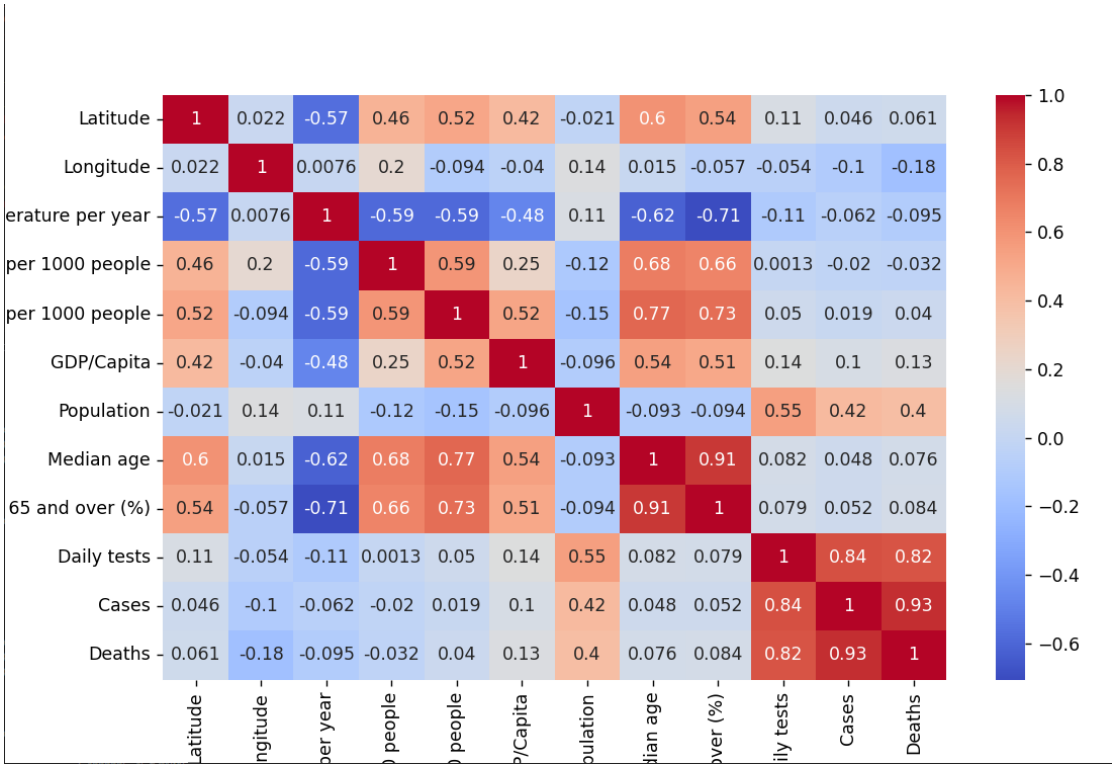
Εξόρυξη Δεδομένων

Σεργιάννης Παρασκευάς - Βασίλειος
AM: 1067467
Έτος: 5ο

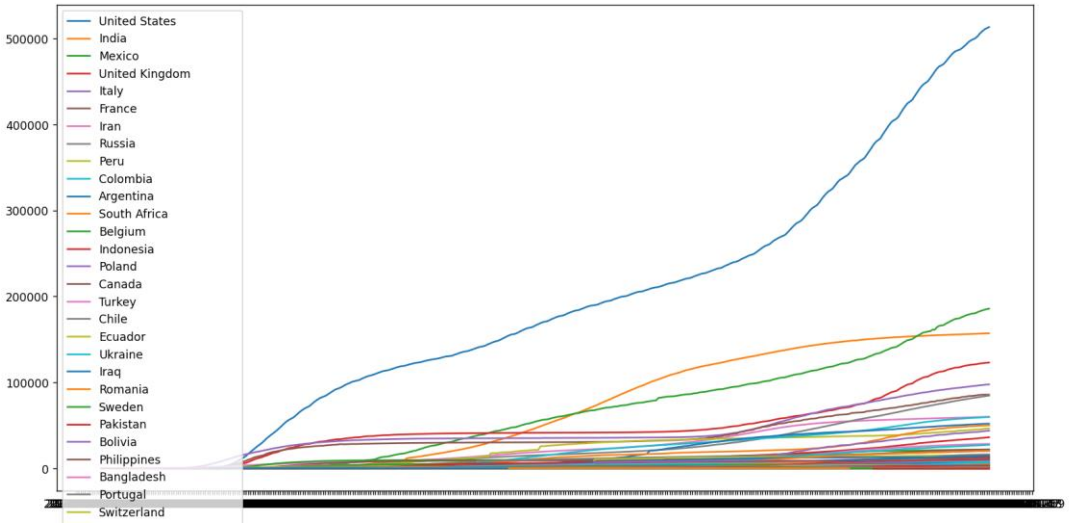
Ερώτημα 1

Για να πραγματοποιήσουμε την πρώτη ανάλυση των δεδομένων φορτώνουμε τα δεδομένα μας σε μια dataframe μεταβλητή (την df) και έπειτα υπολογίζουμε τα συγκεντρωτικά στατιστικά μεγέθη για τις δοθέντες τιμές με την χρήση του df.describe(), το πλήθος των τιμών που λείπουν με την εντολή df.isnull().sum() και τέλος τις συσχετίσεις μεταξύ των features με αριθμητικές τιμές με το df.corr(). Όλα αυτά τα κάνουμε print.
Έπειτα σχεδιάζουμε τις ακόλουθες γραφικές παραστάσεις που ακολουθούν, τις οποίες θα μελετήσουμε για να βγάλουμε τα αρχικά μας συμπεράσματα.
Τα γραφήματα φαίνονται παρακάτω και αφορούν:

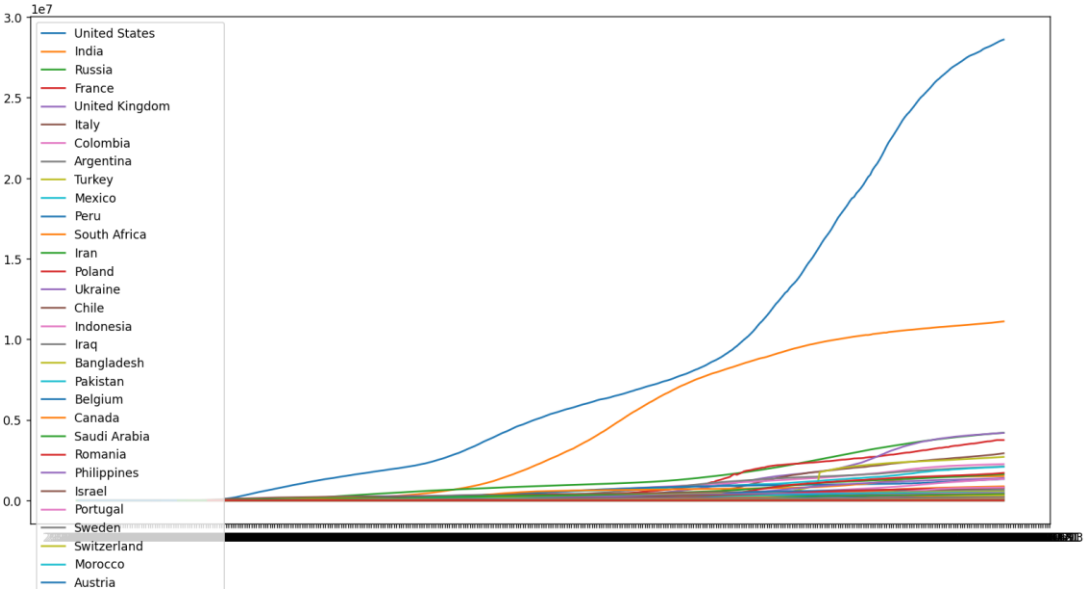
Heatmap με Συσχετίσεις Χαρακτηριστικών

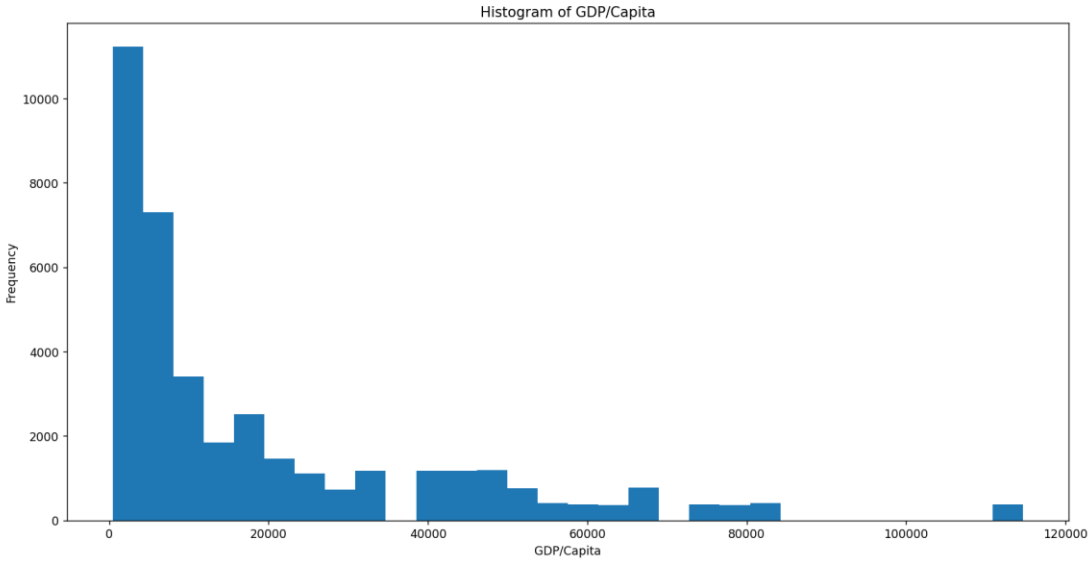


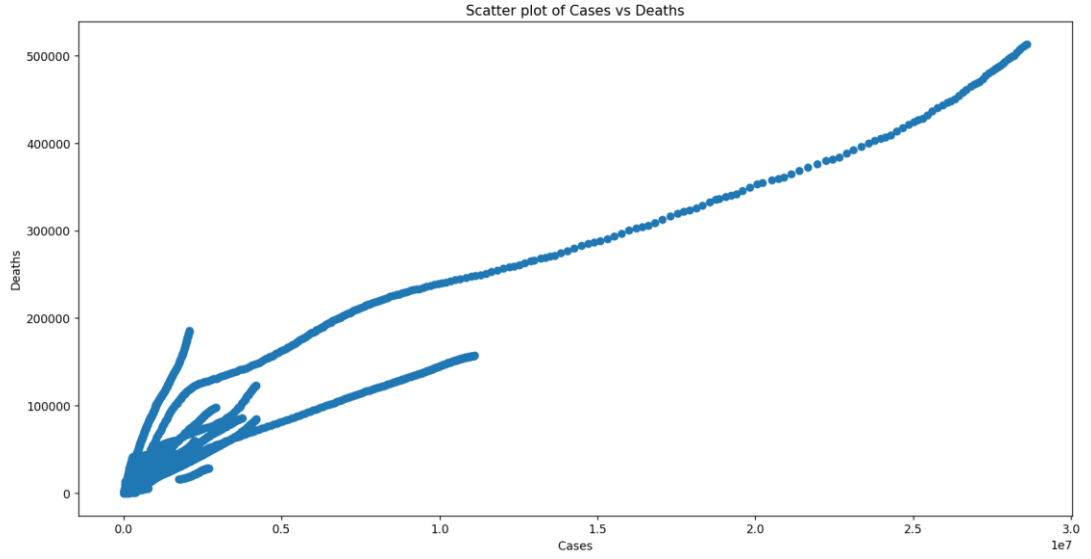
Deaths over time per country

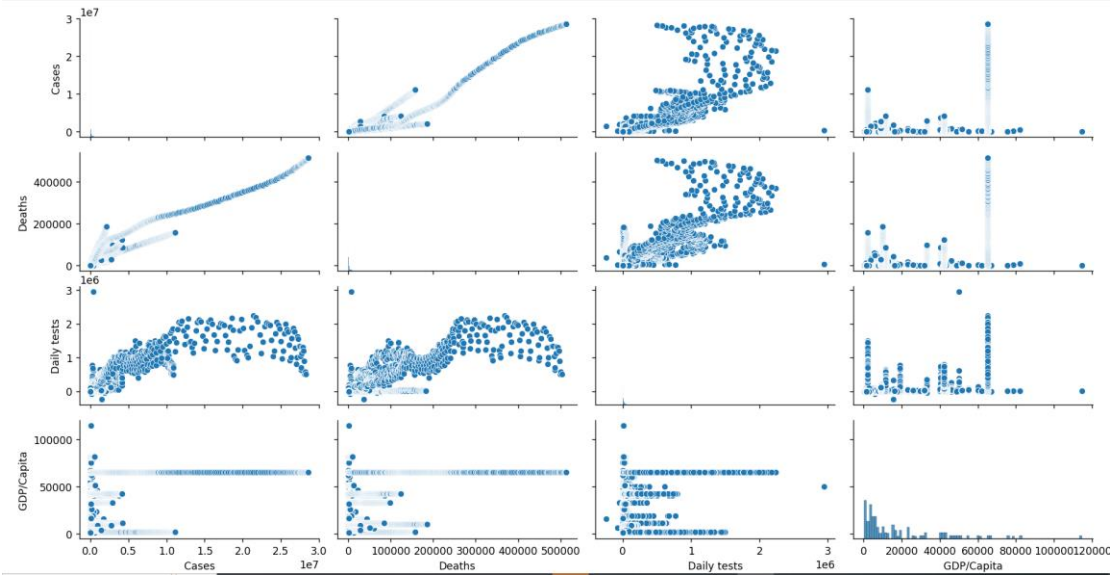


