ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Αξιολόγηση λειτουργικών και κλινικών δεδομένων νοσοκομείων με κατάλληλους δείκτες και χρήση σύγχρονων τεχνολογιών της επιστήμης δεδομένων

Κατσαράκη Βασιλική

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ
Υπεύθυνη
Κατσαράκη Βασιλική

Λαμία,2020-2021

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Αξιολόγηση λειτουργικών και κλινικών δεδομένων νοσοκομείων με κατάλληλους δείκτες και χρήση σύγχρονων τεχνολογιών της επιστήμης δεδομένων

Κατσαράκη Βασιλική

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Επιβλέπων
Χαράλαμπος Καρανίκας
Λέκτορας

Λαμία,2020-2021

Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις [1], που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

1. *Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί* ***χωρίς να τα περικλείω σε εισαγωγικά*** *και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάσθηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.*

2. *Δέχομαι ότι η αυτολεξεί* ***παράθεση χωρίς εισαγωγικά****, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.*

3. *Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια*

4. *Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.*

Ημερομηνία: 3/6/2021

Η Δηλ.

Κατσαράκη Βασιλική

(1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.

Αξιολόγηση λειτουργικών και κλινικών δεδομένων νοσοκομείων με κατάλληλους δείκτες και χρήση σύγχρονων τεχνολογιών της επιστήμης δεδομένων

Κατσαράκη Βασιλική

Τριμελής Επιτροπή:

Καρανίκας Χαράλαμπος, Λέκτορας(επιβλέπων)

Βασίλης Πλαγιανάκος, Καθηγητής

Σωτήρης Τασουλής, Επίκουρος Καθηγητής

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΥΧΑΡΙΣΤΙΕΣ

Ευχαριστώ πολύ τον επιβλέπων καθηγητή μου κ.Καρανίκα Χαράλαμπο για την πολύτιμη βοήθεια του και τις συμβολές του για τη συγγραφή της παρούσας πτυχιακής εργασίας. Επίσης ευχαριστώ πολύ την οικογένεια μου για την στήριξη που μου παρείχε κατά τη διάρκεια των μαθητικών και των φοιτητικών μου χρόνων με στόχο την εκπλήρωση των ονείρων και των φιλοδοξιών μου.

ΠΡΟΛΟΓΟΣ

Η πτυχιακή αυτή με τίτλο *«Αξιολόγηση λειτουργικών και κλινικών δεδομένων νοσοκομείων με κατάλληλους δείκτες και χρήση σύγχρονων τεχνολογιών της επιστήμης δεδομένων (data science)»* έχει βασικό αντικείμενο την ανάλυση και την απεικόνιση δεδομένων των νοσοκομείων της χώρας μας μέσω της γλώσσας προγραμματισμού R και της επιστήμης δεδομένων ώστε να προκύψουν γνώσεις και χρήσιμα συμπεράσματα.

Η *επιστήμη των δεδομένων ή αλλιώς Data Science* είναι ένα πεδίο που χρησιμοποιείται για να παραχθούν πληροφορίες από δεδομένα. Η επιστήμη αυτή αναπτύχθηκε τα τελευταία χρόνια διότι τα δεδομένα που χρησιμοποιεί είναι τύπου big data.

Στην εποχή που ζούμε η ανάλυση και επεξεργασία πληροφοριών και δεδομένων καταλαμβάνουν αναπόσπαστο κομμάτι της καθημερινής μας ζωής. Στον τομέα της υγείας, οι εργαζόμενοι πρέπει να επεξεργαστούν πληροφορίες (π.χ. τη μέση διάρκεια νοσηλείας) ώστε να παράξουν χρήσιμα συμπεράσματα για την αποτελεσματικότερη θεραπεία των ασθενών τους. Η ανάλυση των δεδομένων αυτών είναι δεν είναι ακριβή διότι τα περισσότερα λογισμικά είναι δωρεάν.

Για την εργασία αυτή χρησιμοποιήθηκε η γλώσσα προγραμματισμού R η οποία μας βοηθάει να αναλύσουμε δεδομένα ώστε να οδηγηθούμε σε χρήσιμα συμπεράσματα τα οποία μπορούμε να τα σχεδιάσουμε και να τα απεικονίσουμε με ποικίλους τρόπους.

Στο πρώτο κεφάλαιο της εργασίας αναπτύχθηκαν οι έννοιες των data science και big data αλλά και τα είδη των δεικτών των νοσοκομείων. Στο δεύτερο και στο τρίτο κεφάλαιο παρουσιάζονται τα βασικά χαρακτηριστικά και οι ενέργειες που διέπουν την γλώσσα προγραμματισμού R με στόχο κάποιες από αυτές να χρησιμοποιηθούν μετέπειτα.

Στο επόμενο κεφάλαιο περιγράφονται και αναλύονται οι δείκτες των νοσοκομείων όπως για παράδειγμα το συνολικό μέσο κόστος ανά ασθενή και ανά ημέρα νοσηλείας. Το πέμπτο και τελευταίο κεφάλαιο περιλαμβάνει την ανάλυση των δεδομένων με την υλοποίηση των δεικτών μέσω της R από την οποία εξάγονται κάποια σημαντικά συμπεράσματα.

Κεφάλαιο 1ο- Data Science και Δείκτες Νοσοκομείου

1.1:Data Science-Big Data

Οι έννοιες των *big data* και *data science* (επιστήμη των δεδομένων) είναι άμεσα συνδεδεμένες, διότι η δεύτερη αναπτύχθηκε χάρη στην πρώτη. Πιο συγκεκριμένα, η επιστήμη των δεδομένων χρησιμοποιεί «μεγάλα δεδομένα» (big data) τα οποία μπορούν να έχουν κάποια συγκεκριμένη δομή ή και όχι. Έτσι λοιπόν χάρη στην επιστήμη αυτή αλλά και σε άλλες (όπως στη στατιστική) μπορούν να εξαχθούν συμπεράσματα μεγάλης σημασίας.

Κάποια χαρακτηριστικά των big data είναι τα παρακάτω:

- *Όγκος*:έχουν πολλά δεδομένα.
- *Ποικιλία:* υπάρχει μεγάλη ποικιλία για αυτό συχνά απαιτούν επεξεργασία πριν χρησιμοποιηθούν.
- *Ειλικρίνεια*: είναι απαραίτητο να ανταποκρίνονται στην πραγματικότητα.

Για να επεξεργαστούμε big data θα πρέπει να ακολουθήσουμε τα παρακάτω βήματα:

1 συλλογή δεδομένων
2 αποθήκευση δεδομένων
3 επεξεργασία δεδομένων
4 ανάλυση των δεδομένων και
5 κοινοποίηση των αποτελεσμάτων που προέκυψαν από την μελέτη.

Συμπερασματικά, μέσω του data science μπορούν να υπάρξουν προβλέψεις για το μέλλον, έτσι ώστε κάθε επιχείρηση, σε αυτήν την περίπτωση τα νοσοκομεία, να μπορούν να παράσχουν αποτελεσματικότερες υπηρεσίες προς τους καταναλωτές τους δηλαδή προς τους ασθενείς τους.

