ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Στατιστική στην Πληροφορική (2023-2024)

Διονύσιος Ρηγάτος (P3200262)

Εργασία 1

Άσκηση 1

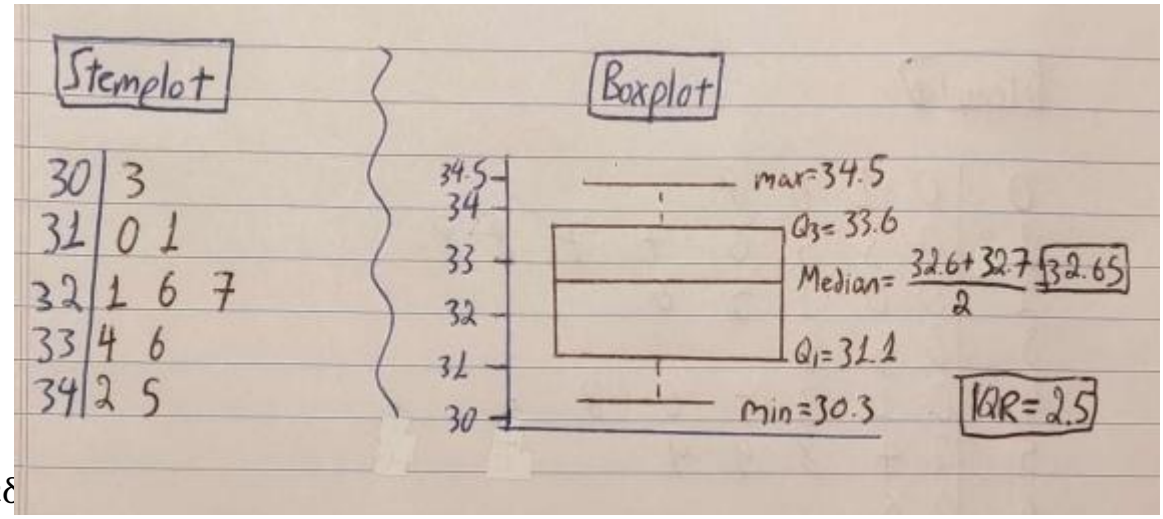
Δεδομένα I									
30.3	31.0	31.1	32.1	32.6	32.7	33.4	33.6	34.2	34.5

Δεδομένα II									
0.0	0.0	0.2	0.8	1.2	1.4	3.2	4.2	6.4	9.0

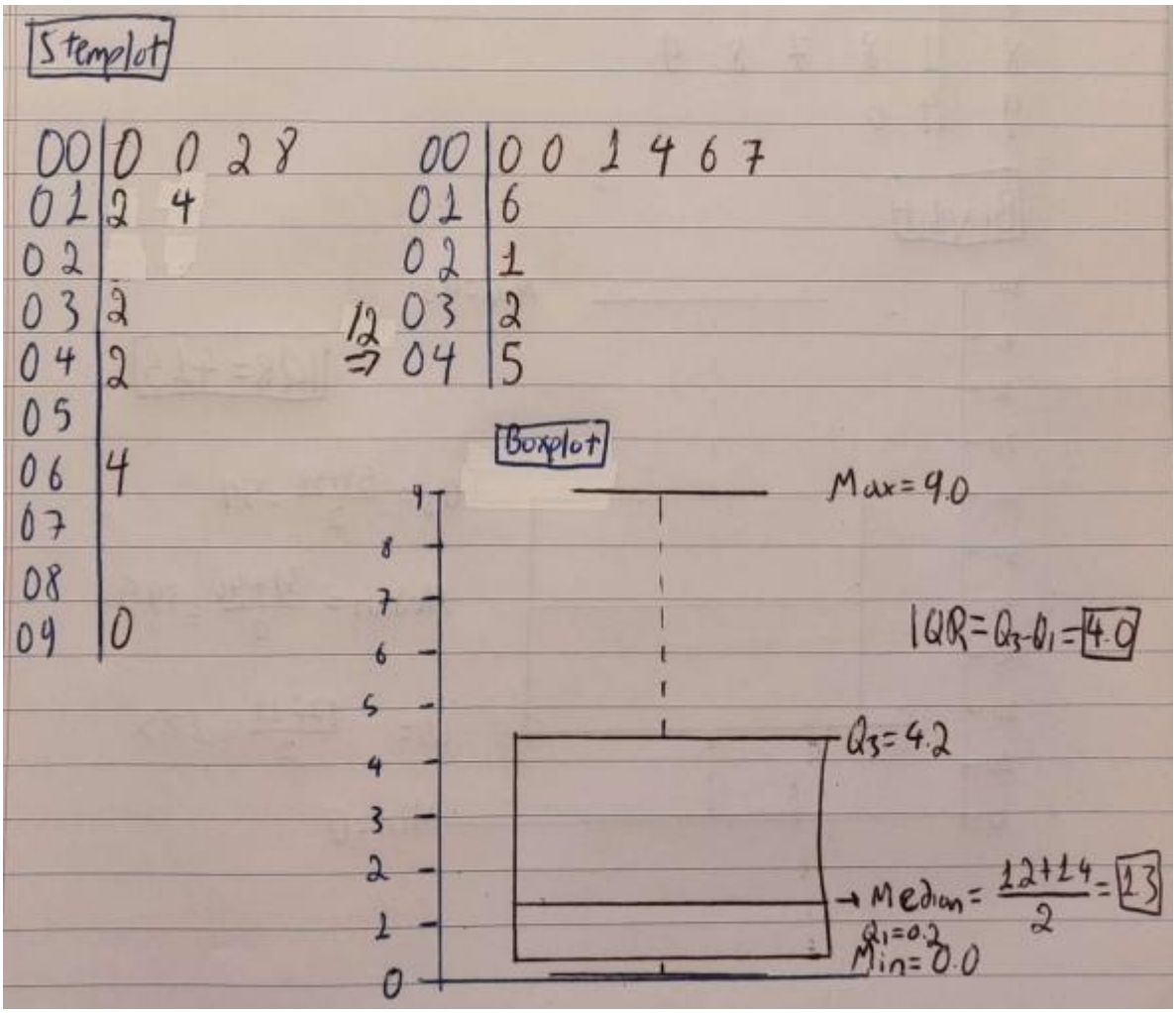
Δεδομένα III									
0	1	6	8	10	13	15	16	17	17
18	18	20	20	21	25	26	30	35	39
40	41	43	44	46	48	52	54	58	59
59	60	66	81	86	87	88	89	94	96

