Τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής
Πολυτεχνική Σχολή - Πανεπιστήμιο Ιωαννίνων

Εξόρυξη Δεδομένων *(MYE012)*

Πρώτη Σειρά Ασκήσεων

Συγγραφείς:

Χουλιάρας Ιωάννης[1] and **Αλεξίου Αλέξανδρος**[2]

[1]**ΑΜ: 2631** - Email: cs02631@uoi.gr
[2]**ΑΜ: 2929** - Email: cs02929@uoi.gr

Δεκέμβριος, 2020

Πληροφορίες

Για την παράδοση της άσκησης χρησιμοποιήθηκε ένα free pass από τα τέσσερα διαθέσιμα.

Ερώτηση 1

Υποερώτημα Α.1

Περιγραφή του τροποποιημένου αλγορίθμου Reservoir Sampling που διαλέγει ένα ομοιόμορφο δείγμα K αντικειμένων από ένα ρεύμα N αντικειμένων.

Έστω ότι έχουμε ένα ρεύμα με N στοιχεία. Η δομή μας (reservoir) θα αρχικοποιηθεί με τα πρώτα K στοιχεία από το ρεύμα. Έστω ότι επεξεργαζόμαστε το i-οστό στοιχείο του ρεύματος με i > K. Η πιθανότητα δειγματοληψίας κάθε στοιχείου από μία ομοιόμορφη κατανομή εάν θέλουμε να επιλέξουμε K στοιχεία, είναι $\frac{K}{i}$. Αυτή είναι και η πιθανότητα το i-οστό στοιχείο να μπει τελικά στην δομή μας. Εάν τελικά μπει, θα αντικαταστήσει τυχαία ένα υπάρχον στοιχείο της δομής με όλα τα στοιχεία να έχουν ίση πιθανότητα $\frac{1}{K}$ να αντικατασταθούν.

Υποερώτημα Α.2

Απόδειξη ότι ο αλγόριθμός παράγει ένα ομοιόμορφα τυχαίο δείγμα, δηλαδή, για κάθε i, $1 \leq i \leq N$, το i-οστό στοιχείο έχει πιθανότητα $\frac{K}{N}$ να εμφανιστεί στο δείγμα.

Απόδειξη:

- Υποθέτουμε ότι τα πρώτα i στοιχεία έχουν επιλεγεί με πιθανότητα $\frac{K}{i}$
- Το $i+1$ στοιχείο επιλέγεται με πιθανότητα $\frac{K}{i+1}$
- Αν τελικά επιλεγεί, κάθε στοιχείο στο reservoir έχει $\frac{1}{K}$ πιθανότητα να αλλαχθεί. Άρα η πιθανότητα ένα στοιχείο στο reservoir να αντικατασταθεί από το i+1 στοιχείο είναι: $P(A) = \frac{1}{K} \cdot \frac{K}{i+1} = \frac{1}{i+1}$
  Ενώ η πιθανότητα να μην αλλαχθεί: $P(A') = 1 - P(A) = 1 - \frac{1}{i+1} = \frac{i}{i+1}$

- Η πιθανότητα ότι ένα στοιχείο από το ρεύμα βρίσκεται στο τελικό reservoir είναι:

$$P(e) = \frac{K}{i} \cdot \left(\frac{i}{i+1}\right) \cdot \left(\frac{i+1}{i+2}\right) \cdot \ldots \cdot \left(\frac{N-1}{N}\right)$$

  Το $\frac{K}{i}$ είναι η πιθανότητα που έχουν επιλεγεί τα πρώτα i στοιχεία. Το $\frac{i}{i+1}$ είναι η πιθανότητα να μην αλλαχθεί το i+1. Το $\frac{i+1}{i+2}$ η πιθανότητα να μην αλλαχθεί το i+2 και ούτω καθεξής. Εάν απλοποιηθούν οι όροι i, i+1, i+2 ... N-1 καταλήγουμε στο:

$$P(e) = \frac{K}{N}$$

Υποερώτημα Α.3 (Python)

```
import sys
import random

K = int(sys.argv[1])
infile_name = sys.argv[2]
reservoir = [""]*K

with open(infile_name) as infile:
    for i, line in enumerate(infile):
        if i < K:
            reservoir[i] = line
            continue
        # pick a random number between 0 and i
        # with probability 1/i for each object to be chosen
        j = random.randrange(i)
        # If j < K then the object at index j
        # will be replaced with the line with probability k/i to be replaced
        if j < K:
            reservoir[j] = line

print(*reservoir, sep="\n")
```

Το πρόγραμμα τρέχει με time complexity O(n) καθώς διατρέχει μια φορά τα δεδομένα και space complexity O(K) καθώς αποθηκεύει μόνο τον πίνακα reservoir μεγέθους K.

Εκτέλεση

```
./sample.py 10 input.txt
```

Όπου 10 το μέγεθος του δείγματος και input.txt το αρχείο εισόδου με τα N αντικείμενα.

Περιγραφή:

Αρχικά ο αλγόριθμος αρχικοποιεί τον πίνακα reservoir με τα πρώτα K αντικείμενα της ροής (0..K-1).
Έπειτα από i=K επιλέγει έναν αριθμό j από το 0 έως i(όπου i ο αριθμός της επόμενης γραμμής του αρχείο ροής) με πιθανότητα K/i και αν αυτός ο αριθμός είναι μικρότερος του K τότε το στοιχείο στην θέση j θα αντικατασταθεί από το στοιχείο i της ροής. Η διαδικασία επαναλαμβάνεται μέχρι το i=N-1.

Υποερώτημα Β

Απόδειξη ορθότητας αλγορίθμου Reservoir Sampling με σταθμισμένη δειγματοληψία.

Περιγραφή σταθμισμένης δειγματοληψίας:

Στην σταθμισμένη δειγματοληψία, κάθε αντικείμενο έχει πιθανότητα να επιλεγεί ανάλογη με το βάρος της. Το βάρος του αντικειμένου i είναι w_i και το συνολικό βάρος των N αντικειμένων είναι W. Τότε η πιθανότητα επιλογής του i αντικειμένου είναι $\frac{w_i}{W}$.

Απόδειξη

Για το πρώτο αντικείμενο (N=1) η πιθανότητα να επιλεχθεί είναι $\frac{w_1}{w_1} = 1$ διότι το πρώτο αντικείμενο θα επιλεχθεί για το reservoir όπως και στον κλασικό αλγόριθμο. Για το δεύτερο αντικείμενο(N=2) η πιθανότητα να επιλεχθεί είναι $\frac{w_2}{w_1+w_2}$ καθώς στο συνολικό βάρος θα προστεθεί και το βάρος του νέου αντικειμένου.

Εφόσον έχουμε πλέον δύο αντικείμενα (N=2) και έχουμε υπολογίσει την πιθανότητα να επιλεχθεί το δεύτερο αντικείμενο, μπορούμε να υπολογίσουμε και την πιθανότητα να μην επιλεχθεί:

$$1 - \frac{w_2}{w_1 + w_2}$$

Μπορούμε να υπολογίσουμε ποια είναι η πιθανότητα να παραμείνει το πρώτο αντικείμενο στο reservoir έπειτα από την έλευση του δεύτερου αντικειμένου. Πολλαπλασιάζουμε την πιθανότητα να επιλεγεί το πρώτο αντικείμενο με την πιθανότητα να μην επιλεγεί το δεύτερο.

$$\frac{w_1}{w_1} \cdot \left(1 - \frac{w_2}{w_1 + w_2}\right) = \frac{w_1}{w_1} \cdot \left(\frac{w_1}{w_1 + w_2}\right)$$

Με τις κατάλληλες απλοποιήσεις καταλήγουμε ότι η πιθανότητα να παραμείνει στο reservoir το w_1 με N=2 αντικείμενα είναι $\frac{w_1}{W}$ όπου ο παρανομαστής είναι το συνολικό βάρος των αντικειμένων $W = w_1 + w_2$.

Με άγνωστο N αποδεικνύεται ότι η πιθανότητα του i-οστού στοιχείου να επιλεχθεί στο τελικό reservoir είναι $\frac{w_i}{W}$:

$$\frac{w_1}{w_1} \cdot \left(\frac{w_1}{w_1+w_2}\right) \cdot \left(\frac{w_1+w_2}{w_1+w_2+w_3}\right) \cdot \ ... \ \cdot \left(\frac{w_1 \ +...+w_{n-1}}{w_1+...+w_n}\right) = \frac{w_1}{W} \ , \ i=1$$

Τέλος, για N αντικείμενα στο ρεύμα, το άθροισμα βαρών των αντικειμένων προς το συνολικό άθροισμα θα πρέπει να είναι:

$$\sum_{i=1}^{N} \frac{w_i}{\sum_{j=1}^{N} w_j} \ = \ 1$$

Αλγόριθμος

Ο αλγόριθμος ξεκινάει αρχικοποιώντας το πρώτο στοιχείο της ροής ως επιλεχθέν και κρατάει το βάρος του w_0 με πιθανότητα $\frac{w_0}{w_0} \ = \ 1$ (Το πρώτο στοιχείο πάντα επιλέγεται). Όταν θα έρθει το επόμενο w_1 θα κρατήσει το συνολικό βάρος των αντικειμένων μέχρι στιγμής $w_0 + w_1$ και θα υπολογίσει μια πιθανότητα για να αποφασίσει αν θα αλλάξει το w_0 με το w_1 ($\frac{w_1}{w_0+w_1}$ να επιλεγεί το w_1). Η διαδικασία επαναλαμβάνεται μέχρι να έρθει και το αντικείμενο w_n.

