Εξόρυξη Δεδομένων και Αλγόριθμοι Μηχανικής Μάθησης

Υλοποιητική Εργασία

Λουδάρος Ιωάννης (1067400) - Χριστίνα Κρατημένου (1059660)

Μπορείτε να δείτε την τελευταία έκδοση του Project εδώ ή σκανάροντας τον κωδικό QR που βρίσκεται στην επικεφαλίδα.

Περιγραφή Αναφοράς

Παρακάτω παραθέτουμε τις απαντήσεις μας στην "Υλοποιητική Εργασία" του μαθήματος "Εξόρυξη Δεδομένων και Αλγόριθμοι Μηχανικής Μάθησης" καθώς και σχόλια τα οποία προέκυψαν κατά την εκπόνηση της.

Περιεχόμενα

Jupyter Notebook

Απαντήσεις

Εισαγωγικά

Περιβάλλον Υλοποίησης

Το Project είναι υλοποιημένο σε python. Δεν χρειάζεται να εγκαταστήσετε τίποτα για να δείτε τον κώδικα και τα αποτελέσματα του. Αν επιθυμείτε να τον εκτελέσετε τοπικά θα χρειαστεί να διαβάσετε το περιεχόμενο του παραλληλογράμμου που ακολουθεί, διαφορετικά, μπορείτε να το προσπεράσετε.

Για την εκτέλεση του κώδικα θα χρειαστείτε ένα περιβάλλον το οποίο να υποστηρίζει Jupiter Notebooks. Ενδεικτικά μπορείτε:

1. Να εγκαταστήσετε το JupyterLab, ή
2. Να εγκαταστήσετε το VSCode και ύστερα, να προσθέσετε το Extension Jupiter

Βιβλιοθήκες που Χρησιμοποιούνται

Για την υλοποίηση του project χρησιμοποιήσαμε τις παρακάτω βιβλιοθήκες. Αν θέλετε να εκτελέσετε τον κώδικα μας τοπικά, θα χρειαστεί να τις εγκαταστήσετε με την βοήθεια του pip.

Διαχείριση Δεδομένων

- Pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool.

Machine Learning Algorithms

- Scikit-learn is an open source machine learning library that supports supervised and unsupervised learning. It also provides various tools for model fitting, data preprocessing, model selection, model evaluation, and many other utilities.

- Keras is the high-level API of the TensorFlow platform. It provides an approachable, highly-productive interface for solving machine learning (ML) problems, with a focus on modern deep learning

---

Visualisation

- Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

- Yellowbrick: Machine Learning Visualization

Δομή Παραδοτέου

Η παρούσα αναφορά συνοδεύεται από ένα Jupyter Notebook το οποίο περιέχει και σχολιάζει λεπτομερώς την υλοποίηση του Project μας. Στο παρόν έγγραφο λοιπόν θα σταθούμε μόνο στην παρουσίαση των αποτελεσμάτων που προκύπτουν, δίνοντας παραπομπές που αναφέρονται στο Notebook ώστε να μπορεί ο αναγνώστης να ανατρέξει στον αντίστοιχο κώδικα. Χρησιμοποιήστε το κουμπί στην πρώτη σελίδα για να οδηγηθείτε στο Notebook.

1. Πρώτες Παρατηρήσεις γύρω από το Dataset

Αρχικά παρατηρούμε πως το dataset μας περιέχει στατιστικά στοιχεία σχετικά με την καθημερινή εξέλιξη της νόσου COVID19 σε 104 χώρες από τον Ιανουάριο του 2020 έως και τον Φεβρουάριο του 2021. Είναι στοιχισμένο πρώτα ανά χώρα και μετά ανά ημερομηνία, κάτι που διευκολύνει την δουλειά μας.