α)

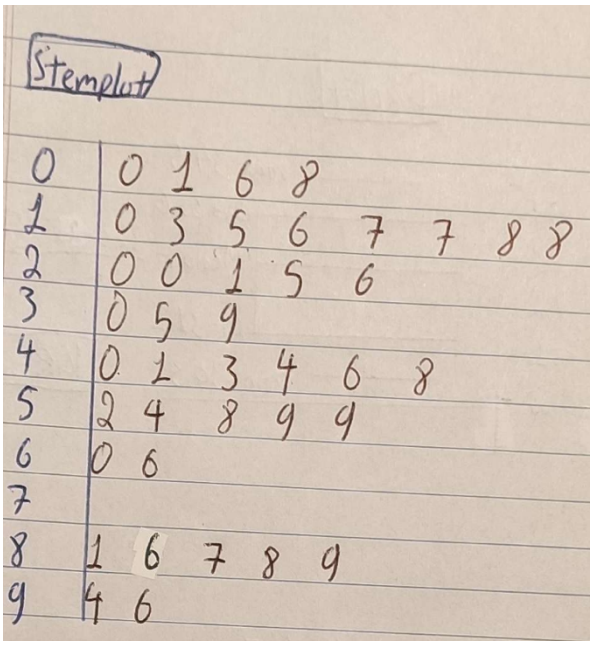
Δεδομένα I

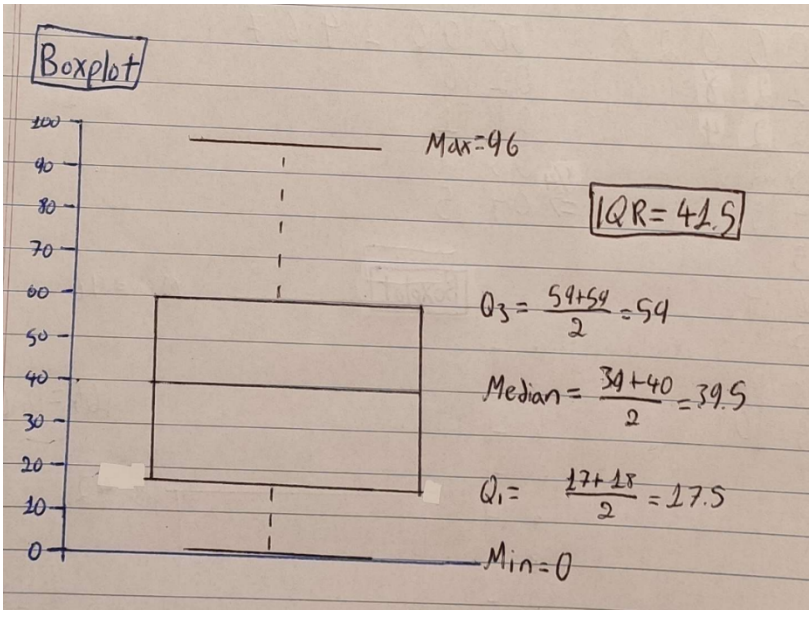


Δεδ



Δεδομένα III





β)

Δεδομένα I

Τα δεδομένα είναι ομοιόμορφα κατανεμημένα γύρω από το μ, και επομένως προτιμούμε την τυπική απόκλιση (σ = 1.41) και τη μέση τιμή (μ = 32.55).

