων, υπάρχουν και άλλα είδη. Τα δεδομένα μπορεί να είναι κείμενα ή αριθμοί γραμμένοι σε χαρτιά, ή byte και bit μέσα στη μνήμη των ηλεκτρονικών συσκευών ή μπορεί να είναι γεγονότα που είναι αποθηκευμένα στο μυαλό ενός ατόμου.

Δεδομένα κυρίως στον τομέα της επιστήμης είναι διαφορετικοί τύποι πληροφοριών που συνήθως μορφοποιούνται με συγκεκριμένο τρόπο. Όλο το λογισμικό χωρίζεται σε δύο μεγάλες κατηγορίες, και αυτές είναι τα προγράμματα και τα δεδομένα. Τα προγράμματα είναι η συλλογή που αποτελείται από οδηγίες που χρησιμοποιούνται για τον χειρισμό δεδομένων.

Η έννοια των δεδομένων επεκτείνεται πέρα από την επεξεργασία δεδομένων σε υπολογιστικές εφαρμογές. Όταν πρόκειται για το τι είναι η επιστήμη των δεδομένων, ένα σώμα που αποτελείται από γεγονότα ονομάζεται επιστήμη

δεδομένων. Κατά συνέπεια, τα οικονομικά, τα δημογραφικά στοιχεία, η υγεία και το μάρκετινγκ έχουν επίσης διαφορετικές έννοιες των δεδομένων, οι οποίες τελικά αποτελούν διαφορετικές απαντήσεις για το τι είναι δεδομένα.

4.1 Οπτικοποίηση δεδομένων

Με τόσες πολλές πληροφορίες που συλλέγονται μέσω της ανάλυσης δεδομένων στον κόσμο σήμερα, πρέπει να υπάρχει έναν τρόπο να απεικονίζονται αυτά τα δεδομένα ώστε να γίνεται να τα ερμηνευτούν. Η οπτικοποίηση δεδομένων δίνει μια σαφή ιδέα για το τι σημαίνει η πληροφορία δίνοντάς τους οπτικό πλαίσιο μέσω χαρτών ή γραφημάτων. Αυτό καθιστά τα δεδομένα πιο φυσικά για την κατανόηση τους από το ανθρώπινο μυαλό και ως εκ τούτου διευκολύνει τον εντοπισμό τάσεων, προτύπων και ακραίων στοιχείων σε μεγάλα σύνολα δεδομένων.

Το ανθρώπινο μάτι ελκύεται από χρώματα και σχέδια. Μπορεί να αναγνωριστεί πιο γρήγορα το κόκκινο από το μπλε, το τετράγωνο από τον κύκλο. Ο πολιτισμός είναι οπτικός, περιλαμβάνει τα πάντα, από τέχνη και διαφημίσεις μέχρι τηλεόραση και ταινίες. Η οπτικοποίηση δεδομένων είναι μια άλλη μορφή εικαστικής τέχνης που κεντρίζει το ενδιαφέρον και κρατά τα μάτια στο μήνυμα. Όταν προβάλλεται ένα γράφημα, γίνονται αντιληπτές γρήγορα οι τάσεις και οι ακραίες τιμές. Αν είναι προφανές κάτι, εσωτερικεύεται γρήγορα. Είναι αφήγηση με σκοπό. Εάν προβληθεί ένα τεράστιο υπολογιστικό φύλλο δεδομένων και δεν γίνεται αντιληπτή μια τάση, είναι γνωστό ότι η οπτικοποίηση μπορεί να γίνει πιο αποτελεσματική.

Μια καλή οπτικοποίηση δίνει την πληροφορία, αφαιρώντας τον θόρυβο από τα δεδομένα και επισημαίνοντας τις χρήσιμες πληροφορίες. Ωστόσο, δεν είναι απλά τόσο εύκολο όσο το να ντυθεί ένα γράφημα για να φανεί καλύτερο. Η αποτελεσματική οπτικοποίηση δεδομένων είναι μια λεπτή πράξη εξισορρόπησης μεταξύ μορφής και λειτουργίας. Το πιο απλό γράφημα μπορεί να είναι πολύ ανεπαρκή για να συλλάβεις οποιαδήποτε πληροφορία. Η πιο εντυπωσιακή απεικόνιση θα μπορούσε να αποτύχει τελείως στη μετάδοση του σωστού μηνύματος ή θα μπορούσε να μιλήσει πολύ. Τα δεδομένα και τα γραφικά πρέπει να

συνεργάζονται και είναι περίπλοκη διαδικασία ο συνδυασμός ανάλυσης με αφήγηση.

Η οπτικοποίηση δεδομένων μπορεί να βοηθήσει παρέχοντας δεδομένα με τον πιο αποτελεσματικό δυνατό τρόπο. Ως ένα από τα βασικά βήματα, η οπτικοποίηση δεδομένων λαμβάνει τα ακατέργαστα δεδομένα, τα μοντελοποιεί και παραδίδει τα δεδομένα έτσι ώστε να μπορούν να εξαχθούν συμπεράσματα. Οι επιστήμονες δεδομένων δημιουργούν αλγόριθμους μηχανικής μάθησης για να συγκεντρώνουν καλύτερα βασικά δεδομένα σε οπτικοποιήσεις που είναι πιο εύκολο να κατανοηθούν και να ερμηνευτούν.

Συγκεκριμένα, η οπτικοποίηση δεδομένων χρησιμοποιεί οπτικά δεδομένα για την επικοινωνία πληροφοριών με τρόπο που είναι καθολικός, γρήγορος και αποτελεσματικός. Αυτή η πρακτική μπορεί να βοηθήσει στην δημιουργία συμπερασμάτων για στοχευμένες αλλαγές που μπορούν να γίνουν. Τα οπτικοποιημένα δεδομένα παρέχουν στους υπεύθυνους λήψης αποφάσεων καλύτερη πρόβλεψη για το αντικείμενό τους.

Η οπτικοποίηση δεδομένων επηρεάζει θετικά τη διαδικασία λήψης αποφάσεων με διαδραστικές οπτικές αναπαραστάσεις δεδομένων. Μπορούν πλέον να αναγνωρίζουν μοτίβα πιο γρήγορα, επειδή μπορούν να ερμηνεύουν δεδομένα σε γραφικές ή εικονογραφικές μορφές.

Ακολουθούν ορισμένοι πιο συγκεκριμένοι τρόποι με τους οποίους η οπτικοποίηση δεδομένων μπορεί να ωφελήσει:

- Συσχετίσεις σχέσεων.
- Τάσεις σύμφωνα με την πάροδο του χρόνου.
- Εντοπισμός ευκαιριών.
- Λήψη αποφάσεων με μετρήσιμα στοιχεία, βασισμένα σε δεδομένα.
- Συχνότητα
- Κίνδυνοι
- Αντιδράσεις.

- Εξετάζοντας αποφάσεις.

4.2 Πηγή δεδομένων

Για την συγκεκριμένη εργασία χρειάστηκε μια γκάμα δεδομένων. Αυτά τα δεδομένα λόγω χρόνου και κόπου δεν συλλέχτηκαν από εμάς αλλά βρέθηκαν από δημοσίευση στο διαδίκτυο.

Το **Our World in Data (OWID)** είναι μια επιστημονική διαδικτυακή δημοσίευση που εστιάζει σε μεγάλα παγκόσμια προβλήματα όπως η φτώχεια, οι ασθένειες, η πείνα , η κλιματική αλλαγή, ο πόλεμος, οι υπαρξιακοί κίνδυνοι και η ανισότητα. Είναι ένα έργο του Global Change Data Lab, μιας εγγεγραμμένης φιλανθρωπικής οργάνωσης στην Αγγλία και την Ουαλία, και ιδρύθηκε από τον Max Roser. Ο Max Roser είναι κοινωνικός ιστορικός και οικονομολόγος. Η ερευνητική ομάδα εδρεύει στο Πανεπιστήμιο της Οξφόρδης.

Το Global Change Data Lab, ο μη κερδοσκοπικός οργανισμός που δημοσιεύει το Our World in Data και τα εργαλεία δεδομένων ανοιχτής πρόσβασης που καθιστούν δυνατή την ηλεκτρονική δημοσίευση, χρηματοδοτείται μέσω ενός συνδυασμού επιχορηγήσεων, χορηγών και δωρεών αναγνωστών.

Η αποστολή του Our World in Data είναι η δημοσίευση έρευνας και δεδομένων για να σημειωθεί πρόοδος ενάντια στα μεγαλύτερα προβλήματα του κόσμου. Η δημοσίευση στο Διαδίκτυο χρησιμοποιεί διαδραστικά γραφήματα και χάρτες την απεικόνιση των ευρημάτων της έρευνας συχνά με μακροπρόθεσμη άποψη για να δείξει πώς έχουν αλλάξει οι παγκόσμιες συνθήκες διαβίωσης με την πάροδο του χρόνου.

Στην περίπτωση της διπλωματικής εργασία αντλήθηκε μια μεγάλη ποσότητα δεδομένων από δεικτών για τις διεκπεραίωση της εργασίας. Αυτά τα δεδομένα βρίσκονται αποθηκευμένα σε αρχεία μορφής csv δίνοντας αναλυτικές πληροφορίες για πολλές χώρες σε βάθος χρόνου.

4.3 Εργαλεία οπτικοποίηση δεδομένων

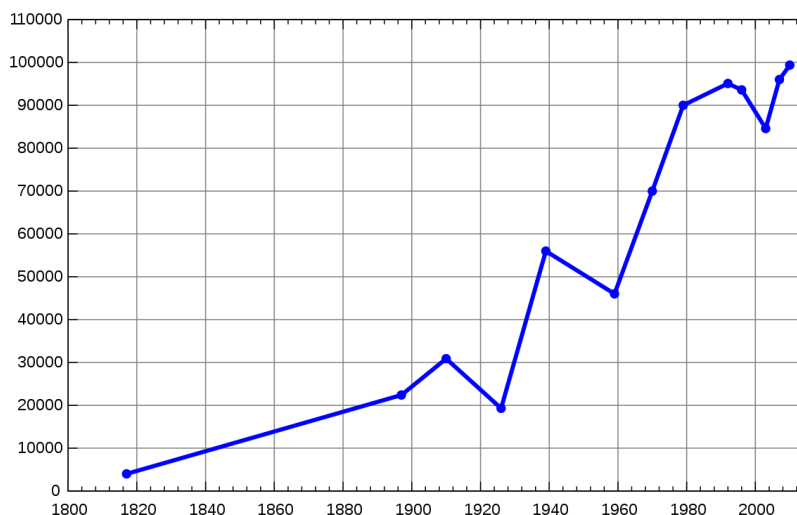
Για την οπτικοποίηση δεδομένων χρησιμοποιούνται διάφορα εργαλεία οπτικοποίηση δεδομένων. Ένα από τα εργαλεία αυτά είναι τα διαγράμματα. Τα διαγράμματα μπορούν να χρησιμοποιηθούν για την καταγραφή πληροφοριών ή την αναπαράσταση δεδομένων. Υπάρχουν πολυάριθμοι τύποι διαγραμμάτων, και ενώ οι όροι διάγραμμα και γράφημα συνδέονται στενά, δεν είναι απολύτως συνώνυμοι.

Μπορείτε να σκεφτείτε ένα γράφημα ως μια εικόνα που χρησιμοποιείται για την αναπαράσταση αριθμητικών δεδομένων. Μπορεί να είναι τόσο απλό όσο μια γραμμή σε μια σελίδα ή πολύ πιο περίπλοκο ανάλογα με το περιβάλλον και τα δεδομένα. Υπάρχουν επίσης πολλοί τύποι γραφημάτων και μερικοί είναι πιο κατάλληλα για την αναπαράσταση ορισμένων δεδομένων.

Τα διαγράμματα, όπως και τα γραφήματα, παρουσιάζουν πληροφορίες με ποικίλες μορφές, συμπεριλαμβανομένης της μορφής γραφημάτων. Ωστόσο, τα διαγράμματα μπορούν να αντιπροσωπεύουν περισσότερα από αριθμητικά δεδομένα. Αυτό σημαίνει ότι ενώ όλα τα γραφήματα είναι διαγράμματα, δεν είναι όλα τα διαγράμματα γραφήματα. Πίνακες, διαγράμματα, διαγράμματα ροής, ακόμη και χάρτες, είναι επίσης τύποι διαγραμμάτων.

4.3.1 Γραμμικό διάγραμμα

Ένα γραμμικό διάγραμμα είναι ένας τύπος διαγράμματος που σχεδιάζεται χρησιμοποιώντας τμήματα γραμμής για τη σύνδεση σημείων δεδομένων. Τα γραφήματα γραμμής είναι ένας από τους πιο συχνά χρησιμοποιούμενους τύπους γραφημάτων. Είναι ιδιαίτερα χρήσιμα για καταστάσεις όπως η απεικόνιση αλλαγών που συμβαίνουν σε μια χρονική περίοδο ή άλλες μεταβλητές.



Σχήμα 4-1 Παράδειγμα γραμμικού διαγράμματος

Υπάρχουν πλεονεκτήματα και μειονεκτήματα για την επιλογή ενός διαγράμματος γραμμής σε σχέση με κάποιο άλλο είδος διαγράμματος ή τεχνική οπτικοποίηση δεδομένων.

Πλεονεκτήματα:

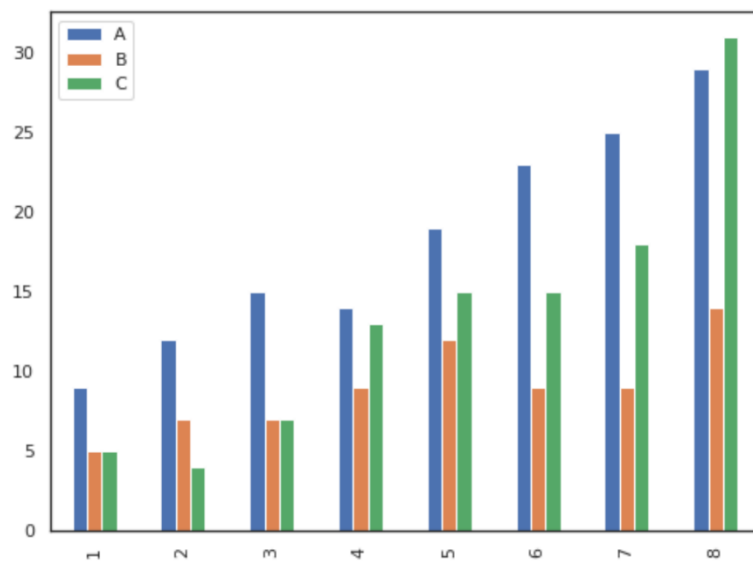
- Χρήσιμο για την αναπαράσταση συνεχών δεδομένων, όπως η αλλαγή με την πάροδο του χρόνου
- Επιτρέπει την πιθανή επέκταση δεδομένων
- Επιτρέπει την δημιουργία εκτίμησης για τα δεδομένα που λείπουν
- Επιτρέπει τη σύγκριση δύο ή περισσότερων στοιχείων για να διαπιστωθεί εάν υπάρχει κάποιου είδους σύνδεση ή σχέση

Μειονεκτήματα:

- Αρκετές φορές γίνεται δύσκολη η ακριβής εκτίμηση τιμής σε ένα δεδομένο σημείου του γραφήματος
- Πολλές γραμμές ή ακόμα και δύο γραμμές που έχουν τιμές που είναι πολύ παρόμοιες, μπορεί να κάνουν τη σύγκριση δεδομένων δύσκολη

4.3.2 Διάγραμμα ράβδων

Ένα διάγραμμα ράβδων, είναι ένας τύπος γραφήματος που χρησιμοποιεί ορθογώνιες ράβδους για να αναπαραστήσει τιμές. Τα μήκη των ράβδων είναι ανάλογα με τις τιμές τους. Όσο μεγαλύτερη είναι η μπάρα, τόσο μεγαλύτερο είναι το μέγεθος. Τα διαγράμματα ράβδων μπορούν να σχεδιαστούν είτε κάθετα είτε οριζόντια. Μπορούν να χρησιμοποιηθούν με πολλούς διαφορετικούς τρόπους και είναι σχετικά εύκολο να διαβαστούν και να κατανοηθούν.



Σχήμα 4-2 Παράδειγμα διαγράμματος ράβδων

Τα διαγράμματα ράβδων χωρίζονται σε δυο κατηγορίες:

- κατακόρυφο ραβδωτό διάγραμμα
- οριζόντιο ραβδωτό διάγραμμα
- ομαδοποιημένο διάγραμμα ράβδων
- στοιβαγμένο σύνθετο ραβδωτό διάγραμμα

Πλεονεκτήματα:

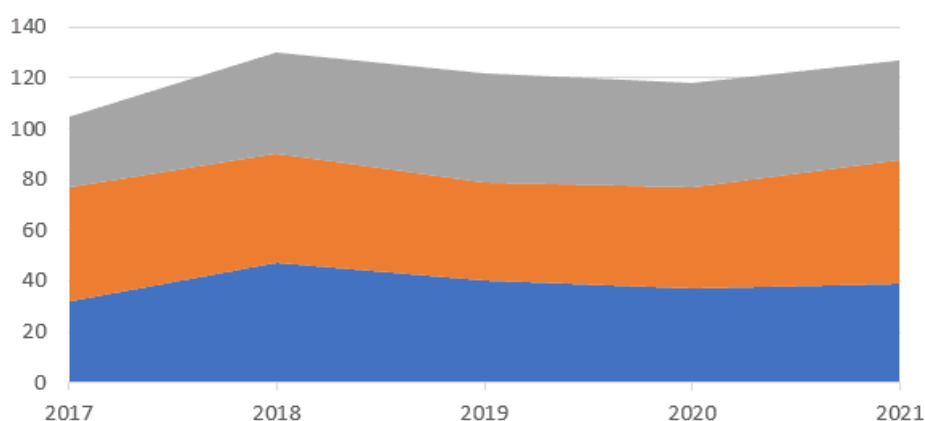
- Συνοψίζει μεγάλα δεδομένα σε οπτική μορφή
- Διευκρινίζει τις τάσεις καλύτερα από τους πίνακες
- Επιτρέπει την εύκολη εκτίμηση των βασικών τιμών
- Εμφανίζει σχετικούς αριθμούς ή αναλογίες πολλαπλών κατηγοριών

Μειονεκτήματα:

- Απαιτεί πρόσθετη γραπτή ή προφορική εξήγηση
- Εύκολη χειραγώγηση για λανθασμένη εντύπωση
- Αποτυγχάνει να αποκαλύψει υποθέσεις, αιτίες και αποτελέσματα
- Δεν είναι κατάλληλο εάν υπάρχει μεγάλος αριθμός κατηγοριών

4.3.3 Διάγραμμα περιοχής

Ένα διάγραμμα περιοχής μοιάζει με ένα γραμμικό διάγραμμα όσον αφορά τον τρόπο με τον οποίο οι τιμές δεδομένων σχεδιάζονται στο διάγραμμα και συνδέονται χρησιμοποιώντας τμήματα γραμμής. Σε ένα διάγραμμα περιοχής ωστόσο, η περιοχή μεταξύ των τμημάτων γραμμής και του άξονα x είναι γεμάτη με χρώμα. Ένα γράφημα περιοχής είναι μια καλή επιλογή όταν θέλετε να δείτε την αλλαγή του όγκου σε μια χρονική περίοδο, χωρίς να εστιάσετε σε συγκεκριμένες τιμές δεδομένων.