Χρόνος

Time complexity: O(N) - διατρέχει μία φορά τα δεδομένα.

Space complexity: O(1) σταθερός χρόνος(2 int μεταβλητές στην μνήμη) και το αρχείο δεν αποθηκεύεται στην μνήμη ολόκληρο.

Ερώτηση 3

1. *Πρόβλεψη βαθμολογίας για την ταινία 6 από τον χρήστη X.*

Figure 1: Ratings

	Movie 1	Movie 2	Movie 3	Movie 4	Movie 5	Movie 6
User X	5	3		1		
User Y	4	2	1			1
User Z	3			1	3	3
User W	2	5	1	5	3	4

Αρχικά υπολογίζουμε το Cosine Similarity του χρήστη X με όλους τους άλλους χρήστες.

$$\cos\left(a,\ b\right) = \frac{a \cdot b}{|a| \cdot |b|}$$

$$\cos\left(X,\ Y\right) = 0.937$$

$$\cos\left(X,\ Z\right) = 0.511$$

$$\cos\left(X,\ W\right) = 0.566$$

Σύμφωνα με τους υπολογισμούς ο χρήστης X μοιάζει περισσότερο με τον Y και W. Άρα ο σταθμισμένος μέσος όρος της βαθμολογίας των 2 πιο όμοιων χρηστών στον X είναι

$$X_6 = \frac{1}{\sum_{i=1}^{2}\ W_i} \cdot \sum_{i=1}^{2} w_i \cdot\ \ r_i\ \ = 2,13$$

2. *Πρόβλεψη βαθμολογίας για την ταινία 6 από τον χρήστη X κανονικοποιώντας τα δεδομένα.*

Αφαιρώντας την μέση τιμή βαθμολογιών από κάθε γραμμή υπολογίζουμε τον κανονικοποιημένο πίνακα.

Figure 2: Normalized Ratings

	Movie 1	Movie 2	Movie 3	Movie 4	Movie 5	Movie 6
User X	6/3	0		-6/3		
User Y	8/4	0	-4/4			-4/4
User Z	2/4			-6/4	2/4	2/4
User W	-4/3	5/3	-7/3	5/3	-1/3	2/3

Υπολογίζουμε ξανά το Cosine Similarity του χρήστη X με όλους τους άλλους χρήστες.

$$\cos\left(X,\ Y\right) = 0.632$$

$$\cos(X,\ Z) = 0.353$$

$$\cos(X,\ W) = -0,580$$

Υπολογίζουμε την απόκλιση από τον μέσο όρο για το κελί X_6 ως τον σταθμισμένο μέσο όρο των 2 πιο όμοιων χρηστών με τον X και την προσθέτουμε στον μέσο όρο βαθμολογιών του X.

$$X_6 = \overline{\mathrm{X}} + \frac{\sum_{\ v \in TopK(\mathrm{X})} \cos\ (\mathrm{X}, v)\ \cdot \left(r_{v_i} - \overline{r_v}\right)}{\sum_{\ v \in\ TopK(\mathrm{X})} \cos\ (\mathrm{X}, v)} = 2,537$$

3. *Παρατηρήσεις.*

Στην πρώτη εκδοχή κάνουμε την παραδοχή ότι οι τιμές που λείπουν είναι μηδενικές. Κατά συνέπεια θεωρούνται αρνητικές βαθμολογίες και τα similarities δείχνουν ότι ο χρήστης X μοιάζει με τον W περισσότερο από τον Z. Κοιτώντας τα δεδομένα βλέπουμε ότι είναι αρκετά διαφορετικοί στο πως βαθμολογούν στην πραγματικότητα. Τα κελιά τα οποία είναι πιο σημαντικά για τον υπολογισμό είναι τα Y,6 και W,6. Η βαθμολογία που υπολογίστηκε για την ταινία 6 από τον χρήστη X είναι 2,13.

Στην δεύτερη εκδοχή κάνουμε την κανονικοποίηση στα δεδομένα για να λύσουμε το πρόβλημα των μηδενικών τιμών. Ακόμα διαχειριζόμαστε τις τιμές που λείπουν ως μηδενικές ωστόσο πλέον κεντρικοποιήσαμε τις βαθμολογίες κάθε χρήστη γύρω από το μηδέν έτσι ώστε η μέση βαθμολογία πλέον είναι η βαθμολογία 0. Άρα θετική βαθμολογία σημαίνει ότι στον χρήστη άρεσε η ταινία περισσότερο από τον μέσο όρο και αρνητική βαθμολογία σημαίνει ότι στον χρήστη άρεσε η ταινία λιγότερο από τον μέσο όρο. Η βαθμολογία τώρα για την ταινία 6 από τον χρήστη X είναι 2,53.

Ερώτηση 4

Σε αυτή την ερώτηση μελετάμε την πανδημία του Covid-19. Θα προσπαθήσουμε να βρούμε τυχόν συσχετίσεις μεταξύ των αριθμών των κρουσμάτων και θυμάτων με χαρακτηριστικά των περιοχών. Για τις ερωτήσεις Α, Β και C κατεβάσαμε τα δεδομένα από το εδώ. Ενώ για το τελευταίο ερώτημα θα εστιάσουμε στις Ηνωμένες Πολιτείες. Κατεβάσαμε από εδώ τα δεδομένα για την τρέχουσα κατάσταση της πανδημίας ανά πολιτεία καθώς και τα αποτελέσματα από τις πρόσφατε εκλογές, όπου για κάθε πολιτεία έχουμε την πληροφορία αν εξέλεξαν Δημοκρατικούς ή Ρεπουμπλικανούς. Θα εξετάσουμε εάν υπάρχει στατιστικά ενδιαφέρουσα διαφορά στον βαθμό της έξαρσης της πανδημίας ανάλογα με το τι ψηφίζει η πολιτεία.

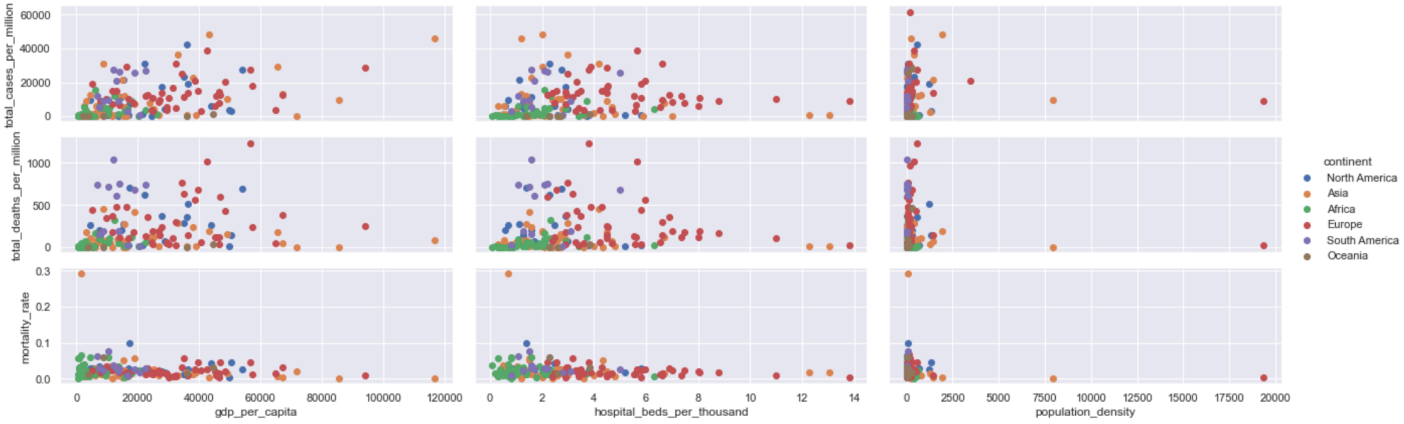
Υποερώτημα Α

Αφού φορτώσουμε σε ένα Pandas DataFrame τα δεδομένα μας, παρατηρούμε ότι υπάρχουν 58698 εγγραφές και 50 στήλες. Στο συγκεκριμένο ερώτημα θα κοιτάξουμε σε μία συγκεκριμένη χρονική στιμή, την 1/11/2020 και θα ασχοληθούμε με τους δείκτες (στήλες) : τον συνολικό αριθμό των κρουσμάτων ανά εκατομμύριο (total_cases_per_million), τον συνολικό αριθμό των θανάτων ανά εκατομμύριο (total_deaths_per_million) και το ποσοστό θνησιμότητας που είναι ο λόγος των δύο δεικτών.