Δεδομένα II

Τα δεδομένα δεν είναι ομοιόμορφα κατανεμημένα γύρω από το μ, καθώς έχουμε πολλές περισσότερες χαμηλές τιμές απ' ότι υψηλές, επομένως η σύνοψη των 5 αριθμών αποτελεί καλύτερη αναπαράσταση καθώς μας δίνει καλύτερη εικόνα για το που πέφτουν οι τιμές μας στα ποσοστημόρια. Παρατηρούμε επίσης ότι σ = 3.05 και μ = 2.64.

Δεδομένα III

Τα δεδομένα θα αναπαρασταθούν καλύτερα από την σύνοψη των 5 αριθμών και την μέση τιμή και τυπική απόκλιση (σ = 28.26 & μ = 41.15 με min = 0, median = 39.5, max = 96). Παρ' όλο που και οι 2 αναπαραστάσεις είναι καλές, θα προτιμήσουμε την σύνοψη των 5 αριθμών καθώς μας δίνει περισσότερες πληροφορίες για τα δεδομένα μας.

γ)

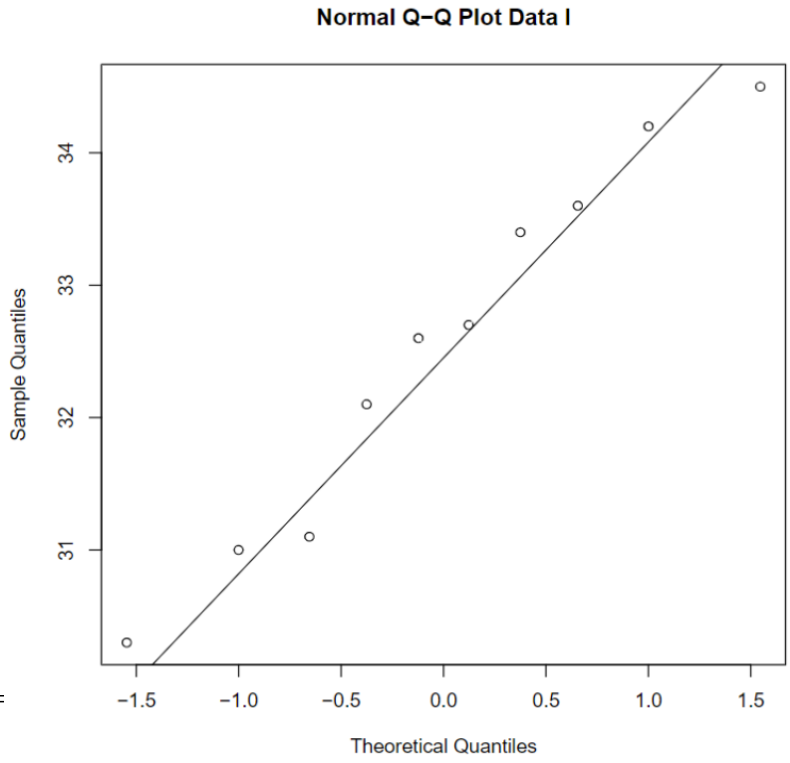
Στην κανονική κατανομή για τα δεδομένα ισχύει ότι:

68% εντός (μ-σ, μ+σ)

95% εντός (μ-2σ, μ+2σ)

99% εντός (μ-3σ, μ+3σ)

Δεδομένα I



μ = 32.55, σ =

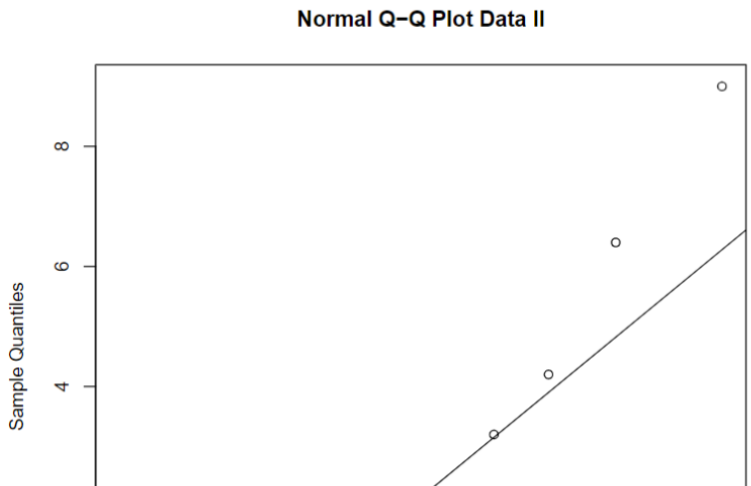
50% εντός (3

100% εντός (29.71, 35.38) (Έπρεπε 95%)

100% εντός (28.29, 36.8) (Έπρεπε 99.5%)

Παρατηρούμε ότι, ειδικά στο (μ-σ, μ+σ) και (μ-2σ, μ+2σ), υπάρχουν μεγάλες αποκλίσεις και συνεπώς η προσέγγιση από την καμπύλης πυκνότητας της Κανονικής Κατανομής δεν θα ήταν ιδιαίτερα ακριβής.