1.2: Δείκτες Νοσοκομείου

Γενικά, ο κυριότερος στόχος κάθε νοσοκομείου είναι η παροχή υπηρεσιών με το χαμηλότερο δυνατό κόστος χωρίς όμως να επηρεάζεται το επίπεδο των παροχών προς τους ασθενείς. Ουσιαστικά, τα κατάλληλα πρόσωπα λαμβάνουν αποφάσεις, αν χρειάζεται, σχετικά με την βελτίωση των υπηρεσιών που προσφέρονται μέσα από μία διαδικασία που ονομάζεται αξιολόγηση νοσοκομείων. Η αξιολόγηση των νοσοκομείων είναι δύο ειδών:

1 είναι η *ποιοτική αξιολόγηση* που σχετίζεται με τις υπηρεσίες του νοσοκομείου. Σύμφωνα, με τον Donabedian η περίθαλψη των ασθενών χωρίζεται σε 3 επίπεδα, στη δομή, στη διαδικασία και στο αποτέλεσμα. Για να υπάρχει λοιπόν ποιότητα στην περίθαλψη θα πρέπει φυσικά να υπάρχει ποιότητα και στα 3 αυτά επίπεδα.
2 και η *οικονομική αξιολόγηση* που σχετίζεται με την αποδοτικότητα των υπηρεσιών που παρέχουν.

Η αξιολόγηση των νοσοκομείων όπως αναφέρθηκε και προηγουμένως πραγματοποιείται μέσω κάποιων προσώπων με τη βοήθεια ορισμένων δεικτών (λειτουργικών και οικονομικών) οι οποίοι πρέπει να έχουν κάποια βασικά χαρακτηριστικά που είναι τα εξής:

*1.αξιοπιστία:*οι δείκτες πρέπει να δίνουν το ίδιο αποτέλεσμα όσες φορές και αν χρησιμοποιηθούν την ίδια χρονική στιγμή,
*2.χρησιμότητα:*θα πρέπει να εξάγεται ένα σημαντικό συμπέρασμα για τους χρήστες (δηλαδή να έχουν χρησιμότητα),
*3.ακρίβεια:*οι δείκτες απαιτούνται να είναι όσο το δυνατόν περισσότερο ακριβείς (δηλαδή να μην έχουν αποκλίσεις),
*4.εγκυρότητα:*οι δείκτες πρέπει να είναι αληθείς δηλαδή θα πρέπει να μελετούν αυτό για το οποίο χρησιμοποιούνται, σε αντίθετη περίπτωση υπάρχει σφάλμα και
5.ευαισθησία: πρέπει να αλλάζουν όταν αλλάζει και η κατάσταση στην οποία εφαρμόζονται.

Εν τέλει, η ανάλυση των δεδομένων από τα νοσοκομεία μπορεί να πραγματοποιηθεί με την χρήση κάποιων δεικτών έτσι ώστε να βελτιωθεί η απόδοση τους και η λειτουργία τους. Το τελευταίο στάδιο το οποίο ακολουθεί την ανάλυση δεδομένων είναι η αξιολόγηση των νοσοκομείων που μπορεί να γίνει σε 3 επίπεδα:

- Νοσοκομειακό Επίπεδο (κάθε νοσοκομείο ξεχωριστά).
- Ομάδες Νοσοκομείων (οι ομάδες των νοσοκομείων μπορεί να γίνει είτε ως προς το είδος κάθε νοσοκομείου είτε ως προς το μέγεθος τους δηλαδή το πλήθος των κλινών τους).
- Εθνικό Επίπεδο (όλα τα νοσοκομεία της χώρας μαζί).

Κεφάλαιο 2ο-Εισαγωγή στην R

2.1: Τι είναι η R- Εγκατάσταση

Η R είναι είναι μία γλώσσα προγραμματισμού η οποία χρησιμοποιείται για την ανάλυση δεδομένων και για το σχεδιασμό γραφημάτων. Για την εγκατάσταση της R πρέπει να επισκεφτείτε όχι μόνο την ιστοσελίδα https://www.r-project.org/ και να εγκαταστήσετε την R αλλά και την ιστοσελίδα https://www.rstudio.com/ για την εγκατάσταση του RStudio.

2.2:Γραφικό περιβάλλον της R και ανάγνωση δεδομένων

Αφού γίνει η εγκατάσταση του προγράμματος και ανοίξετε την R θα παρατηρήσετε ότι το γραφικό περιβάλλον της R χωρίζεται σε 4 τμήματα. Το πάνω δεξιά παράθυρο περιλαμβάνει τις εξής καρτέλες:

- *To environment* το οποίο περιλαμβάνει τα σύνολα των δεδομένων (dataset) τα οποία έχουν τις μεταβλητές και τις παρατηρήσεις,
- Το *history* που έχει όλο το ιστορικό των εντολών,
- Τα *tutorials* και *connections*.

Τα κύρια χαρακτηριστικά του κάτω δεξιά παραθύρου είναι τα αρχεία,τα πακέτα αλλά και τα γραφήματα που μπορούν να δημιουργηθούν. Ορισμένα πακέτα (θα αναλυθούν παρακάτω) είναι ήδη εγκατεστημένα ωστόσο υπάρχει περίπτωση κατά τη διάρκεια ανάλυσης δεδομένων να χρειαστείτε πακέτα τα οποία δεν είναι διαθέσιμα. Τα πακέτα αυτά μπορείτε να τα εγκαταστήσετε εύκολα στο σημείο install των πακέτων. Στο αριστερά κάτω παράθυρο μπορείτε να πληκτρολογήσετε τις εκάστοτε εντολές που θέλετε να εκτελέσετε, ενώ στον πάνω αριστερά χώρο εμφανίζονται λεπτομερώς οι μεταβλητές και οι παρατηρήσεις του dataset που επιλέγετε κάθε φορά.

Τα δεδομένα που μπορούν να καταχωρηθούν στην R μπορεί είτε να υπάρχουν ήδη σε κάποια άλλη μορφή (όπως excel, stata, spss) είτε να τα εισάγετε εσείς κατευθείαν στο πρόγραμμα. Αν τα δεδομένα, είναι σε μορφή excel, stata ή spss τότε μπορείτε να

επισκεφτείτε την κορδέλα των εντολών πάνω-πάνω και να επιλέξετε το File και μετά το Import From.

Για την εμφάνιση των δεδομένων χρησιμοποιείται η συνάρτηση view() όπου μέσα στις παρενθέσεις αναγράφεται το dataset του οποίου θέλετε να εμφανίσετε τα δεδομένα. Για να γίνει αυτό απαιτείται στην αρχή να χρησιμοποιηθεί το πακέτο tidyverse με την εντολή library(tidyverse).