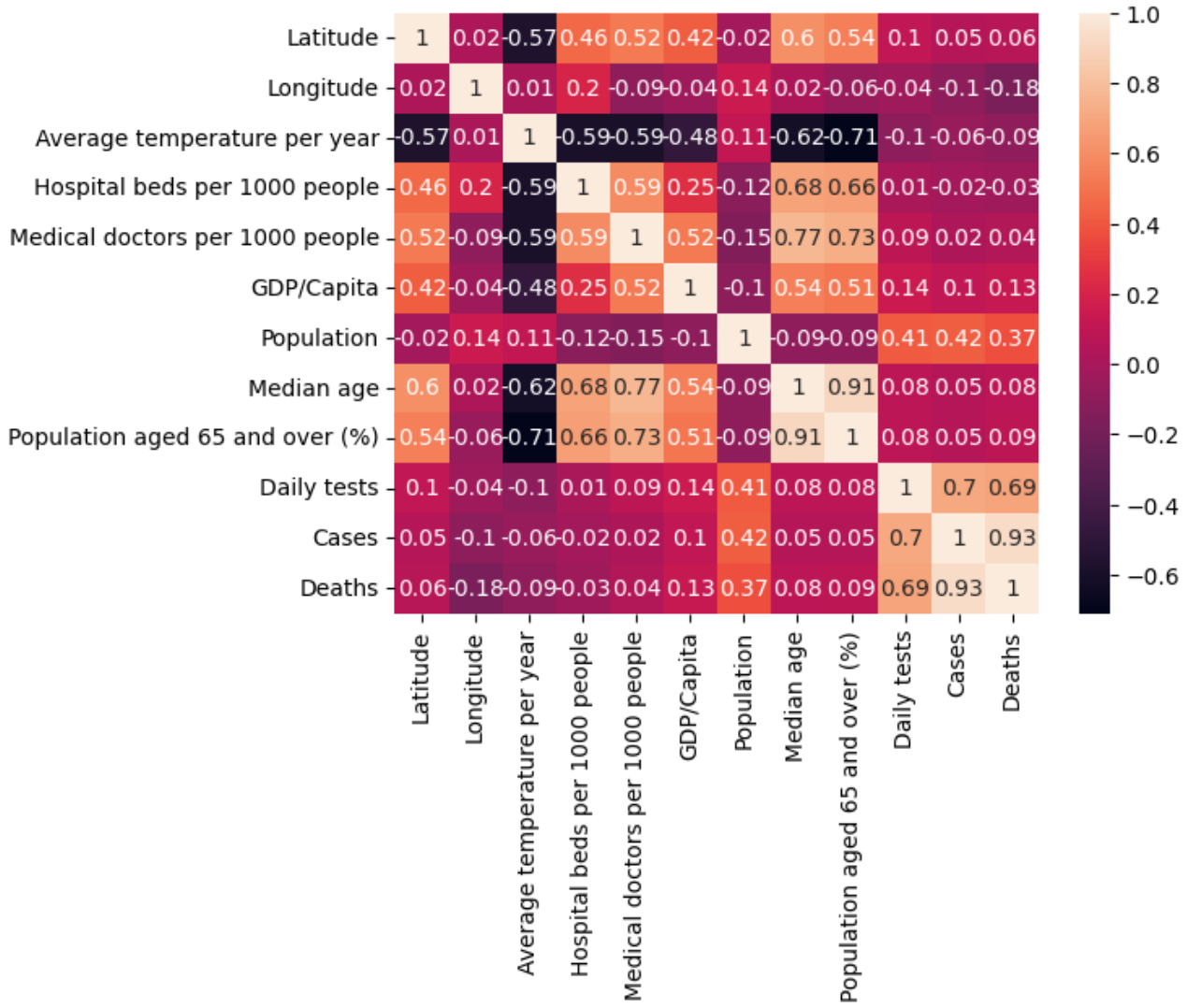
Μετά από το απαιτούμενο preprocessing (αποκαταστήσαμε τις χαμένες τιμές χρησιμοποιώντας γραμμική παλινδρόμηση) προσπαθούμε να εξάγουμε κάποια πρώτα στοιχεία από το dataset. Ξεκινάμε εξάγοντας στατιστικά για ολόκληρο το dataset ανά στήλη. Ύστερα, σχεδιάζουμε ένα heatmap για να εξερευνήσουμε τις συσχετίσεις μεταξύ των σειρών που έχουμε. Επίσης, εξερευνούμε τις επιδόσεις κάθε χώρας. Με τα παραπάνω στοιχεία ελπίζουμε να ανακαλύψουμε κάποια χαρακτηριστικά στοιχεία τα οποία φαίνεται να παίζουν σημαντικό ρόλο στην “επίδοση” της κάθε χώρα στην αντιμετώπιση του ιού.

Αρχικά στατιστικά για ολόκληρο το dataset

	Latitude	Longitude	Average temperature per year	Hospital beds per 1000 people	Medical doctors per 1000 people	GDP/Capita
count	38472.00	38472.00	38472.00	38472.00	38472.00	38472.00
mean	23.74	20.21	17.72	3.17	2.09	19002.33
std	26.06	61.07	8.13	2.56	1.52	22271.11
min	-40.90	-106.35	-2.00	0.20	0.02	411.60
25%	8.62	-3.44	11.00	1.40	0.82	3659.00
50%	27.51	21.82	20.00	2.50	1.89	8821.80
75%	45.94	47.48	25.00	4.49	3.21	25946.20
max	64.96	179.41	29.00	13.05	7.52	114704.60

	Population	Median age	Population aged 65 and over (%)	Daily tests	Cases	Deaths
count	38472.00	38472.00	38472.00	30577.00	38218.00	34862.00
mean	48969830.00	32.75	10.66	39440.59	287902.70	8090.50
std	142725100.00	8.47	6.77	150184.70	1405243.00	29548.75
min	341284.00	16.00	1.00	-239172.00	1.00	1.00
25%	4793900.00	27.00	5.00	1505.00	2074.00	77.00
50%	11484640.00	32.00	8.00	5520.00	21431.00	527.00
75%	42862960.00	41.00	16.00	20382.00	137377.00	3480.50
max	1339180000.0	48.00	28.00	2945871.00	28605670.00	513091.00