Δεδομένα II



μ = 2.64, σ = 3.05

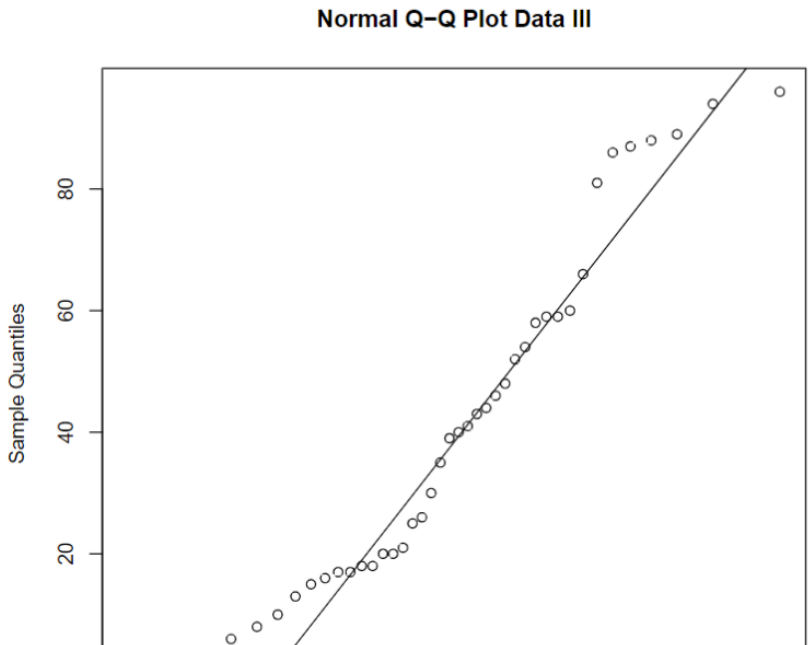
80% εντός (-0.42, 5.70) (Έπρεπε 68%)

90% εντός (-3.47, 8.75) (Έπρεπε 95%)

100% εντός (-6.53, 11.81) (Έπρεπε 99.5%)

Παρατηρούμε ότι, ειδικά στο (μ-σ, μ+σ), υπάρχουν μεγάλες αποκλίσεις και συνεπώς η προσέγγιση από την καμπύλης πυκνότητας της Κανονικής Κατανομής δεν θα ήταν ιδιαίτερα ακριβής.

Δεδομένα III



$\mu = 41.15, \sigma = 28.26$

70% εντός (12.88, 69.41) (Έπρεπε 68%)

100% εντός (-15.38, 97.68) (Έπρεπε 95%)

100% εντός (-43.65, 125.95) (Έπρεπε 99.5%)

Παρατηρούμε ότι υπάρχουν αποκλίσεις σε όλα τα διαστήματα, και συνεπώς η προσέγγιση από την καμπύλης πυκνότητας της Κανονικής Κατανομής δεν θα ήταν ιδιαίτερα ακριβής.

Άσκηση 2

α) Τα δεδομένα αφορούν την εκπομπή διοξειδίου του άνθρακα (σε κιλά) per capita και τις περιπτώσεις καρκίνου per capita στην Αμερική σε βάθος 20ετίας (1999-2019), όπου το κάθε έτος αποτελεί μια περίπτωση.

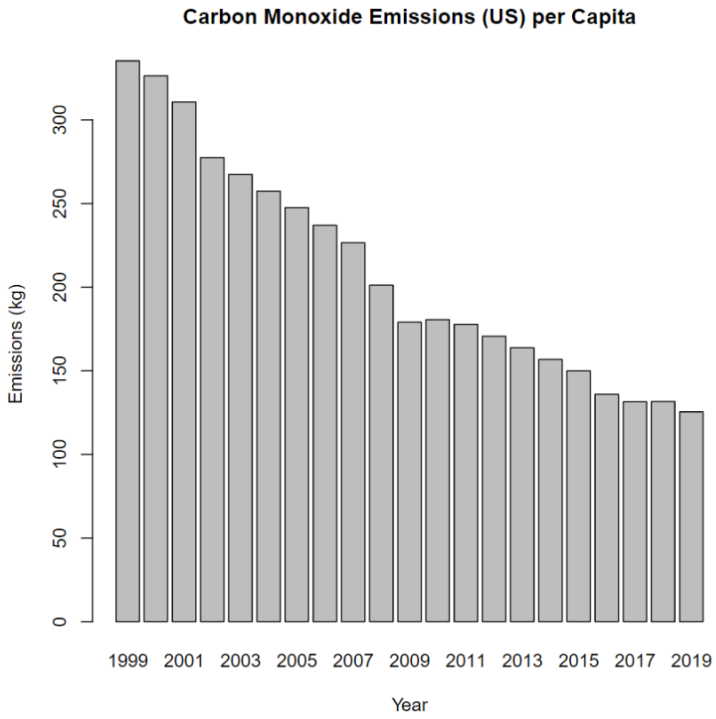
Οι εκπομπές προέκυψαν από dataset του OECD (*stats.oecd.org/Index.aspx?DataSetCode=AIR_EMISSIONS*) και οι περιπτώσεις καρκίνου από το GIS CDC (*gis.cdc.gov/Cancer/USCS/#/Trends/*).