2.3:Βοήθεια στην R

Για να λάβετε βοήθεια για κάποια συνάρτηση που δεν γνωρίζετε πληκτρολογήστε help και το (όνομα συνάρτησης) ή ?όνομα συνάρτησης και κάτω αριστερά στο γραφικό περιβάλλον της R στην καρτέλα Help θα εμφανιστούν πληροφορίες για την συνάρτηση.

```
> help(mean)
> ?mean
```

Εικόνα:Help για τη συνάρτηση mean()

2.4:Βασικά στοιχεία της R

2.4.1: Εντολές

Στην R πρέπει να σημειωθεί στο σημείο αυτό ότι υπάρχει διάκριση μεταξύ μικρών και κεφαλαίων γραμμάτων ως εκ τούτο το A και το a είναι δύο διαφορετικά σύμβολα. Ακόμη, η R έχει δύο είδη εντολών:

Α)Εκφράσεις: όπου υπολογίζεται μία τιμή η οποία εμφανίζεται στην οθόνη και στην συνέχεια χάνεται (δεν αποθηκεύεται κάπου).

```
> 4+3
[1] 7
```

Εικόνα:Έκφραση

Β)Εκχωρήσεις: όπου μία έκφραση καταχωρείται σε μία μεταβλητή (με ←) και έτσι διατηρείται στο πάνω δεξιά παράθυρο της R.

```
> a<-4+3
```

Εικόνα:Εκχώρηση

2.4.2:Πακέτα-Εισαγωγή σχολίων

Στην R όλες οι συναρτήσεις είναι αποθηκευμένες σε συγκεκριμένα πακέτα. Για τον λόγο αυτό για να γίνει χρήση μιας συνάρτησης πρέπει να χρησιμοποιηθεί το κατάλληλο πακέτο με την εντολή library και (όνομα πακέτου). Επίσης, στην R μπορούμε να εισάγουμε στον κώδικα σχόλια τοποθετώντας το σύμβολο # πριν εισάγουμε τα σχόλια.

2.4.3:Συναρτήσεις και αντικείμενα

Βασικά στοιχεία της R είναι τα αντικείμενα και οι συναρτήσεις. Τα αντικείμενα στην R μπορεί να είναι μια μεταβλητή για παράδειγμα το a αλλά και ένας ολόκληρος πίνακας με τιμές όπως για παράδειγμα ένα dataset.

```
> library(tidyverse)
> mydata<-tibble(
+ id=c(0,2,4,6)
+ ,sex=c("Male","Male","Male","Female"),
+ var1=c(1,2,3,4),
+ var2=c(NA,NA,4,NA),
+ var3=c(2,NA,NA,NA))
```

Εικόνα:Δημιουργία ενός dataset με συγκεκριμένα στοιχεία

Πέρα από το σύμβολο ← χρησιμοποιούνται συχνά και άλλα σύμβολα που είναι το = , το %>% και το $. Αρχικά το σύμβολο %>% χρησιμοποιείται για εκχωρηθούν αντικείμενα σε συναρτήσεις, ενώ το $ για να επιλεχθεί μια συγκεκριμένη στήλη από έναν πίνακα δεδομένων (dataset). Τέλος, το σύμβολο = χρησιμοποιείται για να γίνει ο καθορισμός ενός ορίσματος σε μία συνάρτηση.

```
>
> mydata$sex
[1] "Male"   "Male"   "Male"   "Female"
>
```

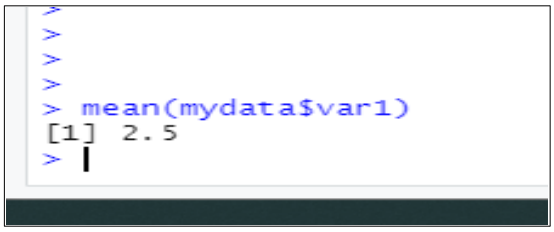
Εικόνα:Επιλέγεται να εμφανιστεί η στήλη sex-φύλο από το mydata με το σύμβολο $

Όπως αναφέρθηκε και στην αρχή η R πέρα από τα αντικείμενα έχει και συναρτήσεις. Η συνάρτηση στην R όπως και σε άλλες γλώσσες προγραμματισμού έχει την είσοδο που περικλείεται από () και μία έξοδο.

Χρήσιμες συναρτήσεις της R:

Mean():

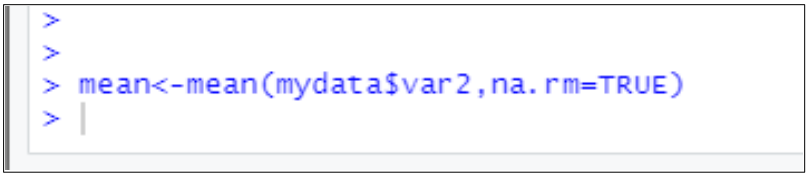
Μια κοινή συνάρτηση που χρησιμοποιείται συχνά είναι η συνάρτηση της μέσης τιμής mean(). Έτσι αν θέλουμε να βρούμε τη μέση τιμή της στήλης var1του mydata τότε χρησιμοποιούμε την εξής εντολή:



Εικόνα:Μέση τιμή της var1 του mydata

Na.rm():

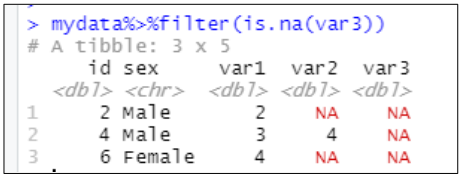
Η μέση τιμή της μεταβλητής var2 η οποία έχει missing values(NA) είναι ίση με NA. Για να αποφύγετε το πρόβλημα αυτό εισάγετε στη συνάρτηση mean το na.rm=TRUE έτσι ώστε να υπολογιστεί η μέση τιμή των τιμών χωρίς να συμπεριληφθούν οι missing values(NA).



Εικόνα:Μέση τιμή για το var2 που έχει missing values

Is.na():

Συχνό πρόβλημα των χρηστών της R είναι η διαχείριση των missing values (NA). Για να φιλτράρουμε τις τιμές αυτές χρησιμοποιούμε την συνάρτηση is.na().



Εικόνα:Εμφάνιση μόνο των παρατηρήσεων(γραμμών) που στο var3 υπάρχουν missing values

Max-Min():

Αντίστοιχες συναρτήσεις με την mean() είναι η max() και η min() οι οποίες υπολογίζουν τη μέγιστη και την ελάχιστη τιμή αντίστοιχα.

```
>
> max(mydata$id)
[1] 6
> min(mydata$id)
[1] 0
>
```

Εικόνα:Μέγιστη και ελάχιστη τιμή του id

Seq():

Άλλη μια χρήσιμη συνάρτηση είναι η seq() η οποία παίρνει σαν όρισμα ένα εύρος αριθμών και εμφανίζει όλους τους αριθμούς που περιλαμβάνονται στο εύρος αυτό.

```
> example<-seq(1,10)
> example
 [1]  1  2  3  4  5  6  7  8  9 10
```

Εικόνα:Εμφάνιση των αριθμών από το 1 μέχρι το 10

C():

Άλλη μια συνάρτηση είναι η c() που συνδυάζει διάφορες τιμές και επιστρέφει ένα διάνυσμα.

```
> k<-c(1,2)
> k
[1] 1 2
>
```

Εικόνα:Η συνάρτηση c()

Nrow():

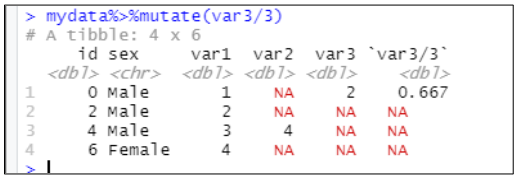
Ακόμα η συνάρτηση nrow() εμφανίζει το πλήθος των γραμμών.

```
> nrow(mydata)
[1] 4
```

Εικόνα:Εμφάνιση του πλήθους των γραμμών του mydata

Mutate():

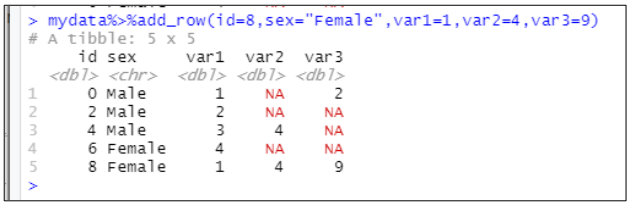
Μια σημαντική λειτουργία της R είναι η πρόσθεση νέων στηλών στον πίνακα των δεδομένων. Για να προσθέσουμε λοιπόν νέες στήλες ή για να τροποποιήσουμε τις ήδη υπάρχουσες χρησιμοποιούμε την συνάρτηση mutate().