Σχήμα 4-3 Παράδειγμα διαγράμματος περιοχής

Πλεονεκτήματα:

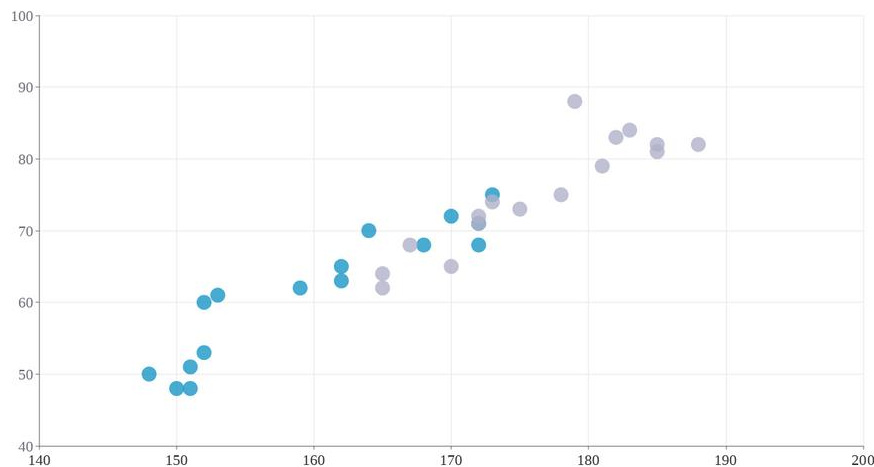
- Προσθέτει όγκο στα δεδομένα σας
- Για την απεικόνιση η σχέση ανάμεσα σε ολόκληρο και σε ένα μέρος μεταξύ των ομάδων
- Ανάλυση τάσης μεγέθους ποσοτικών δεδομένων
- Σύγκριση τάσης/αναλογίας κάθε κατηγορίας

Μειονεκτήματα:

- Χρήσιμο μόνο για σύγκριση τάσεων, όχι ακριβών τιμών
- Λειτουργεί αποτελεσματικά μόνο για μικρό αριθμό ομάδων

4.3.4 Διάγραμμα διασποράς

Η σχεδίαση ενός διαγράμματος διασποράς είναι ο απλούστερος τρόπος για την διαγραμματική αναπαράσταση δεδομένων. Για μια κατανομή (x,y) εάν οι τιμές των μεταβλητών x και y απεικονίζονται κατά μήκος του άξονα x και του άξονα y αντίστοιχα στο επίπεδο XY . Το διάγραμμα των κουκίδων που εμφανίζεται είναι γνωστό ως διάγραμμα διασποράς.



Σχήμα 4-4 Παράδειγμα διαγράμματος διασποράς

Πλεονεκτήματα:

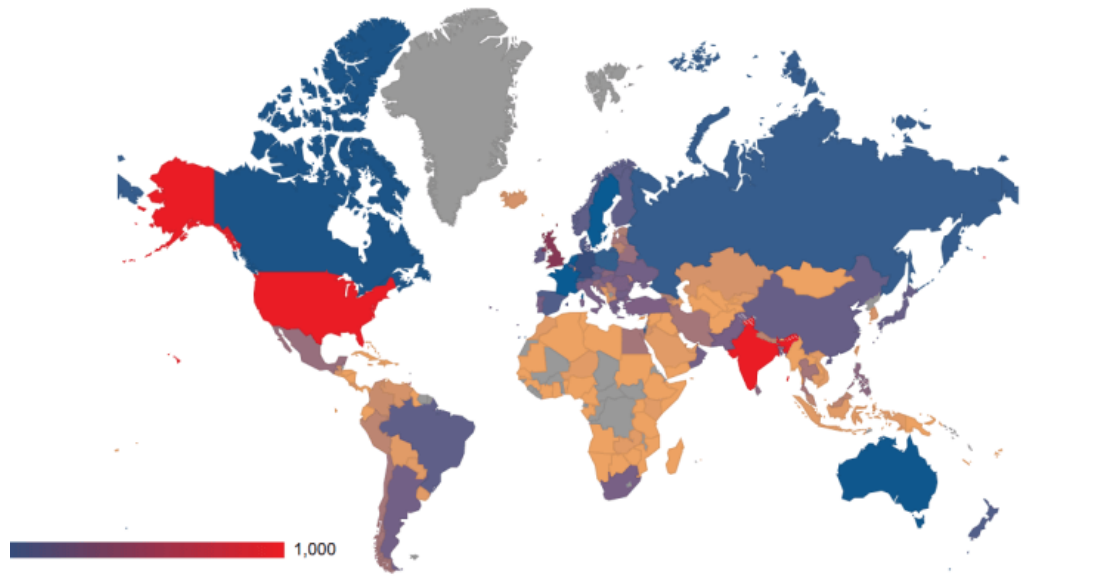
- Εύκολη κατασκευή και κατανόηση
- Οι ακραίες τιμές είναι μεμονωμένες και δεν επηρεάζουν τα αποτελέσματα
- Προσδιορισμός συνδέσεων ανάμεσα σε δύο μεταβλητές

Μειονεκτήματα:

- Απεικονίζει μόνο την κατεύθυνση συσχέτισης και όχι τον βαθμό
- Χρήσιμη μόνο για μικρό αριθμό δεδομένων
- Απεικόνιση μέχρι δύο μεταβλητών

4.3.5 Γεωγραφικό διάγραμμα

Τα γεωγραφικά διαγράμματα (γνωστά και ως γεωγραφήματα ή χάρτες) είναι ιδανικά εργαλεία οπτικοποίηση για δεδομένα που είναι διαφορετικά για κάθε πολιτεία, χώρα ή περιοχή. Ένα γεωγραφικό διάγραμμα δείχνει έναν θερμικό χάρτη τιμών κατανεμημένων σε διάφορες περιοχές ενός χάρτη.



Σχήμα 4-5 Παράδειγμα γεωγραφικού διαγράμματος

Πλεονεκτήματα:

- Ξεκάθαρη οπτικοποίηση δεδομένων
- Εύκολη δημιουργία συνδέσεων ανάμεσα σε δύο ή περισσότερες περιοχές
- Δημιουργία τρίτης διάστασης

Μειονεκτήματα:

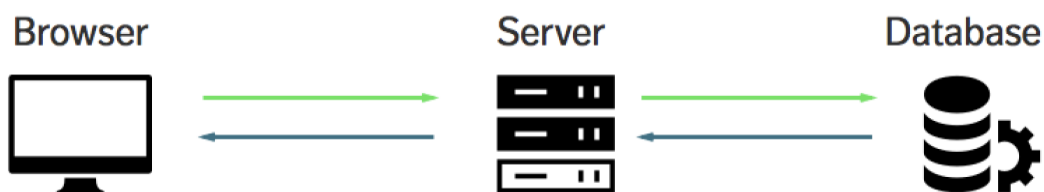
- Γεωγραφική παραμόρφωση
- Εύρεση κατάλληλων χρωμάτων
- Περιορισμός στους δείκτες που μπορούν να απεικονιστούν

Κεφάλαιο 5: Backend

4.1 Βάση δεδομένων

4.2 Σύστημα και επικοινωνία βάσης δεδομένων με τον πελάτη

Η ανάπτυξη back-end αναφέρεται στην ανάπτυξη λογικής από την πλευρά του διακομιστή που τροφοδοτεί ιστότοπους και εφαρμογές από τα παρασκήνια. Περιλαμβάνει όλο τον κώδικα που απαιτείται για τη δημιουργία της βάσης δεδομένων, του διακομιστή και της εφαρμογής. Μπορεί να είναι μια σύνδεση λογαριασμού ή μια αγορά από ένα ηλεκτρονικό κατάστημα. Ο κώδικας που γράφτηκε από προγραμματιστές back-end βοηθά τα προγράμματα περιήγησης να επικοινωνούν με πληροφορίες της βάσης δεδομένων.



Σχήμα 5-1 Παράδειγμα Δομής Εφαρμογής

Ένας προγραμματιστής backend εργάζεται μαζί με προγραμματιστές frontend, προγραμματιστές full stack, προγραμματιστές ή ειδικούς UX για τη δημιουργία ολοκληρωμένων ψηφιακών λύσεων. Διασφαλίζουν ότι ο ιστότοπος είναι επεκτάσιμος και ότι εξακολουθεί να μπορεί να λειτουργεί όταν χτυπηθεί από μεγάλα φορτία. Τα μεγάλα φορτία μπορεί να προκαλούνται είτε από μεγάλο αριθμό επισκέψεων είτε από απαιτητικά σενάρια. Είναι επίσης επιφορτισμένοι με τη

συντήρηση και τη δοκιμή των υπαρχόντων στοιχείων πίσω άκρου για να διασφαλίσουν ότι είναι όσο το δυνατόν πιο γρήγορα και αποτελεσματικά. Η αποθήκευση δεδομένων εμπίπτει επίσης στην αρμοδιότητα τους, κάτι που απαιτεί από αυτούς να έχουν καλή γνώση της ασφάλειας και της συμμόρφωσης των δεδομένων.

5.1 Βάση δεδομένων

Μια βάση δεδομένων είναι μια οργανωμένη συλλογή δομημένων πληροφοριών ή δεδομένων, που συνήθως αποθηκεύονται ηλεκτρονικά σε ένα σύστημα υπολογιστή. Μια βάση δεδομένων ελέγχεται συνήθως από ένα σύστημα διαχείρισης βάσεων δεδομένων (DBMS) . Τα δεδομένα και το DBMS, μαζί με τις εφαρμογές που σχετίζονται με αυτά, αναφέρονται ως σύστημα βάσης δεδομένων, ή σε απλή μορφή βάση δεδομένων.

Τα δεδομένα στους πιο συνηθισμένους τύπους βάσεων δεδομένων που λειτουργούν σήμερα μοντελοποιούνται συνήθως σε γραμμές και στήλες σε μια σειρά πινάκων για να καταστήσουν αποτελεσματική την επεξεργασία και την αναζήτηση δεδομένων. Στη συνέχεια, τα δεδομένα μπορούν να έχουν εύκολη πρόσβαση, διαχείριση, τροποποίηση, ενημέρωση, έλεγχο και οργάνωση. Οι περισσότερες βάσεις δεδομένων χρησιμοποιούν δομημένη γλώσσα ερωτημάτων (SQL) για τη σύνταξη και την αναζήτηση δεδομένων. Η SQL είναι μια γλώσσα προγραμματισμού που χρησιμοποιείται από σχεδόν όλες τις σχεσιακές βάσεις δεδομένων για την αναζήτηση, το χειρισμό και τον ορισμό δεδομένων και για την παροχή ελέγχου πρόσβασης.

5.1.1 Περιβάλλον και εργαλεία

Για την κατασκευή της βάσης δεδομένων χρησιμοποιήθηκε η MySQL όπου είναι ένα ανοιχτού κώδικα σχεσιακό σύστημα διαχείρισης βάσεων δεδομένων (RDBMS). Ένα RDBMS όπως το MySQL έχει ένα λειτουργικό σύστημα για την υλοποίηση μιας σχεσιακής βάσης δεδομένων στο σύστημα αποθήκευσης ενός υπολογιστή όπου διαχειρίζεται τους χρήστες, επιτρέπει την πρόσβαση στο δίκτυο και διευκολύνει τον

έλεγχο της ακεραιότητας της βάσης δεδομένων και τη δημιουργία αντιγράφων ασφαλείας.

Το MySQL Workbench είναι ένα οπτικό εργαλείο σχεδίασης βάσεων δεδομένων που ενσωματώνει την ανάπτυξη, τη διαχείριση, το σχεδιασμό, τη δημιουργία και τη συντήρηση της βάσης δεδομένων SQL σε ένα ενιαίο ολοκληρωμένο περιβάλλον ανάπτυξης για το σύστημα βάσεων δεδομένων MySQL. Είναι ένα από τα εργαλεία που βοήθησαν στην κατασκευή της βάση δεδομένων.

Χρησιμοποιήθηκε επίσης το εργαλείο βάσης δεδομένων του IntelliJ IDEA. Μέσω του εργαλείου δίνεται η δυνατότητα προβολής πληροφοριών για τους πίνακες της βάσης, όπως παράμετροι και δεδομένα, καθώς και η δημιουργία τροποποιήσεων πάνω σε αυτούς.

5.1.2 Σχεδιασμός βάσης δεδομένων

Ο σχεδιασμός της βάσης δεδομένων είναι η οργάνωση δεδομένων σύμφωνα με ένα μοντέλο βάσης δεδομένων. Αρχικά πρέπει να λάβουμε υπόψη πώς είναι δομημένα τα δεδομένα csv. Ένα CSV (comma separated values) είναι ένα οριοθετημένο αρχείο κειμένου που χρησιμοποιεί κόμμα για να διαχωρίσει τιμές. Ένα αρχείο CSV αποθηκεύει δεδομένα πίνακα (αριθμούς και κείμενο) σε απλό κείμενο. Το σύνολο δεδομένων που έχουμε περιλαμβάνει διάφορα δεδομένα με δείκτες που κυμαίνονται από εκπαίδευση έως περιβάλλον και υγεία.

Τα αρχεία csv του [Our World in Data \(OWID\)](#) έχουν συγκεκριμένη μορφή. Η πρώτη γραμμή ονομάζεται κεφαλίδα. Η κεφαλίδα μας δίνει πληροφορίες για την κάθε στήλη που περιέχει το αρχείο. Οι υπόλοιπες γραμμές αναφέρονται σε μια χώρα για μια συγκεκριμένη χρονιά. Κάθε στήλη δίνει την δική της πληροφορία. Η πρώτη και δεύτερη στήλη δείχνει την ονομασία της χώρας που αναφέρεται η γραμμή, καθώς και τον τριψήφιο κωδικό της. Στην τρίτη στήλη αναφέρεται η σχετική χρονιά και στις υπόλοιπες η τιμή που αντιπροσωπεύει το δείκτη που αναγράφεται στην επικεφαλίδα. Μπορεί στο αρχείο να υπάρχουν περισσότεροι από ένα δείκτη.

Οι παρακάτω πίνακες εξηγούν τις δύο μορφές που μπορούν να έχουν τα CSV αρχεία, με ένα δείκτη ή παραπάνω.

	A	B	C	D
1	Entity	Code	Year	Urban population (%) long-run with 2050 projections (OWID)
2	Afghanistan	AFG	1950	6
3	Afghanistan	AFG	1951	6.208
4	Afghanistan	AFG	1952	6.422
5	Afghanistan	AFG	1953	6.643
6	Afghanistan	AFG	1954	6.872
7	Afghanistan	AFG	1955	7.107

Σχήμα 5-2 Παράδειγμα μορφή Δεδομένα CSV με ένα Δείκτη

	A	B	C	D	E
1	Entity	Code	Year	Urban population (%)	Rural population (%)
2	Afghanistan	AFG	1950	6	94
3	Afghanistan	AFG	1951	6.208	93.792
4	Afghanistan	AFG	1952	6.422	93.578003
5	Afghanistan	AFG	1953	6.643	93.357002
6	Afghanistan	AFG	1954	6.872	93.127998
7	Afghanistan	AFG	1955	7.107	92.892998

Σχήμα 5-3 Παράδειγμα μορφής Δεδομένων CSV με περισσότερους από ένα Δείκτη

Όπως είναι γνωστό έχουμε περισσότερα από ένα csv αρχεία με έναν ή περισσότερους δείκτες. Αυτό μας αναγκάζει να αλλάξουμε τον σχεδιασμό της βάσης και να ομαλοποιήσουμε τα δεδομένα μας. Πριν από αυτό θα μιλήσουμε για το σχήμα της βάσης δεδομένων. Ο όρος σχήμα αναφέρεται στην οργάνωση των δεδομένων ως ένα προσχέδιο του τρόπου κατασκευής της βάσης δεδομένων (διαίρεται σε πίνακες βάσεων δεδομένων στην περίπτωση σχεσιακών βάσεων δεδομένων).

Κατά τη σχεδίαση του σχήματος της βάσης δεδομένων υπάρχει πάντα μια ανταλλαγή. Θα μπορούσαμε να χρησιμοποιήσουμε πολλούς πίνακες για τους δείκτες που θα κάνουν τα ερωτήματά μας πιο γρήγορα, αλλά αυτό θα έχει ως

αποτέλεσμα πιο πολλές ερωτήσεις σχεσιακής άλγεβρας. Για τον λόγο αυτό θα χρησιμοποιηθεί έναν ενιαίο πίνακα που σίγουρα θα επιβραδύνει την αναζήτηση, αλλά θα κάνει τις ερωτήσεις σχεσιακής άλγεβρας πιο απλές.

Η βάση δεδομένων θα περιέχει πέντε πίνακες. Ο πρώτος πίνακας θα ονομάζεται `country` και όπως υποδηλώνει το όνομα αποθηκεύει δεδομένα για κάθε χώρα. Ο παρακάτω πίνακας δηλώνει την δομή του `country`.

id	country_code_two_digits	country_name	country_code_three_digits	region	sub_region
1	AF	Afghanistan	AFG	Asia	Southern Asia
2	AX	Åland Islands	ALA	Europe	Northern Europe
3	AL	Albania	ALB	Europe	Southern Europe
4	DZ	Algeria	DZA	Africa	Northern Africa
5	AS	American Samoa	ASM	Oceania	Polynesia
6	AD	Andorra	AND	Europe	Southern Europe
7	AO	Angola	AGO	Africa	Sub-Saharan Africa
8	AI	Anguilla	AIA	Americas	Latin America and the Caribbean
9	AQ	Antarctica	ATA	<null>	<null>
10	AG	Antigua and Barbuda	ATG	Americas	Latin America and the Caribbean

Σχήμα 5-4 Παράδειγμα Δομής Πίνακα `country` της Βάσης Δεδομένων

Τα έξι πεδία που απεικονίζονται στον πίνακα είναι το `id` που είναι η ταυτότητα κάθε χώρας, το `country_name` που όπως υποδηλώνει το όνομα του είναι η ονομασία της χώρας. Ακολουθούν τα `country_code_two_digits` και `country_code_three_digits` όπου είναι οι κωδικοί των χωρών με δυο και τρία γράμματα αντίστοιχα. Τέλος υπάρχουν τα πεδία `region` και `sub_region` όπου είναι οι περιοχές που ανήκουν.

Ακόμα στην βάση δεδομένων έχουμε τον πίνακα `indicator` όπου περιέχει όλες τις πληροφορίες για καθένα από τους δείκτες που χρησιμοποιούμε. Ο παρακάτω πίνακας μας δηλώνει την δομή του `indicator`.

id	indicator_label	indicator_name	description
1	Clean fuels and techno...	Access to clean fuels a...	Clean cooking fuels and technologies represent non-solid fuels ...
2	Age at marriage women	Estimated average age a...	Estimated average age at marriage.
3	Number of births	Estimates, 1950 - 2020:...	Estimates, 1950 - 2020: Annually interpolated demographic indic...
4	Deaths from executions	Number of executions (A...	Deaths - Executions (Amnesty International) - Sex: Both - Age: ...
5	Deaths from meningitis	Deaths - Meningitis - S...	Deaths - Meningitis - Sex: Both - Age: All Ages (Number).
6	Deaths from neoplasms	Deaths - Neoplasms - Se...	Deaths - Neoplasms - Sex: Both - Age: All Ages (Number).
7	Deaths from fire, heat...	Deaths - Fire, heat, an...	Deaths - Fire, heat, and hot substances - Sex: Both - Age: All ...
8	Deaths from malaria	Deaths - Malaria - Sex:...	Deaths - Malaria - Sex: Both - Age: All Ages (Number).
9	Deaths from drowning	Deaths - Drowning - Sex...	Deaths - Drowning - Sex: Both - Age: All Ages (Number).
10	Deaths from interperso...	Deaths - Interpersonal ...	Deaths - Interpersonal violence - Sex: Both - Age: All Ages (Nu...

Σχήμα 5-5 Παράδειγμα Δομής Πίνακα `indicator` της Βάσης Δεδομένων

Τα τέσσερα πεδία που υπάρχουν στον πίνακα αναφέρονται στην ταυτότητα του δείκτη, όπου είναι το id, την περιγραφή του description, την επιγραφή του όπου είναι το indicator_label και το όνομα όπως βρέθηκε στην κεφαλίδα από τα CSV αρχεία indicators_name.

Ο κάθε δείκτης ανήκει σε μια ή περισσότερες κατηγορίες οι οποίες αναγράφονται σε ένα πίνακα βάσης δεδομένων που ονομάζεται category. Ο παρακάτω πίνακας δηλώνει την δομή του category.

id	category_name	filter
1	Demographic changes	main
2	Energy	main
3	Environment	main
4	Health	main
5	Poverty and Economic Development	main
6	Living conditions, Community and well being	main
7	Education and Knowledge	main
8	Human rights and Democracy	main
9	War	main
10	Population	sub

Σχήμα 5-6 Παράδειγμα Δομής Πίνακα category της Βάσης Δεδομένων

Τα τρία πεδία που απεικονίζονται στον πίνακα είναι το id που είναι η μοναδική ταυτότητα της κατηγορίας, το category_name που είναι η ονομασία της και το filter όπου είναι η κατηγορία φίλτρων στην οποία ανήκει η κατηγορία.