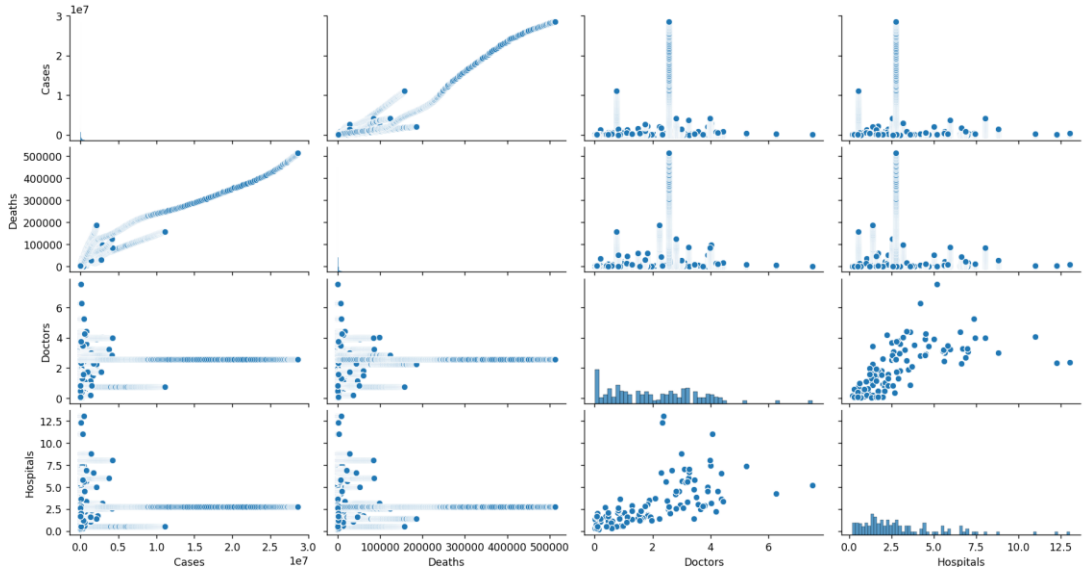
Cases over time per country




Histogram of GDP/Capita
Frequency
GDP/Capita
0
2000
4000
6000
8000
10000
0
20000
40000
60000
80000
100000
120000


Scatter plot of Cases vs Deaths
Deaths
Cases
1e7
0
100000
200000
300000
400000
500000
0.0
0.5
1.0
1.5
2.0
2.5
3.0


Cases
Deaths
Daily tests
GDP/Capita
1e7
1e6
0
1
2
3
0
200000
400000
0
50000
100000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
100000
200000
300000
400000
500000
0
1
2
3
0
20000
40000
60000
80000
100000
120000