Correlation Heatmap



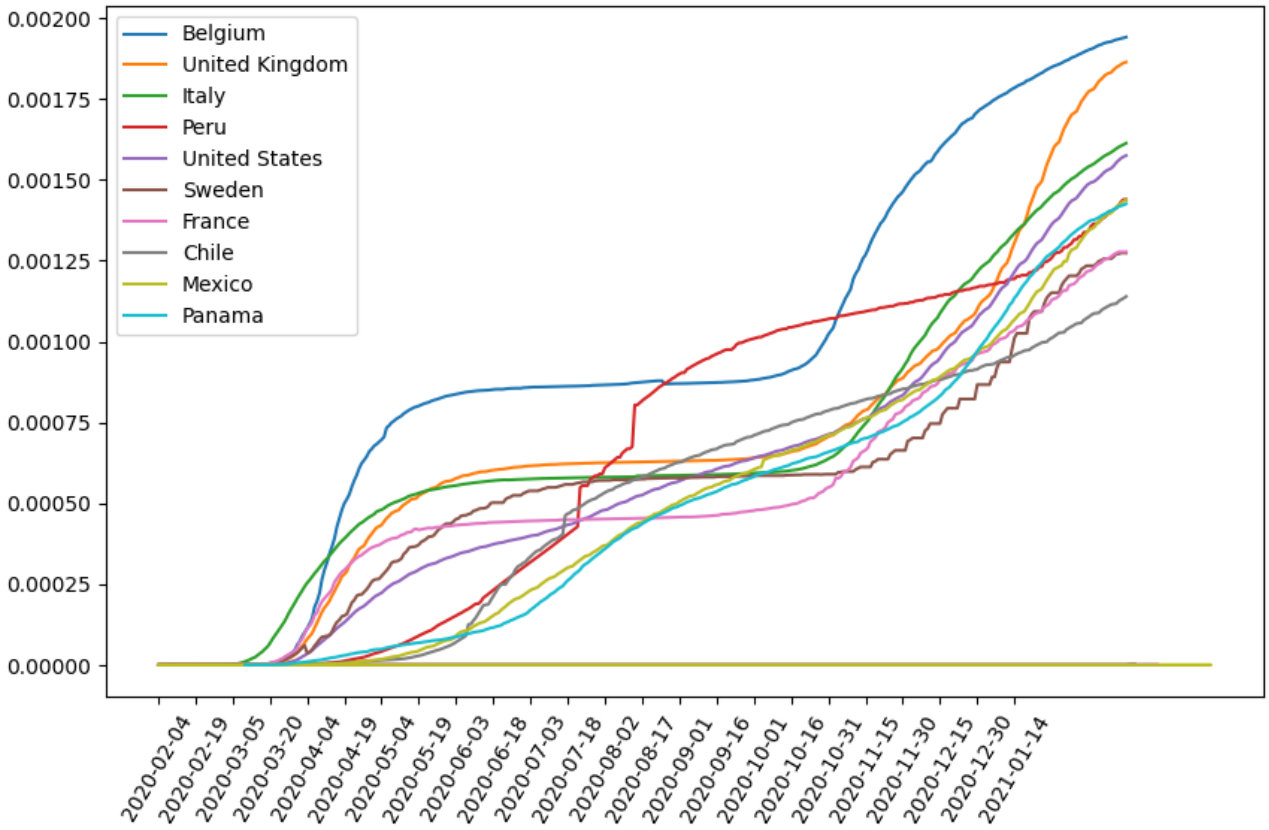
Μπορούμε να κάνουμε κάποια σχόλια ήδη για τις συσχετίσεις που προκύπτουν από τα δεδομένα.

- Όπως είναι αναμενόμενο, οι θάνατοι έχουν πολύ σημαντική θετική συσχέτιση με το πλήθος των κρουσμάτων.
- Πιο γερασμένοι πληθυσμοί φαίνεται να παρουσιάζουν περισσότερους θανάτους.
- Το Γεωγραφικό Μήκος έχει υψηλή συσχέτηση με την θερμοκρασία.
- Οι χώρες που κάνουν περισσότερα τεστ, έχουν περισσότερα κρούσματα (αφού μπορούν να τα αναφέρουν)
- Οι πλουσιότερες χώρες τείνουν να έχουν περισσότερα νοσοκομεία και γιατρούς κατά κεφαλήν και να έχουν πιο γερασμένους πληθυσμούς.

Έχοντας ανακαλύψει τις παραπάνω συσχετίσεις αποφασίζουμε να κάνουμε γραφικές παραστάσεις για να ανακαλύψουμε περισσότερα για αυτές.