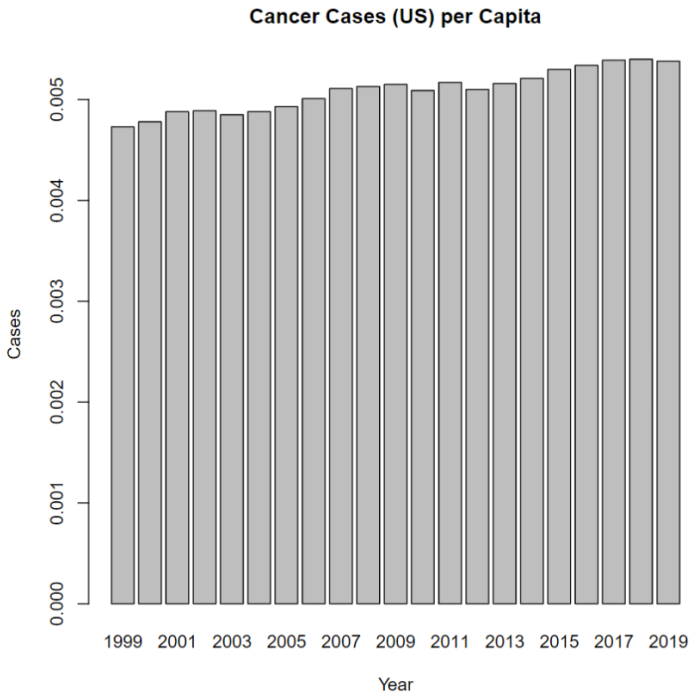
β) **Κατηγορική μεταβλητή** είναι η χρονολογία (year) και παίρνει τιμές 1999 έως και 2019.

Οι **ποσοτικές μεταβλητές** είναι η κατά κεφαλήν ποσότητα κιλών μονοξειδίου του άνθρακα emitted (Carbon Monoxide in kg per capita) και οι κατά κεφαλήν περιπτώσεις καρκίνου (Cancer Cases per capita) στην Αμερική.

Στην ποσότητα μονοξειδίου του άνθρακα χρησιμοποιήθηκε το σύνολο των εκπομπών της Αμερικής δια τον πληθυσμό της. Αντίστοιχα στις περιπτώσεις καρκίνου χρησιμοποιήθηκε το σύνολο των περιπτώσεων δια τον πληθυσμό.

γ)