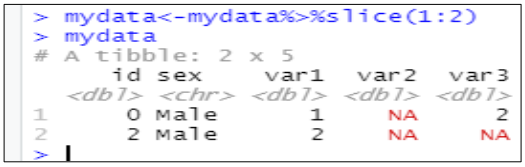
Εικόνα:Δημιουργία μιας νέας στήλης var3/3

Add_row()-Slice():

Πέρα από την προσθήκη στηλών μπορούμε στην R να προσθέσουμε μια νέα γραμμή δηλαδή μια νέα παρατήρηση με την εντολή add_row. Επίσης με τη συνάρτηση slice() επιτυγχάνεται η εμφάνιση ορισμένων γραμμών από τον πίνακα δεδομένων:



Εικόνα:Εισαγωγή νέας γραμμής με id=8,sex=Female,var1=1,var=4 και var3=9



Εικόνα:Αφαιρούνται οι γραμμές 3 και 4

Select()-paste():

Με τη συνάρτηση select() εμφανίζονται οι στήλες που εισάγουμε μέσα στην παρένθεση,ενώ με τη συνάρτηση paste() τοποθετούνται οι χαρακτήρες μαζί.

```
> mydata%>%mutate(plot_label= paste(id,"the sex is",sex,"and the values were'
var1,var2,var3))%>%select(plot_label)
# A tibble: 2 x 1
  plot_label
  <chr>
1 0 the sex is Male and the values were 1 NA 2
2 2 the sex is Male and the values were 2 NA NA
>
```

Εικόνα:Η χρήση των εντολών select-paste

Η συνάρτηση select() χρησιμοποιείται όχι μόνο για την εμφάνιση συγκεκριμένων στηλών αλλά και για την μετονομασία και την αλλαγή σειράς των στηλών.

```
> mydata%>%select(id,var3=var2)
# A tibble: 4 x 2
     id  var3
  <dbl> <dbl>
1     0    NA
2     2    NA
3     4     4
4     6    NA
```

Εικόνα:Μετονομασία του var2 σε var3

Join():

Επιπροσθέτως, στην R μπορούν να συνδυαστούν δεδομένα που βρίσκονται σε διαφορετικούς πίνακες. Αυτό επιτυγχάνεται με την συνάρτηση join().

```
> mydata1<-tibble(id=c(1,3,5,7),sex=c("Male","Female","Female","Female"),var1=
c(3,2,4,5),var2=c(NA,4,5,0),var3=c(NA,NA,1,1))
> mydata1
# A tibble: 4 x 5
     id sex     var1  var2  var3
  <dbl> <chr>  <dbl> <dbl> <dbl>
1     1 Male       3    NA    NA
2     3 Female     2     4    NA
3     5 Female     4     5     1
4     7 Female     5     0     1
>
```

Εικόνα:Νέος πίνακας δεδομένων mydata1

Για την εμφάνιση όλων των στοιχείων του mydata και του mydata1 χρησιμοποιείται η συνάρτηση full_join().

```
> full_join(mydata,mydata1)
Joining, by = c("id", "sex", "var1", "var2", "var3")
# A tibble: 8 x 5
     id sex     var1  var2  var3
  <dbl> <chr>  <dbl> <dbl> <dbl>
1     0 Male       1    NA     2
2     2 Male       2    NA    NA
3     4 Male       3     4    NA
4     6 Female     4    NA    NA
5     1 Male       3    NA    NA
6     3 Female     2     4    NA
7     5 Female     4     5     1
8     7 Female     5     0     1
```

Εικόνα:Η εντολή full join

Αν θέλουμε να εμφανίσουμε μόνο τα κοινά τους στοιχεία με βάση το id χρησιμοποιούμε την συνάρτηση inner_join().

```
> inner_join(mydata,mydata1)
Joining, by = c("id", "sex", "var1", "var2", "var3")
# A tibble: 0 x 5
# ... with 5 variables: id <dbl>, sex <chr>, var1 <dbl>, var2 <dbl>,
#   var3 <dbl>
>
```

Εικόνα: Η εντολή inner join(δεν έχουν κοινά στοιχεία)

Τέλος για την εμφάνιση των στοιχείων μόνο ενός πίνακα είτε του αριστερά είτε του δεξιά χρησιμοποιείται το left_join() ή το right_join() αντίστοιχα.

```
> left_join(mydata,mydata1)
Joining, by = c("id", "sex", "var1", "var2", "var3")
# A tibble: 4 x 5
     id sex     var1  var2  var3
  <dbl> <chr>  <dbl> <dbl> <dbl>
1     0 Male       1    NA     2
2     2 Male       2    NA    NA
3     4 Male       3     4    NA
4     6 Female     4    NA    NA
>
```

Εικόνα:Η εντολή left join

Η R πέρα από αντικείμενα και συναρτήσεις έχει και πολλούς τελεστές οι οποίοι χρησιμοποιούνται για να συνδυαστούν διάφορα αντικείμενα ή δεδομένα. Αρχικά υπάρχουν οι συγκριτικοί τελεστές που είναι οι ==, <, >, <=, >= με τους οποίους ελέγχεται αν μια τιμή είναι ίση, μικρότερη, μεγαλύτερη, μικρότερη ή ίση και μεγαλύτερη ίση από μια άλλη τιμή. Οι τελεστές αυτοί χρησιμοποιούνται μέσα στη συνάρτηση filter.

```
> mydata%>%filter(id<4)
# A tibble: 2 x 5
     id sex     var1  var2  var3
  <dbl> <chr>  <dbl> <dbl> <dbl>
1     0 Male       1    NA     2
2     2 Male       2    NA    NA
>
```

Εικόνα:Εμφάνιση των παρατηρήσεων που το id είναι μικρότερο από 4

Η R επίσης έχει και δύο λογικούς τελεστές το AND(&),και το OR(|) οι οποίοι μπορούν να συνδυάσουν δεδομένα.

```
> mydata%>%filter(var1>3 & id>2)
# A tibble: 1 x 5
     id sex     var1  var2  var3
  <dbl> <chr>  <dbl> <dbl> <dbl>
1     6 Female     4    NA    NA
>
```

Εικόνα:Εμφάνιση των παρατηρήσεων όταν το var1>3 και το id>2

Τέλος, η R υποστηρίζει και συνθήκες ελέγχου με τις εντολές if-else.

```
> mydata%>%mutate(above_threshold = if_else(var1>2,"Above two","Below two"))
# A tibble: 4 x 6
     id sex     var1  var2  var3 above_threshold
  <dbl> <chr>  <dbl> <dbl> <dbl> <chr>
1     0 Male       1    NA     2 Below two
2     2 Male       2    NA    NA Below two
3     4 Male       3     4    NA Above two
4     6 Female     4    NA    NA Above two
> |
```

Εικόνα:Δημιουργία νέας στήλης αν το var1>2 τότε γραφεί Above two αλλιώς Below two

2.4.4:Τύποι δεδομένων

Στην R μπορούν να χρησιμοποιηθούν τρεις τύποι δεδομένων που είναι οι παρακάτω:

1 τα συνεχή δεδομένα όπως είναι ακέραιοι αριθμοί και οι δεκαδικοί,
2 τα κατηγορικά δεδομένα όπως είναι οι χαρακτήρες αλλά και οι λογικοί τελεστές true/false και
3 τα δεδομένα που αφορούν ημερομηνία και ώρα.