Scatter plots και Pearson Correlation Coefficient

Το πρώτο που θα εξετάσουμε είναι αν υπάρχει συσχέτιση μεταξύ των παραπάνω δεικτών και χαρακτηριστικά των διαφορετικών χωρών. Τα τρία χαρακτηριστικά που θα εστιάσουμε είναι: ΑΕΠ (gdp_per_capita), αριθμό κλινών (hospital_beds_per_thousand) και πυκνότητα πληθυσμού (population_density).
Ο πρώτος τρόπος που θα μελετήσουμε τα δεδομένα μας είναι μέσω οπτικοποίησης. Φτιάχνουμε scatter plots για κάθε ζευγάρι δείκτη και χαρακτηριστικού.

Figure 3: Scatter plot για κάθε ζευγάρι δείκτη-χαρακτηριστικό. Τα δεδομένα μας δεν έχουν τροποποιηθεί

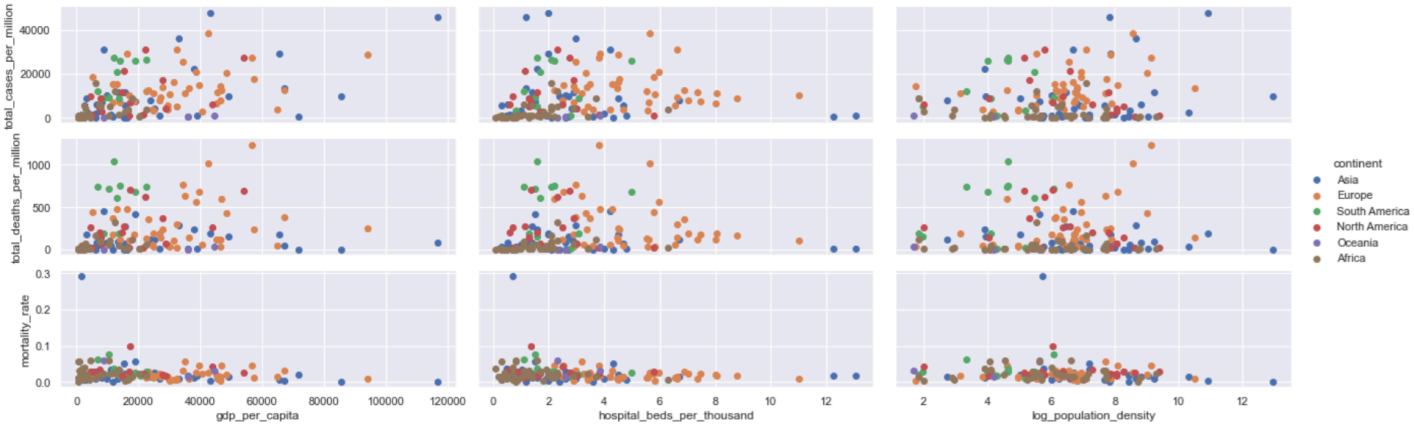


Παρατηρούμε ότι οι Ήπειροι που έχουν χαμηλό GDP έχουν και χαμηλό αριθμό από κλίνες και το αντίστροφο, ωστόσο για την Βόρεια Αμερική βλέπουμε ότι ενώ έχει αυξημένο GDP δεν έχει τόσες κλίνες. Η Αφρική έχει μεγαλύτερο ποσοστό θνησιμότητας και πολύ λιγότερες κλίνες σε σχέση με την Ευρώπη η οποία έχει από τα μικρότερα ποσοστά

θνησιμότητας. Το ίδιο παρατηρούμε και στα δεδομένα της Νότιας Αμερικής συγκρίνοντας τα με την Ευρώπη. Στα δεδομένα της Ασίας παρατηρούμε ότι εκεί που υπάρχει έξαρση της πανδημίας αντιστοιχεί χαμηλότερος αριθμός κλινών σε σχέση με τον πληθυσμό. Ενδιαφέρον είναι ότι η Ευρώπη παρουσιάζει διακύμανση στις μετρήσεις που εξετάζονται. Βλέπουμε ότι το ποσοστό θνησιμότητας των περισσότερων χωρών είναι αρκετά χαμηλό. Η Ωκεανία βλέπουμε ότι δεν παρουσιάζει έξαρση στην πανδημία, έχουμε λίγα κρούσματα και λίγους θανάτους. Στο γράφημα 5 παρατηρούμε ότι για την πυκνότητα πληθυσμού τα δεδομένα είναι όλα συγκεντρωμένα στην αρχή του άξονα x και είναι δύσκολο να βγάλει κανείς συμπεράσματα. Στην συνέχεια θα προχωρήσουμε σε λογαρίθμιση του άξονα.

Υπολογισμός Pearson Correlation Coefficient Θα καθαρίσουμε τα δεδομένα μας από null τιμές και θα υπολογίσουμε το Peason Correlation Coefficient και τα p-values. Υπολογίζουμε την λογαριθμική τιμή της πυκνότητας του πληθυσμού ώστε να βγάλουμε κάποιο συμπέρασμα.

Figure 4: Pearson Correlation Coefficient, p-values και λογαριθμική τιμή της πυκνότητας πληθυσμού



Με την λογαρίθμιση του άξονα στο γράφημα 4 με τα συνολικά κρούσματα ανά πυκνότητα πληθυσμού παρατηρούμε μια αύξηση ως προς τα κρούσματα όσο αυξάνεται η πυκνότητα, στην Ευρώπη και στην Ασία. Επίσης στο γράφημα των συνολικών θανάτων παρατηρούμε μια αύξηση των θανάτων σε σχέση με την αύξηση της πυκνότητας του πληθυσμού. Ωστόσο στο γράφημα του ποσοστού θνησιμότητας βλέπουμε ότι δεν εμφανίζεται κάποια συσχέτιση μεταξύ πυκνότητας πληθυσμού και ποσοστού θνησιμότητας.

Figure 5: PCC και p-values για κάθε δείκτη-χαρακτηριστικό για όλες τις Ηπείρους

	gdp_per_capita		hospital_beds_per_thousand		population_density	
indicators	PCC	p-value	PCC	p-value	PCC	p-value
total_cases_per_million	0,498	$0.015*10^{-8}$	0,152	0,066	0,084	0,311
total_deaths_per_million	0,216	0,009	0,080	0,336	-0,065	0,435
mortality_rate	-0,167	0,044	-0,171	0,039	-0,111	0,181

Στατιστική σημασία

Αποδεχόμαστε τα p-values τα οποία είναι μικρότερα από το 0.05. Μέσω του PCC φαίνεται ότι με την αύξηση του gdp παρατηρούνται και περισσότερα κρούσματα και περισσότεροι θάνατοι αλλά παρατηρείται ότι το ποσοστό θνησιμότητας μειώνεται. Όσων αφορά το χαρακτηριστικό hospital beds η μόνη στατιστικά σημαντική συσχέτιση

παρατηρείται στο mortality rate. Αποδεχόμαστε ότι υπάρχει στατιστικά σημαντική συσχέτιση (p-value < 0.05) μεταξύ του αριθμού των κλινών και του ποσοστού θνησιμότητας και συγκεκριμένα φαίνεται ότι η σχέση αυτή είναι αντιστρόφως ανάλογη σε έναν βαθμό δηλαδή όσο χαμηλότερος είναι ο αριθμός των κλινών τόσο αυξάνεται το ποσοστό θνησιμότητας κάτι το οποίο είναι αναμενόμενο (κορεσμός συστήματος υγείας). Το population density δεν δείχνει ούτε εδώ κάποια σχέση με κάποιον δείκτη όπως παρατηρήθηκε αρχικά στα scatter plots.

Αφαίρεση Αφρικής

Με την σύγκριση των scatter plots πριν και μετά την αφαίρεση της Αφρικής μπορούμε να παρατηρήσουμε ότι οι χώρες της παρουσιάζουν μια ομοιογένεια σε όλους τους δείκτες σε σχέση με τις άλλες ηπείρους όπου οι διάφορες χώρες παρουσιάζουν μια ποικιλομορφία.