Αυτούς τους δύο πίνακες τους ενώνει ένα πίνακας συνένωσης (join table) όπου ονομάζεται categorized_indicators. Ο λόγος ύπαρξης αυτού του πίνακα είναι για να καταλάβουμε ποιος δείκτης ανήκει σε ποιες κατηγορίες. Ο παρακάτω πίνακας δηλώνει την δομή του categorized_indicator.

indicator_id	categories_id
3	1
37	1
172	1
220	1
221	1
271	1
272	1
273	1
1	2
224	2

Σχήμα 5-7 Παράδειγμα Δομής Πίνακα categorized_indicators της Βάσης Δεδομένων

Ο categorized_indicators έχει για πεδία τις μοναδικές ταυτότητες των πινάκων category και indicator.

Ο Τελευταίος πίνακας της βάσης δεδομένων ονομάζεται metric. Ο πίνακας αυτός περιέχει πληροφορίες για κάθε μέτρηση ξεχωριστά. Η κάθε μέτρηση αναφέρεται σε μια χώρα για μια συγκεκριμένη χρονιά. Ο παρακάτω πίνακας δηλώνει την δομή του metric.

id	value	year	country_id	indicator_id
1	8.8	2000	1	1
2	9.51	2001	1	1
3	10.39	2002	1	1
4	11.46	2003	1	1
5	12.43	2004	1	1
6	13.49	2005	1	1
7	14.81	2006	1	1
8	15.99	2007	1	1
9	17.44	2008	1	1
10	18.84	2009	1	1

Σχήμα 5-8 Παράδειγμα Δομής metric Πίνακα της Βάσης Δεδομένων

Ο πίνακας metric περιέχει πέντε πεδία. Το Πρώτο πεδίο ονομάζεται id και αναφέρεται στην μοναδική ταυτότητα της μέτρησης. Το δεύτερο πεδίο δηλώνει την τιμή που έχει η μέτρηση για τον δείκτη. Το τρίτο πεδίο αναφέρεται στην χρονιά και το τέταρτο και πέμπτο στην μοναδικές ταυτότητες της χώρας και του δείκτη αντίστοιχα για την συγκεκριμένη μέτρηση.

5.1.3 Δημιουργία βάσης δεδομένων

Η κατασκευή της βάσης δεδομένων έγινε μέσω της γλώσσας περιγραφής δεδομένων (DDL). Στο πλαίσιο της SQL ,η γλώσσα περιγραφής δεδομένων (DDL) είναι μια σύνταξη για τη δημιουργία και την τροποποίηση αντικειμένων βάσης δεδομένων όπως πίνακες, δείκτες και χρήστες. Οι δηλώσεις DDL είναι παρόμοιες με μια γλώσσα προγραμματισμού υπολογιστή για τον ορισμό δομών δεδομένων, ειδικά σχημάτων βάσεων δεδομένων . Τα κοινά παραδείγματα δηλώσεων DDL περιλαμβάνουν CREATE, ALERT και DROP.

Γλώσσα περιγραφής δεδομένων (DDL) για τον πίνακα country της βάσης δεδομένων.

```

1 CREATE TABLE `country` (
2   `id` bigint NOT NULL AUTO_INCREMENT,
3   `country_code_three_digits` varchar(3) NOT NULL,
4   `country_code_two_digits` varchar(2) NOT NULL,
5   `country_name` varchar(100) NOT NULL,
6   `region` varchar(25) DEFAULT NULL,
7   `sub_region` varchar(100) DEFAULT NULL,
8   PRIMARY KEY (`id`),
9   UNIQUE KEY `UK_d8wb65eemw2oo3dq14nl6qrqq` (`country_code_three_digits`),
10  UNIQUE KEY `UK_63mnb77k24jy9jui610p92yq` (`country_code_two_digits`),
11  UNIQUE KEY `UK_qrkn270121aljmucrdbnmd35p` (`country_name`)
12 ) ENGINE=InnoDB AUTO_INCREMENT=250 DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_0900_ai_ci

```

Σχήμα 5-9 DDL του Πίνακα country της Βάσης Δεδομένων

Γλώσσα περιγραφής δεδομένων (DDL) για τον πίνακα indicator της βάσης δεδομένων.

```

1 CREATE TABLE `indicator` (
2   `id` bigint NOT NULL AUTO_INCREMENT,
3   `description` longtext,
4   `indicator_label` varchar(255) DEFAULT NULL,
5   `indicator_name` varchar(255) NOT NULL,
6   PRIMARY KEY (`id`),
7   UNIQUE KEY `UK_33sk0c9mw7ux61yu042dwr946` (`indicator_label`)
8 ) ENGINE=InnoDB AUTO_INCREMENT=276 DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_0900_ai_ci

```

Σχήμα 5-10 DDL του Πίνακα indicator της Βάσης Δεδομένων

Γλώσσα περιγραφής δεδομένων (DDL) για τον πίνακα category της βάσης δεδομένων.

```

1 CREATE TABLE `category` (
2   `id` bigint NOT NULL AUTO_INCREMENT,
3   `category_name` varchar(255) NOT NULL,
4   `filter` varchar(4) DEFAULT NULL,
5   PRIMARY KEY (`id`),
6   UNIQUE KEY `UK_lroeo5fvfdeg4hpicn4lw7x9b` (`category_name`)
7 ) ENGINE=InnoDB AUTO_INCREMENT=42 DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_0900_ai_ci

```

Σχήμα 5-11 DDL του Πίνακα category της Βάσης Δεδομένων

Γλώσσα περιγραφής δεδομένων (DDL) για τον πίνακα categorized_indicators της βάσης δεδομένων.

```

1 CREATE TABLE `categorized_indicators` (
2   `indicator_id` bigint NOT NULL,
3   `categories_id` bigint NOT NULL,
4   PRIMARY KEY (`indicator_id`,`categories_id`),
5   KEY `FK8ibtwnhib3o8lmiufejl1w736` (`categories_id`),
6   CONSTRAINT `FK8ibtwnhib3o8lmiufejl1w736` FOREIGN KEY (`categories_id`) REFERENCES `category` (`id`),
7   CONSTRAINT `FKym26j3gkoahudekwesych6o` FOREIGN KEY (`indicator_id`) REFERENCES `indicator` (`id`)
8 ) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_0900_ai_ci

```

Σχήμα 5-12 DDL του Πίνακα `categorized_indicators` της Βάσης Δεδομένων

Γλώσσα περιγραφής δεδομένων (DDL) για τον πίνακα `metric` της βάσης δεδομένων.

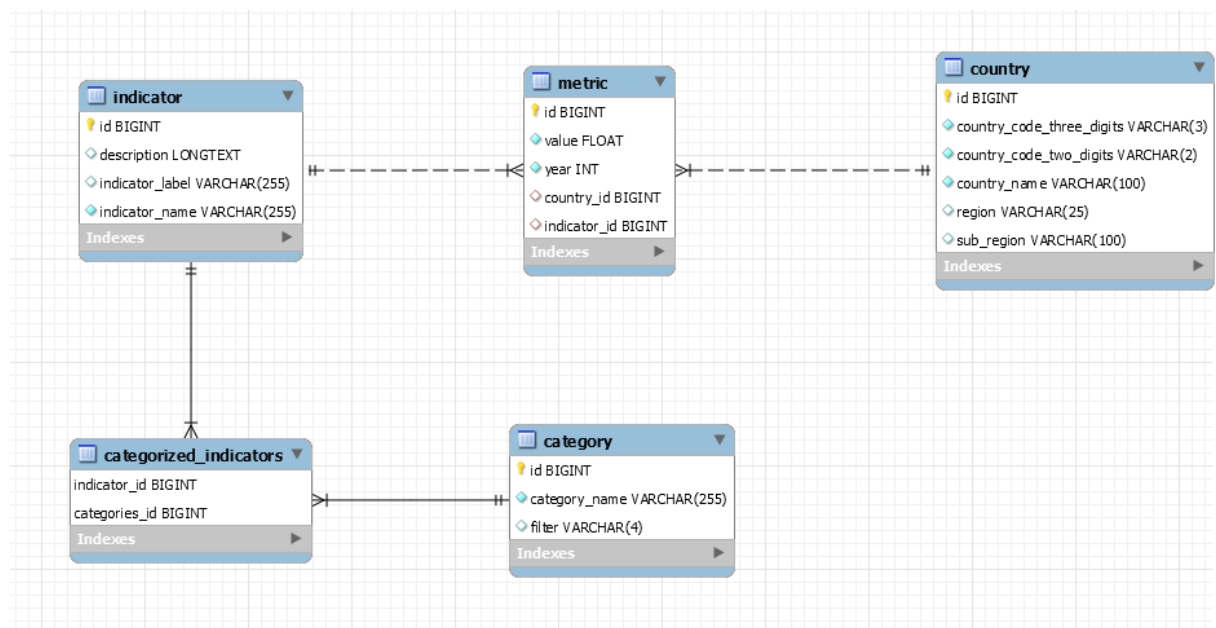
```

1 CREATE TABLE `metric` (
2   `id` bigint NOT NULL AUTO_INCREMENT,
3   `value` float NOT NULL,
4   `year` int NOT NULL,
5   `country_id` bigint DEFAULT NULL,
6   `indicator_id` bigint DEFAULT NULL,
7   PRIMARY KEY (`id`),
8   KEY `FKhslywdl9laq4qfh46sila1il5` (`country_id`),
9   KEY `FKlr724jlqlwhyajlbqddumcx1` (`indicator_id`),
10  CONSTRAINT `FKhslywdl9laq4qfh46sila1il5` FOREIGN KEY (`country_id`) REFERENCES `country` (`id`),
11  CONSTRAINT `FKlr724jlqlwhyajlbqddumcx1` FOREIGN KEY (`indicator_id`) REFERENCES `indicator` (`id`)
12 ) ENGINE=InnoDB AUTO_INCREMENT=1749520 DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_0900_ai_ci

```

Σχήμα 5-13 DDL του Πίνακα `metric` της Βάσης Δεδομένων

Η τελική μορφή της βάσης δεδομένων είναι:

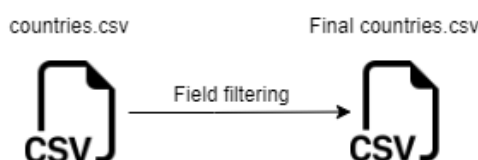


Σχήμα 5-14 Σχήμα Βάσης Δεδομένων

5.1.4 Προετοιμασία δεδομένων

Για την φόρτωση δεδομένων στην βάση χρειάζονται κατάλληλες τροποποιήσεις. Αν υπάρχουν περισσότερα από ένα csv αρχεία, όπως στην περίπτωση με τους δείκτες, χρειάζεται να πραγματοποιηθεί συνένωσή τους σε ένα τελικό αρχείο. Επίσης πρέπει τα δεδομένα των αρχείων να βρίσκονται ακριβώς στην ίδια μορφή με αυτή που έχουν οι πίνακες της βάσης δεδομένων γιατί αλλιώς δεν μπορεί να επιτευχθεί η φόρτωση.

Αρχικά πρέπει να βρεθούν δεδομένα για τις χώρες που δεν υπάρχουν μέσα στα csv αρχεία των δεικτών όπως κωδικοί και περιοχές. Για να αντληθούν αυτά υπάρχουν έτοιμες βάσεις σε csv αρχείο που περιέχουν πληθώρα δεδομένων για κάθε χώρα.



Σχήμα 5-15 Διάγραμμα επεξεργασίας countries csv αρχείου

```
Country,alpha-2,alpha-3,country-code,iso_3166-2:region,sub-region,intermediate-region,region-code,sub-region-code,intermediate-region-code
Afghanistan,AF,AFG,004,ISO 3166-2:AF,Asia,Southern Asia,"",142,034,""
Åland Islands,AX,ALA,248,ISO 3166-2:AX,Europe,Northern Europe,"",150,154,""
Albania,AL,ALB,008,ISO 3166-2:AL,Europe,Southern Europe,"",150,039,""
Algeria,DZ,DZA,012,ISO 3166-2:DZ,Africa,Northern Africa,"",002,015,""
American Samoa,AS,ASM,016,ISO 3166-2:AS,Oceania,Polynesia,"",009,061,""
Andorra,AD,AND,020,ISO 3166-2:AD,Europe,Southern Europe,"",150,039,""
Angola,AO,AO,024,ISO 3166-2:AO,Africa,Sub-Saharan Africa,Middle Africa,002,202,017
Anguilla,AI,AIA,660,ISO 3166-2:AI,Americas,Latin America and the Caribbean,Caribbean,019,419,029
Antarctica,AQ,ATA,010,ISO 3166-2:AQ,"","",,"",,"",,""
```

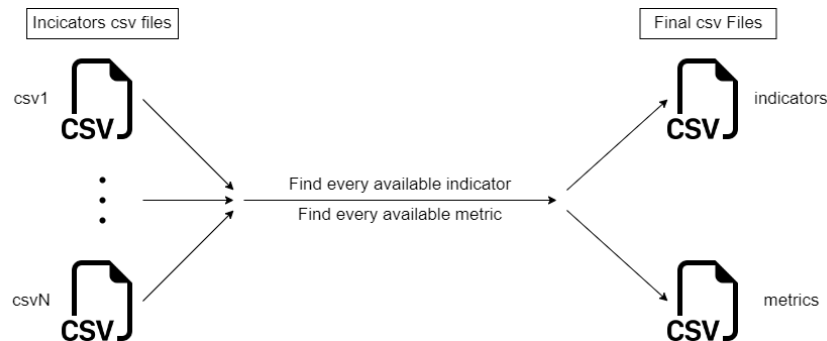
Σχήμα 5-16 Παράδειγμα έτοιμης βάσης χωρών

Από αυτά μέσω αλγόριθμου φιλτραρίστηκαν τα χρήσιμα δεδομένα και δημιουργήθηκε το countries.csv αρχείο.

```
Country,CountryCodeTwoDigits,CountryCodeThreeDigits,Region,SubRegion
Afghanistan,AF,AFG,Asia,Southern Asia
Åland Islands,AX,ALA,Europe,Northern Europe
Albania,AL,ALB,Europe,Southern Europe
Algeria,DZ,DZA,Africa,Northern Africa
American Samoa,AS,ASM,Oceania,Polynesia
Andorra,AD,AND,Europe,Southern Europe
Angola,AO,AO,Africa,Sub-Saharan Africa
Anguilla,AI,AIA,Americas,Latin America and the Caribbean
Antarctica,AQ,ATA,,
```

Σχήμα 5-17 Παράδειγμα τελικής μορφής countries αρχείου έτοιμο για φόρτωση

Στην συνέχεια θα πρέπει να διαβαστούν όλα τα csv αρχεία που αφορούν τους δείκτες. Αυτό θα έχει ως αποτέλεσμα την δημιουργία δυο καινούργιων αρχείων.



Σχήμα 5-18 Διάγραμμα επεξεργασίας indicators csv αρχείων

Το πρώτο αρχείο ονομάζεται indicators.csv. Αυτό θα περιέχει μόνο την ονομασία του όπως βρέθηκε στις κεφαλίδες των αρχείων.

```

Indicator
Access to clean fuels and technologies for cooking (% of population)
Estimated average age at marriage", women
Estimates", 1950 - 2020: Annually interpolated demographic indicators - Births (thousands)
Number of executions (Amnesty International)
Deaths - Meningitis - Sex: Both - Age: All Ages (Number)
Deaths - Neoplasms - Sex: Both - Age: All Ages (Number)
Deaths - Fire", heat", and hot substances - Sex: Both - Age: All Ages (Number)
Deaths - Malaria - Sex: Both - Age: All Ages (Number)
Deaths - Drowning - Sex: Both - Age: All Ages (Number)
  
```

Σχήμα 5-19 Παράδειγμα τελικής μορφής indicators αρχείου έτοιμο για φόρτωση

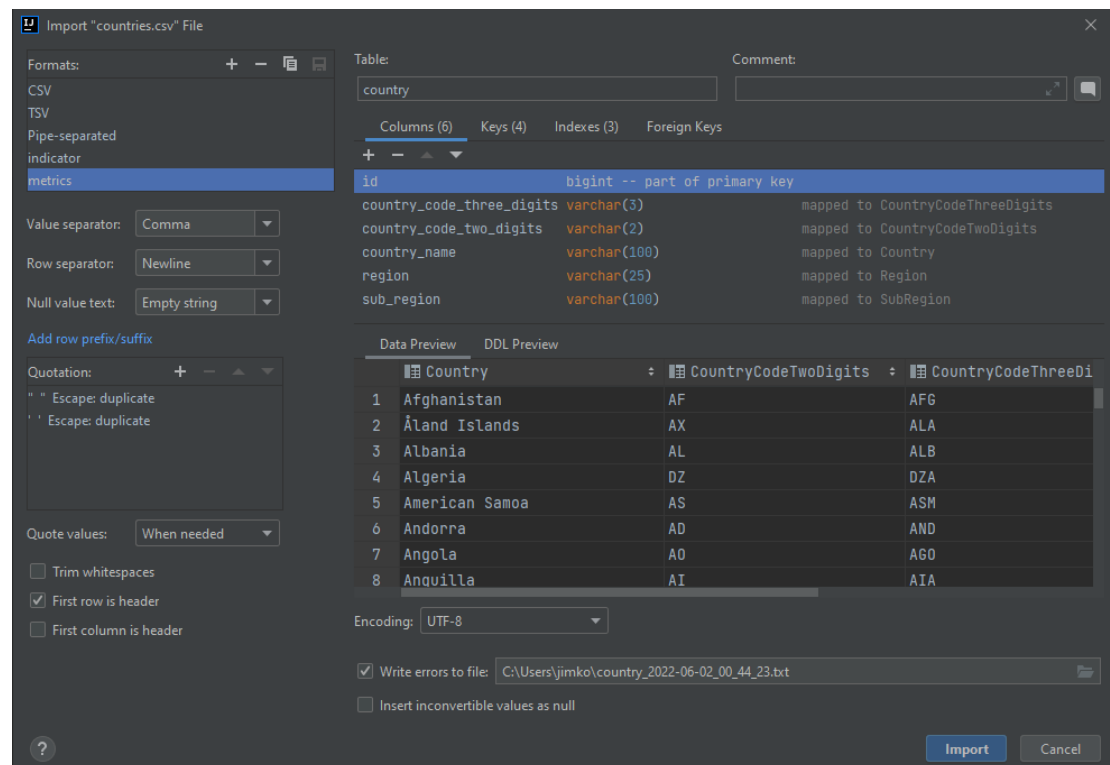
Το δεύτερο αρχείο ονομάζεται metrics.csv και θα περιέχει όλες τις μετρήσεις που βρέθηκαν μέσα στα αρχεία δεικτών. Η κάθε μέτρηση θα αποτελείται από τις μοναδικές ταυτότητες των χωρών και των δεικτών, για να γίνεται αντιληπτό που αναφέρεται. Επίσης θα υπάρχει η χρονιά της μέτρησης αλλά και η τιμή της.

```

CountryId,Year,IndicatorId,value
1,2000,1,8.8
1,2001,1,9.51
1,2002,1,10.39
1,2003,1,11.46
1,2004,1,12.43
1,2005,1,13.49
1,2006,1,14.81
1,2007,1,15.99
1,2008,1,17.44
  
```

Σχήμα 5-20 Παράδειγμα τελικής μορφής metrics αρχείου έτοιμο για φόρτωση

Βασική παραμετροποίηση είναι η επιλογή της ύπαρξης κεφαλίδας όπου θα αποφύγει την εισαγωγή της ως δεδομένα. Στην συνέχεια πρέπει να γίνει αντιστοίχιση των πεδίων του πίνακα της βάσης με την στήλη που βρίσκονται τα αντίστοιχα δεδομένα. Τέλος μετά την ολοκλήρωση, τα δεδομένα βρίσκονται στην βάση δεδομένων μας.