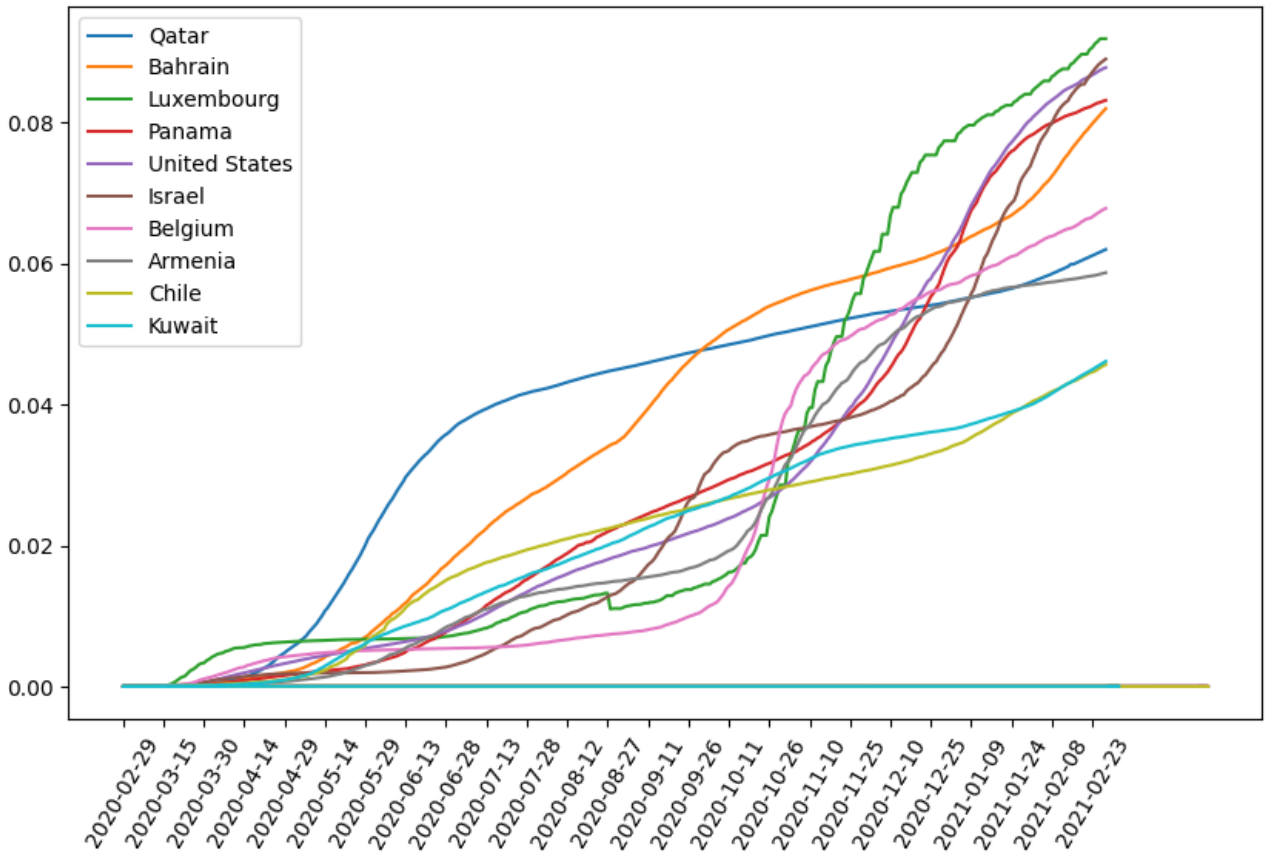
Εξέλιξη θανάτων ανά χώρα (κανονικοποιημένο με τον εκάστοτε πληθυσμό)

Παρουσιάζουμε τις χειρότερες 10 χώρες.