Σε κάθε στήλη του πίνακα είναι απαραίτητο να περιλαμβάνονται δεδομένα ίδιου τύπου, ωστόσο ένας πίνακας δεδομένων μπορεί να περιέχει στήλες διαφορετικού τύπου. Με την εντολή read η R μας επιστρέφει τον τύπο δεδομένων που περιλαμβάνει κάθε στήλη του πίνακα.

Τα *συνεχή δεδομένα* μπορεί να είναι ακέραιοι αριθμοί (integer), δηλαδή αριθμοί χωρίς δεκαδικό μέρος. Αξίζει να σημειωθεί ότι το πλήθος των δεκαδικών ψηφίων σε ένα *δεκαδικό αριθμό* που μπορεί να εμφανίσει η R είναι έξι, ωστόσο αυτό δεν σημαίνει ότι δεν υπάρχουν άλλα ψηφία στον αριθμό αυτόν.

Τα *κατηγορικά δεδομένα* μπορεί να περιλαμβάνουν χαρακτήρες, γράμματα αλλά και ολόκληρη πρόταση τα οποία εμπεριέχονται σε “ ” ή ‘ ’.

Εκτός από συνεχή και κατηγορικά δεδομένα,η R μπορεί να αναγνωρίσει και *ημερομηνίες και ώρες*. Με την συνάρτηση Sys.time() εμφανίζεται η τρέχουσα ώρα και η ημερομηνία του υπολογιστή.

```
> current_date<-Sys.time()
> current_date
[1] "2020-11-25 15:47:08 +03"
```

Εικόνα:Εμφάνιση της τρέχουσας ημερομηνίας και ώρας

Με την εντολή *my_datetime ← "2020-11-9 11:00"* εκχωρείται η συγκεκριμένη ημερομηνία και η ώρα στη μεταβλητή my_datetime.

```
> my_datetime <- "2020-11-9 11:00"
```

Εικόνα:Εμφάνιση συγκεκριμένης ημερομηνίας και ώρας

Ωστόσο στην R όταν εκχωρείται μια συγκεκριμένη ημερομηνία και ώρα την αντιλαμβάνεται ως χαρακτήρα και όχι σαν ημερομηνία. Για τον λόγο αυτό, χρησιμοποιείται η συνάρτηση ymd_hm (απαιτείται η βιβλιοθήκη lubridate) η οποία δείχνει στην R ότι το my_datetime περιέχει ημερομηνία και ώρα και όχι χαρακτήρες.

```
> library(lubridate)
> my_datetime_converted <- ymd_hm(my_datetime)
> my_datetime_converted
[1] "2020-11-09 11:00:00 UTC"
```

Εικόνα:Χρήση της εντολής ymd_hm

2.5:Ομαδοποίηση δεδομένων

Στο σημείο αυτό θα αναλυθούν διάφορες συναρτήσεις που χρησιμοποιούνται στην R για τη ομαδοποίηση των δεδομένων. Η πρώτη συνάρτηση που θα εξετασθεί είναι η *summarise()*. Η *summarise()* μοιάζει με την συνάρτηση *sum()* ωστόσο η *sum()* βρίσκει το άθροισμα των τιμών που τοποθετούνται σαν όρισμα ενώ η *summarise()* δημιουργεί νέο πίνακα με τα αθροίσματα. Για την ανάλυση των συναρτήσεων αυτών χρησιμοποιήθηκε το προηγούμενο dataset.

```
>
> mydata$var1%>%sum()
[1] 10
> mydata%>%summarise(sum(var1))
# A tibble: 1 x 1
  `sum(var1)`
        <dbl>
1          10
>
```

Εικόνα:Η συνάρτηση sum και summarise()

Πολύ συχνά η summarise() χρησιμοποιείται μαζί με τη συνάρτηση group by(). Τοποθετώντας την group by() καθορίζεται στην summarise() σε ποια ή ποιες υποομάδες θα γίνουν οι υπολογισμοί. Έστω ότι το group_by() γίνεται με βάση το φύλο(sex):

```
> mydata%>%group_by(sex)%>%summarise(sum(var1))
`summarise()` ungrouping output (override with `.groups` argument)
# A tibble: 2 x 2
  sex    `sum(var1)`
  <chr>        <dbl>
1 Female           4
2 Male             6
>
```

Εικόνα:Η συνάρτηση summarise() με την group by()

Ακόμα μία συνάρτηση είναι η *percent()* που υπολογίζει το ποσοστό. Η συνάρτηση αυτή απαιτεί την βιβλιοθήκη *scales*.

```
> mydata<-mydata%>%mutate(per=percent(var1/var2))
> mydata
# A tibble: 4 x 6
     id sex     var1  var2  var3 per
  <dbl> <chr>  <dbl> <dbl> <dbl> <chr>
1     0 Male       1    NA     2 NA
2     2 Male       2    NA    NA NA
3     4 Male       3     4    NA 75%
4     6 Female     4    NA    NA NA
```

Εικόνα:Η συνάρτηση percent()

Στην R οι πιο συχνές αριθμητικές συναρτήσεις πέρα από την *sum()* που είδαμε προηγουμένως είναι η *median(), η mean() και η sd()*.

```
> examples <- c(10, 20, 7)
```

Εικόνα:Η μεταβλητή examples έχει τις τιμές 10,20 και 7

Mean():

```
> mean(examples)
[1] 12.33333
```

Εικόνα:Η μέση τιμή είναι 12.333

Median():

```
> median(examples)
[1] 10
```

Εικόνα:Ο μέσος αριθμός των 7,10,20 είναι το 10

Sd():

```
> sd(examples)
[1] 6.806859
```

Εικόνα:Τυπική απόκλιση

Μία ακόμα συνάρτηση είναι η *arrange()* με την οποία εμφανίζονται κατά αύξουσα σειρά:

```
> mydata%>%arrange(sex)
# A tibble: 4 x 6
     id sex     var1  var2  var3 per
  <dbl> <chr>  <dbl> <dbl> <dbl> <chr>
1     6 Female     4    NA    NA NA
2     0 Male       1    NA     2 NA
3     2 Male       2    NA    NA NA
4     4 Male       3     4    NA 75%
> |
```

Εικόνα:Η συνάρτηση arrange

Για την εμφάνιση των κατηγορικών δεδομένων κατά ελαττωμένη σειρά χρησιμοποιείται πέρα από την συνάρτηση *arrange()* και η συνάρτηση *desc()* :

```
> mydata%>%arrange(desc(sex))
# A tibble: 4 x 6
     id sex     var1  var2  var3 per
  <dbl> <chr>  <dbl> <dbl> <dbl> <chr>
1     0 Male       1    NA     2 NA
2     2 Male       2    NA    NA NA
3     4 Male       3     4    NA 75%
4     6 Female     4    NA    NA NA
>
```

Εικόνα:Χρήση των συναρτήσεων arrange() και desc()

Η τελευταία συνάρτηση του κεφαλαίου αυτού είναι η *levels()* η οποία εμφανίζει τα στάδια της εκάστοτε κατηγορικής μεταβλητής.

```
>
>
>
> x <- factor(c("male", "male", "male", "female"));
> levels(x)
[1] "female" "male"
>
```

Εικόνα:Συνάρτηση levels()

2.6:Δημιουργία διαγραμμάτων

Υπάρχουν διάφορα πακέτα που χρησιμοποιούνται για τη δημιουργία διαγραμμάτων ωστόσο αυτό που είναι πιο σύνηθες είναι το ggplot. Το συγκεκριμένο πακέτο περιέχει μόνο δύο στοιχεία, το ένα είναι οι μεταβλητές που αναλύει και το άλλος είναι ο τρόπος.