Σχήμα 5-22 Παράδειγμα παραμετροποιήσεων για εισαγωγή αρχείου σε Πίνακα της Βάσης Δεδομένων

5.1.6 Μη αυτοματοποιημένες παρεμβάσεις

Πέρα από την φόρτωση δεδομένων χρειάστηκαν επιπλέον επεμβάσεις για να φτάσουν τα δεδομένα της βάση στην τελική τους μορφή. Αυτές οι παρεμβάσεις δεν γινόταν να γίνουν με αυτοματοποιημένο τρόπο ή μέσω κάποιου αλγορίθμου. Πραγματοποιήθηκαν χειροκίνητα από τις αλλαγές που επιτρέπει να γίνονται μέσα από την προέκταση της βάσης δεδομένων του περιβάλλοντος IntelliJ IDEA της JetBrains. Η προέκταση μας επιτρέπει πέρα από την εμφάνιση των δεδομένων, που έχουν αποθηκευτεί στους πίνακες της βάσης δεδομένων, την επεξεργασία πεδίων αλλά και την προσθήκη καινούργιων γραμμών σε αυτούς.

Το πρώτο σημείο που έγιναν οι παρεμβάσεις είναι στον πίνακα `indicator` της βάσης. Σε αυτό τον πίνακα υπάρχουν πεδία που δεν μπορούν να συμπληρωθούν αυτόματα από τα δεδομένα που υπάρχουν στα `csv` αρχεία τους. Αυτά τα δύο πεδία είναι το `indicators_label` και το `description`. Το πρώτο αναφέρεται όπως μας προϋδεάζει το όνομά του στην επιγραφή του δείκτη και το δεύτερο στην περιγραφή που αναφέρεται στον δείκτη αυτό.

Στην συνέχεια αυτοματοποιημένα δεν μπορεί να γίνει η δημιουργία των δεδομένων για τον πίνακα `category` της βάσης δεδομένων. Αυτός ο πίνακας αποθηκεύει δεδομένα για κατηγορίες που ανήκουν οι δείκτες. Ο λόγος που δεν μπορούσε να γίνει αυτοματοποιημένα η εύρεση των κατηγοριών είναι γιατί δημιουργήθηκαν από ανθρώπινο παράγοντα.

Τέλος ο πίνακας της βάσης `categorized_indicators` είναι ένας πίνακας συνένωσης του πίνακα `indicator` και `category`. Η κατανομή των κατηγοριών στους δείκτες έγινε και αυτή με μη αυτοματοποιημένο τρόπο, και όχι από κάποια πηγή δεδομένων.

5.1.7 Αντίγραφο ασφαλείας δεδομένων

Μετά και τις τελευταίες παρεμβάσεις κατέληξαν τα δεδομένα στην τελική τους μορφή. Για να μην ξαναχρηαστεί όλη η παραπάνω χρονοβόρα διαδικασία έγινε η εξαγωγή τους και από τους πέντε πίνακες της βάσης δεδομένων σε πέντε διαφορετικά `csv` αρχεία. Σε περίπτωση βλάβης ή για κάποιον άλλο λόγο συντήρησης, αφού κατασκευαστεί η βάση μπορεί να επιτευχθεί η φόρτωση αυτών των πέντε `csv` αρχείων όπως αναφέρεται στην υποενότητα **4.1.5**.

5.2 Σύστημα και επικοινωνία βάσης δεδομένων με τον πελάτη

Ο τρόπος με τον οποίο γίνεται η ανάπτυξη ιστού `backend` είναι ότι όταν ο πελάτης επισκέπτεται μια σελίδα δημιουργεί αιτήσεις για περιεχόμενο. Ο διακομιστής επεξεργάζεται αυτά τα αιτήματα και δημιουργεί μια απάντηση που αποστέλλεται στο πρόγραμμα περιήγησης.

Όταν ένας ιστότοπος αποδίδει από την πλευρά του διακομιστή, όλες οι διαδικασίες που εμπλέκονται στη δημιουργία μιας σελίδας HTML που μπορεί να κατανοήσει το πρόγραμμα περιήγησής στον ιστό γίνονται σε έναν απομακρυσμένο διακομιστή που φιλοξενεί τον ιστότοπο ή την εφαρμογή Ιστού. Αυτό περιλαμβάνει την αναζήτηση βάσεων δεδομένων για πληροφορίες και την επεξεργασία οποιασδήποτε λογικής που απαιτεί η εφαρμογή Ιστού.

Ενώ ο απομακρυσμένος διακομιστής είναι απασχολημένος στη δουλειά, το πρόγραμμα περιήγησής είναι αδρανές, περιμένοντας από τον διακομιστή να ολοκληρώσει την επεξεργασία του αιτήματος και να στείλει μια απάντηση. Όταν λαμβάνεται η απάντηση, τα προγράμματα περιήγησης ιστού την ερμηνεύουν και εμφανίζουν το περιεχόμενο στην οθόνη.

Για την κατασκευή του διακομιστή χρησιμοποιήθηκε ένα RESTful service με Spring και η επικοινωνία διακομιστή με πελάτη μέσω μιας διεπαφής API.

5.2.1 REST

Το Representational State Transfer (REST) είναι ένα αρχιτεκτονικό στυλ λογισμικού που δημιουργήθηκε για να καθοδηγήσει το σχεδιασμό και την ανάπτυξη της αρχιτεκτονικής για τον Παγκόσμιο Ιστό . Το REST ορίζει ένα σύνολο περιορισμών για το πώς πρέπει να συμπεριφέρεται η αρχιτεκτονική ενός κατανεμημένου συστήματος υπερμέσων σε κλίμακα Διαδικτύου , όπως ο Ιστός. Το αρχιτεκτονικό στυλ REST δίνει έμφαση στην επεκτασιμότητα των αλληλεπιδράσεων μεταξύ των στοιχείων, στις ομοιόμορφες διεπαφές, στην ανεξάρτητη ανάπτυξη των στοιχείων και στη δημιουργία μιας αρχιτεκτονικής με στρώσεις για να διευκολύνει την αποθήκευση στοιχείων στην κρυφή μνήμη για τη μείωση της αντιληπτής από τον χρήστη καθυστέρησης, επιβολή ασφάλειας και ενθυλάκωση παλαιών συστημάτων .

Το REST έχει χρησιμοποιηθεί σε όλη τη βιομηχανία λογισμικού και είναι ένα ευρέως αποδεκτό σύνολο οδηγιών για τη δημιουργία αξιόπιστων API ιστού διατηρώντας την κατάσταση από προηγούμενα αιτήματα. Ένα Web API που υπακούει στους περιορισμούς REST περιγράφεται ανεπίσημα ως RESTful. Τα RESTful

Web API βασίζονται συνήθως σε μεθόδους HTTP για πρόσβαση σε πόρους μέσω παραμέτρων κωδικοποιημένων με URL και στη χρήση JSON ή XML για τη μετάδοση δεδομένων.

Τα Web resources ορίστηκαν για πρώτη φορά στον Παγκόσμιο Ιστό ως έγγραφα ή αρχεία που προσδιορίζονται από τις διευθύνσεις URL τους . Σήμερα, ο ορισμός είναι πολύ πιο γενικός και αφηρημένος και περιλαμβάνει κάθε πράγμα, οντότητα ή ενέργεια που μπορεί να αναγνωριστεί, να ονομαστεί, να απευθυνθεί, να χειριστεί ή να εκτελεστεί με οποιονδήποτε τρόπο στον Ιστό. Σε μια υπηρεσία RESTful Web, τα αιτήματα που γίνονται στο URI ενός πόρου προκαλούν μια απάντηση με ένα ωφέλιμο φορτίο μορφοποιημένο σε HTML , XML , JSON ή κάποια άλλη μορφή. Για παράδειγμα, η απάντηση μπορεί να επιβεβαιώσει ότι η κατάσταση του πόρου έχει αλλάξει. Η απάντηση μπορεί επίσης να περιλαμβάνει υπερκείμενο συνδέσμων προς σχετικούς πόρους. Το πιο κοινό πρωτόκολλο για αυτά τα αιτήματα και απαντήσεις είναι το HTTP. Παρέχει λειτουργίες (μέθοδοι HTTP) όπως GET, POST, PUT και DELETE. Χρησιμοποιώντας ένα πρωτόκολλο χωρίς κατάσταση και τυπικές λειτουργίες, τα συστήματα RESTful στοχεύουν σε γρήγορη απόδοση, αξιοπιστία και ικανότητα ανάπτυξης με επαναχρησιμοποίηση στοιχείων που μπορούν να διαχειρίζονται και να ενημερώνονται χωρίς να επηρεάζεται το σύστημα στο σύνολό του, ακόμη και όταν εκτελείται.

Ο στόχος του REST είναι η αύξηση της απόδοσης, της επεκτασιμότητας, της απλότητας, της δυνατότητας τροποποίησης, της ορατότητας, της φορητότητας και της αξιοπιστίας. Αυτό επιτυγχάνεται ακολουθώντας τις αρχές REST, όπως η αρχιτεκτονική πελάτη-διακομιστή, η ανικανότητα, η δυνατότητα αποθήκευσης στην κρυφή μνήμη, η χρήση ενός πολυεπίπεδου συστήματος, η υποστήριξη για κώδικα κατά παραγγελία και η χρήση ομοιόμορφης διεπαφής. Αυτές οι αρχές πρέπει να ακολουθούνται για να ταξινομηθεί το σύστημα ως RESTful.

5.2.2 Spring Boot

Spring Boot είναι ένα πλαίσιο ανοιχτού κώδικα που βασίζεται σε Java που χρησιμοποιείται για τη δημιουργία micro Service. Αναπτύχθηκε από την Pivotal Team και χρησιμοποιείται για την κατασκευή αυτόνομων και έτοιμων για παραγωγή εφαρμογών spring .

Micro Service είναι μια αρχιτεκτονική που επιτρέπει στους προγραμματιστές να αναπτύσσουν υπηρεσίες ανεξάρτητα. Κάθε υπηρεσία που εκτελείται έχει τη δική της διαδικασία και αυτό επιτυγχάνει το ελαφρύ μοντέλο υποστήριξης επιχειρηματικών εφαρμογών. Οι υπηρεσίες Micro προσφέρουν τα ακόλουθα πλεονεκτήματα:

- Εύκολη ανάπτυξη
- Απλή επεκτασιμότητα
- Συμβατό με containers
- Ελάχιστη διαμόρφωση
- Μικρότερος χρόνος παραγωγής

Το Spring Boot παρέχει μια καλή πλατφόρμα για προγραμματιστές Java για να αναπτύξουν μια αυτόνομη εφαρμογή spring ποιότητας παραγωγής που μπορείτε απλώς να εκτελεστεί . Μπορεί να ξεκινήσει με ελάχιστες διαμορφώσεις χωρίς να χρειάζεται μια ολόκληρη ρύθμιση παραμέτρων Spring. Το Spring Boot προσφέρει τα ακόλουθα πλεονεκτήματα:

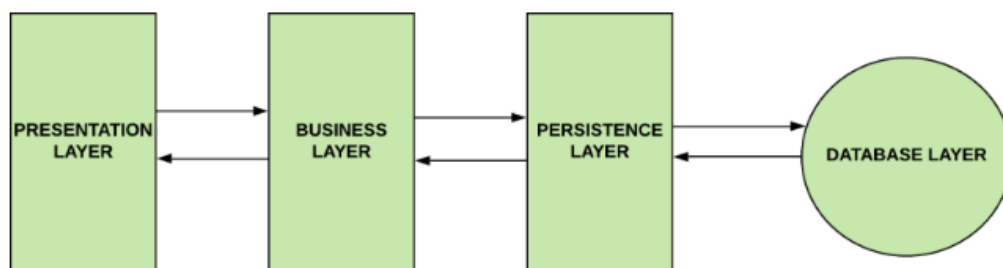
- Εύκολη κατανόηση και ανάπτυξη εφαρμογών Spring
- Αυξάνει την παραγωγικότητα
- Μειώνει τον χρόνο ανάπτυξης

Ο λόγος που επιλέχτηκε το Spring Boot είναι λόγω των δυνατοτήτων και των πλεονεκτημάτων που προσφέρει. Κάποιες από αυτές είναι ότι παρέχει έναν ευέλικτο τρόπο ρύθμισης παραμέτρων Java Beans, διαμορφώσεων XML και συναλλαγών βάσης δεδομένων, διαχειρίζεται τα τελικά σημεία REST, τα πάντα

ρυθμίζονται αυτόματα, δηλαδή δεν χρειάζονται χειροκίνητες ρυθμίσεις και προσφέρει εφαρμογή spring που βασίζεται σε σχολιασμούς (annotation).

Η αρχιτεκτονική της Spring Boot έχει τέσσερα επίπεδα.

- Presentation Layer
- Business Layer
- Persistence Layer
- Database Layer



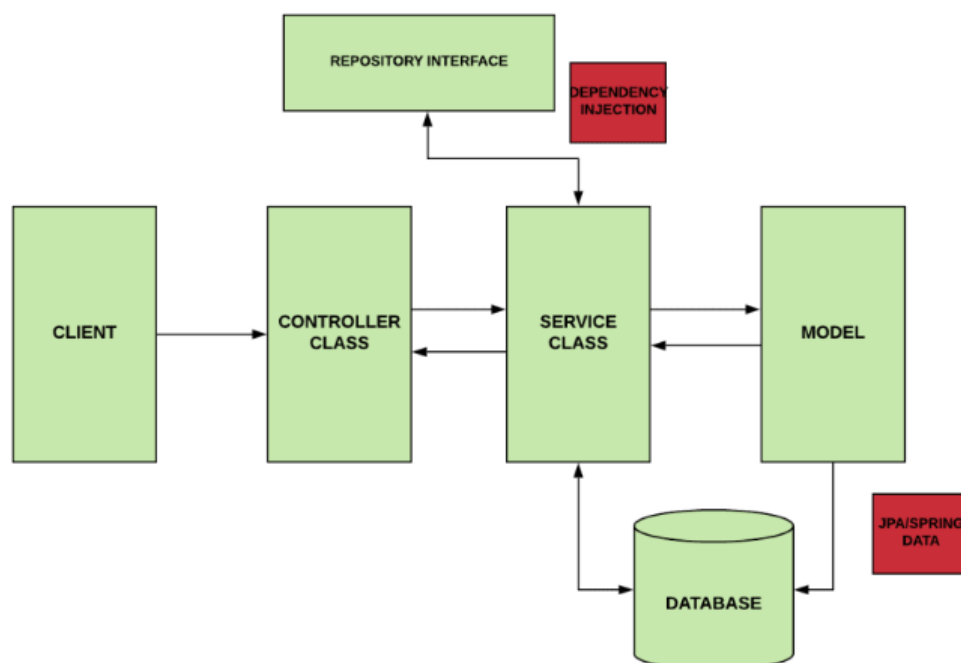
Σχήμα 5-23 αρχιτεκτονική Spring Boot

Το Presentation layer βρίσκεται στην κορυφή της αρχιτεκτονικής βαθμίδα είναι υπεύθυνη για την εκτέλεση ελέγχου ταυτότητας, την μετατροπή δεδομένων σε JSON αντικείμενο (και αντίστροφα), τον χειρισμό των HTTP αιτημάτων και την μεταφορά ελέγχου ταυτότητας στο επιχειρηματικό επίπεδο. Το επίπεδο παρουσίασης είναι το ισοδύναμο της κλάσης Controller. Η κλάση Controller χειρίζεται όλα τα εισερχόμενα αιτήματα REST API (GET, POST, PUT, DELETE, PATCH) από τον πελάτη.

Το Business layer είναι υπεύθυνο για την εκτέλεση της επικύρωσης, την εκτέλεση της εξουσιοδότησης και τον χειρισμό της επιχειρηματικής λογικής. Αυτό το επίπεδο είναι ισοδύναμο με την κλάση Service. Η επιχειρηματική λογική στη μηχανική λογισμικού είναι το σημείο που αποφασίζει το λογισμικό τι πρέπει να κάνει. Ένα παράδειγμα αυτού είναι η επικύρωση. Το Business layer επικοινωνεί τόσο με το Presentation layer όσο και με το Persistence layer.

Το Persistence layer είναι υπεύθυνο για την λογική αποθήκευσης αλλά και την ανάκτηση και μετάφραση δεδομένων σε σειρές βάσης δεδομένων. Αυτό το επίπεδο είναι το ισοδύναμο της διεπαφής Repository. Μέσα σε αυτή την διεπαφή γράφονται τα ερωτήματα βάσης δεδομένων. Το επίπεδο Persistence είναι το μόνο επίπεδο που επικοινωνεί με το Business layer και το Database layer.

Το Database layer είναι υπεύθυνο για την εκτέλεση λειτουργιών δεδομένων, κυρίως εντολές CRUD (create, read, update, delete). Αυτό το επίπεδο είναι απλώς η πραγματική βάση δεδομένων που χρησιμοποιείτε για την δημιουργία της εφαρμογής.



Σχήμα 5-24 Ροή εργασίας Spring Boot

Ένα παράδειγμα ροής εργασίας είναι:

- Βήμα 1. Ο πελάτης κάνει ένα αίτημα HTTP.
- Βήμα 2. Η κλάση Controller λαμβάνει το αίτημα HTTP.
- Βήμα 3. Ο ελεγκτής κατανοεί τι είδους αίτηση θα επεξεργαστεί και, στη συνέχεια, το αντιμετωπίζει.
- Βήμα 4. Εάν χρειάζεται, καλεί την κλάση υπηρεσιών.

Βήμα 5. Η κλάση υπηρεσιών θα χειριστεί την επιχειρηματική λογική. Αυτό το κάνει στα δεδομένα από τη βάση δεδομένων.

Βήμα 6. Εάν όλα πάνε καλά, επιστρέφουμε μια αποτέλεσμα.

5.2.3 API

Τα API παρέχουν έναν δομημένο τρόπο πρόσβασης μιας εφαρμογής στις δυνατότητες μιας άλλης εφαρμογής. Συνήθως, αυτή η επικοινωνία πραγματοποιείται μέσω του Διαδικτύου μέσω ενός διακομιστή API. Μια εφαρμογή πελάτη (όπως μια εφαρμογή για κινητά) στέλνει ένα αίτημα στον διακομιστή και μετά την επεξεργασία του αιτήματος, ο διακομιστής επιστρέφει μια απάντηση στον πελάτη.

Ένα αίτημα περιλαμβάνει τη διεύθυνση URL του τερματικού σημείου API και μια μέθοδο αιτήματος HTTP. Η μέθοδος υποδεικνύει την ενέργεια που θέλετε να εκτελέσει το API. Τα τέσσερα αιτήματα που χρησιμοποιήθηκαν για την εργασία είναι τα εξής:

- GET .../group-countries/indicators/selectPage
- GET .../group-countries/categories/{filter}
- POST .../group-countries/metrics/getMetrics
- GET .../group-countries/process/processValue

Οι λειτουργίες των αιτημάτων θα αναλυθούν στην επόμενη ενότητα μαζί με όλη την διαδικασία που ακολουθεί το σύστημα για την επιστροφή αποτελεσμάτων.

5.2.4 Λειτουργίες

Στην συγκεκριμένη εργασία υπάρχουν τέσσερα είδη λειτουργιών. Στην πρώτη λειτουργία ο πελάτης δημιουργεί ένα αίτημα που ζητάει από το σύστημα να επιστρέψει τις κατηγορίες δεικτών που ανήκουν στην κατηγορία κατηγοριών που υπάρχουν για τα φίλτρα. Η διαδικασία που ακολουθείται είναι ότι το αίτημα GET φτάνει στον controller και από εκεί μέσω του repository πραγματοποιείται η ερώτηση στην βάση. Η ερώτηση στην βάση γίνεται στον πίνακα category που δημιουργήθηκε στην ενότητα **4.1.3**. Για να καταλάβει η εφαρμογή μας σε ποια

οντότητα αναφέρεται έχουμε δημιουργήσει ένα αντικείμενο με το annotation @Entity.

Στην δεύτερη λειτουργία ο πελάτης δημιουργεί ένα αίτημα GET που ζητάει από το σύστημα να επιστρέψει όλους του δείκτες που υπάρχουν στη βάση μας. Η διαδικασία που ακολουθεί η λειτουργία είναι ακριβώς ίδια με αυτή που ακολουθήθηκε στην πρώτη περίπτωση.

Στην τρίτη λειτουργία ο πελάτης δημιουργεί ένα αίτημα POST που ζητάει από το σύστημα να επιστρέψει τις μετρήσεις για τους δείκτες που ζήτησε. Αυτή η διαδικασία είναι παρόμοια με τις δύο προηγούμενες. Η διαφορά είναι ότι αφού φτάσει το αίτημα στον controller θα συνεχίσει το Business layer (service). Ο λόγος που γίνεται αυτό είναι γιατί οι μετρήσεις πρέπει να επεξεργαστούν πριν επιστραφούν από το σύστημα.