Figure 6: PCC, p-values και λογαριθμική τιμή της πυκνότητας πληθυσμού χωρίς τις χώρες της Αφρικής

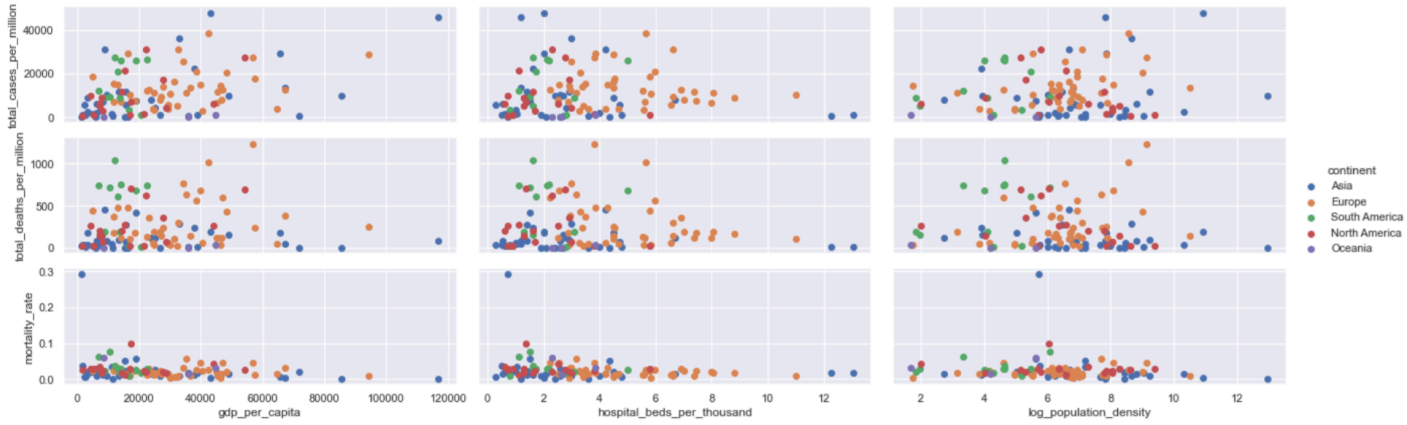


Figure 7: PCC και p-values για κάθε δείκτη-χαρακτηριστικό για όλες τις Ηπείρους χωρίς την Αφρική

	gdp_per_capita		hospital_beds_per_thousand		population_density	
indicators	PCC	p-value	PCC	p-value	PCC	p-value
total_cases_per_million	0,398	$0.017*10^{-3}$	-0,009	0,924	0,059	0,540
total_deaths_per_million	0,073	0,446	-0,066	0,492	-0,100	0,296
mortality_rate	-0,207	0,030	-0,193	0,043	-0,115	0,233

Στατιστική σημασία

Αφαιρώντας τα δεδομένα της Αφρικής φαίνεται ότι δεν υπάρχει συσχέτιση στο gdp με τον δείκτη total deaths και το PCC έχει σχεδόν μηδενιστεί. Ωστόσο το mortality rate είναι αποδεκτό όπως και ο δείκτης total cases. Για τις κλίνες παρατηρείται ότι συνεχίζεται η ίδια συσχέτιση του αριθμού κλινών με το ποσοστό θνησιμότητας. Το population density δεν δείχνει ούτε εδώ κάποια σχέση με κάποιον δείκτη.

Ευρώπη

Παρατηρώντας τα scatter plots της Ευρώπης για το χαρακτηριστικό gdp σε σχέση με τους δείκτες δεν διακρίνεται κάποιο μοτίβο που να οδηγεί σε κάποιο συμπέρασμα. Σχετικά με τον αριθμό κλινών μπορούμε να δούμε ότι όσο περισσότερες κλίνες έχουμε όλοι οι δείκτες μειώνονται. Για το χαρακτηριστικό της πυκνότητας πληθυσμού παρατηρούμε ότι όσο αυξάνονται οι τιμές των χαρακτηριστικών υπάρχει αύξηση στους δείκτες.

Figure 8: PCC, p-values και λογαριθμική τιμή της πυκνότητας πληθυσμού για τις χώρες της Ευρώπης

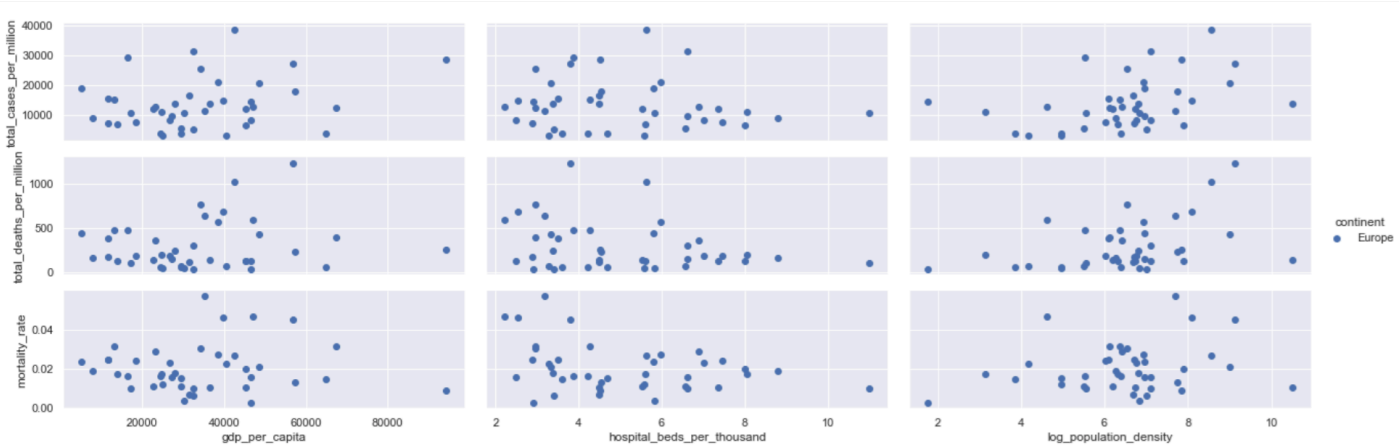


Figure 9: PCC και p-values για κάθε δείκτη-χαρακτηριστικό για την Ευρώπη

	gdp_per_capita		hospital_beds_per_thousand		population_density	
indicators	PCC	p-value	PCC	p-value	PCC	p-value
total_cases_per_million	0,250	0,115	-0,104	0,518	0,265	0,094
total_deaths_per_million	0,162	0,313	-0,269	0,089	0,242	0,128
mortality_rate	0,030	0,853	-0,321	0,041	0,057	0,725

Στατιστική σημασία

Στα δεδομένα της Ευρώπης βλέπουμε ότι πλέον δεν υπάρχει κάποια συσχέτιση του gdp με τους δείκτες που μελετούμε. Αυτό πιθανώς οφείλεται στο γεγονός ότι στην Ευρώπη υπάρχει διακύμανση στο gdp ανά των χωρών. Η πανδημία έχει έξαρση τόσο σε χώρες με υψηλό gdp όσο και σε χώρες με χαμηλό. Επίσης το mortality rate δεν δείχνει κάποια συσχέτιση με το gdp. Ενδεχομένως η Ευρώπη να έχει ένα καλό σύστημα υγείας σε σχέση με άλλες ηπείρους. Ωστόσο, βλέπουμε ότι ακόμα υπάρχει συσχέτιση μεταξύ κλινών και ποσοστού θνησιμότητας. Όσων αφορά το population density, μέσω του PCC φαίνεται να υπάρχει μια θετική σχέση με τα συνολικά κρούσματα και τους θανάτους αλλά δεν θεωρείται στατιστικά σημαντική αν παρατηρήσουμε τα p-values.

Υποερώτημα Β

Στη συνέχεια εξετάζουμε τις διαφορές μεταξύ των δεικτών για τις διαφορετικές ηπείρους. Δημιουργούμε bar plots με τις average τιμές ανά ήπειρο, με error bars.