Για το κεφάλαιο αυτό χρησιμοποιείται η συλλογή δεδομένων gapminder (*R for Health Data ScienceEwen Harrison and Riinu Pius 2021-01-15*) η οποία υπάρχει ήδη μέσα στην R για αυτό αρκεί να γίνει η χρήση της βιβλιοθήκης library (gapminder). Ο πίνακας αυτός έχει 6 μεταβλητές (στήλες) και 1704 παρατηρήσεις (γραμμές). Άρα,

```
> library(gapminder)
> gapminder
# A tibble: 1,704 x 6
   country     continent  year lifeExp      pop gdpPercap
   <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
 1 Afghanistan Asia       1952    28.8  8425333      779.
 2 Afghanistan Asia       1957    30.3  9240934      821.
 3 Afghanistan Asia       1962    32.0 10267083      853.
 4 Afghanistan Asia       1967    34.0 11537966      836.
 5 Afghanistan Asia       1972    36.1 13079460      740.
 6 Afghanistan Asia       1977    38.4 14880372      786.
 7 Afghanistan Asia       1982    39.9 12881816      978.
 8 Afghanistan Asia       1987    40.8 13867957      852.
 9 Afghanistan Asia       1992    41.7 16317921      649.
10 Afghanistan Asia       1997    41.8 22227415      635.
# ... with 1,694 more rows
>
```

Εικόνα:Το σύνολο δεδομένων gapminder

Η δημιουργία γραφημάτων χωρίζεται σε 2 μέρη:

1 *Καθορισμός των μεταβλητών του γραφήματος δηλαδή το x και το y*. Η συνάρτηση aes είναι μία συνάρτηση μέσα στην οποία μπαίνουν πάντα οι μεταβλητές που απεικονίζονται.
2 Στο δεύτερο βήμα καθορίζεται *ο τρόπος απεικόνισης των δεδομένων*.

Διάγραμμα 1ο-Geom_point():

Με βάση την προηγούμενη θεωρία δημιουργούμε ένα νέο πίνακα με τα στοιχεία που αφορούν μόνο το 2002.

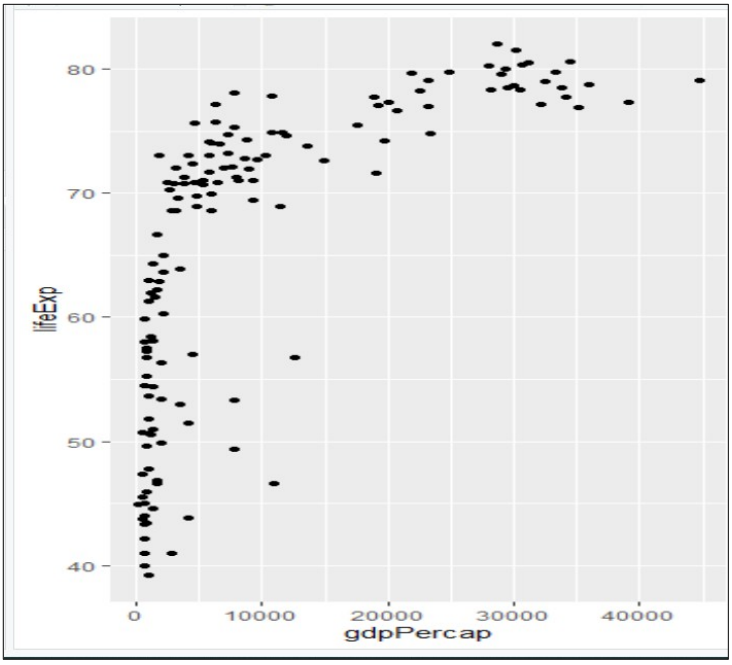
```
> gapminder2002<-gapminder%>%filter(year==2002)
> gapminder2002
# A tibble: 142 x 6
   country     continent  year lifeExp       pop gdpPercap
   <fct>       <fct>     <int>   <dbl>     <int>     <dbl>
 1 Afghanistan Asia       2002    42.1  25268405       727.
 2 Albania     Europe     2002    75.7   3508512      4604.
 3 Algeria     Africa     2002    71.0  31287142      5288.
 4 Angola      Africa     2002    41.0  10866106      2773.
 5 Argentina   Americas   2002    74.3  38331121      8798.
 6 Australia   Oceania    2002    80.4  19546792     30688.
 7 Austria     Europe     2002    79.0   8148312     32418.
 8 Bahrain     Asia       2002    74.8    656397     23404.
 9 Bangladesh  Asia       2002    62.0 135656790      1136.
10 Belgium     Europe     2002    78.3  10311970     30486.
# ... with 132 more rows
> |
```

Εικόνα:Δεδομένα για το 2002

Θεωρούμε ότι το y=lifeExp. το x=gdpPercap και ότι ο τρόπος απεικόνισης είναι geom_point():

```
>
> gapminder2002%>%ggplot(aes(x=gdpPercap,y=lifeExp))+geom_point()
>
```

Εικόνα:Απεικόνιση των δεδομένων μέσω της συνάρτησης geom_point()

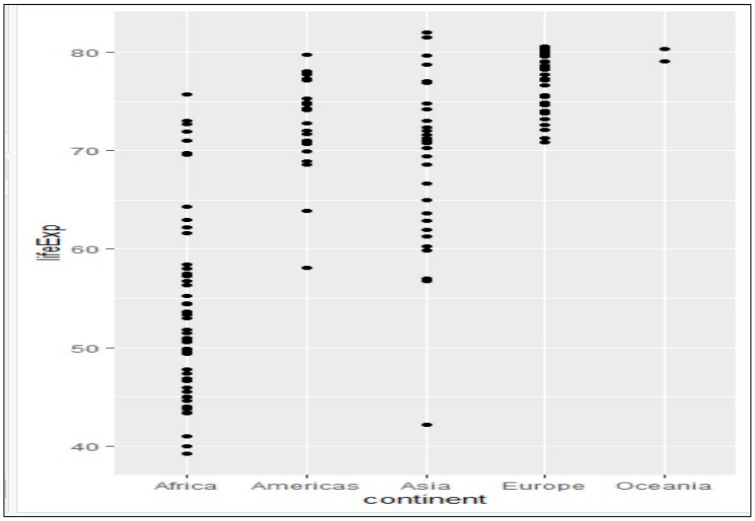


Εικόνα:Τελικό αποτέλεσμα

Ένα διάγραμμα geom_point() μπορεί να έχει όμως και μια συνεχή μεταβλητή (lifeExp) και μία κατηγορική (continent):

```
> gapminder2002%>%ggplot(aes(x=continent,y=lifeExp))+geom_point()
>
```