Τέλος στην τέταρτη και τελευταία λειτουργία ο χρήστη δημιουργεί ένα αίτημα GET που ζητάει από το σύστημα να τον ενημερώσει σε ποιο σημείο βρίσκεται η τροποποίηση των δεδομένων που αναφέρθηκε στην τρίτη λειτουργία.

Κεφάλαιο 6: Website

6.1 Κατασκευή ιστοσελίδας

6.2 Οδηγός κατασκευής εφαρμογής σε React

6.3 Περιγραφή ιστοσελίδας

Ένας ιστότοπος (Website) είναι μια συλλογή ιστοσελίδων σχετικού περιεχομένου που προσδιορίζεται από ένα κοινό όνομα τομέα και δημοσιεύεται σε τουλάχιστον έναν διακομιστή ιστού.

Όλοι οι ιστότοποι που είναι προσβάσιμοι στο κοινό αποτελούν συλλογικά τον Παγκόσμιο Ιστό. Υπάρχουν επίσης ιδιωτικοί ιστότοποι στους οποίους είναι δυνατή η πρόσβαση μόνο σε ιδιωτικό δίκτυο, όπως ο εσωτερικός ιστότοπος μιας εταιρείας για τους υπαλλήλους της.

Οι ιστότοποι είναι συνήθως αφιερωμένοι σε ένα συγκεκριμένο θέμα ή σκοπό, όπως ειδήσεις, εκπαίδευση, εμπόριο, ψυχαγωγία ή κοινωνική δικτύωση. Η υπερσύνδεση μεταξύ ιστοσελίδων καθοδηγεί την πλοήγηση στον ιστότοπο, η οποία συχνά ξεκινά με μια αρχική σελίδα.

Οι χρήστες μπορούν να έχουν πρόσβαση σε ιστότοπους σε μια σειρά συσκευών, όπως επιτραπέζιους υπολογιστές, φορητούς υπολογιστές, tablet και smartphone. Η εφαρμογή που χρησιμοποιείται σε αυτές τις συσκευές ονομάζεται πρόγραμμα περιήγησης ιστού.

6.1 Κατασκευή ιστοσελίδας

Κάθε σελίδα στον ιστό δημιουργείται χρησιμοποιώντας μια σειρά ξεχωριστών οδηγιών, η μία μετά την άλλη. Το πρόγραμμα περιήγησής είναι ένας σημαντικός παράγοντας στη μετάφραση κώδικα σε κάτι που απεικονίζεται ακόμη και για την αλληλεπίδραση. Όταν ανοίγει μια ιστοσελίδα, το πρόγραμμα περιήγησής ανακτά την HTML και άλλες γλώσσες προγραμματισμού που εμπλέκονται και τις ερμηνεύει. Οι γλώσσες και τα εργαλεία που εμπλέκονται για την κατασκευή μιας σελίδας είναι οι HTML, CSS και η JavaScript.

Η HTML βρίσκεται στον πυρήνα κάθε ιστοσελίδας, ανεξάρτητα από την πολυπλοκότητα ενός ιστότοπου ή τον αριθμό των εμπλεκόμενων τεχνολογιών. HTML σημαίνει γλώσσα σήμανσης υπερκειμένου. Γλώσσα σήμανσης σημαίνει ότι, αντί να χρησιμοποιεί μια γλώσσα προγραμματισμού για την εκτέλεση συναρτήσεων, η HTML χρησιμοποιεί ετικέτες για να προσδιορίσει διαφορετικούς τύπους περιεχομένου και τους σκοπούς που εξυπηρετεί το καθένα στην ιστοσελίδα.

CSS σημαίνει Cascading Style Sheets. Αυτή η γλώσσα προγραμματισμού υπαγορεύει πώς πρέπει να εμφανίζονται πραγματικά τα στοιχεία HTML ενός ιστότοπου στο frontend της σελίδας. Η HTML παρέχει τα ακατέργαστα εργαλεία που απαιτούνται για τη δομή του περιεχομένου σε έναν ιστότοπο. Το CSS, από την άλλη πλευρά, βοηθά στο στυλ αυτού του περιεχομένου, ώστε να φαίνεται στον χρήστη με τον τρόπο που προοριζόταν να φαίνεται. Αυτές οι γλώσσες διατηρούνται ξεχωριστές για να διασφαλιστεί ότι οι ιστότοποι έχουν κατασκευαστεί σωστά πριν επαναδιαμορφωθούν.

Επίσης JavaScript είναι μια πιο περίπλοκη γλώσσα από την HTML ή την CSS. Σήμερα, η JavaScript υποστηρίζεται από όλα τα σύγχρονα προγράμματα περιήγησης ιστού και χρησιμοποιείται σχεδόν σε κάθε τοποθεσία στον Ιστό για πιο ισχυρή και πολύπλοκη λειτουργικότητα. Η JavaScript είναι μια γλώσσα προγραμματισμού που βασίζεται στη λογική και μπορεί να χρησιμοποιηθεί για να τροποποιήσει το περιεχόμενο του ιστότοπου και να τον κάνει να συμπεριφέρεται με διαφορετικούς τρόπους ως απάντηση στις ενέργειες ενός χρήστη. Οι συνήθεις χρήσεις της

JavaScript περιλαμβάνουν πλαίσια επιβεβαίωσης, παροτρύνσεις για δράση και προσθήκη νέων ταυτοτήτων σε υπάρχουσες πληροφορίες.

Πέρα από τα προηγούμενα τρία εργαλεία που χρησιμοποιήθηκαν για την κατασκευή της ιστοσελίδας χρησιμοποιήθηκε και η React. Η React (γνωστή και ως React.js ή ReactJS) είναι μια δωρεάν και ανοιχτού κώδικα βιβλιοθήκη JavaScript front-end για τη δημιουργία διεπαφών χρήστη που βασίζονται σε στοιχεία διεπαφής χρήστη. Συντηρείται από τη Meta (πρώην Facebook) και μια κοινότητα μεμονωμένων προγραμματιστών και εταιρειών. Η React μπορεί να χρησιμοποιηθεί ως βάση για την ανάπτυξη εφαρμογών μιας σελίδας, κινητών ή απόδοσης από διακομιστή με πλαίσια όπως το Next.js.

6.2 Οδηγός κατασκευής εφαρμογής σε React

Η React δίνει τον τρόπο με τον οποίο προτείνει αυτή να γίνει η κατασκευή μια ιστοσελίδας που την χρησιμοποιεί. Αυτός ο τρόπος περιγράφεται με συγκεκριμένα βήματα. Τα δύο πράγματα που χρειάζεται να έχουμε για να ξεκινήσει η κατασκευή της ιστοσελίδας είναι ένα παράδειγμα σχεδίασης (mock) και την μορφή των δεδομένων που θα ληφθούν από το API.

Το πρώτο βήμα που πρέπει να γίνει είναι το σπάσιμο της διεπαφής χρήστη σε μια ιεραρχία στοιχείων. Αυτό απαιτεί να γίνει η σχεδίαση πλαισίων γύρο από κάθε στοιχείο και υποστοιχείο (component και subcomponent) και να δοθούν όλα τα ονόματά τους. Τα ονόματα θα καταλήξουν να γίνουν τα ονόματα των στοιχείων (component) στην React. Πρέπει να δοθεί μεγάλη προσοχή στον διαχωρισμό των στοιχείων για την απλοποίηση την διαδικασίας και την αποφυγή δημιουργίας εμποδίων πολυπλοκότητας. Στο τέλος αυτού του βήματος θα έχουμε όλα τα στοιχεία της σχεδίασης μέσα σε μια ιεραρχία.

Επόμενο βήμα είναι η δημιουργίας στατικής έκδοσης στη React. Αυτό το βήμα προβλέπει την δημιουργία μιας έκδοσης της ιστοσελίδας που θα λαμβάνει το μοντέλο δεδομένων και θα αποδίδει την διεπαφή χρήστη, αλλά δεν θα έχει διαδραστικότητα. Είναι καλός ο διαχωρισμός των δύο διαδικασιών με την

δημιουργία μιας στατικής μορφής, γιατί η διαδραστικότητα απαιτεί πιο πολλή σκέψη.

Στην συνέχεια θα πρέπει να προσδιοριστεί η ελάχιστη αλλά πλήρη αναπαράσταση της κατάστασης διεπαφής χρήστη. Για να γίνει η διεπαφή διαδραστική πρέπει να μπορούν να πραγματοποιηθούν αλλαγές στο υποκείμενο μοντέλο δεδομένων. Η `react` αυτό το πετυχαίνει με το `state`. Για να δημιουργηθεί σωστά η εφαρμογή πρέπει να σκεφτούμε το ελάχιστο σύνολο που χρειάζεται η εφαρμογή.

Ακόμα ένα από τα πιο σημαντικά και δύσκολα βήματα για την κατασκευή μιας σελίδας είναι ο προσδιορισμός δραστηριοποίησης του κάθε `state`. Αφού έχουμε βρει το ελάχιστο σύνολο `state` της εφαρμογής πρέπει να εντοπίσουμε ποιο στοιχείο μεταλλάσσει ή κατέχει καθένα από αυτά.

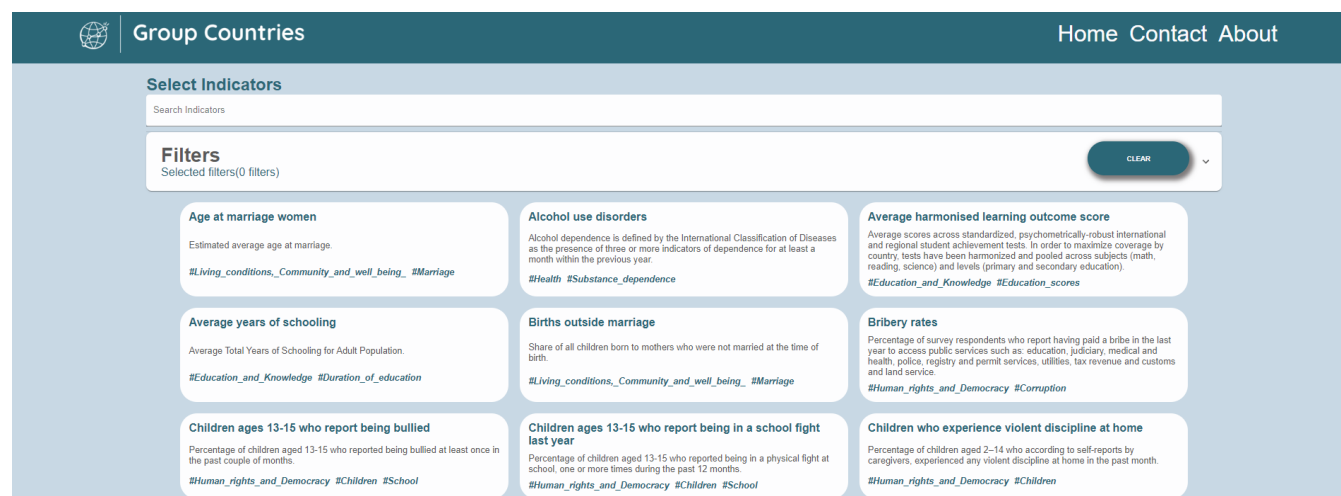
Τέλος, το βήμα που πρέπει να εκτελεστεί για την ολοκλήρωση την διαδικασίας κατασκευής ιστοσελίδας είναι η προσθήκη αντίστροφης ροής δεδομένων. Μέχρι στιγμής, έχει δημιουργηθεί μια εφαρμογή που αποδίδεται σωστά ως συνάρτηση των στηρίξεων και της κατάστασης που ρέουν προς τα κάτω στην ιεραρχία. Τώρα είναι ώρα να υποστηριχθεί η ροή δεδομένων αντίστροφα.

6.3 Περιγραφή ιστοσελίδας

Η ιστοσελίδα που έχει κατασκευαστεί για τις ανάγκες την εργασίας έχει δύο βασικές σελίδες (pages) με διαφορετικές λειτουργίες. Η διαδικασία που ακολουθήθηκε για την κατασκευή τους περιγράφεται στο προηγούμενο κεφάλαιο **6.2**.

Η πρώτη σελίδα αφορά την επιλογή δεικτών από τον χρήστη (Select Indicators Page). Μέσα σε αυτήν πέρα από την απεικόνιση των δεικτών και τις

λειτουργίες επιλογής δίνονται στο δείκτη διάφορα βοηθήματα για την πιο αποδοτική εξυπηρέτηση του. Η αρχική μορφή την σελίδας απεικονίζεται παρακάτω.



Group Countries Home Contact About

Select Indicators

Search Indicators

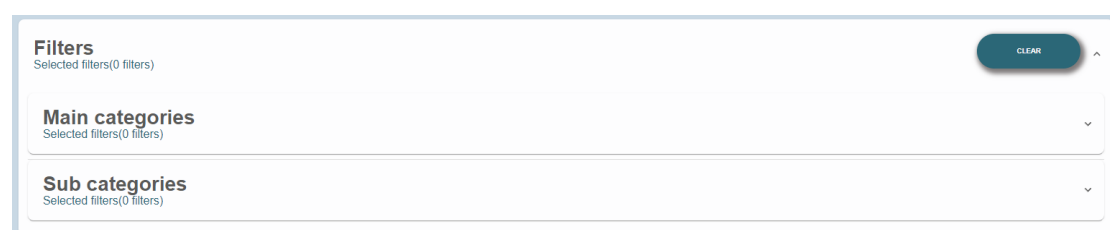
Filters
Selected filters(0 filters) CLEAR

Age at marriage women Estimated average age at marriage. #Living_conditions_Community_and_well_being_ #Marriage	Alcohol use disorders Alcohol dependence is defined by the International Classification of Diseases as the presence of three or more indicators of dependence for at least a month within the previous year. #Health #Substance_dependence	Average harmonised learning outcome score Average scores across standardized, psychometrically-robust international and regional student achievement tests. In order to maximize coverage by country, tests have been harmonized and pooled across subjects (math, reading, science) and levels (primary and secondary education). #Education_and_Knowledge #Education_scores
Average years of schooling Average Total Years of Schooling for Adult Population. #Education_and_Knowledge #Duration_of_education	Births outside marriage Share of all children born to mothers who were not married at the time of birth. #Living_conditions_Community_and_well_being_ #Marriage	Bribery rates Percentage of survey respondents who report having paid a bribe in the last year to access public services such as: education, judiciary, medical and health, police, registry and permit services, utilities, tax revenue and customs and land service. #Human_rights_and_Democracy #Corruption
Children ages 13-15 who report being bullied Percentage of children aged 13-15 who reported being bullied at least once in the past couple of months. #Human_rights_and_Democracy #Children #School	Children ages 13-15 who report being in a school fight last year Percentage of children aged 13-15 who reported being in a physical fight at school, one or more times during the past 12 months. #Human_rights_and_Democracy #Children #School	Children who experience violent discipline at home Percentage of children aged 2-14 who according to self-reports by caregivers, experienced any violent discipline at home in the past month. #Human_rights_and_Democracy #Children

Σχήμα 6-1 Αρχική μορφή σελίδας επιλογής δεικτών

Όπως φαίνεται παραπάνω υπάρχει μια μπάρα αναζήτησης όπου ο χρήστης μπορεί να αναζητήσει έναν δείκτη σύμφωνα με την ονομασία του ή ακόμα και από τα ονόματα των κατηγοριών στα οποία ανήκει.

Στην συνέχεια υπάρχει ένα πλαίσιο που ονομάζεται Filters. Αυτό το πλαίσιο ασχολείται με την επιλογή κατηγοριών δεικτών για την πιο εύκολη αναζήτηση δεικτών. Για να ξεκινήσει η επιλογή δεικτών ο χρήστης πρέπει να πατήσει πάνω στο πλαίσιο για την επέκταση του και την απεικόνιση των επιλογών που του δίνονται.



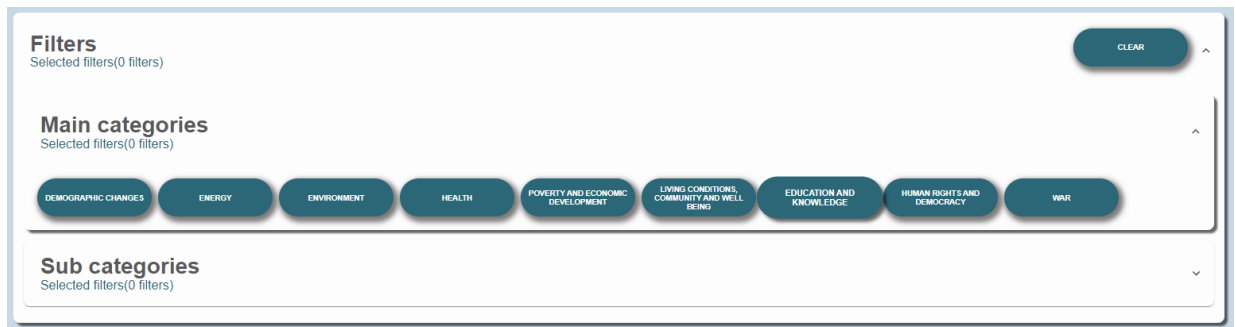
Filters
Selected filters(0 filters) CLEAR

Main categories
Selected filters(0 filters)

Sub categories
Selected filters(0 filters)

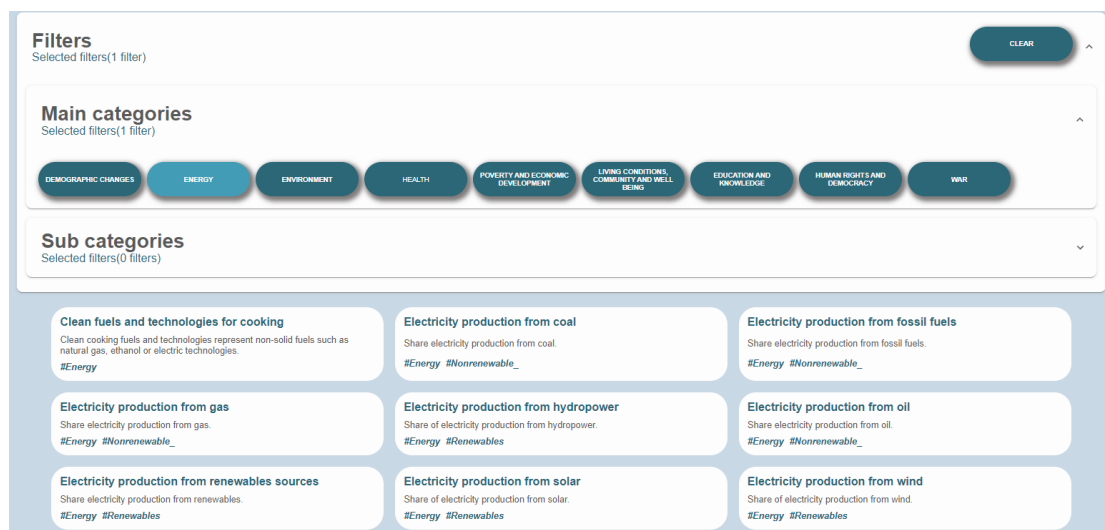
Σχήμα 6-2 Απεικόνιση πλαισίου κατηγοριών δεικτών

Όπως φαίνεται στην εικόνα παραπάνω απεικονίζονται δύο επιπλέον πλαίσια. Τα πλαίσια αυτά αφορούν τις δύο κατηγορίες που διαχωρίσαμε τις κατηγορίες των δεικτών. Για την απεικόνιση των επιλογών πρέπει για ακόμα μια φορά να πατηθεί το πλαίσιο της κατηγορίας που επιθυμούμε.



Σχήμα 6-3 Απεικόνιση κατηγοριών δεικτών

Όπως φαίνεται έχουν απεικονιστεί όλες οι κατηγορίες που ανήκουν στην κατηγορία main categories. Πλέον ο χρήστης μπορεί να επιλέξει μια ή περισσότερες επιλογές από όποια κατηγορία επιθυμεί. Κατά την επιλογή κατηγοριών φιλτράρονται οι δείκτες και απεικονίζονται μόνο αυτοί που ανήκουν στις αντίστοιχες κατηγορίες.