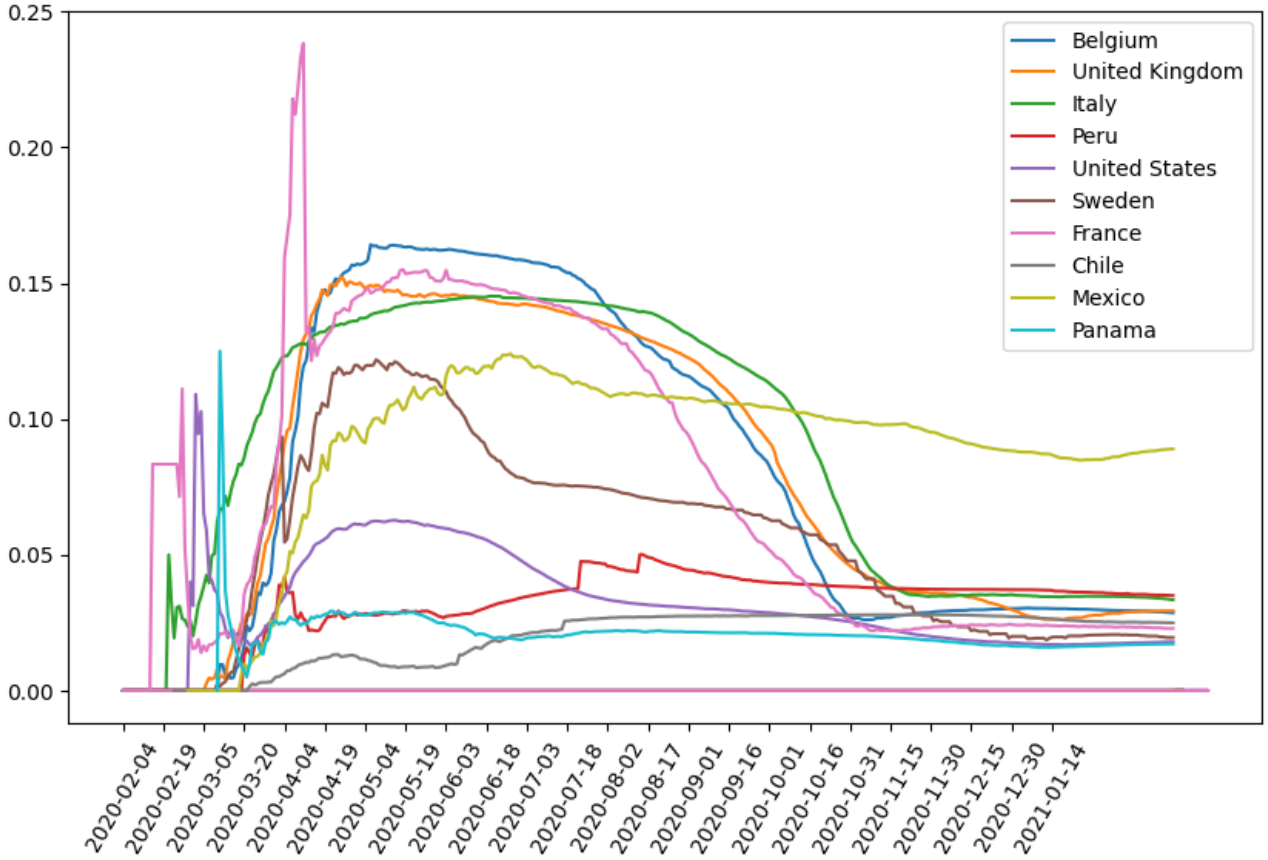


Εξέλιξη κρουσμάτων ανά χώρα (κανονικοποιημένο)

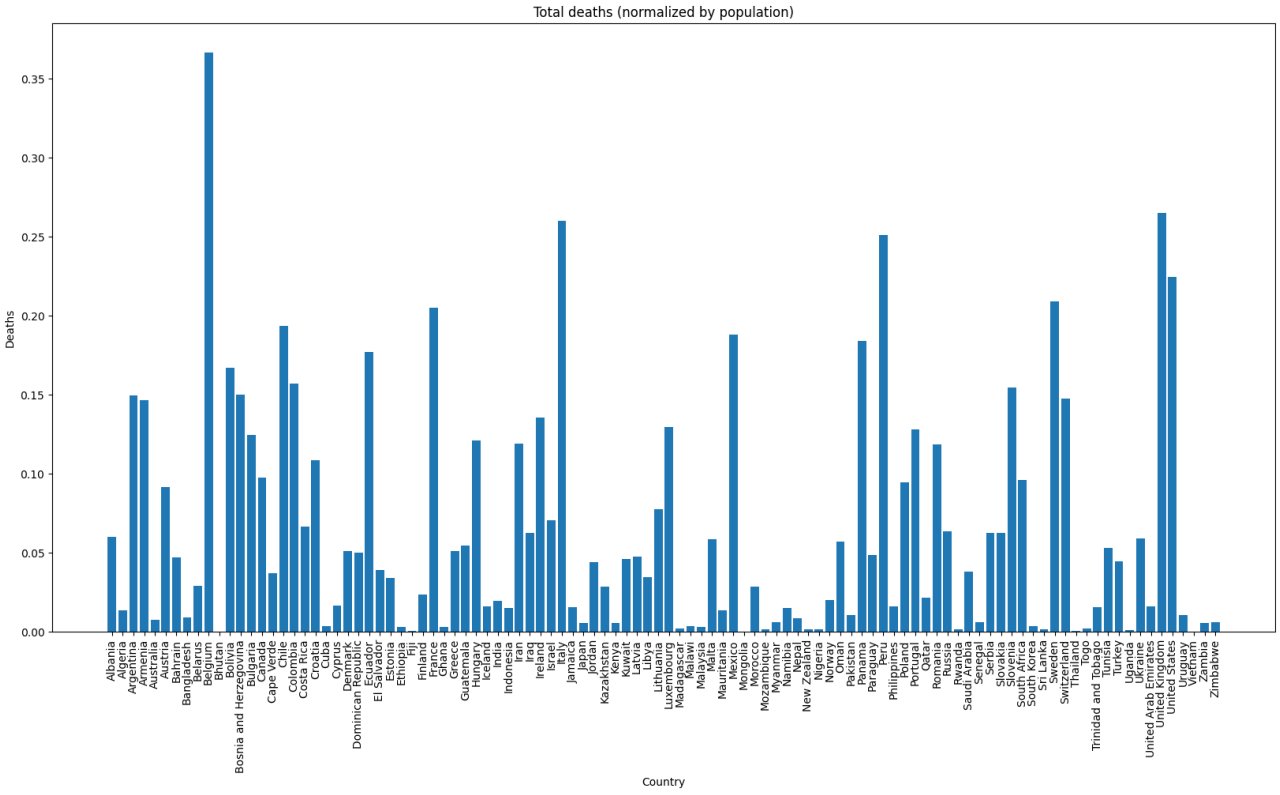
Παρουσιάζουμε τις χειρότερες 10 χώρες.