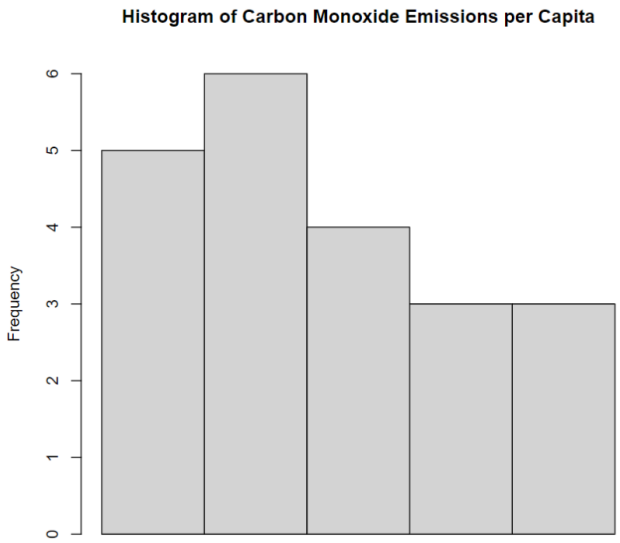


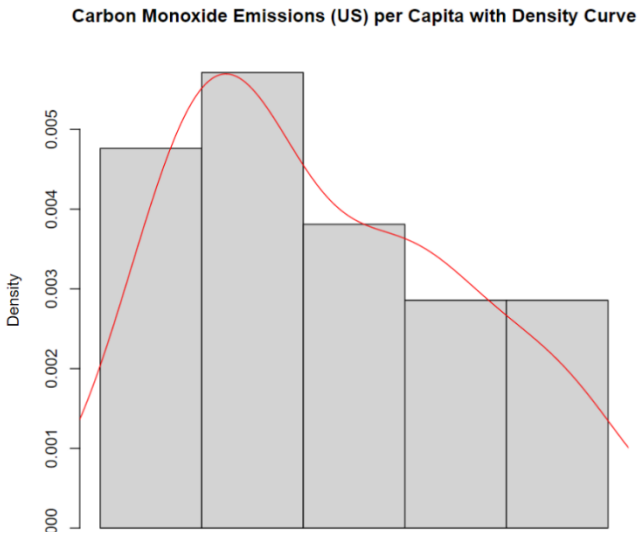


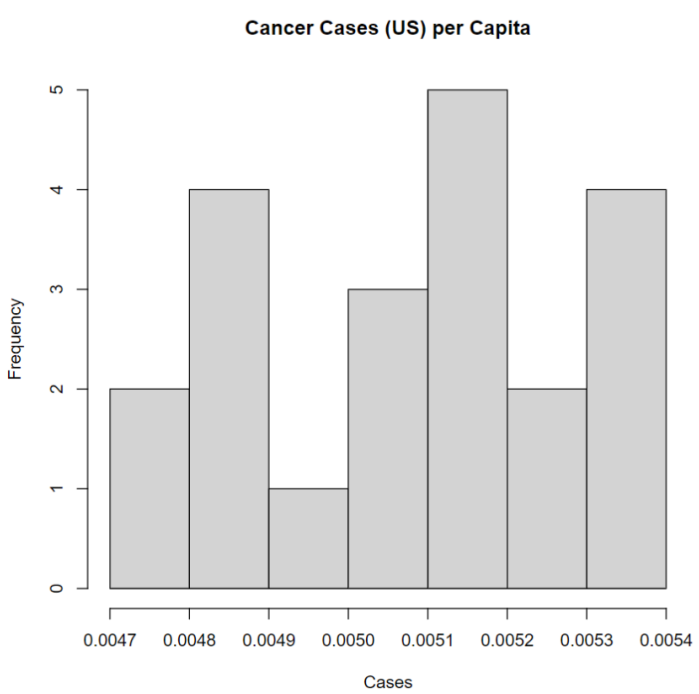
Στα παραπάνω barcharts βλέπουμε την συστηματική μείωση κατά κεφαλήν εκπομπών διοξειδίου του άνθρακα στο πέρασμα του χρόνου, αλλά την αύξηση των κατά κεφαλήν περιπτώσεων καρκίνου.

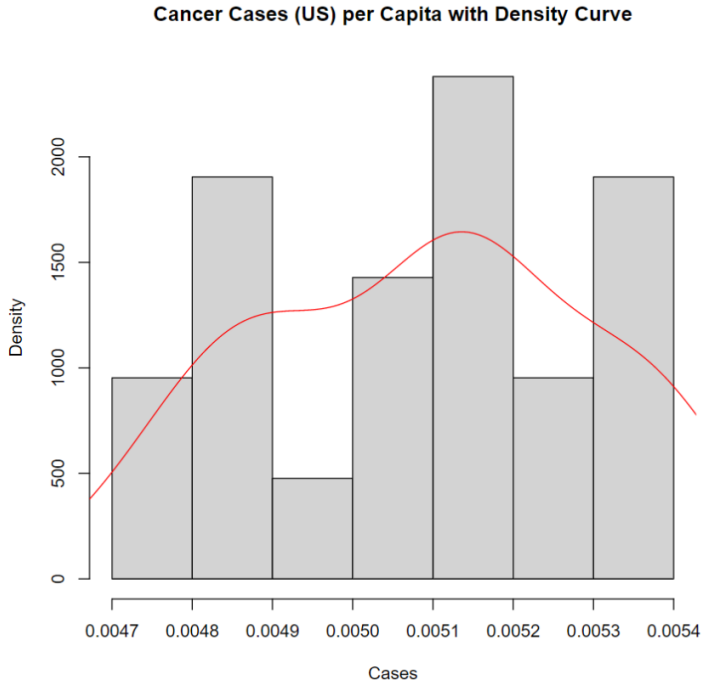
Η μείωση των εκπομπών σε βάθος 20ετιας οφείλεται στην λήψη μέτρων κατά του φαινόμενου του θερμοκηπίου μέσω νόμων.

Η ελαφριά αύξηση των περιπτώσεων καρκίνου είναι αμελητέα και βρίσκεται στα φυσιολογικά πλαίσια εάν ληφθεί υπόψιν ο δυτικός (Αμερική) τρόπος ζωής (κάπνισμα, fast food κλπ).