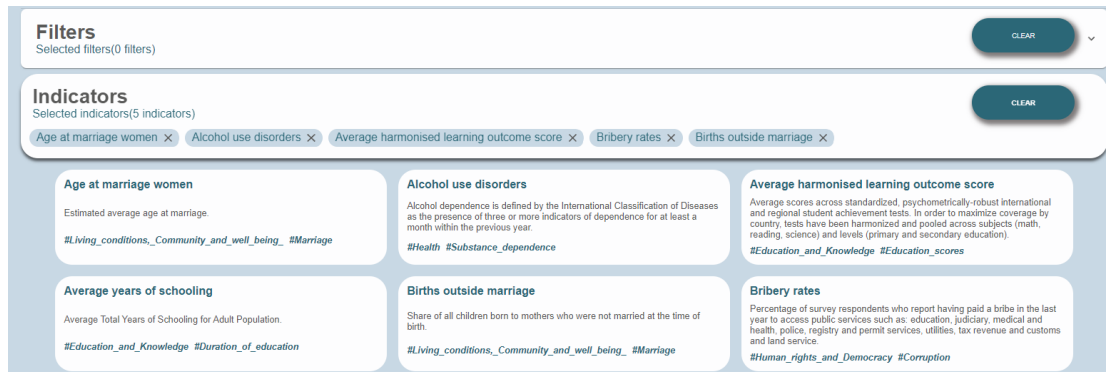


Σχήμα 6-4 Επιλογή κατηγορίας και απεικόνιση κατάλληλων δεικτών

Μετά από αυτές τις κινήσεις παρατηρούμε ότι έχουν αλλάξει χρώμα οι επιλεγμένες κατηγορίες και αναγράφεται κάτω από κάθε κεφαλίδα ο ακριβής αριθμός επιλεγμένων δεικτών. Με την επιλογή του κουμπιού clear που βρίσκεται πάνω αριστερά στο πλαίσιο, δεν υπάρχουν πλέον επιλεγμένοι δείκτες και ξανά εμφανίζονται όλες οι επιλογές δεικτών.

Στην συνέχεια, κάτω από το πλαίσιο επιλογής δεικτών υπάρχουν οι δείκτες προς επιλογή. Κάθε δείκτης έχει μέσα στο πλαίσιο που του αντιστοιχεί τον τίτλο

του, την περιγραφή του και μια λίστα από τις κατηγορίες στις οποίες ανήκει. Κατά την επιλογή ενός ή περισσότερων δεικτών εμφανίζεται ένα καινούργιο πλαίσιο πάνω από αυτούς.



Σχήμα 6-5 Παράδειγμα επιλεγμένων δεικτών

Μέσα σε αυτό το πλαίσιο απεικονίζονται οι τίτλοι από τους δείκτες που επιλέχθηκαν. Δίπλα από κάθε τίτλο υπάρχει η επιλογή διαγραφής του από τους επιλεγμένους δείκτες. Επιπλέον υπάρχει κάτω από τον τίτλο του πλαισίου ένας μετρητής που απεικονίζει τον αριθμό των δεικτών που έχουν επιλεχθεί. Σε περίπτωση που ο χρήστης θέλει να διαγράψει όλες τις επιλογές του μαζί μπορεί να επιλέξει την επιλογή clear που υπάρχει στο πάνω δεξιό μέρος του πλαισίου όπως και στην περίπτωση με το Filters.

Όπως παρατηρήθηκε στην αρχική μορφή της σελίδας το πλαίσιο επιλεγμένων δεικτών δεν υπήρχε. Αυτό συμβαίνει επειδή δεν υπάρχουν επιλεγμένοι δείκτες. Πέρα από την εμφάνιση αυτού του πλαισίου, όταν υπάρχει έστω και ένας επιλεγμένος δείκτης εμφανίζεται στο κάτω δεξιό μέρος της οθόνης μια επιλογή με το όνομα submit. Η επιλογή αυτή λειτουργεί σαν παράγοντας μετάβασης ανάμεσα στην συγκεκριμένη σελίδα επιλογής δεικτών και την επόμενη που αφορά την απεικόνιση των αποτελεσμάτων τους.

Group Countries Home Contact About

Select Indicators

Search Indicators

Filters
Selected filters(0 filters)

Indicators
Selected indicators(5 indicators)

Age at marriage women X Alcohol use disorders X Average harmonised learning outcome score X Bribery rates X Births outside marriage X

Age at marriage women
Estimated average age at marriage
#Living_conditions_Community_and_well_being_Marriage

Average years of schooling
Average Total Years of Schooling for Adult Population
#Education_and_Knowledge #Duration_of_education

Children ages 13-15 who report being bullied
Percentage of children aged 13-15 who reported being bullied at least once in the past couple of months
#Human_rights_and_Democracy #Children #School

Civil liberties
The variable denotes the best estimate of the extent to which extent to which citizens enjoy physical integrity rights, freedoms of religion, of movement, and of assembly and association
#Human_rights_and_Democracy #Civil Liberties

Alcohol use disorders
Alcohol dependence is defined by the International Classification of Diseases as the presence of three or more indicators of dependence for at least a month within the previous year.
#Health #Substance_dependence

Births outside marriage
Share of all children born to mothers who were not married at the time of birth
#Living_conditions_Community_and_well_being_Marriage

Children ages 13-15 who report being in a school fight last year
Percentage of children aged 13-15 who reported being in a physical fight at school, one or more times during the past 12 months
#Human_rights_and_Democracy #Children #School

Civil liberties lower bound
The variable denotes the lower bound estimate of the extent to which extent to which citizens enjoy physical integrity rights, freedoms of religion, of movement, and of assembly and association
#Human_rights_and_Democracy #Civil Liberties

Average harmonised learning outcome score
Average scores across standardized, psychometrically-robust international and regional student achievement tests. In order to maximize coverage by country, tests have been harmonized and pooled across subjects (math, reading, science) and levels (primary and secondary education)
#Education_and_Knowledge #Education_scores

Bribery rates
Percentage of survey respondents who report having paid a bribe in the last year to access public services such as: education, judiciary, medical and health, police, registry and permit services, utilities, tax revenue and customs and land services
#Human_rights_and_Democracy #Corruption

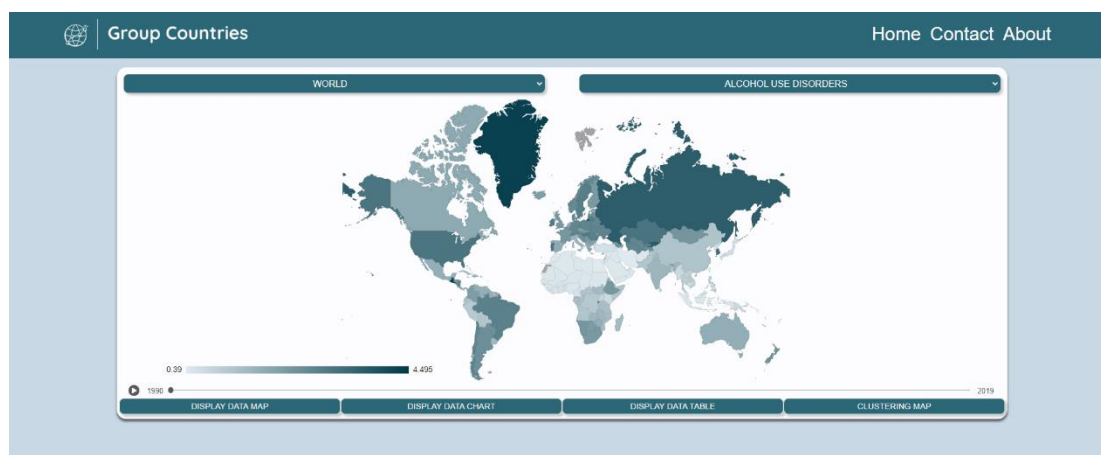
Children who experience violent discipline at home
Percentage of children aged 2-14 who according to self-reports by caregivers, experienced any violent discipline at home in the past month
#Human_rights_and_Democracy #Children

Civil liberties upper bound
The variable denotes the upper bound estimate of the extent to which extent to which citizens enjoy physical integrity rights, freedoms of religion, of movement, and of assembly and association
#Human_rights_and_Democracy #Civil Liberties

SUBMIT

Σχήμα 6-6 Εμφάνιση επιλογής Submit

Η σελίδα στην οποία απεικονίζεται μετά την ενεργοποίηση της επιλογής submit αφορά την απεικόνιση των δεδομένων (Display Data Page). Σε αυτή την σελίδα δίνονται τα δεδομένα με διάφορες μορφές απεικόνισης. Η μορφή της σελίδας φαίνεται παρακάτω.

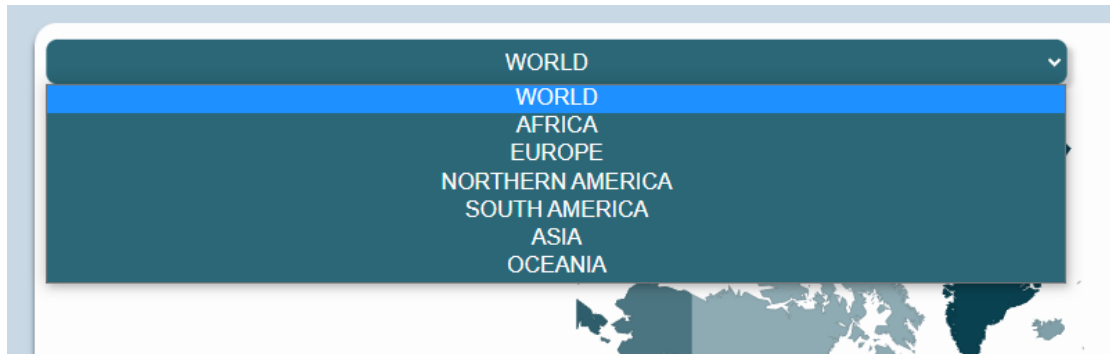


Σχήμα 6-7 Αρχική μορφή σελίδας απεικόνισης δεδομένων

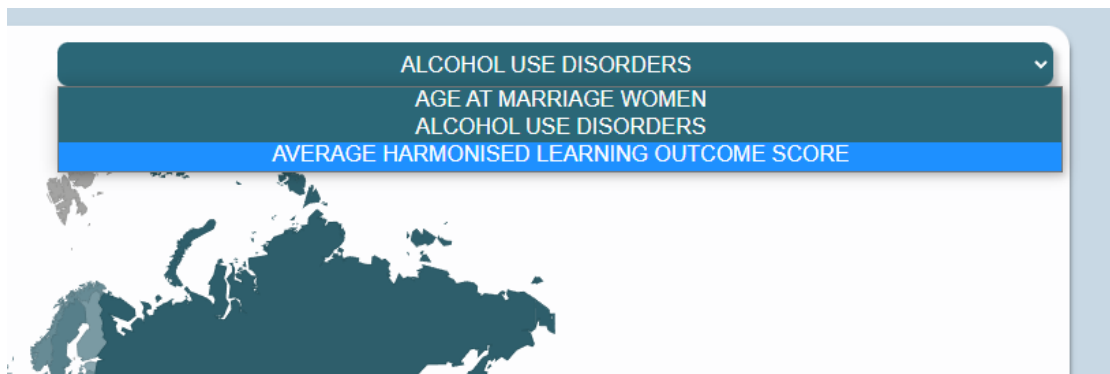
Στο κάτω μέρος της σελίδας υπάρχουν τέσσερις επιλογές. Οι επιλογές αυτές αντιστοιχούν σε τέσσερις διαφορετικούς τρόπους απεικόνισης δεδομένων. Οι τρεις από αυτούς του τρόπους αφορούν την απλή απεικόνιση των δεδομένων χωρίς κάποια επεξεργασία ενώ ο τελευταίος αφορά την ομαδοποίηση.

Πρώτη περίπτωση είναι η απεικόνιση δεδομένων σε μορφή χάρτη, όπως φαίνεται στην εικόνα πάνω. Αυτή η επιλογή ονομάζεται Display Data Map. Σε κάθε

τρόπο απεικόνισης δεδομένων υπάρχουν διαφορετικές παράμετροι που μπορεί να επιλέξει ο χρήστης για τα αποτελέσματα που επιθυμεί. Στην συγκεκριμένη περίπτωση υπάρχουν δύο κατηγορίες λιστών. Στην πρώτη λίστα υπάρχουν επιλογές περιοχών που μπορεί να επιλέξει ο χρήστης για την απεικόνιση δεδομένων μόνο για την συγκεκριμένη. Η δεύτερη λίστα περιέχει όλους τους δείκτες που επιλέχτηκαν για να επιλεγθεί ποιόν θέλουμε να απεικονίσουμε.

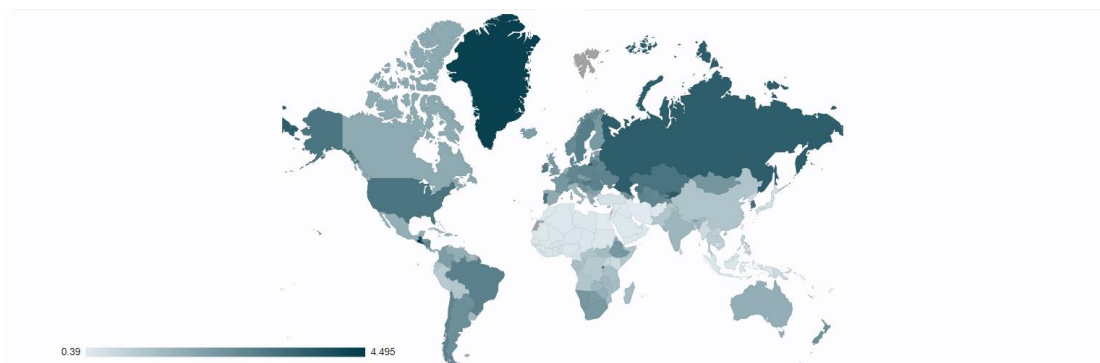


Σχήμα 6-8 Επιλογές περιοχών



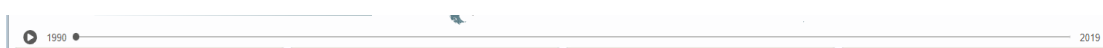
Σχήμα 6-9 Παράδειγμα επιλογών δεικτών

Στο κέντρο απεικονίζεται ο χάρτης όπου χρωματίζεται σύμφωνα με την τιμή που αντιστοιχεί σε κάθε χώρα. Σε περίπτωση που ο χρήστης μετακινήσει το βελάκι πάνω σε κάποια από τις χρωματισμένες χώρες θα εμφανιστεί η τιμή που αντιστοιχεί στην συγκεκριμένη χώρα. Επιπλέον κάτω αριστερά εμφανίζεται μια μπάρα χρωμάτων που δείχνουν τα χρώματα ανάλογα με την τιμή του δείκτη.



Σχήμα 6-10 Παράδειγμα απεικόνισης δεδομένων σε μορφή χάρτη

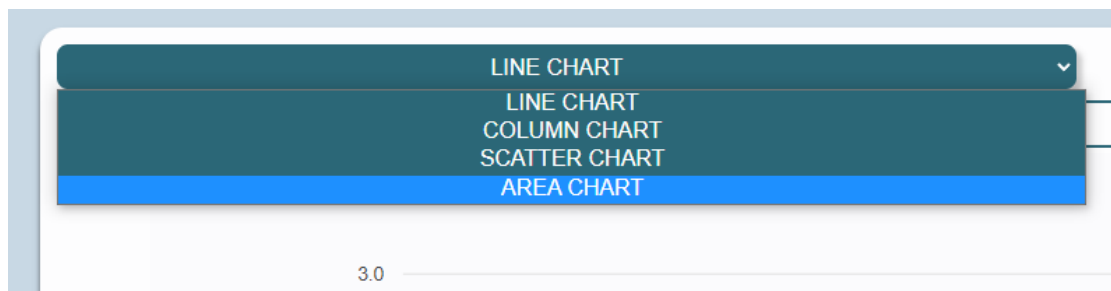
Ακόμα όσο αφορά την απεικόνιση δεδομένων δεικτών με χάρτη, υπάρχει μια μπάρα πάνω από τις τέσσερις βασικές επιλογές. Σε αυτή την μπάρα υπάρχουν δεξιά και αριστερά δύο τιμές. Αυτές οι δύο τιμές αντιστοιχούν στην μικρότερη και μεγαλύτερη χρόνια αντίστοιχα που έχουμε έστω και ένα δεδομένο για τον επιλεγμένο δείκτη. Αριστερά επίσης υπάρχει μια εικόνα παίξε (play) η οποία με την ενεργοποίησή της ξεκινάει την προσπέλαση των χρονιών από αυτήν που απεικονίζεται αριστερά μέχρι και εκείνη που βρίσκεται δεξιά. Η εναλλαγή των χρόνων γίνεται μετά από προκαθορισμένο χρονικό διάστημα που έχει καθορίσει η εφαρμογή. Στην χρονική περίοδο που αναγράφεται στο αριστερό μέρος αντιστοιχούν και τα δεδομένα που απεικονίζονται στο χάρτη. Τέλος ο χρήστης μπορεί να επιλέξει ανεξάρτητα την επιλογή παίξε μία τυχαία χρόνια, μέσα στο εύρος, που επιθυμεί για την απεικόνιση δεδομένων.



Σχήμα 6-11 Μπάρα απεικόνισης χρονιών

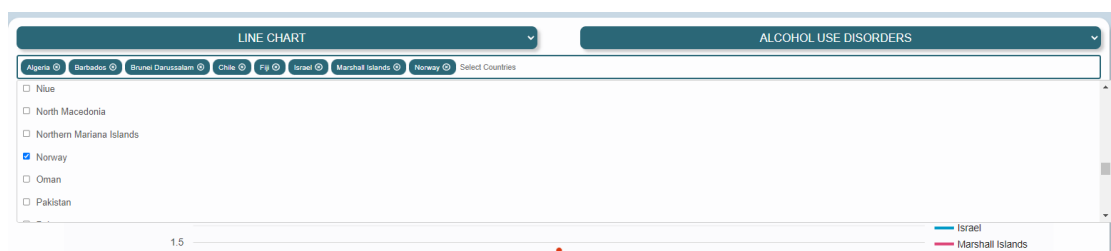
Επόμενη περίπτωση απεικόνισης δεδομένων είναι τα διαγράμματα. Την επιλογή αυτή η εφαρμογή την ονομάζει Display Data Chart. Σε αυτή την επιλογή τα δεδομένα απεικονίζονται στον χρήστη μέσω διαγραμμάτων. Αυτή την φορά οι παράμετροι που υπάρχουν για να επιλέξει ο χρήστης είναι διαφορετικές. Πέρα από την επιλογή δεικτών που βρίσκεται στο πάνω αριστερά μέρος με την μορφή λίστας, όπως και στον χάρτη, υπάρχουν άλλες δύο επιλογές. Σε αυτή την περίπτωση αντί για την επιλογή περιοχών υπάρχει η επιλογή διαγράμματος. Οι τέσσερις επιλογές

που δίνονται από την εφαρμογή είναι το διάγραμμα γραμμής, διάγραμμα ράβδων, διάγραμμα περιοχής και διάγραμμα διασποράς.



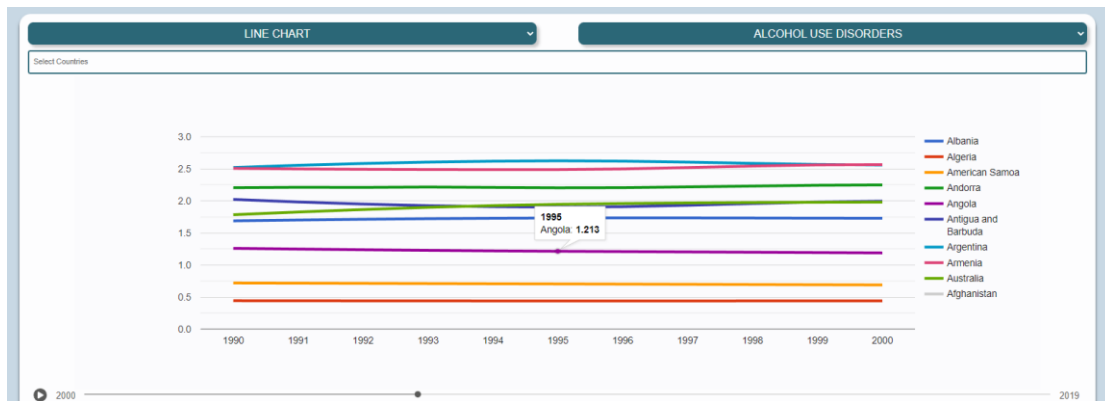
Σχήμα 6-12 Επιλογές τύπων διαγραμμάτων

Συμπληρωματικά υπάρχει μια ακόμη παράμετρο που μπορεί να επιλέξει ο χρήστης. Αυτή η παράμετρο αφορά την επιλογή χωρών προς απεικόνιση. Ο αριθμός των χωρών που μπορεί να επιλέξει ο χρήστης είναι μέχρι δέκα. Αυτό συμβαίνει επειδή αν απεικονιστούν όλες οι υπάρχουσες χώρες πάνω σε ένα διάγραμμα, τα αποτελέσματα δεν θα είναι ευανάγνωστα. Σε περίπτωση που δεν επιλεγεί κάποια χώρα τότε η εφαρμογή προεπιλέγει τις δέκα πρώτες αλφαβητικά.