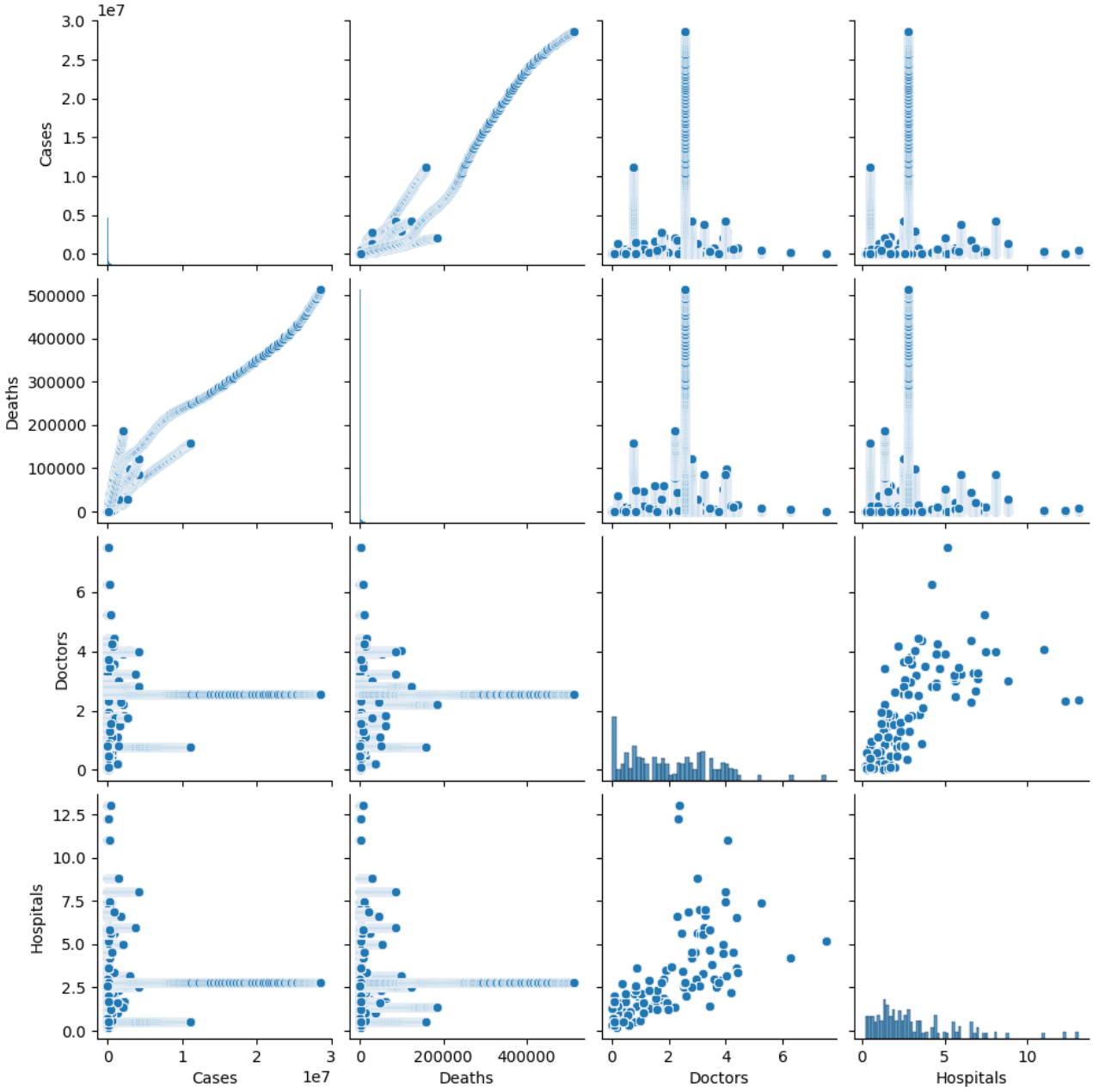
Εξέλιξη θανάτων ανά κρούσμα σε κάθε χώρα



Total Deaths (normalized)



Pairplot



Μια ενδιαφέρουσα παρατήρηση από το παραπάνω pairplot είναι ότι οι περισσότεροι θάνατοι είναι συγκεντρωμένοι στα αριστερά μισά των γραφικών παραστάσεων των γιατρών και των νοσοκομείων. Αυτή η παρατήρηση μπορεί να υποδηλώνει ότι το πλήθος των γιατρών και των νοσοκομείων σε μια χώρα παίζει σημαντικότερο ρόλο από ότι είχαμε εκτιμήσει στην αρχή μέσω το correlation matrix.

2. Συσταδοποίηση

Για την Συσταδοποίηση των χωρών χρησιμοποιήσαμε κάποιες μετρικές που θεωρήσαμε χρήσιμες όπως:

- Θετικότητα στα Test
- Θνητότητα
- Λόγος Κρουσμάτων στον συνολικό πληθυσμό
- Λόγος Θανάτων στον συνολικό πληθυσμό
- Λόγος Test στον συνολικό πληθυσμό

Αρχικά θεωρήσαμε απαραίτητο να κάνουμε scaling στα δεδομένα μας. Για να αναδείξουμε τη χρησιμότητα του, θα παρατηρήσετε ότι έχουμε εκτελέσει τον αλγόριθμο Means δύο φορές. Την πρώτη, χρησιμοποιήσαμε raw δεδομένα, ενώ την δεύτερη, χρησιμοποιήσαμε τα scaled δεδομένα.

Χρησιμοποιώντας το elbow method καταλήξαμε ότι χρειαζόμαστε 4 clusters.