Figure 10: Bar plot για τα τα συνολικά κρούσματα ανά ήπειρο

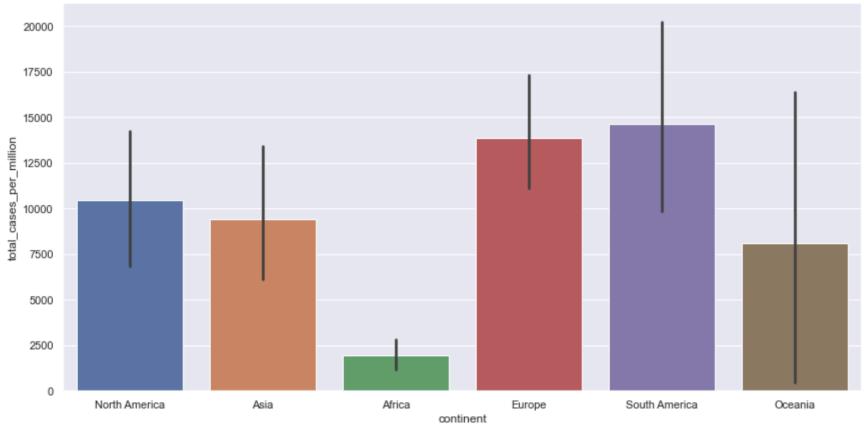


Figure 11: Bar plot για τους συνολικούς θανάτους ανά ήπειρο

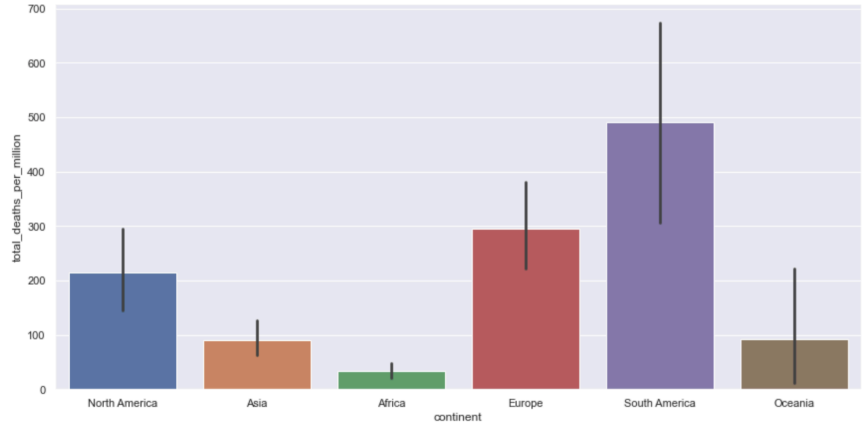
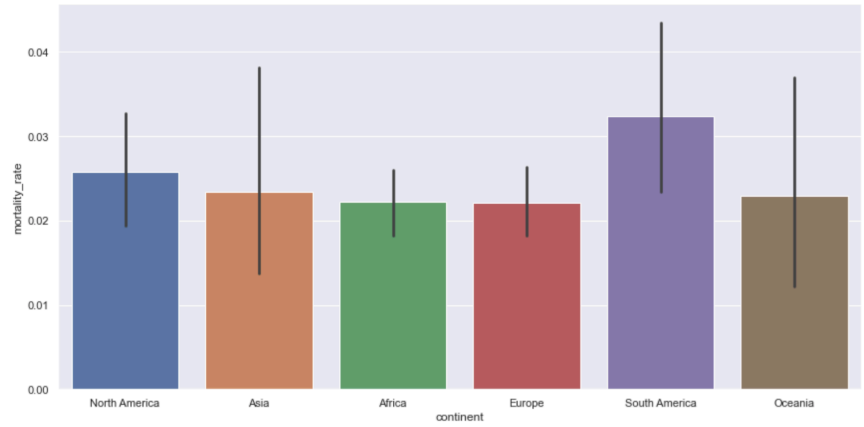


Figure 12: Bar plot για το ποσοστό θνησιμότητας ανά ήπειρο



Παρατηρήσεις

Στο γράφημα 10 μπορούμε να δούμε ότι τα υψηλότερα κρούσματα τα έχει η Νότια Αμερική ενώ τα χαμηλότερα η Αφρική. Μετά την Νότια Αμερική ακολουθούν η Ευρώπη, η Βόρεια Αμερική και η Ασία. Για την Ωκεανία δεν μπορούμε να βγάλουμε συμπέρασμα καθώς τα δεδομένα είναι ελλειπή όπως φαίνεται από το error bar. Διακρίνεται ότι η Αφρική διαφέρει σε πολύ μεγάλο βαθμό από τις υπόλοιπες ηπείρους.

Στο γράφημα 11 φαίνεται πως η Νότια Αμερική έχει πολύ μεγαλύτερο αριθμό συνολικών θανάτων σε σύγκριση με την Ευρώπη ενώ οι δύο αυτοί ήπειροι παρουσιάζουν παρόμοιο βαθμό συνολικών κρουσμάτων.

Στο γράφημα 12 φαίνεται πως η Νότια Αμερική έχει το μεγαλύτερο ποσοστό θνησιμότητας (0.03). Οι υπόλοιπες ήπειροι έχουν σχεδόν το ίδιο ποσοστό θνησιμότητας με την Βόρεια Αμερική να είναι περίπου στο 0.025. Για την Νότια Αμερική ίσως οφείλεται στο ότι το επίπεδο διαβίωσης είναι χαμηλότερο από την Βόρεια Αμερική η οποία εμφανίζει αρκετά μεγάλο ποσοστό θνησιμότητας ενώ είναι περισσότερο αναπτυγμένη. Κοιτώντας τα error bars η Ασία και η Νότια Αμερική φαίνεται να έχουν μεγάλη διακύμανση στις τιμές των χωρών τους σε αντίθεση με την Αφρική και την Ευρώπη.

Ενδιαφέρον είναι το γεγονός ότι σύμφωνα με τις μετρήσεις ενώ η Ευρώπη έχει από τις πιο υψηλές τιμές συνολικών κρουσμάτων το ποσοστό θνησιμότητας είναι από τα πιο χαμηλά.

Διεξαγωγή t-test

Στο notebook της μελέτης αυτής δημιουργήθηκε συνάρτηση σε Python η οποία κάνει διεξαγωγή t-test για μια ήπειρο με όλες τις ηπείρους σε όλους τους δείκτες. Από τα αποτελέσματα των t-test θα ελέγξουμε αν οι παραπάνω παρατηρήσεις επιβεβαιώνονται από τα p-values ως στατιστικά σημαντικές.

Figure 13: p-values από το t-test Africa με όλες τις χώρες στον συνολικό αριθμό κρουσμάτων.

Africa	total_cases_per_million				
	North America	Asia	Europe	South America	Oceania
	$0.019*10^{-2}$	$0.054*10^{-2}$	$0.093*10^{-12}$	0,001	0,251

Figure 14: p-values από το t-test South America με Europe στους συνολικούς θανάτους.

South America	total_deaths_per_million
	Europe
	0,092

Figure 15: p-values από το t-test South America με όλες τις χώρες στο ποσοστό θνησιμότητας.

South America	mortality_rate				
	North America	Asia	Europe	Africa	Oceania
	0,324	0,316	0,098	0,098	0,301

Παρατηρήσεις μετά από το t-test

Στο γράφημα 13 επιβεβαιώνεται ότι η Αφρική όντως διαφέρει κατά πολύ από τις υπόλοιπες ηπείρους στον συνολικό αριθμό κρουσμάτων. Η κύρια διαφορά που παρατηρήθηκε στα bar plots ήταν ο αυξημένος αριθμός θανάτων της Νότιας Αμερικής συγκριτικά με την Ευρώπη ενώ και οι δύο ήπειροι παρουσιάζουν παρόμοιο αριθμό συνολικών κρουσμάτων. Το t-test τελικά έδειξε (πίνακας 14) ότι δεν υπάρχει κάποια στατιστικά σημαντική διαφορά. Παρατηρώντας τα

αποτελέσματα του t-test στον πίνακα 15 όλες οι τιμές p-value των υπόλοιπων ηπείρων συγκριτικά με την Νότια Αμερική είναι μεγαλύτερες του 0.05. Επομένως η υπόθεση ότι η θνησιμότητα της Νότιας Αμερικής διαφέρει στατιστικά σημαντικά από τις υπόλοιπες απορρίπτεται.

Υποερώτημα C

Σε αυτό το ερώτημα θα ασχοληθούμε συνολικά για κάθε Ήπειρο. Για κάθε μέρα θα υπολογίσουμε τον συνολικό αριθμό νέων κρουσμάτων, τον συνολικό αριθμό θανάτων και τον λόγο τους. Θα κάνουμε plot με τις μέσες τιμές ανά μήνα με confidence intervals. Θα μελετήσουμε δηλαδή την εξέλιξη αυτών των δεικτών στο χρόνο.

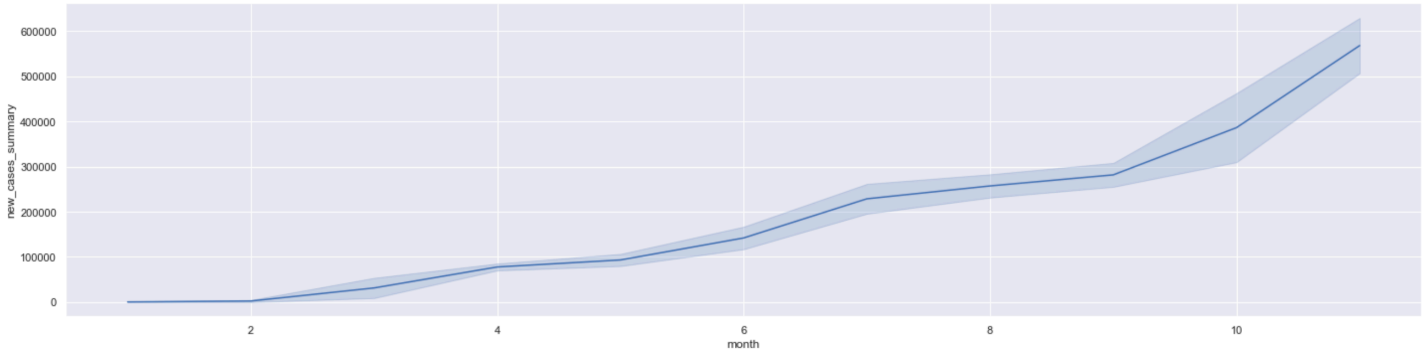
Από τα αρχικά δεδομένα μας θα αφαιρέσουμε τις null τιμές από την στήλη με τις ηπείρους. Θα κρατήσουμε τις τέσσερις στήλες που μας ενδιαφέρουν: continent, date, new_cases και new_deaths. Κάνουμε την μετατροπή του τύπου της στήλης date σε datetime. Στην συνέχεια θα φιλτράρουμε τα δεδομένα μας ώστε να ξεκινάνε από την ημερομηνία 1/1/2020.