Εικόνα:Καθορισμός μεταβλητών μαζί με τον τρόπο απεικόνισης



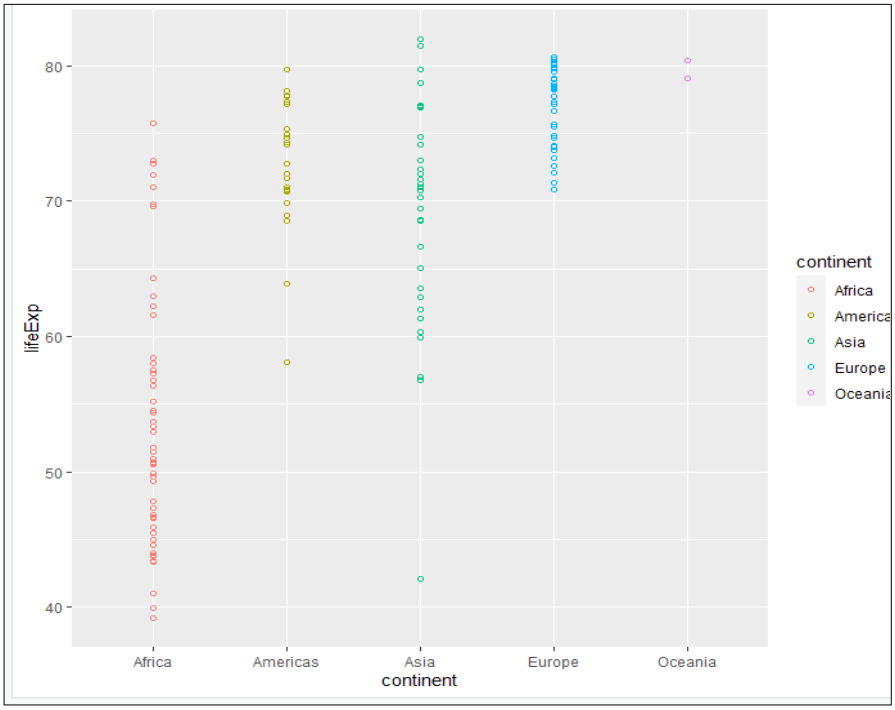
Εικόνα:Τελικό αποτέλεσμα

Στο πρώτο διάγραμμα κάθε κουκκίδα απεικονίζει μια διαφορετική παρατήρηση όπως και σε αυτό το διάγραμμα. Ωστόσο παρατηρείτε ότι οι κουκκίδες είναι κάθετες διότι η μεταβλητή x είναι κατηγορική.

Οι χρήστες μπορούν να αλλάξουν το χρώμα κάθε κουκκίδας τοποθετώντας την εντολή (colour) μέσα στην συνάρτηση aes() αλλά και το σχήμα με την εντολή shape μέσα στην συνάρτηση geom_point() :

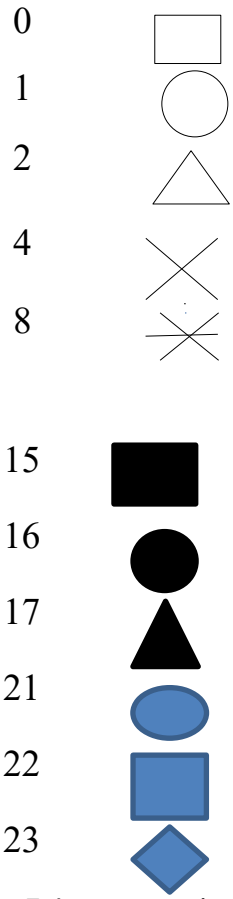
```
>
>
> gapminder2002%>%ggplot(aes(x=continent,y=lifeExp,colour=continent))+geom_point(shape=1)
>
```

Εικόνα:Εντολή colour και shape



Εικόνα:Τελικό αποτέλεσμα

Παρακάτω παρατίθεται ο πίνακας με τα σχήματα που περιλαμβάνει η R :

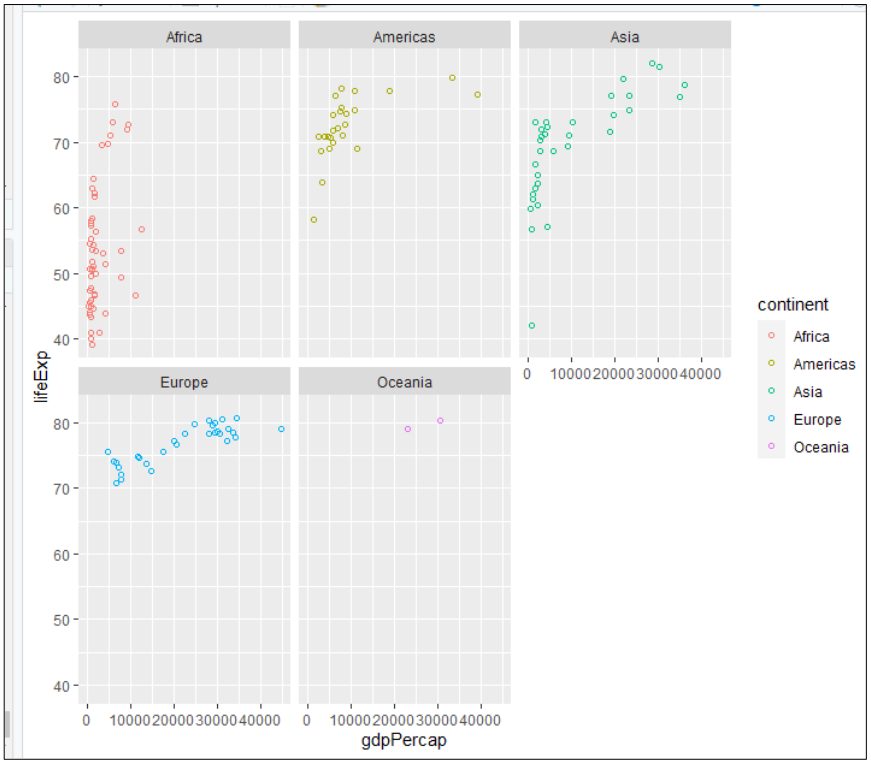


Πίνακας:Σχήματα που περιλαμβάνει η R

Τα διαγράμματα μπορούν να χωριστούν σε μικρότερα. Πιο συγκεκριμένα στο παράδειγμα αυτό μπορεί να χωριστεί το διάγραμμα σε μικρότερα με βάση την ήπειρο με τη συνάρτηση facet_wrap.

```
>
>
>
> gapminder2002%>%ggplot(aes(x=gdpPercap,y=lifeExp,colour=continent))+geom_point(shape=1)+facet_wrap(~continent)
>
```