Από τα γραφήματα και την πρώτη ανάλυση του συνόλου δεδομένων, παρατηρούμε οτι:
- Οι χώρες με υψηλότερο "Median age" τείνουν να έχουν μεγαλύτερο "Population aged 65 and over (%)" που είναι πολύ λογικό καθώς και τα δύο έχουν να κάνουν με την ηλικειακή κατανομήτης κάθε χώρας.
- Οι χώρες που βρίσκονται πιο κοντά στους πόλους (υψηλότερες τιμές "latitude") τείνουν να έχουν χαμηλότερες μέσες θερμοκρασίες και υψηλότερο ποσοστό ηλικιωμένων.
- Οι χώρες που πραγματοποιούν περισσότερα τεστ αναφέρουν περισσότερα κρούσματα και, κατά συνέπεια και περισσότερους θανάτους.
- Οι χώρες με γηραιότερο πληθυσμό μπορεί να έχουν καλύτερες υποδομές υγειονομικής περίθαλψης (υψηλή θετική συχέτιση μεταξύ "Hospital beds per 1000 people", "Medical doctors per 1000 people", "Median age" και "Population aged 65 and over (%)").
- Οι πλουσιότερες χώρες τείνουν να έχουν μεγαλύτερους πληθυσμούς και περισσότερους γιατρούς κατά κεφαλήν (θετική συχέτιση μεταξύ "GDP/Capita" , "Median age" (0.541829), "Population aged 65 and over (%)" (0.508596), και "Medical doctors per 1000 people" (0.523766))
- Οι μεγαλύτερες χώρες τείνουν να πραγματοποιούν περισσότερα τεστ και να αναφέρουν περισσότερα κρούσματα και θανάτους
- Οι χώρες με τους περισσότερους θανάτους είναι η Αμερική, η Ινδία και το Μεξικό με διαφορά. Οι πρώτες δύο είχαν περισσότερα κρούσματα από το Μεξικό.
- Οι περισσότερες χώρες έχουν χαμηλό GDP/ Capita.
- Εγγραφές λείπουν μόνο από τα πεδία “Daily tests” (7895) και “ Cases” (254) που δείχνει οτι λογικά τα συνολικά τεστ ήταν αρκετά δύσκολο να καταγραφούν σε σχέση με τα υπόλοιπα πεδία, όπως και τα “ Cases ” αλλά σε αρκετά μικρότερο βαθμό.

Ερώτημα 2

Χωρίσαμε τις χώρες σε συστάδες με βάση τις επιδόσεις τους στην αντιμετώπιση του ιού. Προς αυτόν τον σκοπό χρησιμοποιήσαμε τα εξής κριτήρια:

'Cases per Capita' = Συνολικά κρούσματα ανά τον πληθισμό ('Cases' / 'Population')
Αυτό μπορεί να δώσει μια ιδέα για τον συνολική μετάδοτικότητα της νόσου στη χώρα.

'Deaths per Capita' = Συνολικοί θάνατοι ανά τον πληθισμό ('Deaths' / 'Population')
Ένα υψηλό ποσοστό θνησιμότητας θα μπορούσε να υποδηλώνει ζητήματα όπως η έλλειψη πρόσβασης στην υγειονομική περίθαλψη ή οι καθυστερήσεις στη θεραπεία.

'Tests per Capita' = Συνολικά τεστ ανά τον πληθισμό ('Daily tests' / 'Population')
Αυτό μπορεί να δώσει μια ιδέα για το πόσο συνειδητοποιημένοι ήταν οι πολίτες στη χώρα ή πόσο η κάθε χώρα φρόντιζε να παρέχει αρκετά τεστ και τις κατάλληλες υποδομές για αυτά στους πολίτες της.

'Positivity Rate' = Συνολικά κρούσματα ανά ημερίσια τεστ ('Cases' / 'Daily tests')
Αυτό θα μπορούσε να υποδηλώνει ευρεία μετάδοση στην κοινότητα ή/και ανεπαρκείς δοκιμές.