Για τον υπολογισμό του αθροίσματος θα ομαδοποιήσουμε τις δύο στήλες new_cases και new_deaths ανά μήνα και ανά μέρα. Τα αποτελέσματα θα τα αποθηκεύσουμε σε δύο μεταβλητές. Η εντολή είναι:

```
nc = data_c.groupby([(data_c.date.dt.month),(data_c.date.dt.day)])['new_cases'].sum()
```

με τον τρόπο αυτό μπορούμε να αποθηκεύσουμε σε μία νέα στήλη τα αθροίσματα χρησιμοποιώντας την .apply του pandas.DataFrame και με μία lambda συνάρτηση όπου ανάλογα την εγγραφή που επεξεργαζόμαστε, αντλεί τις τιμές της ομαδοποιημένης μεταβλητής. Τέλος υπολογίζουμε το ποσοστό θνησιμότητας.

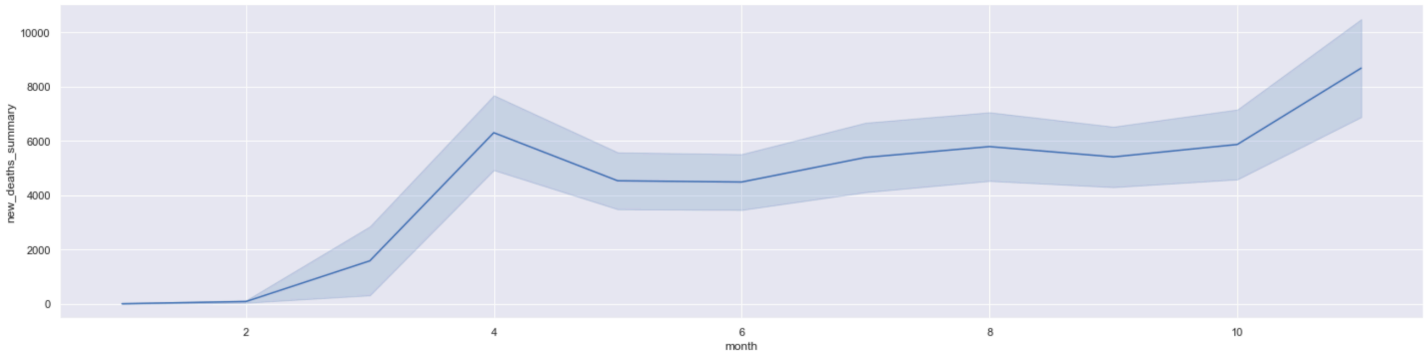
Figure 16: Μέση τιμή ανά μήνα για τον δείκτη new_cases



Στον άξονα x του γραφήματος 16 έχουμε τους μήνες και στον άξονα y την μέση τιμή ανά μήνα του συνολικού αθροίσματος για τα νέα κρούσματα. Τον Ιανουάριο και Φεβρουάριο τα κρούσματα ήταν σχετικά λίγα. Ο αριθμός ήταν κάποιες δεκάδες χιλιάδες παγκοσμίως. Από τον Φεβρουάριο μέχρι τον Απρίλιο βλέπουμε μία σημαντική αύξηση με τον αριθμό κρουσμάτων να φτάνει περίπου τις εκατό χιλιάδες. Την περίοδο εκείνη ξεκίνησε η πρώτη καραντίνα παγκοσμίως. Η καμπύλη από τον Απρίλιο μέχρι και την λήξη της καραντίνας, τον Μάιο βλέπουμε να σταθεροποιείται. Αμέσως μετά την λήξη της πρώτης καραντίνας και μέχρι το τέλος του καλοκαιριού παρατηρείται μια σημαντική αύξηση με συνολικό

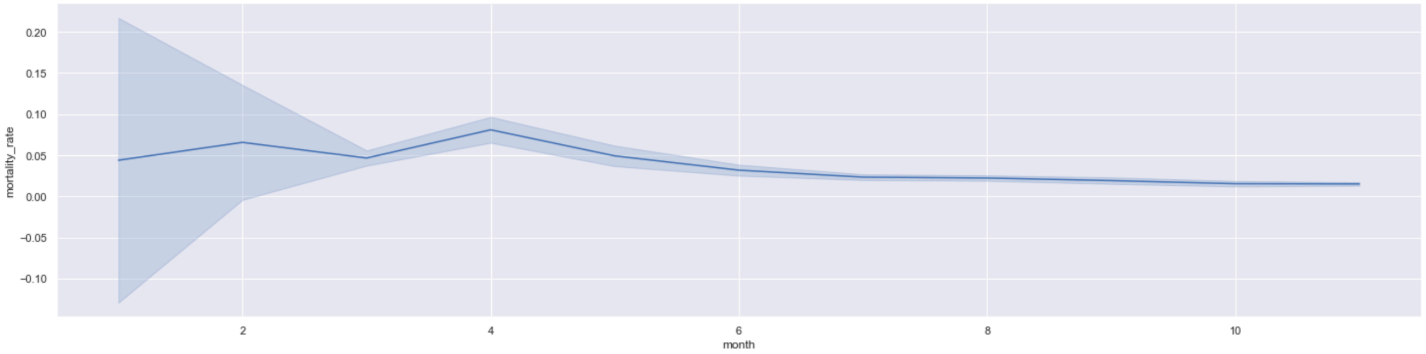
αριθμό τις τριακόσιες χιλιάδες. Η σημαντική όμως αύξηση ξεκίνησε τον Σεπτέμβριο, όπου παρατηρείται μία έξαρση όπου συνεχίζει να αυξάνεται με εκθετικούς ρυθμούς έως και το τέλος του Νοεμβρίου με συνολικό αριθμό, περίπου στις εξακόσιες χιλιάδες. Για την πορεία των συνολικών θανάτων μέσα στο 2020 μπορούμε να δούμε το γράφημα 17

Figure 17: Μέση τιμή του αθροίσματος των ημερών ανά μέρα για τον δείκτη new_deaths



Μέχρι τον Φεβρουάριο η μέση τιμή των συνολικών θανάτων είναι σχεδόν μηδενική. Από τον Φεβρουάριο μέχρι τον Μάρτιο παρατηρείται μία αύξηση που φτάνει περίπου τις δύο χιλιάδες. Σημαντική αύξηση όμως παρατηρείται στο διάστημα του Μαρτίου και Απριλίου όπου ο αριθμός αυξήθηκε κατά τέσσερις χιλιάδες. Είδαμε στο γράφημα 16 ότι τα κρούσματα και οι θάνατοι σε αυτό το διάστημα άρχισαν να αυξάνονται. Οπότε φαίνεται λογική η απόφαση της πρώτης καραντίνας. Μετά την ολοκλήρωση της, βλέπουμε τον αριθμό να έχει μειωθεί και να είναι σταθερός εκτός από μικρές αυξήσεις μέσα στο καλοκαίρι. Τον Οκτώβρη όμως, όπως και τα κρούσματα, ο αριθμός των συνολικών θανάτων κάθε μέρα όλο και αυξάνεται με μέγιστη τιμή περίπου τις εννιά χιλιάδες. Ο λόγος των θανάτων προς τα κρούσματα φαίνεται στο γράφημα 18.

Figure 18: μέση τιμή του ποσοστού θνησιμότητας ανά μήνα

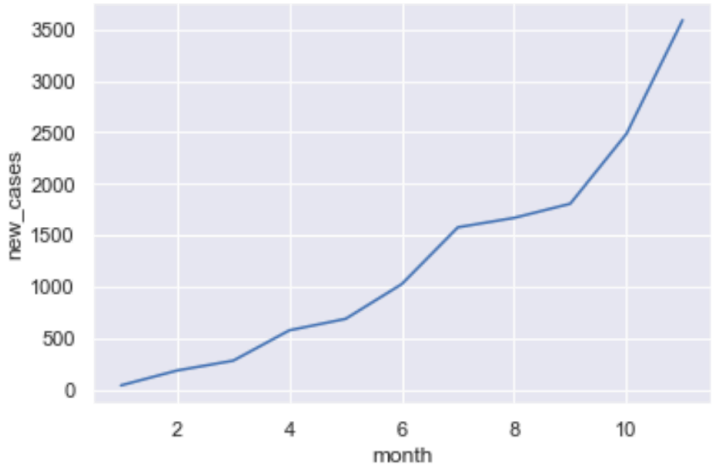


Το **confidence interval** είναι ένας τύπος εκτίμησης που υπολογίζεται από τα στατιστικά στοιχεία των παρατηρούμενων δεδομένων και προτείνει ένα διάστημα από τιμές για την μέση τιμή ενός δείκτη. Η πραγματική παράμετρος θα βρίσκεται μέσα σε αυτό το διάστημα.

