Εξόρυξη Γνώσης από Δεδομένα

Εργασία

Άνοιξε: Δευτέρα, 31 Ιανουαρίου 2022, 12:00 πμ
Λήγει: Κυριακή, 27 Φεβρουαρίου 2022, 11:59 μμ

Εξόρυξη Γνώσης από Δεδομένα 2021-2022. Εργασία (ατομική)

Χρονικός προγραμματισμός

- ΕΚΦΩΝΗΣΗ: 31 Ιανουαρίου 2022
- DUE DATE: 27 Φεβρουαρίου 2022, 23:59

Α. Εισαγωγή

Στο πλαίσιο της εργασίας του μαθήματος, καλείστε να αναλύσετε μια μεγάλη συλλογή από δεδομένα για την πανδημία του κορονοϊού SARS-Cov-2, που αφορούν όλο τον κόσμο και για συγκεκριμένο χρονικό διάστημα, το οποίο καθορίζεται από το τελευταίο ψηφίο του αριθμού μητρώου σας, σύμφωνα με τον παρακάτω πίνακα

Τελευταίο ψηφίο Α.Μ.	**0**	**1**	**2**	**3**	**4**
Χρονικό διάστημα	Μάιος - Ιούν 2020	Ιούλ - Αύγ 2020	Σεπτ - Οκτ 2020	Νοε - Δεκ 2020	Ιαν - Φεβρ 2021
Τελευταίο ψηφίο Α.Μ.	**5**	**6**	**7**	**8**	**9**
Χρονικό διάστημα	Μαρτ - Απρ 2021	Μάι - Ιουν 2021	Ιουλ - Αυγ 2021	Σεπτ - Οκτ 2021	Νοε - Δεκ 2021

Τα δεδομένα προέρχονται από τον βρετανικό μη-κερδοσκοπικό οργανισμό Our World in Data και είναι διαθέσιμα μέσω του παρακάτω github repository

https://github.com/owid/covid-19-data/blob/master/public/data/README.md

Και πιο συγκεκριμένα μέσω του ακόλουθου συνδέσμου (CSV αρχείο, περίπου 50 MB)

https://covid.ourworldindata.org/data/owid-covid-data.csv

Σκοπός της εργασίας είναι να απαντήσετε στα ακόλουθα ερωτήματα, για το χρονικό διάστημα που σας αντιστοιχεί

Β. 1ο Ερώτημα: Επισκόπηση των Δεδομένων

Σκοπός αυτού του ερωτήματος είναι να μελετήσετε και να κατανοήσετε τα χαρακτηριστικά του συνόλου δεδομένων σας. Πιο συγκεκριμένα, θα πρέπει να απαντήσετε στα παρακάτω ερωτήματα:

1. Ποιος είναι ο αριθμός των δειγμάτων & των χαρακτηριστικών του συνόλου δεδομένων;
2. Ποια είναι τα είδη των χαρακτηριστικών του συνόλου δεδομένων;
3. Υπάρχουν μη διατεταγμένα χαρακτηριστικά και ποια είναι αυτά;
4. Υπάρχουν απουσιάζουσες τιμές; Ταξινομήστε και παρουσιάστε, κατά φθίνουσα σειρά τα χαρακτηριστικά που έχουν απουσιάζουσες τιμές.

Γ. 2ο Ερώτημα: Προεπεξεργασία του συνόλου δεδομένων

Σε συνέχεια του πρώτου ερωτήματος, αναφέρετε αν παρατηρήσατε τα παρακάτω φαινόμενα και σε περίπτωση θετικής απάντηση, πως τα αντιμετωπίσατε

1. Τις απουσιάζουσες τιμές
2. Τις έκτοπες τιμές (outliers)
3. Τις διπλοεγγραφές (αν υπήρχαν)

Δ. 3ο Ερώτημα: Διερευνητική Ανάλυση Δεδομένων

Κατόπιν, θα πρέπει να απαντήσετε στα ακόλουθα υποερωτήματα για τη χρονική περίοδο που μελετάτε. Όταν το χαρακτηριστικό που ζητείται δεν προσδιορίζεται με απόλυτη ακρίβεια, επιλέξτε αυτό που εσείς θεωρείτε πιο κατάλληλο και τεκμηριώστε συνοπτικά την επιλογή σας.