'Death Rate' = Συνολικοί θάνατοι ανά τον συνολικά κρούσματα ('Deaths' / 'Cases')
Αυτό μπορεί να δώσει μια ιδέα για τον συνολική θνησιμότητα της νόσου στη χώρα.

Όπου τα 'Cases', 'Deaths' και 'Daily tests' είναι επί της ουσίας τα Total 'Cases', 'Deaths' και 'Daily tests' που προκύπτουν από το άθροισμα όλων των εγγραφών για αυτά τα features ανά χώρα. Να σημειωθεί ότι γι αυτό το ερώτημα έχουμε κάνει merge όλες τις εγγραφές της κάθε χώρας ξεχωριστά για να μπορούμε να τις μελετήσουμε μεμονωμένα.

Με βάση αυτά τα κριτήρια και την χρήση της silhouette score καταλήξαμε σε 4 κλάσεις που ήταν και ο αριθμός των κλάσων που μου εμφάνισε τις περισσότερες φορές η μέθοδος τις οποίες αργότερα διατήρησα σταθερές σε αριθμό.

Με βάση αυτήν την κατηγοριοποίηση οι επόμενες χώρες ανήκουν στις αντίστοιχες κλάσεις:

Cluster 0 countries:

['Australia' 'Austria' 'Bahrain' 'Belarus' 'Croatia' 'Cuba' 'Cyprus'

'Denmark' 'Estonia' 'Finland' 'Greece' 'Iceland' 'Israel' 'Japan'

'Kazakhstan' 'Kuwait' 'Latvia' 'Lithuania' 'Luxembourg' 'Malta'

'Mongolia' 'New Zealand' 'Norway' 'Poland' 'Portugal' 'Qatar' 'Russia'

'Saudi Arabia' 'Serbia' 'Slovakia' 'Slovenia' 'South Korea' 'Switzerland'

'Ukraine' 'United Arab Emirates' 'Uruguay']

Αυτό το cluster έχει χαμηλότερο μέσο όρο περιπτώσεων κατά κεφαλήν (4,16) και θανάτων κατά κεφαλήν (0,05) σε σύγκριση με το cluster 2.
Έχει υψηλότερο tests per capita (0,90), γεγονός που υποδηλώνει οτι έκαναν καλύτερη χρήση τεστ και πρόληψη.
Το positivity rate είναι πολύ χαμηλότερο (μέσος όρος: 8,72), υποδηλώνοντας καλύτερη διαχείριση του ιού.
Έχει επίσης το υψηλότερο κατά κεφαλήν ΑΕΠ (μέσος όρος: 33355,86) και τους πιο σημαντικούς πόρους υγειονομικής περίθαλψης (νοσοκομειακές κλίνες ανά 1000 άτομα κατά μέσο όρο: 5,07, γιατροί ανά 1000 άτομα κατά μέσο όρο: 3,42) μεταξύ όλων των ομάδων.
Χώρες όπως η Αυστραλία, η Αυστρία και η Ιαπωνία εμπίπτουν σε αυτό το σύμπλεγμα, αντανακλώντας μια καλά διαχειριζόμενη κατάσταση δεδομένων των πόρων τους.

Cluster 1 countries:

['Albania' 'Bangladesh' 'Bhutan' 'Cape Verde' 'Costa Rica'

'Dominican Republic' 'El Salvador' 'Ethiopia' 'Fiji' 'Ghana' 'Guatemala'

'India' 'Indonesia' 'Iraq' 'Jamaica' 'Jordan' 'Kenya' 'Libya' 'Malawi'

'Malaysia' 'Morocco' 'Mozambique' 'Myanmar' 'Namibia' 'Nepal' 'Nigeria'

'Pakistan' 'Paraguay' 'Philippines' 'Rwanda' 'Senegal' 'South Africa'

'Sri Lanka' 'Thailand' 'Togo' 'Trinidad and Tobago' 'Turkey' 'Uganda'

'Vietnam' 'Zambia' 'Zimbabwe']

Αυτό το cluster έχει τον χαμηλότερο μέσο όρο κρουσμάτων κατά κεφαλήν (1,19) και θανάτων κατά κεφαλήν (0,02).

Τα κατά κεφαλήν τεστ είναι επίσης χαμηλά σε αριθμό (μέσος όρος: 0,09), και το ποσοστό θετικότητας είναι σχετικά μέτριο (μέσος όρος: 15,23).