Σχήμα 6-13 Επιλογή χωρών προς απεικόνιση

Στο κέντρο απεικονίζεται το επιλεγμένο διάγραμμα. Ο χρήστης πηγαίνοντας το βελάκι πάνω στο διάγραμμα, απεικονίζεται η τιμή, χρονιά και χώρα που αφορά την συγκεκριμένη μέτρηση. Επιπλέον δεξιά από την προβολή των διαγραμμάτων υπάρχει μια λίστα η οποία απεικονίζει τις χώρες και τα χρώματα που απεικονίζονται στο διάγραμμα. Τέλος όσο αφορά την επιλογή χρονικής περιόδους είναι ίδια με αυτή της απεικόνισης του χάρτη.



Σχήμα 6-14 Παράδειγμα απεικόνισης δεδομένων σε μορφή διαγράμματος

Στην τρίτη περίπτωση έχουμε την απεικόνιση δεδομένων σε μορφή πίνακα. Σε αυτή την περίπτωση η μόνη παράμετρος που υπάρχουν για να επιλέξει ο χρήστης αφορά την επιλογή δείκτη.

ALCOHOL USE DISORDERS						
Country Name	Starting Year	Current Year	Starting Year Value	Current Year Value	Absolute Change	Relative Change
Afghanistan	1990	2000	0.444	0.44	0.004	-0.82%
Albania	1990	2000	1.688	1.728	0.041	2.42%
Algeria	1990	2000	0.442	0.44	0.002	-0.55%
American Samoa	1990	2000	0.718	0.689	0.03	-4.13%
Andorra	1990	2000	2.205	2.248	0.043	1.96%
Angola	1990	2000	1.259	1.188	0.071	-5.62%
Antigua and Barbuda	1990	2000	2.022	1.994	0.028	-1.38%
Argentina	1990	2000	2.519	2.558	0.039	1.54%
Armenia	1990	2000	2.505	2.564	0.059	2.36%
Australia	1990	2000	1.783	1.98	0.196	11.01%

2000 2019

DISPLAY DATA MAP DISPLAY DATA CHART DISPLAY DATA TABLE CLUSTERING MAP

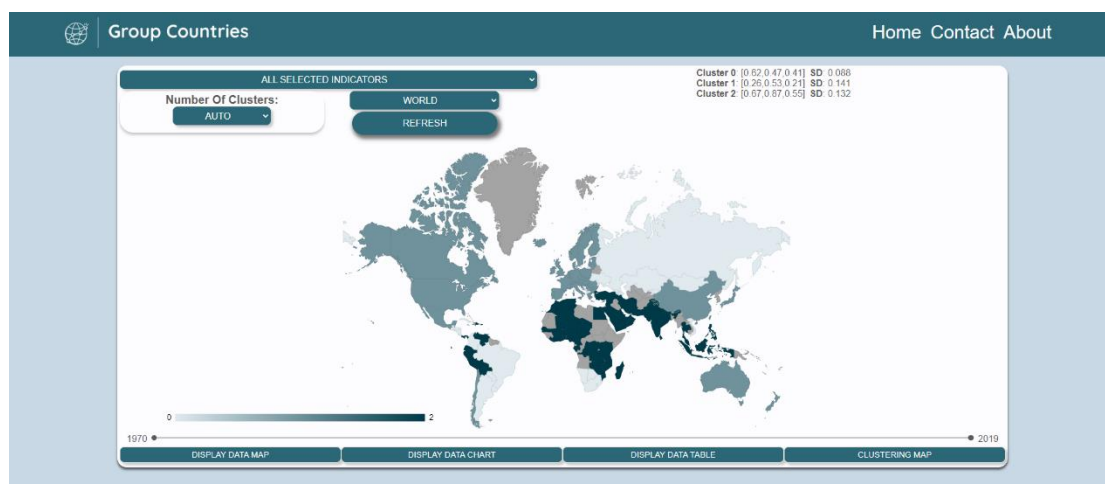
Σχήμα 6-15 Παράδειγμα απεικόνισης δεδομένων σε μορφή πίνακα

Όπως φαίνεται παραπάνω ο πίνακας απεικονίζει επτά κατηγορίες δεδομένων. Η πρώτη κατηγορία ονομάζεται country name και αφορά την ονομασία τη χώρα. Η δεύτερη ονομάζονται starting year και απεικονίζουν την αρχική χρονιά που εμφανίστηκε τιμή για τον δείκτη. Η επόμενη κατηγορία current year αφορά την τρέχουσα χρονιά που απεικονίζεται. Οι επόμενες δύο κατηγορίες starting year value και current year value απεικονίζουν τις αντίστοιχες τιμές που έχει ο δείκτης τις συγκεκριμένες χρονικές περιόδους. Τέλος υπάρχουν οι κατηγορίες absolute change και relative change όπου αφορούν τις μεταβολές που έχει κάνει στην τιμή ο δείκτης από την αρχική χρονιά μέχρι την τρέχουσα.

Ο συγκεκριμένος πίνακας μας δίνει ακόμα την δυνατότητα ταξινόμησης με αύξοντα ή και φθίνοντα τρόπο τα δεδομένα, σύμφωνα με την κατηγορία που θα

επιλεχθεί. Επίσης όσο αφορά την επιλογή χρονικής περιόδους είναι ίδια με αυτές των προηγούμενων δύο περιπτώσεων.

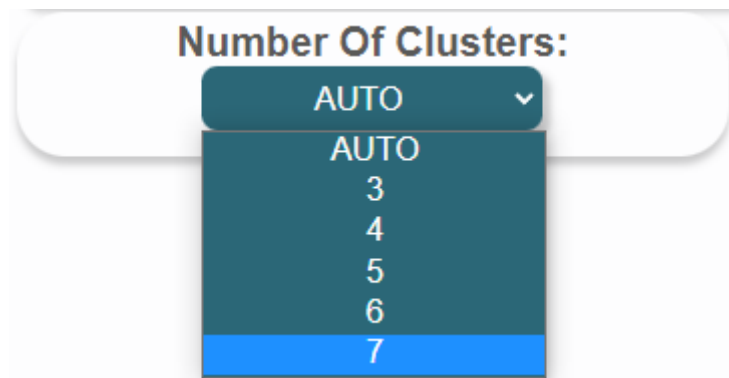
Η τελευταία και σημαντικότερη περίπτωση απεικόνισης δεδομένων διαφέρει από τις προηγούμενες τρεις. Σε αυτή την περίπτωση τα δεδομένα παρόλο που απεικονίζονται ξανά σε χάρτη, έχουν υποστεί κάποιου είδους επεξεργασία. Η επεξεργασία αυτή αφορά την ομαδοποίηση με τον αλγόριθμο k-means όπως αναφέρθηκε στην υποενότητα **2.3**. Η μορφή αυτής της απεικόνισης δεδομένων ονομάζεται από την εφαρμογή clustering map.



Σχήμα 6-16 Παράδειγμα απεικόνισης αποτελεσμάτων ομαδοποίησης σε μορφή χάρτη

Στην συγκεκριμένη περίπτωση, όπως φαίνεται στο πάνω αριστερά μέρος της παραπάνω εικόνας, οι παράμετροι που μπορεί να επιλέξει ο χρήστης διαφέρουν με τις προηγούμενες τρεις περιπτώσεις. Σε αυτή την περίπτωση υπάρχουν οι δύο παράμετροι σε μορφή λίστας όπως υπήρχαν και στο χάρτη. Αυτές οι δύο περιπτώσεις είναι η επιλογή περιοχής προς απεικόνιση και η επιλογή δείκτη. Στην συγκεκριμένη περίπτωση όμως, υπάρχει η δυνατότητα προβολής αποτελεσμάτων ομαδοποίησης για όλους τους επιλεγμένους δείκτες μαζί πέρα από τον κάθε δείκτη ξεχωριστά. Εκτός από αυτές τις δύο παραμέτρους υπάρχει μια ακόμα επιλογή που ονομάζεται number of clusters. Σε αυτή την περίπτωση ο χρήστης μπορεί να επιλέξει τον αριθμό των ομάδων (clusters) που επιθυμεί να ομαδοποιήσει τις χώρες. Ως προεπιλογή, όπως φαίνεται παραπάνω, υπάρχει μια επιλογή auto. Αυτή η επιλογή δίνει στην εφαρμογή την δυνατότητα να διαλέξει μόνη της τον αριθμό των

ομάδων (clusters) που θα χωρίσει τις χώρες. Ο αριθμός αυτός επιλέγεται σύμφωνα με τον αλγόριθμο Elbow που αναλύθηκε στην υποενότητα **2.3.3**.



Σχήμα 6-17 Επιλογές αριθμών ομάδων

Στο πάνω δεξιά μέρος της οθόνης η εφαρμογή προβάλλει πληροφορίες για τα κέντρα των ομάδων. Στις πληροφορίες αυτές υπάρχουν τα κέντρα και η τυπική απόκλιση αυτών. Τα κέντρα εμφανίζονται σε μορφή πίνακα όπου ανάλογα με τις διαστάσεις που έχουν τα δεδομένα μας, δηλαδή τον αριθμό των δεικτών προς ομαδοποίηση, τόσες είναι και οι τιμές των κέντρων που υπάρχουν μέσα σε αυτόν. Η τυπική απόκλιση εμφανίζεται με τον τίτλο SD. Ο υπολογισμός της γίνεται μέσω του τύπου:

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

σ : Population standard deviation

x : Datapoint value

μ : Population mean

N : Population size

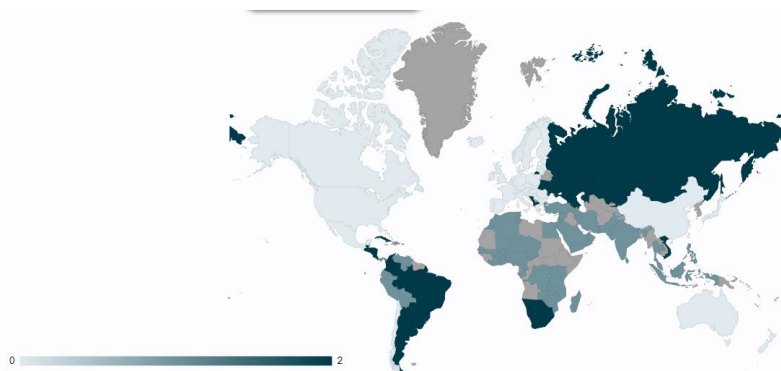
Σχήμα 6-18 Τύπος υπολογισμού τυπικής απόκλισης

Στον παραπάνω τύπο το σ αναφέρεται στην τυπική απόκλιση, το χ στον αριθμό των κέντρων των δεικτών, το μ στην μέση τιμή από τα κέντρα των δεικτών και τέλος το N στον αριθμό των κέντρων.

Cluster 0:	[0.26,0.53,0.21]	SD: 0.141
Cluster 1:	[0.67,0.87,0.55]	SD: 0.132
Cluster 2:	[0.62,0.47,0.41]	SD: 0.088

Σχήμα 6-19 Παράδειγμα αποτελεσμάτων για τρία κέντρα και τρεις επιλεγμένους δείκτες

Η απεικόνιση των αποτελεσμάτων στον χάρτη γίνεται μέσα από τις τιμές που δίνει το label, όπως αναφέρθηκε στο κεφάλαιο 3. Η κάθε χώρα έχει μια τιμή η οποία αναφέρεται στην ομάδα που οποία ανήκει. Οι χώρες με τα ίδια χρώματα ανήκουν στην ίδια ομάδα.



Σχήμα 6-20 Παράδειγμα αποτελεσμάτων ομαδοποίησης σε μορφή χάρτη

Ακόμα για την επιλογή της χρονικής περιόδου από τον χρήστη για την εκτελεστεί του αλγορίθμου ομαδοποίησης, υπάρχει μια μπάρα χρόνου. Αυτή η μπάρα είναι διαφορετική από αυτή των προηγούμενων περιπτώσεων. Σε αυτή την περίπτωση η μπάρα δεξιά και αριστερά αναγράφει τις διαθέσιμες χρονιές για τους επιλεγμένους δείκτες. Αύτη την φορά υπάρχει η δυνατότητα αντί να επιλέξουμε μόνο μια χρονιά να επιλεχθεί ένα εύρος χρονιών.



Σχήμα 6-21 Παράδειγμα επιλογής χρονικής περιόδου για τον αλγόριθμο ομαδοποίησης

Τέλος κάθε φορά που αλλάζει μια παράμετρο ομαδοποίησης, πέρα από την επιλογή δεικτών, δεν αλλάζουν αυτόματα τα δεδομένα. Για εκτελεστεί ξανά ο αλγόριθμος, στο πάνω δεξιά μέρος της οθόνης, μαζί με τις παραμέτρους, υπάρχει μια επιλογή Refresh.

Κεφάλαιο 7: Απεικόνιση αποτελεσμάτων προσομοιώσεων

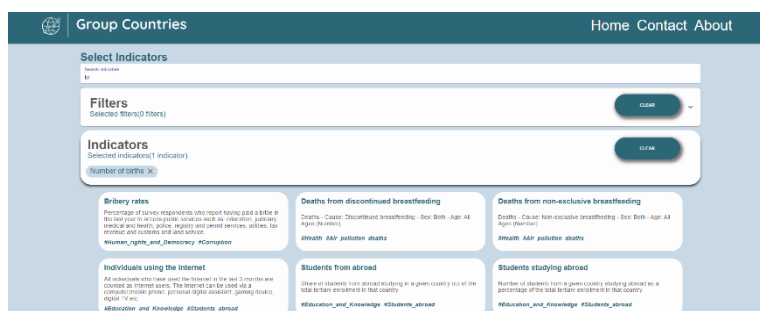
7.1 Αποτελέσματα προσομοίωσης χωρίς ομαδοποίηση

7.2 Αποτελέσματα προσομοίωσης με ομαδοποίηση

Σε αυτό το κεφάλαιο θα εκτελεστούν μερικά παραδείγματα προσομοίωσης της ιστοσελίδας. Σε αυτά τα παραδείγματα η επιλογή δεικτών θα γίνει με τυχαίο τρόπο προσπαθώντας να καλύψουμε όσο το δυνατόν την μεγαλύτερη γκάμα διαφορετικών περιπτώσεων. Οι προσομοιώσεις θα περιέχουν επιλογές ενός αλλά και περισσότερων δεικτών.

7.1 Αποτελέσματα προσομοίωσης χωρίς ομαδοποίηση

Ως πρώτη περίπτωση προσομοίωση θα γίνει η επιλογή ενός μόνο δείκτη από τις διαθέσιμες επιλογές. Ο δείκτης αυτός ονομάζεται αριθμός γεννήσεων (number of births). Όπως αναφέρεται και στην περιγραφή ο δείκτης αφορά τον αριθμό γεννήσεων σύμφωνα με δημογραφικούς δείκτες.



Σχήμα 7-1 Παράδειγμα επιλογής ενός δείκτη

7.1.1 Οπτικοποίηση δεδομένων σε μορφή χάρτη

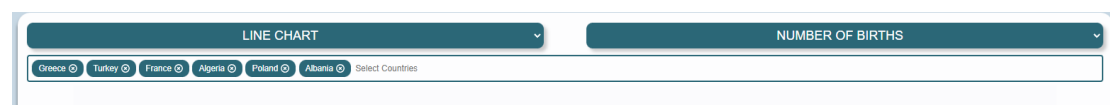
Στις παρακάτω εικόνες εμφανίζονται τα δεδομένα του επιλεγμένου δείκτη σε μορφή χάρτη. Οι εικόνες αναφέρονται σε τέσσερις διαφορετικές χρονικές περιόδους. Η πρώτη και η τελευταία περίοδος αναφέρεται στην πρώτη και τελευταία χρονιά που υπήρχαν διαθέσιμα δεδομένα. Οι υπόλοιπες δύο αφορούν δύο τυχαίες χρονιές ανάμεσα στις προηγούμενες δύο.



Σχήμα 7-2 Αποτελέσματα προσομοίωσης ενός δείκτη σε μορφή χάρτη

7.1.2 Οπτικοποίηση δεδομένων σε μορφή διαγράμματος

Σύμφωνα με την υποενότητα 6.3, υπάρχουν τέσσερις τύποι διαγραμμάτων για την οπτικοποίηση δεδομένων. Όπως είναι γνωστό τα διαγράμματα εμφανίζουν μέχρι δέκα χώρες. Για το παράδειγμα αυτό επιλέχθηκαν οι χώρες που φαίνονται παρακάτω.



Σχήμα 7-3 Επιλεγμένες χώρες

Για το γραμμικό διάγραμμα και το διάγραμμα περιοχής επιλέξαμε την πιο πρόσφατη διαθέσιμη χρονική περίοδο. Για τα υπόλοιπα δύο επιλέχθηκαν κάποιες από τις πρώτες χρονικές περιόδους.



Σχήμα 7-4 Οπτικοποίηση δεδομένων ενός δείκτη σε μορφή διαγραμμάτων

7.1.3 Οπτικοποίηση δεδομένων σε μορφή πίνακα

Τελευταία μορφή οπτικοποίησης δεδομένων χωρίς ομαδοποίηση είναι αυτή του πίνακα. Οι εικόνες αναφέρονται σε τέσσερις διαφορετικές χρονικές περιόδους. Η πρώτη και η τελευταία περίοδος αναφέρεται στην πρώτη και τελευταία χρονιά που υπήρχαν διαθέσιμα δεδομένα. Οι υπόλοιπες δύο αφορούν δύο τυχαίες χρονιές ανάμεσα στις προηγούμενες δύο.

Country Name	Starting Year	Current Year	Starting Year (Base)	Current Year (Base)	Relative Change	Relative Change
Albania	1993	1993	1993	1993	0	0.00%
Armenia	1993	1993	1993	1993	0	0.00%
Azerbaijan	1993	1993	1993	1993	0	0.00%
Bulgaria	1993	1993	1993	1993	0	0.00%
Georgia	1993	1993	1993	1993	0	0.00%
Turkey	1993	1993	1993	1993	0	0.00%
Albania	1993	1993	1993	1993	0	0.00%
Armenia	1993	1993	1993	1993	0	0.00%
Azerbaijan	1993	1993	1993	1993	0	0.00%
Bulgaria	1993	1993	1993	1993	0	0.00%
Georgia	1993	1993	1993	1993	0	0.00%
Turkey	1993	1993	1993	1993	0	0.00%

Σχήμα 7-5 Οπτικοποίηση δεδομένων ενός δείκτη σε μορφή πινάκων

Σε αυτές τις περιπτώσεις προβολής των δεδομένων, απεικονίζονται αποτελέσματα αποκλειστικά για τον πρώτο δείκτη. Σε περίπτωση που επιλεχτούν περισσότεροι από έναν δείκτες, εμφανίζονται τα δεδομένα του πρώτου δείκτη. Αν ο χρήστη επιθυμεί να διαλέξει κάποιον από τους υπόλοιπους επιλεγμένους δείκτες,

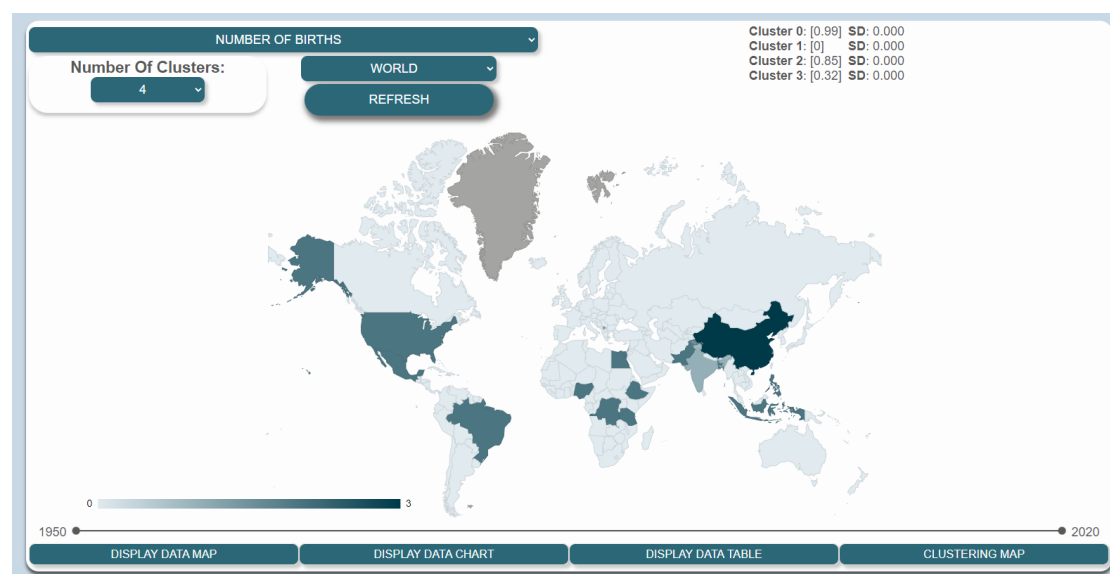
πρέπει να γίνει η επιλογή από τις παραμέτρους όπως αναφέρθηκε στην υποενότητα 6.3.