Αποτελέσματα Συσταδοποίησης

Η συσταδοποίηση, μας επέστρεψε τις ακόλουθες συστάδες

Cluster 0

Albania, Algeria, Argentina, Armenia, Bulgaria, Canada, Cape Verde, Colombia, Costa Rica, Croatia, Dominican Republic, Guatemala, Hungary, Iran, Iraq, Kuwait, Lithuania, Oman, Paraguay, Poland, Romania, Serbia, South Africa, Tunisia, Ukraine

Cluster 1

Austria, Bahrain, Cyprus, Denmark, Israel, Luxembourg, Malta, Qatar, Slovakia, United Arab Emirates

Cluster 2

Australia, Bangladesh, Belarus, Bhutan, Cuba, El Salvador, Estonia, Ethiopia, Fiji, Finland, Ghana, Greece, Iceland, India, Indonesia, Jamaica, Japan, Jordan, Kazakhstan, Kenya, Latvia, Libya, Madagascar, Malawi, Malaysia, Mauritania, Mongolia, Morocco, Mozambique, Myanmar, Namibia, Nepal, New Zealand, Nigeria, Norway, Pakistan, Philippines, Russia, Rwanda, Saudi Arabia, Senegal, South Korea, Sri Lanka, Thailand, Togo, Trinidad and Tobago, Turkey, Uganda, Uruguay, Vietnam, Zambia, Zimbabwe

Cluster 3

Bolivia, Bosnia and Herzegovina, Ecuador, Mexico, Peru

Cluster 4

Belgium, Chile, France, Ireland, Italy, Panama, Portugal, Slovenia, Sweden, Switzerland, United Kingdom, United States

Σχολιασμός Συσταδοποίησης

Cluster 0

Βλέπουμε ότι στο Cluster 0 συγκεντρώνονται χώρες οι οποίες δεν βρίσκονται σε καλή οικονομική κατάσταση. Φαίνεται να έχουν αρκετά κρούσματα, αλλά να αντεπεξήλθαν ικανοποιητικά αφού δεν έχουν τραγικά υψηλή θνητότητα.

Cluster	GDP/Capita	Cases per Capita	Death Rate
0	11183.94	4.04	0.02

Cluster 1

Εδώ βλέπουμε πλούσιες χώρες οι οποίες κατάφεραν να κρατήσουν χαμηλά την θνητότητα τους, παρόλο που είχαν αυξημένα κρούσματα.

Cluster	GDP/Capita	Cases per Capita	Death Rate
1	47424.54	6.99	0.01

Cluster 2

Εδώ έχουμε χώρες μεσαίας οικονομικής κατάστασης, στις οποίες, τόσο τα κρούσματα, όσο και η θνητότητα ήταν περιορισμένα

Cluster	GDP/Capita	Cases per Capita	Death Rate
2	12037.92	1.00	0.02

Cluster 3

Φτωχές χώρες που δεν κατάφεραν να ανταπεξέλθουν στον ιό και είχαν πολλά κρούσματα και υψηλή θνητότητα.

Cluster	GDP/Capita	Cases per Capita	Death Rate
3	6553.62	3.73	0.06

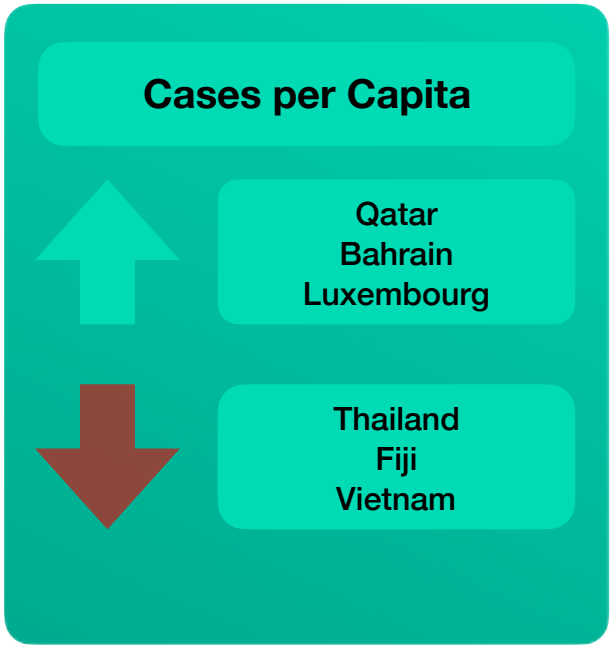
Cluster 4

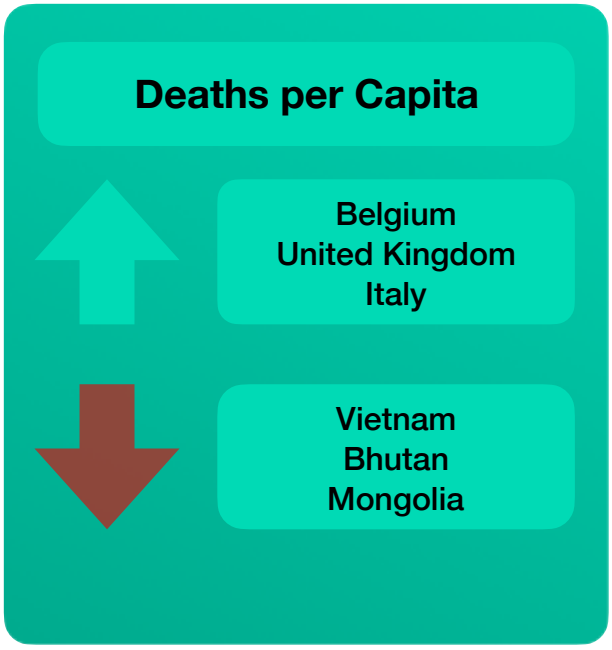
Εδώ εντοπίζουμε χώρες, που παρόλο τον πλούτο τους, είχαν κακές επιδόσεις στην αντιμετώπιση του ιού.