Το κατά κεφαλήν ΑΕΠ είναι χαμηλό (μέσος όρος: 4223,69), παρόμοιο με το Cluster 3, και οι πόροι υγειονομικής περίθαλψης είναι περιορισμένοι (νοσοκομειακές κλίνες ανά 1000 άτομα κατά μέσο όρο: 1,52, γιατροί ανά 1000 άτομα κατά μέσο όρο: 0,77).

Χώρες όπως η Ινδία, το Μπαγκλαντές και η Νιγηρία ανήκουν σε αυτό το σύμπλεγμα, υποδεικνύοντας μικρότερο αντίκτυπο του ιού, που μπορεί να οφείλεται σε χαμηλά τεστ, νεότερα δημογραφικά στοιχεία ή άλλους άγνωστους παράγοντες.

Cluster 2 countries:

['Argentina' 'Belgium' 'Bolivia' 'Bosnia and Herzegovina' 'Bulgaria'

'Canada' 'Chile' 'Colombia' 'Ecuador' 'France' 'Hungary' 'Iran' 'Ireland'

'Italy' 'Mexico' 'Panama' 'Peru' 'Romania' 'Sweden' 'United Kingdom'

'United States']

Αυτές οι χώρες χαρακτηρίζονται από σχετικά υψηλό αριθμό κρουσμάτων κατά κεφαλήν (μέσος όρος: 5,24) και θανάτων κατά κεφαλήν (μέσος όρος: 0,18).
Έχουν επίσης υψηλό positivity rate (μέσος όρος: 24,68) και GDP/Capita (μέσος όρος: 25190,82).
Οι χώρες σε αυτό το σύμπλεγμα έχουν αξιοπρεπή αριθμό νοσοκομειακών κλινών ανά 1000 άτομα (μέσος όρος: 3,37) και γιατρούς ανά 1000 άτομα (μέσος όρος: 2,49).
Χώρες όπως οι Ηνωμένες Πολιτείες, το Ηνωμένο Βασίλειο και η Γαλλία ανήκουν σε αυτό το σύμπλεγμα, γεγονός που υποδηλώνει σημαντικό αντίκτυπο του ιού, παρά το γεγονός ότι έχουν μια λογική υποδομή υγειονομικής περίθαλψης και το κατά κεφαλήν ΑΕΠ.

Cluster 3 countries:

['Armenia' 'Madagascar' 'Mauritania' 'Oman' 'Tunisia']

Αυτό το cluster έχει σχετικά χαμηλό αριθμό κρουσμάτων κατά κεφαλήν (μέσος όρος: 3,30) και θανάτους κατά κεφαλήν (μέσος όρος: 0,05).

Ωστόσο, έχει τον χαμηλότερο μέσο όρο τεστ κατά κεφαλήν (0,02) που υποδηλώνει πιθανή έλλειψη τεστ.

Το ποσοστό θετικότητας στον ιό είναι σημαντικά υψηλό (μέσος όρος: 143,60), γεγονός που υποδηλώνει ότι η κατάσταση μπορεί να είναι χειρότερη από ό,τι υποδηλώνει η ακατέργαστη καταμέτρηση των περιπτώσεων.

Οι χώρες σε αυτό το cluster έχουν το χαμηλότερο κατά κεφαλήν ΑΕΠ (μέσος όρος: 5097,22) και περιορισμένους πόρους υγειονομικής περίθαλψης (νοσοκομειακές κλίνες ανά 1000 άτομα κατά μέσο όρο: 1,74, γιατροί ανά 1000 άτομα κατά μέσο όρο: 1,24).

Χώρες όπως η Αρμενία, η Μαδαγασκάρη και η Τυνησία βρίσκονται σε αυτό το cluster, το οποίο θα μπορούσε να αντιμετώπιζε πρόβλημα λόγω των περιορισμένων πόρων τους.