1. Ποιες είναι οι δέκα πρώτες χώρες ως προς το ποσοστό θνησιμότητας και πως εξελίχθηκε χρονικά η σχετική κατάταξη
2. Ποιες είναι οι δέκα πρώτες χώρες ως προς το κρούσματα ανά εκατομμύριο και πως εξελίχθηκε χρονικά η σχετική κατάταξη
3. Ποιες είναι οι δέκα πρώτες χώρες ως προς τους θανάτους ανά εκατομμύριο και πως εξελίχθηκε χρονικά η σχετική κατάταξη
4. Ποιο είναι το καθημερινό ποσοστό θετικότητας ανά χώρα
5. Να σχεδιάσετε τις καμπύλες των νοσηλευόμενων ασθενών καθώς και των διασωληνωμένων σε Μονάδες Εντατικής Θεραπείας (intensive care units - ICUs)
6. Να κατασκευάσετε γεωγραφικό θερμοχάρτη (heatmap) του απόλυτου αριθμού κρουσμάτων
7. Να εντοπίσετε παραδείγματα γειτονικών, μεταξύ τους, χωρών, που παρουσιάζουν μέγιστη συσχέτιση υπερβάλλουσας θνησιμότητας (excess mortality)
8. Εκτιμήστε την εξάπλωση της λοίμωξης μέσω του ρυθμού αναπαραγωγής (reproduction rate) σε όλες τις Ηπείρους πλην Ανταρκτικής
9. Εμφανίστε τους καθημερινούς διαγνωστικούς ελέγχους ανά χώρα.
10. Να μελετήσετε τη συσχέτιση μεταξύ του πλήθους των ελέγχων για κορονοϊό και της υπερβάλλουσας θνησιμότητας και να εμφανίσετε τις χώρες με τη μεγαλύτερη και τη μικρότερη συσχέτιση
11. Συσταδοποιείστε τις χώρες σε ομάδες ως προς τον αριθμό κρουσμάτων ανά εκατομμύριο και ως προς της υπερβάλλουσας θνησιμότητας ανά εκατομμύριο. Αναλύστε συνοπτικά τα ποιοτικά χαρακτηριστικά των συστάδων που λαμβάνετε.
12. Να υπολογίσετε τη συσχέτιση μεταξύ πορείας εμβολιασμών (αν υπάρχουν την χρονική περίοδο που μελετάτε) και της υπερβάλλουσας θνησιμότητας και να εμφανίσετε τις χώρες με τη μεγαλύτερη και τη μικρότερη συσχέτιση
13. Καθορίστε ποια είναι τα χαρακτηριστικά των χωρών με τη μεγαλύτερη και τη μικρότερη υπερβάλλουσα θνησιμότητα λόγω Covid, όσον αφορά τη γενική υγεία του πληθυσμού τους (πχ κάπνισμα, διαβήτης, καρδιοπάθειες κλπ)
14. Να εξετάσετε την πορεία εξάπλωσης του ιού ως προς τον δείκτη ανθρώπινης ανάπτυξης (human development index) και να σχηματίσετε συστάδες χωρών με παρόμοια χαρακτηριστικά. Αναλύστε συνοπτικά τα ποιοτικά χαρακτηριστικά των συστάδων που λαμβάνετε.

Ε. 4ο Ερώτημα (bonus): Εξόρυξη γνώσης

Στο συγκεκριμένο ερώτημα καλείστε να αναπτύξετε τα συμπεράσματα που προέκυψαν από τη φάση της διερευνητικής ανάλυσης δεδομένων (όχι υποχρεωτικά από την εξέταση των παραπάνω ερωτήσεων), τα οποία εσείς τα κρίνετε ενδιαφέροντα, και τα οποία οδηγούν στην εξαγωγή νέας γνώσης που δεν είναι ούτε τετριμμένη ούτε εύκολο να φανεί εκ των προτέρων. Αν το κρίνετε σκόπιμο, και προς επίρρωση των ισχυρισμών σας, μπορείτε να εμπλουτίσετε τη συλλογή δεδομένων που σας δίνουμε με δημόσια διαθέσιμες πηγές δεδομένων, αναφέροντας την πηγή τους. Σε κάθε περίπτωση θα πρέπει να τεκμηριώσετε επαρκώς την ανάλυσή σας. Η απάντηση σε αυτό το ερώτημα είναι προεραιτική και δίνει ένα μικρό βαθμολογικό bonus.

ΣΤ. Οδηγίες εκπόνησης της εργασίας και παράδοσης της αναφοράς

Θα εργαστείτε σε περιβάλλον Jupyter Notebook με τη γλώσσα προγραμματισμού Python και το PySpark. Το περιβάλλον ανάπτυξης μπορεί να είναι είτε στην εικονική σας μηχανή στον Ωκεανό, ή τοπικά στον υπολογιστή σας ή σε οποιαδήποτε άλλη σχετική υπηρεσία (πχ Google Colaboratory). Θα παραδώσετε το Jupyter notebook εντός του οποίου θα υπάρχει το ονοματεπώνυμό σας, ο Αριθμός Μητρώου σας, η ιδιότητάς σας (ΥΔ, ΕΔΕΜΜ, κλπ) καθώς και το email σας. Στο notebook θα υπάρχουν επίσης οι ερωτήσεις, ο κώδικας σας ως απάντηση και ένας σύντομος σχολιασμός της κάθε απάντησης. Το notebook θα πρέπει να έχει ήδη τρέξει στο δικό σας περιβάλλον και να φαίνεται η απάντηση. Επίσης, το αρχείο των δεδομένων θα πρέπει να το αντιμετωπίσετε με ενιαίο τρόπο (δλδ δεν επιτρέπεται να το “τεμαχίσετε” για να εξάγετε τις ημερομηνίες που αντιστοιχούν στον ΑΜ σας πριν ξεκινήσετε να απαντάτε στα ερωτήματα). Να τονίσουμε επίσης ότι η χρήση άλλων βιβλιοθηκών της python επιτρέπεται εντελώς επικουρικά (πχ matplotlib, numpy), μιας και οι κύριοι υπολογισμοί θα πρέπει να γίνουν στο περιβάλλον του PySpark. Η χρήση άλλων αντίστοιχων βιβλιοθηκών (πχ pandas) δεν επιτρέπεται. Τέλος στον παρακάτω σύνδεσμο, ο οποίος θα ανανεώνεται συνεχώς μέχρι και την προθεσμία υποβολής, μπορείτε να βρίσκετε απαντήσεις σε συχνά ερωτήματα.

https://docs.google.com/document/d/1PbkljVp7VNzhkNvEfgy-Uqrs_1CQ9rs0h-Smkm9cNBY/edit?usp=sharing

Κατάσταση Υποβολής