δ)

Κατά Κεφαλήν Ποσότητα Κιλών Μονοξειδίου του Άνθρακα (US)

Τυπική Απόκλιση: 67.0899

Μέση Τιμή: 209.0651

Σύνοψη Πέντε Αριθμών:

Min	Q1	Median	Q3	Max
125.479	156.858	180.576	257.456	335.294

Οι τιμές είναι κατανεμημένες με ανομοιόμορφο τρόπο με αποτέλεσμα οι τιμές που βρίσκονται πάνω από τη διάμεσο να αποκλίνουν περισσότερο μεταξύ τους σε αντίθεση με τις τιμές κάτω από αυτή, κάτι που φαίνεται από την διαφορά διαμέσου – μέσης τιμής. Συνεπώς η χρήση της σύνοψης των πέντε αριθμών θα ήταν καλύτερη στην αναπαράσταση των δεδομένων.

Κατά Κεφαλήν Περιπτώσεις Καρκίνου (US)

Τυπική απόκλιση: 0.0002063608

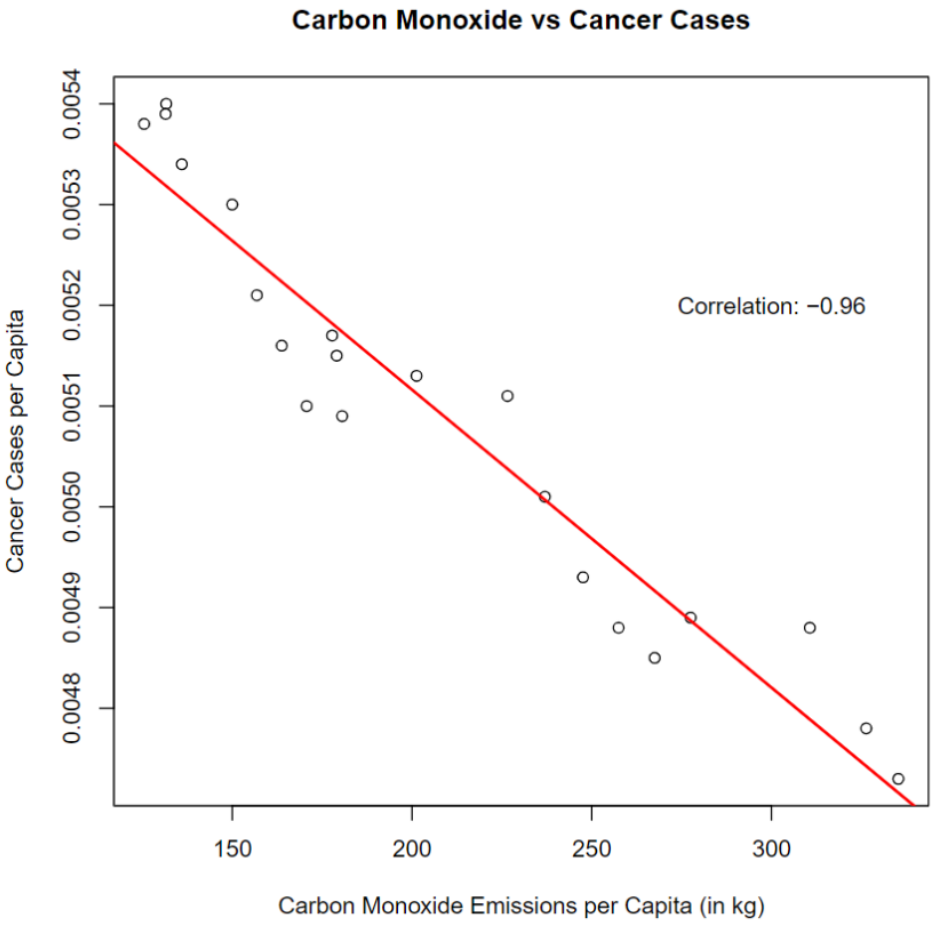
Μέση Τιμή: 0.005089

Σύνοψη Πέντε Αριθμών:

Min	Q1	Median	Q3	Max
0.00473	0.00489	0.00511	0.00521	0.00540

Οι τιμές παρουσιάζουν ανομοιόμορφη κατανομή, οι χαμηλότερες τιμές των δεδομένων έχουν μεγαλύτερες αποκλίσεις μεταξύ τους και κατανέμονται σε μεγαλύτερο διάστημα από ότι το ανώτερο 50% των τιμών. Παρατηρούμε επίσης από την καμπύλη πυκνότητας ότι δεν έχουμε κανονική κατανομή. Η σύνοψη των πέντε αριθμών θα ήταν καλύτερη στην αναπαράσταση των δεδομένων.