Στο γράφημα 19 έχουμε την αύξηση των κρουσμάτων ανά μήνα από τον Ιανουάριο. Με την καραντίνα τον μήνα Απρίλιο η καμπύλη ήρθε σε μια ισορροπία για το χρονικό διάστημα αυτό. Από την λήξη της έως και τον Ιούλιο παρατηρούμε

μια εκθετική αύξηση στην καμπύλη. Τους μήνες του καλοκαιριού παρατηρούμε μια ισορροπία στην καμπύλη. Το φθινόπωρο που ξεκίνησε το νέο κύμα παρατηρούμε η καμπύλη άρχισε ξανά να αυξάνεται εκθετικά.

Figure 19: μέση τιμή κρουσμάτων ανά μήνα



Στο γράφημα 20 βλέπουμε την αύξηση των κρουσμάτων ανά ήπειρο. Τον Απρίλιο τα περισσότερα κρούσματα φαίνεται να τα έχουν οι χώρες στην Ευρώπη και στην Βόρεια Αμερική. Από τον Απρίλιο έως τα μέσα Ιουλίου, βλέπουμε τα κρούσματα στην Ευρώπη να μειώνονται αρκετά με την Νότια Αμερική, την Ασία και Βόρεια Αμερική να αυξάνονται. Όμως, τον Σεπτέμβριο, η εκτόξευση των τιμών της Ευρώπης είναι ανησυχητικές. Η μέση τιμή των κρουσμάτων στην Ευρώπη αυτή την στιγμή αγγίζει τις 250.000. Σε αυτό το σημείο ήρθε και η απόφαση της δεύτερης καραντίνας.

Figure 20: Μέση τιμή συνολικού αριθμού νέων κρουσμάτων ανά ήπειρο

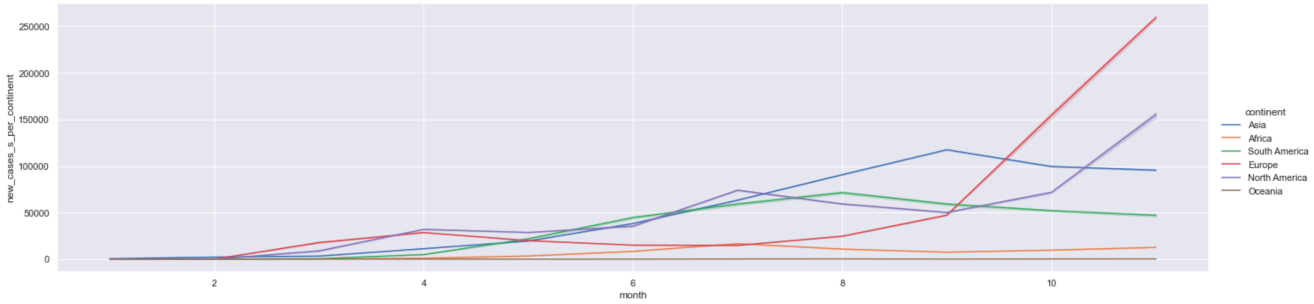
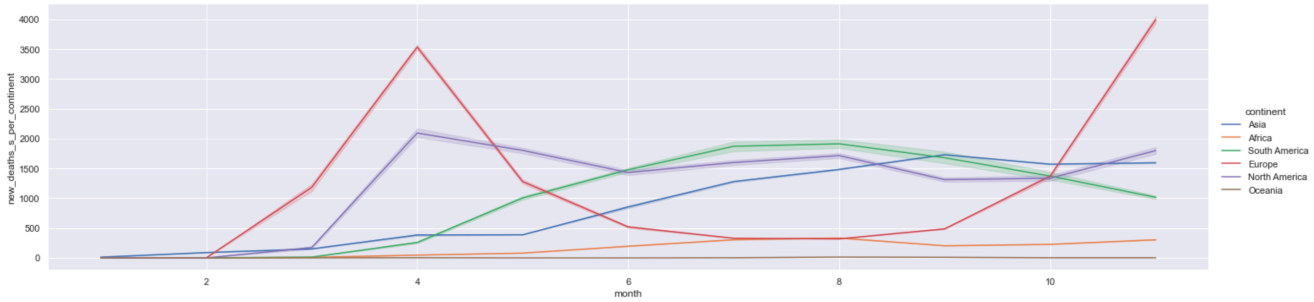


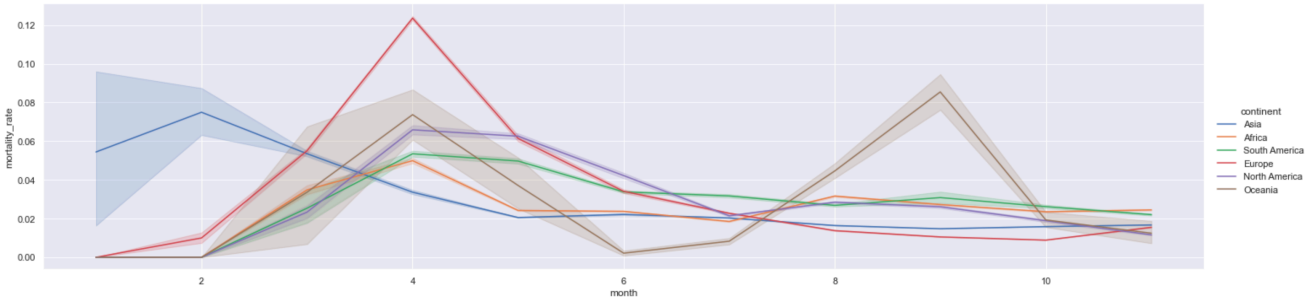
Figure 21: Μέση τιμή συνολικού αριθμού νέων θανάτων ανά ήπειρο



Στο γράφημα 21 έχουμε τον συνολικό αριθμό νέων θανάτων ανά ήπειρο. Τον Φεβρουάριο μέχρι και τον Απρίλιο η τιμή για την Ευρώπη είναι αρκετά υψηλή με 1500 μονάδες πιο κάτω από την Βόρεια Αμερική. Με την έναρξη της καραντίνας

και την μείωση των νέων κρουσμάτων, οι μονάδες εντατικής θεραπείας κατάφεραν να περιθάλψουν περισσότερους ασθενείς με αποτέλεσμα την μείωση των θανάτων. Στην περίοδο του καλοκαιριού, στην Νότια Αμερική και στην Ασία παρατηρείται μία αύξηση των θανάτων η οποία μειώθηκε αμέσως μετά την λήξη του. Όμως, η αύξηση των κρουσμάτων στην Ευρώπη έφερε την μέση τιμή στις 4000.

Figure 22: Ποσοστό θνησιμότητας ανά ήπειρο



Στο γράφημα 22 βλέπουμε το ποσοστό θνησιμότητας. Παρατηρείται η αύξηση σε όλες τις ηπείρους με την Ευρώπη να έχει την μεγαλύτερη τιμή. Η μείωση του ποσοστού έρχεται μετά την πρώτη καραντίνα όπου και σταθεροποιείται το καλοκαίρι έως και τον Νοέμβριο με μόνο μία αύξηση στις χώρες της Ωκεανίας τον Σεπτέμβριο. Πλέον το ποσοστό θνησιμότητας είναι περίπου στο 0.03 πράγμα το οποίο έχει επιβεβαιωθεί ήδη παγκοσμίως.

Υποερώτημα D

Στο τελευταίο ερώτημα της άσκησης θα μελετήσουμε την πανδημία στις Ηνωμένες Πολιτείες. Κατεβάσαμε τα δύο αρχεία για την πανδημία και για τις πρόσφατες εκλογές από τους συνδέσμους και τα φορτώσαμε σε δύο dataframes. Στο αρχείο για τα αποτελέσματα των εκλογών έχουμε την πληροφορία για κάθε πολιτεία αν εξέλεξαν Δημοκρατικούς ή Ρεπουμπλικανούς. Ο συνδυασμός των δύο αρχείων έγινε με την εντολή pandas.merge και αντιστοίχηση στην στήλη State για να πάρουμε τα δεδομένα ως σύνολο καθώς τα states δεν ήταν ίδια στα δύο αρχεία.

Στον άξονα x του γραφήματος 23 έχουμε τις πολιτείες ενώ στον άξονα y τον συνολικό αριθμό κρουσμάτων όπου κάθε τιμή έχει χρωματιστεί ανάλογα με τι εξέλεξαν. Η Καλιφόρνια και το Τέξας έχουν τα περισσότερα κρούσματα με παραπάνω από ένα εκατομμύριο.