Οι χώρες που ξεχώρισαν είναι οι:

<u>Category: Deaths per Capita</u>

Top 3 Countries:

Belgium 0.366624

United Kingdom 0.265057

Italy 0.259883

Bottom 3 Countries:

Vietnam 0.000073

Bhutan 0.000064

Mongolia 0.000037

Category: Tests per Capita

Top 3 Countries:

Luxembourg 3.488961

United Arab Emirates 3.218473

Denmark 2.828744

Bottom 3 Countries:

Vietnam 0.003493

Malawi 0.003061

Madagascar 0.001132

Category: Positivity Rate

Top 3 Countries:

Entity Positivity Rate

Armenia 244.539294

Oman 134.493037

Madagascar 132.563247

Bottom 3 Countries:

Entity Positivity Rate

New Zealand 0.412753

Bhutan 0.202590

Mongolia 0.128347

Category: Death Rate

Top 3 Countries:

Mexico 0.094959

Ecuador 0.068574

Bolivia 0.052753

Bottom 3 Countries:

Qatar 0.001580

Mongolia 0.000557

Bhutan 0.000505

Category: GDP/Capita

Top 3 Countries:

Luxembourg 114704.6

Switzerland 81993.7

Ireland 78661.0

Bottom 3 Countries:

Madagascar 523.4

Mozambique 503.6

Malawi 411.6

Category: Hospital beds per 1000 people

Top 3 Countries:

Japan 13.05

South Korea 12.27

Belarus 11.00

Bottom 3 Countries:

Senegal 0.3

Ethiopia 0.3

Madagascar 0.2

Category: Medical doctors per 1000 people

Top 3 Countries:

Cuba 7.52

Greece	6.26
Austria	5.23

Bottom 3 Countries:

Togo	0.05
Ethiopia	0.02
Malawi	0.02

Οι περισσότερες από αυτές τις χώρες που ξεχωρίζουν σε κάθε κατηγορία ανήκουν στο ίδιο cluster ή στο αμέσως επόμενο cluster που έχει τα πιο κοινά χαρακτηριστικά με αυτές.

Ερώτημα 3

Έχοντας υλοποιήσει πλέν και τα δύο μοντέλα και χρησιμοποιόντας το MSE σαν μετρική σύγκρισης, βγάλαμε τα ακόλουθα αποτελέσματα:

SVM MSE: 8.944866335722077 και RNN MSE: 8.299623151710382

Σε αυτή την περίπτωση, το μοντέλο RNN είναι πιο περίπλοκο και χρειάζεται περισσότερο χρόνο για να εκπαιδευτεί από το μοντέλο SVM.
Παρ' όλα αυτά το RNN μοντέλο έβγαλε καλύτερα αποτελέσματα επειδή έχει μικρότερο τετραγωνικό σφάλμα.

Δεδομένου αυτού του σεναρίου, θα έκλινα προς την επιλογή του μοντέλου Recurrent Neural Network (RNN) για τους ακόλουθους λόγους:

Χρονική εξάρτηση: Δεδομένης της φύσης του προβλήματός σας, όπου προσπαθούμε να προβλέψουμε μια μελλοντική τιμή με βάση προηγούμενες τιμές, τα RNN είναι συνήθως πιο κατάλληλα καθώς έχουν σχεδιαστεί για να χειρίζονται διαδοχικά δεδομένα και μπορούν να καταγράφουν χρονικές δυναμικές, κάτι που ένα μοντέλο όπως το SVM δεν μπορεί.

Απόδοση: Με βάση τις τιμές MSE που παρέχονται, το μοντέλο RNN αποδίδει ελαφρώς καλύτερα από το μοντέλο SVM. Αν και η διαφορά μπορεί να φαίνεται μικρή, σε πραγματικό περιβάλλον, ακόμη και μια μικρή βελτίωση στην ακρίβεια πρόβλεψης μπορεί να κάνει σημαντική διαφορά.

Πολυπλοκότητα και χρόνος εκπαίδευσης: Ενώ τα RNN είναι συνήθως πιο περίπλοκα και χρειάζονται περισσότερο χρόνο για να εκπαιδευτούν από τα SVM, η διαφορά στους χρόνους εκπαίδευσης μπορεί να μην είναι τόσο μεγάλη εάν το σύνολο δεδομένων δεν είναι πολύ μεγάλο. Στην συγκεκριμένη περίπτωση ήταν αισθητό αλλά όχι πολύ αργό.