ε)



Από το scatterplot και το r = -0.96, παρατηρούμε ότι υπάρχει ισχυρή γραμμική αρνητική συσχέτιση στα δεδομένα μας, δηλαδή οι κατά κεφαλήν περιπτώσεις καρκίνου μειώνονται όσο αυξάνονται οι κατά κεφαλήν εκπομπές διοξειδίου του άνθρακα (σε κιλά) (στην Αμερική πάντα).

Η σχέση αυτή μπορεί να απατήσει, καθώς οι εκπομπές διοξειδίου του άνθρακα δεν έχουν σημαντική επίδραση στις περιπτώσεις καρκίνου, και αποτελεί καλό παράδειγμα ότι correlation does not equal causation.

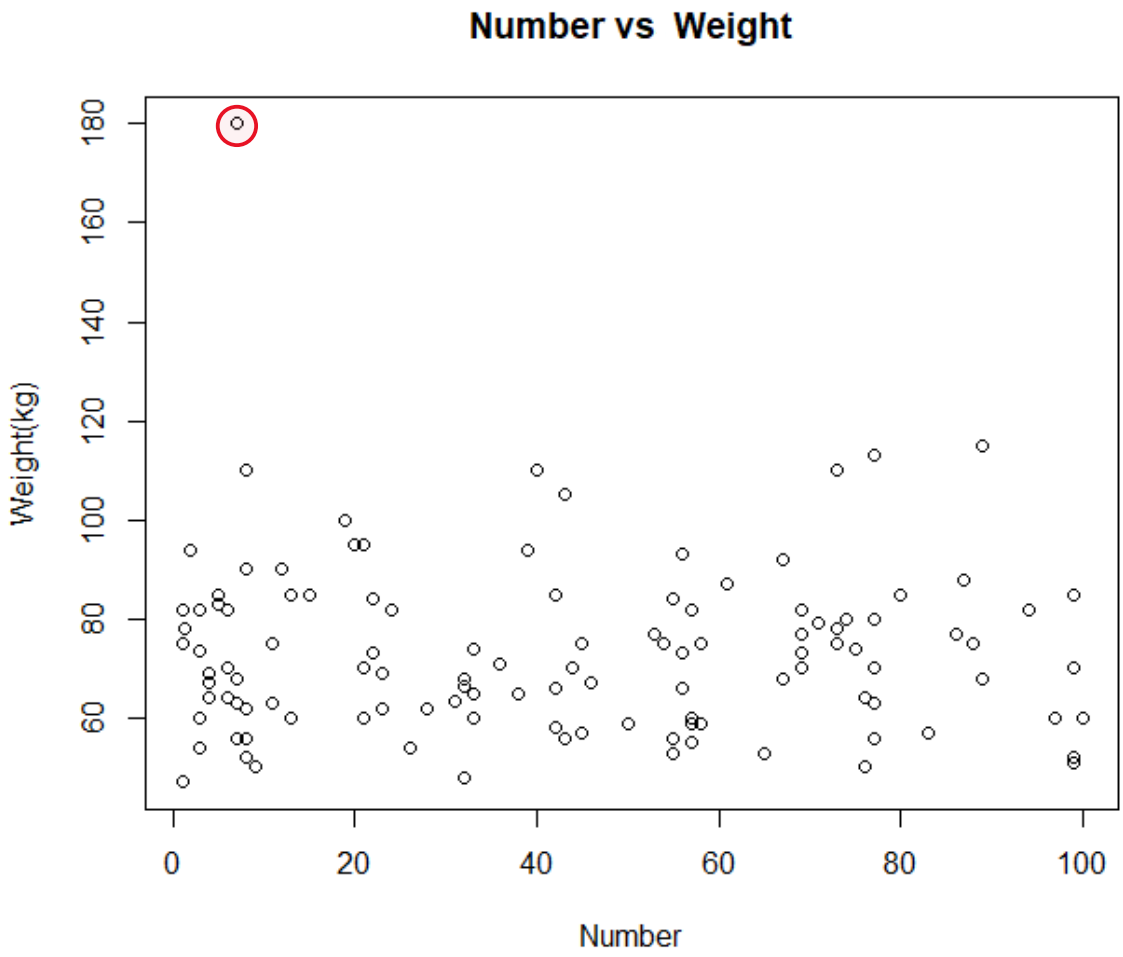
Άσκηση 3

α)

Το παρακάτω διάγραμμα είναι το scatterplot ανάμεσα στην μεταβλητή βάρους (weight) και στον αριθμό που επέλεξε ο αντίστοιχος φοιτητής (number).

Η συσχέτιση είναι γραμμική φθίνουσα με καθόλου ισχύ (r = -0.03231416), γεγονός που πιθανώς οφείλεται στο ότι το βάρος δεν καθορίζεται από τον αριθμό που επιλέγουν οι φοιτητές.

Υπάρχει ένα ατυπικό σημείο.



β)

Γραμμική Παλινδρόμηση Ελαχίστων Τετραγώνων

Συντελεστής Συσχέτισης r: -0.03231416