Παρατηρήσεις

Όσων αφορά τα γραφήματα 23 και 24 δεν παρατηρείται κάποια συσχέτιση μεταξύ των συνολικών κρουσμάτων και θανάτων σε σχέση με τις ψήφους ανά πολιτεία. Ωστόσο στο γράφημα 25 μπορούμε να διακρίνουμε ότι οι πολιτείες με το υψηλότερο ποσοστό θνησιμότητας είναι εκείνες που έχουν εκλέξει το κόμμα των Δημοκρατικών. Θα προχωρήσουμε σε διεξαγωγή t-test για να επιβεβαιώσουμε ή να απορρίψουμε τους ισχυρισμούς των παρατηρήσεων.

Figure 23: Συνολικά κρούσματα ανά πολιτεία

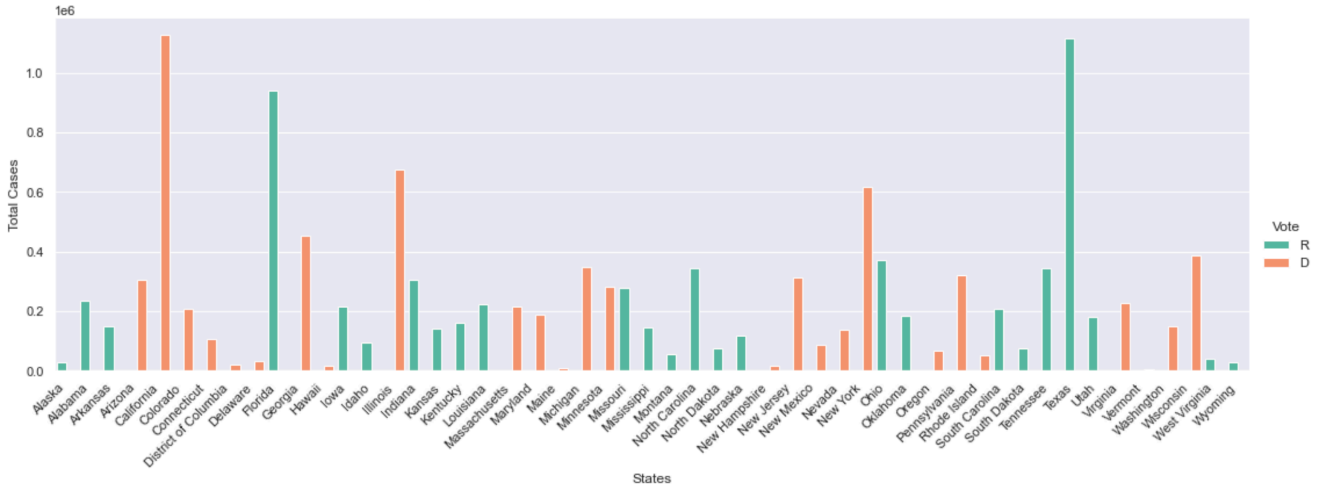


Figure 24: Συνολικοί θάνατοι ανά πολιτεία

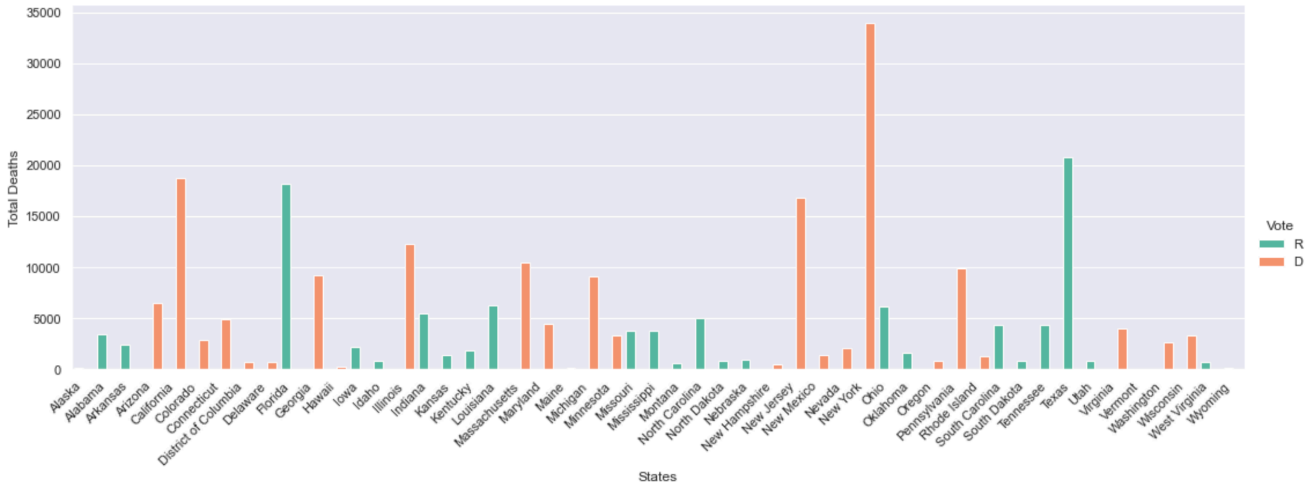


Figure 25: Ποσοστό θνησιμότητας ανά πολιτεία (συνολικοί θάνατοι προς συνολικά κρούσματα)

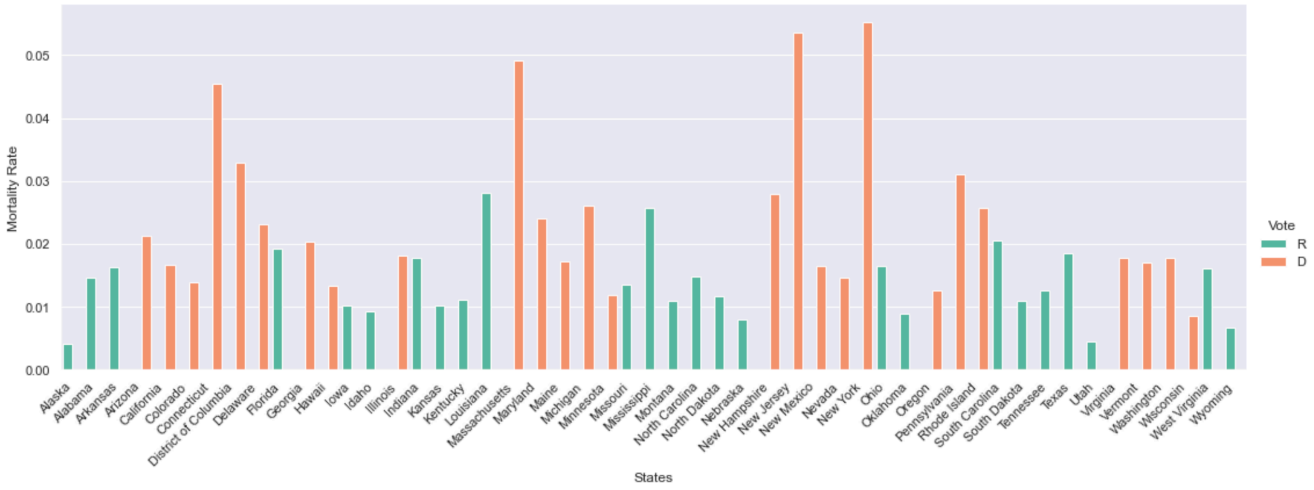


Figure 26: Αποτελέσματα t-test για τους δείκτες των δύο groups

Republicans - Democrats	total_cases	total_deaths	mortality_rate
	0,978	0,211	$0.059*10^{-2}$

Στατιστική σημασία

Έπειτα από την διεξαγωγή του t-test παρατηρούμε στον πίνακα 26 ότι το p-value για το ποσοστό θνησιμότητας είναι μικρότερο του 0.05 άρα θεωρείται στατιστικά σημαντική η συσχέτιση μεταξύ υψηλού ποσοστού θνησιμότητας και εκλογής του κόμματος των Δημοκρατικών. Στα συνολικά κρούσματα και τους συνολικούς θανάτους δεν βλέπουμε κάποια συσχέτιση όπως παρατηρήθηκε και στα bar plots. Μία πιθανή εξήγηση του αποτελέσματος ίσως οφείλεται στην πολιτική που ακολούθησε η προηγούμενη κυβέρνηση των Ρεπουμπλικάνων. Ίσως να μην δόθηκε η απαραίτητη προσοχή, με αποτέλεσμα τα κρούσματα και οι θάνατοι να αυξηθούν απότομα και να φτάσουν σε υψηλά επίπεδα όπως παρατηρήσαμε και σε προηγούμενο ερώτημα. Για τον λόγο αυτό, οι πολιτείες που επλήγησαν περισσότερο ίσως να προτίμησαν τους δημοκρατικούς.