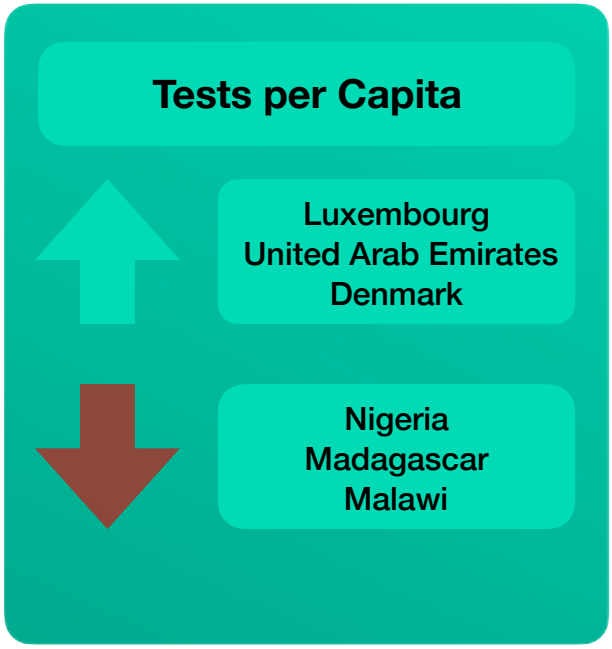
Cluster	GDP/Capita	Cases per Capita	Death Rate
4	43322.16	7.02	0.03

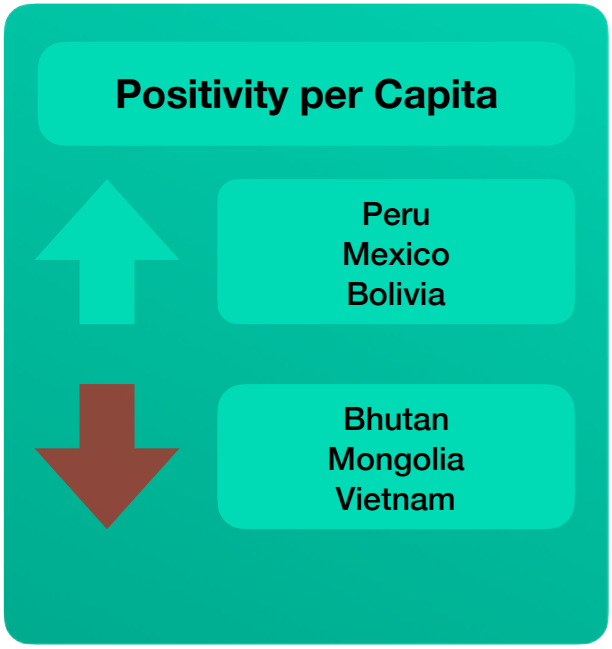
Χώρες που ξεχωρίζουν

Παραθέτουμε τις 3 καλύτερες και τις 3 χειρότερες χώρες σε κάθε μετρική. Όπως είναι αναμενόμενο, συχνότερα από ότι όχι, οι χώρες στις καλύτερες ή στις χειρότερες θέσεις ανήκουν στις ίδιες συστάδες.









3. Εκπαίδευση Παλινδρομητών

Χρησιμοποιήσαμε δύο πηγές[1][2] που βρήκαμε στο διαδύκτιο ώστε να υλοποιήσουμε τους παλινδρομητές μας.

Η διαδικασία που ακολουθήσαμε είναι η παρακάτω:

1. Χωρίσαμε το dataset σε δύο μέρη (πριν και μετά την 1/1/2021). Χρησιμοποιήσαμε όλο το πρώτο μέρος ως training set και όλο το δεύτερο ως testing.
2. Κάναμε Scaling στα δεδομένα μας. (Στο SVM χρησιμοποιήσαμε Standard Scaler ενώ στο RNN χρησιμοποιήσαμε MinMax με range(0, 1).
3. Εκπαιδεύσαμε το μοντέλο μας.
4. Αξιολογήσαμε τα αποτελέσματα.

Νικητής φαίνεται να είναι το RNN για αυτή την εφαρμογή. Χρησιμοποιήσαμε την μετρική του μέσου τετραγωνικού σφάλματος και τα αποτελέσματα φαίνονται παρακάτω:

SVM

MSE : 0.0071

RNN

MSE : 0.0042

Να σημειωθεί πως σίγουρα υπάρχει χώρος για την βελτίωση και των δύο υλοποιήσεων.

[1] Πηγή για RNNs

[2] Πηγή για SVMs