7.2 Αποτελέσματα προσομοίωσης με ομαδοποίηση

Για την προσομοίωση με ομαδοποίηση θα εξεταστούν δυο περιπτώσεις. Η πρώτη περίπτωση θα αφορά την ομαδοποίηση ενός επιλεγμένου δείκτη ενώ η δεύτερη περισσότερων από έναν.

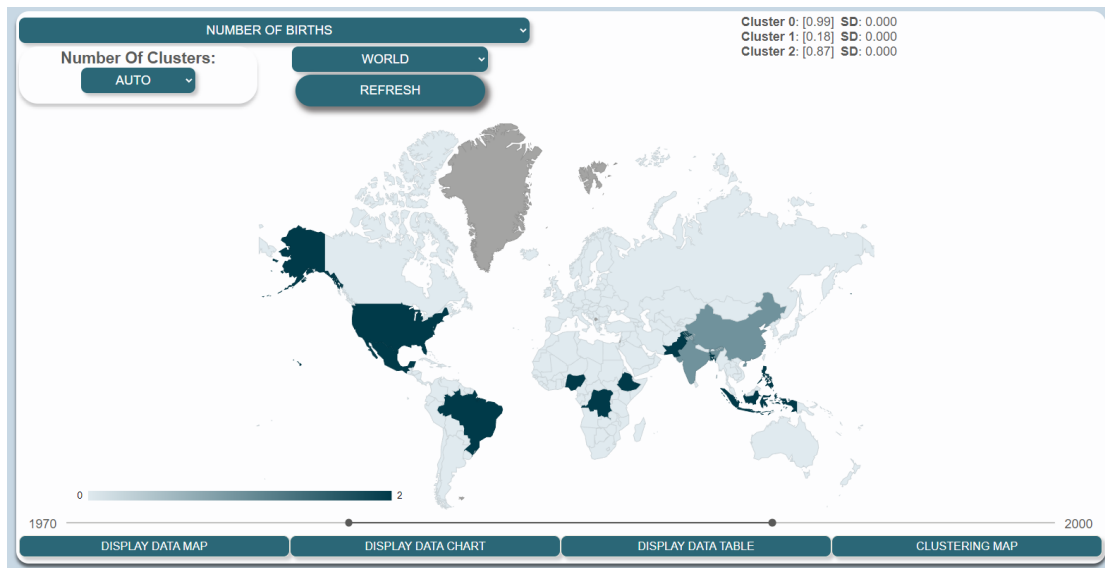
7.2.1 Απεικόνιση αποτελεσμάτων ομαδοποίηση ενός δείκτη

Για την απεικόνιση αποτελεσμάτων ενός δείκτη, επιλέχθηκε ο δείκτης με τον οποίο ασχολήθηκε η υποενότητα 7.1. Για την εκτέλεση του αλγορίθμου ομαδοποίησης επιλέχθηκε ο αριθμός των τεσσάρων ομάδων (clusters) στις οποίες θα χωριστούν οι χώρες. Η χρονική περίοδος που επιλέχθηκε είναι από την πρώτη χρονική διαθέσιμη χρονική περίοδο μέχρι την πιο πρόσφατη.



Σχήμα 7-6 Απεικόνιση αποτελεσμάτων ομαδοποίησης με τέσσερις ομάδες

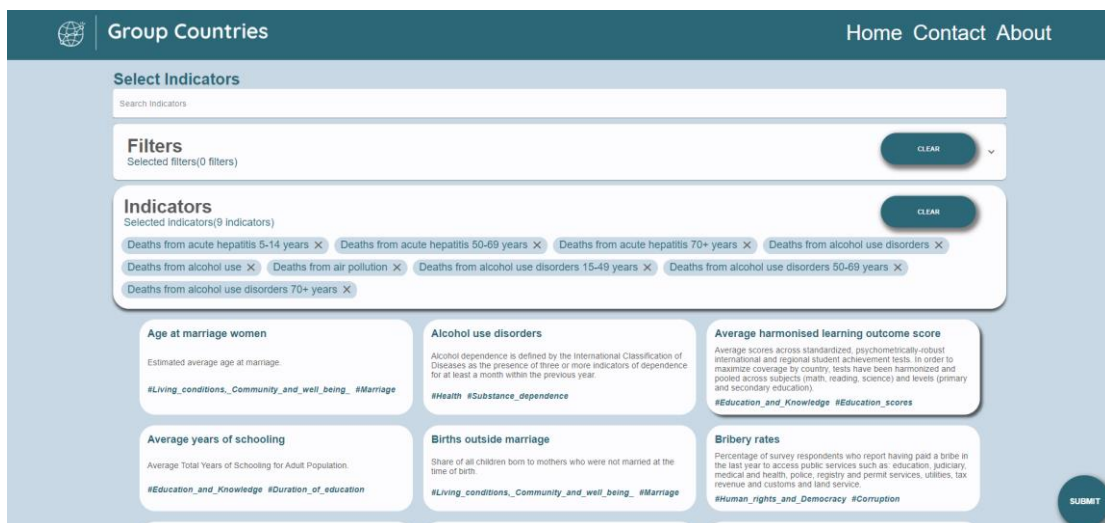
Την ίδια προσομοίωση θα εκτελεστεί με διαφορά στις παραμέτρους. Αυτή την φορά ο αριθμός των ομάδων (clusters) θα είναι ίσος με αυτο, για την εκτέλεση του αλγορίθμου Elbow. Η χρονική περίοδος θα τεθεί ίση με δύο τυχαίες χρονικές περιόδους ανάμεσα στις διαθέσιμες.



Σχήμα 7-7 Απεικόνιση αποτελεσμάτων ομαδοποίησης με επιλογή auto για ομάδες

7.2.2 Απεικόνιση αποτελεσμάτων ομαδοποίησης περισσότερων από ένα δείκτες

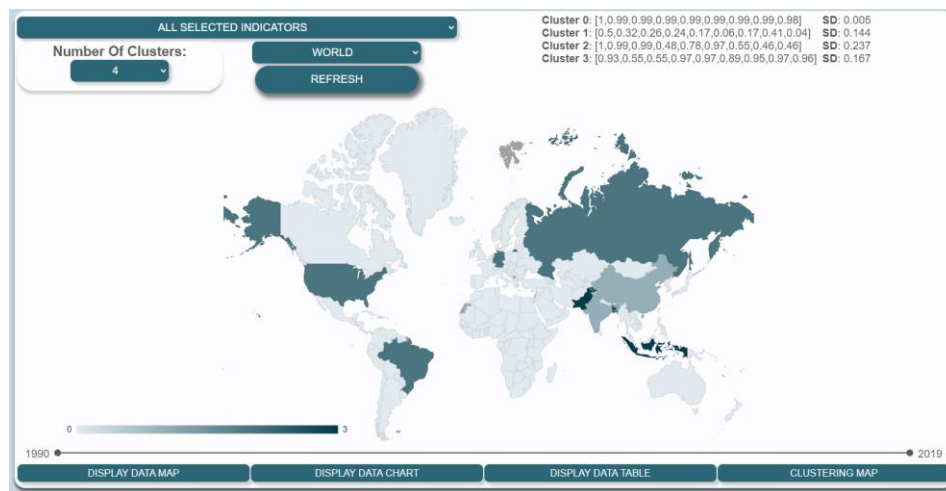
Για την δεύτερη περίπτωση, απεικόνιση αποτελεσμάτων για περισσότερους από έναν δείκτες, έχουν επιλεχθεί οι δείκτες που φαίνονται στην εικόνα που ακολουθεί.



Σχήμα 7-8 Παράδειγμα επιλογής περισσότερων από έναν δείκτη

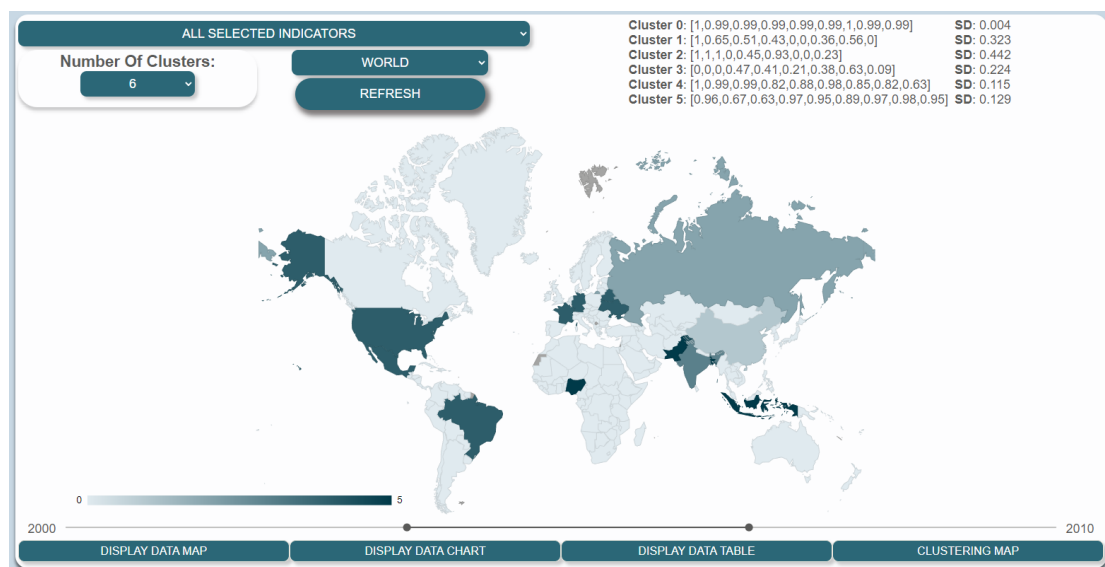
Για την εκτέλεση του αλγορίθμου ομαδοποίησης επιλέχθηκε ο αριθμός των τεσσάρων ομάδων (clusters) στις οποίες θα χωριστούν οι χώρες. Η χρονική περίοδος

που επιλέχθηκε είναι από την πρώτη χρονική διαθέσιμη χρονική περίοδο μέχρι την πιο πρόσφατη.



Σχήμα 7-9 Απεικόνιση αποτελεσμάτων ομαδοποίησης με τέσσερις ομάδες και περισσότερους από έναν δείκτες

Την ίδια προσομοίωση θα εκτελεστεί με διαφορά στις παραμέτρους. Αυτή την φορά ο αριθμός των ομάδων (clusters) θα είναι ίσος έξι και η χρονική περίοδο θα τεθεί ίση με δύο τυχαίες χρονικές περιόδους ανάμεσα στις διαθέσιμες.



Σχήμα 7-10 Απεικόνιση αποτελεσμάτων ομαδοποίησης με έξι ομάδες και περισσότερους από έναν δείκτες

Στην περίπτωση της ομαδοποίησης, μπορούν να απεικονιστούν δεδομένα για περισσότερους από έναν δείκτες μαζί. Σε περίπτωση που ο χρήστης επιθυμεί την επιλογή αποκλειστικά ενός από τους επιλεγμένους δείκτες μπορεί να τον

επιλέξει από τις παραμέτρους όπως αναφέρθηκε στην υποενότητα **6.3**. Αν δεν πραγματοποιηθεί η επιλογή κάποιου δείκτη η εφαρμογή έχει ως προεπιλογή την ομαδοποίηση όλων των επιλεγμένων δεικτών μαζί.

Κεφάλαιο 8: Μελλοντικές επεκτάσεις

8.1 Μεταφόρτωση δεδομένων

8.2 Εξαγωγή αρχείων δεδομένων

Ανακεφαλαιώνοντας έχει δημιουργηθεί μια εφαρμογή η οποία προβάλλει δεδομένα δεικτών στον χρήστη με διάφορες τεχνικές οπτικοποίησης με ομαδοποίηση δεδομένων ή χωρίς. Η κατασκευή της εφαρμογής έχει πραγματοποιηθεί με τέτοιο τρόπο ώστε η οποιαδήποτε μελλοντική επέκταση να χρειαστεί το μικρότερο δυνατό κόστος για την επίτευξή της. Ο λόγος που έχει επιτευχθεί αυτό είναι επειδή έχει δοθεί μεγάλη προσοχή στο να είναι όσο το δυνατόν πιο απλοποιημένη η δομή του κώδικα σε όλα τα επίπεδα. Ακολουθούν παρακάτω μερικές από τις επεκτάσεις που έχουν προταθεί.

8.1 Μεταφόρτωση δεδομένων

Ο χρήστης μπορεί να επιλέξει την απεικόνιση δεδομένων μόνο από μία περιορισμένη γκάμα δεικτών που διαθέτει η εφαρμογή. Θα γινόταν πολύ πιο χρήσιμη η εφαρμογή αν του δινόταν η επιλογή να προβάλλει τους δικούς του δείκτες για την εξαγωγή συμπερασμάτων. Μια σημαντική και πολύ χρήσιμη επέκταση θα ήταν η δημιουργία δυνατότητας προβολής δεδομένων δεικτών που επιθυμεί ο ίδιος. Για να επιτευχθεί αυτό θα πρέπει να δημιουργηθεί η επιλογή της μεταφόρτωσης αρχείων.

Με αυτή την επιλογή, ο χρήστης θα μπορεί να δίνει στην εφαρμογή τα δικά του δεδομένα με την κατάλληλη μορφή αρχείου που θα του υποδεικνύεται. Τα δεδομένα αυτά θα φορτώνονται στην βάση δεδομένων. Πέρα από το αρχείο αυτό θα χρειαστεί να δώσει όνομα, περιγραφή και τις κατηγορίες στις οποίες ανήκει ο

δείκτης. Αφού ολοκληρωθεί η διαδικασία, ο δείκτης θα βρίσκεται στους διαθέσιμους δείκτες για την επιλογή του.

8.2 Εξαγωγή αρχείων δεδομένων

Η εφαρμογή δίνει την δυνατότητα στον χρήστη προβολής των δεδομένων που επέλεξε. Χρήσιμη προσθήκη θα ήταν η δημιουργία δυνατότητας αποθήκευση των δεδομένων για την μελλοντική επαναχρησιμοποίηση. Αυτό θα γινόταν εφικτό με την δημιουργία αντιγράφου σε μορφή csv αρχείου των δεδομένων από την βάση δεδομένων. Ακόμα θα μπορούσε να ήταν δυνατή η εξαγωγή αρχείου csv που θα περιέχει τα αποτελέσματα την ομαδοποίησης. Μέσα σε αυτά τα δεδομένα θα καταγράφονται ο αριθμός των ομάδων που χωρίστηκαν οι χώρες καθώς και οι ομάδα που ανήκει η κάθε χώρα. Τέλος το αρχείο αυτό θα περιέχει πληροφορίες για τα κέντα των ομάδων μαζί με την τυπική απόκλισή τους.

Γλωσσάρι

WCSS: Within Cluster Sum of Squares

OWID: Our World in Data

CSV: Comma Separated Values

SQL: Structured Query Language

DBMS: Database Management System

DDL: Data Definition Language

REST: Representational State Transfer

URL: Uniform Resource Locator

API: Application Programming Interface

HTML: HyperText Markup Language

CSS: Cascading Style Sheets

SD: Standard Deviation

Βιβλιογραφία

Advantages and Disadvantages of Bar Graphs [Ηλεκτρονικό] // All Things Statistics. - 21 Μαΐος 2022. - <https://allthingsstatistics.com/miscellaneous/bar-graphs-advantages-disadvantages/>.

Advantages and Disadvantages of Scatter Diagrams [Ηλεκτρονικό] // All Things Statistics. - 21 Μάιος 2022. - <https://allthingsstatistics.com/miscellaneous/scatter-diagram-advantages-disadvantages/>.

Area Charts: A guide for beginners [Ηλεκτρονικό] // FusionCharts. - 2022. - <https://www.fusioncharts.com/area-charts>.

Clustering in Machine Learning [Ηλεκτρονικό] // developers google. - 13 Ιανουάριος 2021. - <https://developers.google.com/machine-learning/clustering/algorithm/advantages-disadvantages>.

K-Means Clustering Algorithm [Ηλεκτρονικό] // javaTpoint. - <https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning>.

Line graph [Ηλεκτρονικό] // Math.net. - <https://www.math.net/line-graph>.

Maddy Spring Boot Architecture [Ηλεκτρονικό] // dev. - 14 Νοέμβριος 2021. - <https://dev.to/maddy/spring-boot-architecture-547i>.

Martin Matthew What is Backend Developer? Skills Need for Web Development [Ηλεκτρονικό] // Guru99. - 7 Μάιος 2022. - <https://www.guru99.com/what-is-backend-developer.html#3>.

oracle [Ηλεκτρονικό] // What Is a Database?. - 2022. - <https://www.oracle.com/database/what-is-database/>.

Priy Surya Clustering in Machine Learning [Ηλεκτρονικό] // GeeksforGeeks. - 18 Μάιος 2022. - <https://www.geeksforgeeks.org/clustering-in-machine-learning/>.

reactjs [Ηλεκτρονικό] // Thinking in React. - <https://reactjs.org/docs/thinking-in-react.html>.

Representational state transfer [Ηλεκτρονικό] // Wikipedia. - 30 Ιούνιος 2022. - https://en.wikipedia.org/wiki/Representational_state_transfer.

sklearn.cluster.KMeans [Ηλεκτρονικό] // sklearn.cluster. - Μάιος 2021. - <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>.

Spring Boot - Introduction [Ηλεκτρονικό] // tutorialspoint. - 2022. - https://www.tutorialspoint.com/spring_boot/spring_boot_introduction.htm#:~:text=Why%20Spring%20Boot%3F&text=It%20provides%20a%20flexible%20way,no%20manual%20configurations%20are%20needed..

Thakur Nitish Kumar Medium [Ηλεκτρονικό] // k-Means Clustering: Comparison of Initialization strategies.. - 11 Απρίλιος 2020. - <https://medium.com/analytics-vidhya/comparison-of-initialization-strategies-for-k-means-d5ddd8b0350e>.

What is a Geographical Chart? [Ηλεκτρονικό] // tibco. - 2022. - <https://www.tibco.com/reference-center/what-is-a-geographical-chart>.

What Is Data Visualization? Definition, Examples, And Learning Resources [Ηλεκτρονικό] // tableau. - <https://www.tableau.com/learn/articles/data-visualization>.

What Is Machine Learning: Definition, Types, Applications And Examples [Ηλεκτρονικό] // Potentia Analytics. - 8 Νοέμβριος 2021. - <https://www.potentiaco.com/what-is-machine-learning-definition-types-applications-and-examples/#:~:text=These%20are%20three%20types%20of,unsupervised%20learning%2C%20and%20reinforcement%20learning..>