Εικόνα:Διαχωρισμός διαγραμμάτων με βάση την ήπειρο



Εικόνα:Τελικό αποτέλεσμα

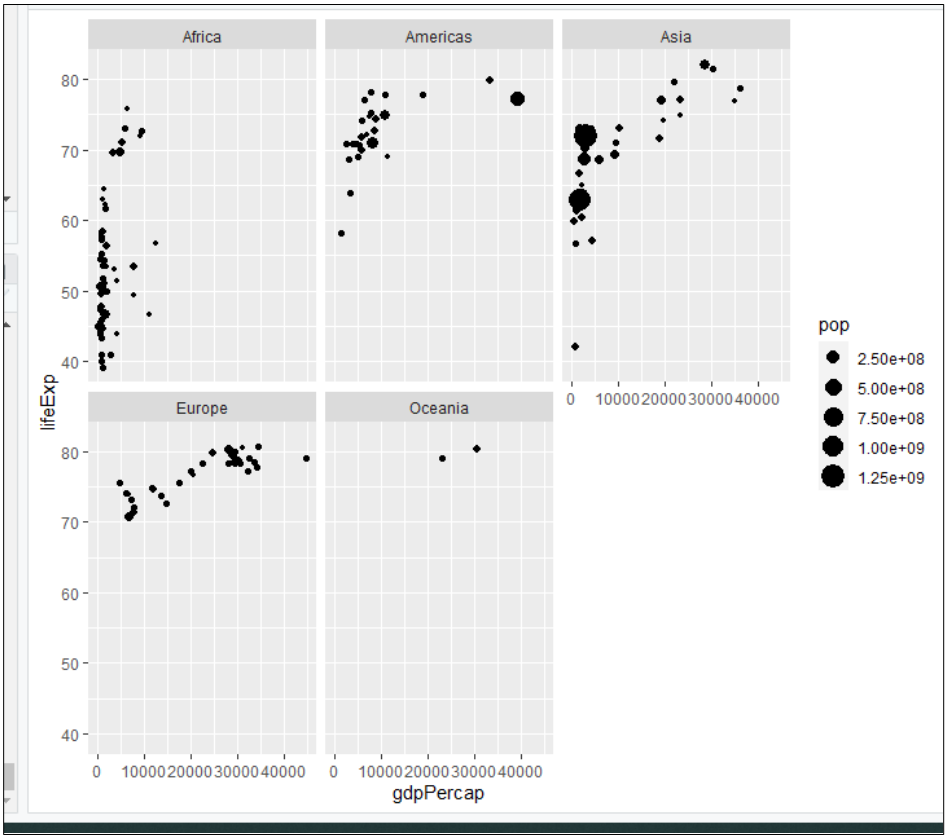
Τέλος, με τις εντολές theme_bw(), theme_dark() και theme classic() αλλάζει η μορφολογία του διαγράμματος που δημιουργείται.

Διάγραμμα 2ο:Bubble plots

Άλλο ένα διάγραμμα είναι το διάγραμμα φυσαλίδων ή αλλιώς τα bubble plots. Για να δημιουργηθεί το διάγραμμα αυτό χρησιμοποιείται η εντολή size=pop μέσα στη συνάρτηση ggplot():

```
>
>
>
>
>
> gapminder2002%>%ggplot(aes(x=gdpPercap,y=lifeExp,size=pop))+geom_point()+facet_wrap(~continent)
>
```

Εικόνα:Η εντολή size

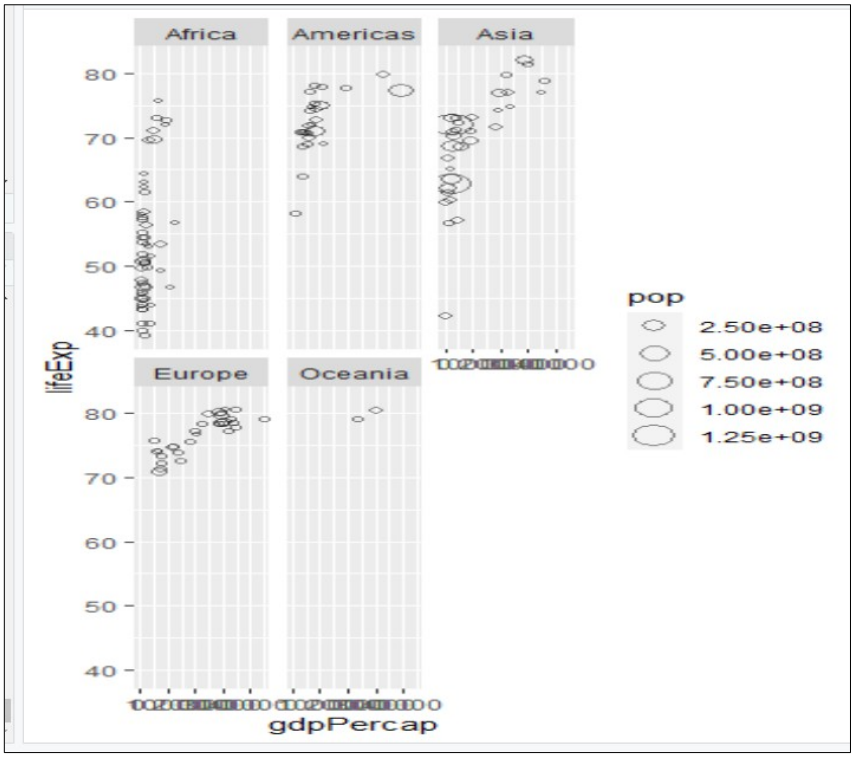


Εικόνα:Bubble plot

Ωστόσο παρατηρείτε ότι σε κάποια σημεία υπάρχουν φυσαλίδες πάνω σε φυσαλίδες, για να αντιμετωπιστεί το φαινόμενο αυτό τοποθετούμε το alpha=0.5 (οι τιμές κυμαίνονται από το 0 όπου οι φυσαλίδες είναι τελείως διαφανείς μέχρι το 1 που δεν είναι καθόλου διαφανείς) έτσι ώστε οι φυσαλίδες να είναι διαφανείς και έτσι να είναι πιο ευδιάκριτες στους χρήστες:

```
>
>
>
>
>
> gapminder2002%>%ggplot(aes(x=gdpPercap,y=lifeExp,size=pop))+geom_point(shape=1,alpha=0.5)+facet_wrap(~continent)
>
```

Εικόνα:Εντολή alpha



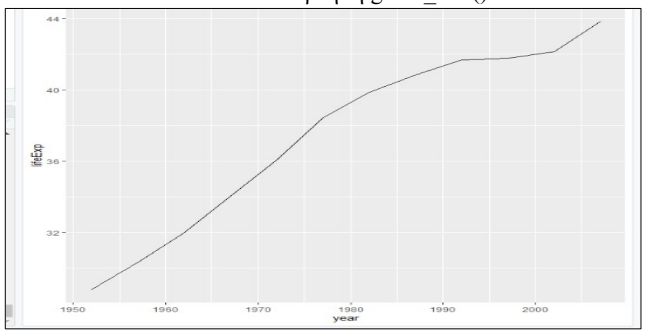
Εικόνα:Bubble plot

Διάγραμμα 3ο:Line plots

Ένα ακόμα διάγραμμα που μπορούμε να κάνουμε στην R είναι το line plot στο οποίο σχηματίζεται μια συνεχής γραμμή. Πιο συγκεκριμένα, αν θέλουμε να εμφανίσουμε το προσδόκιμο ζωής στο Αφγανιστάν από το 1950 μέχρι το 2000 με συνεχή γραμμή πληκτρολογούμε τις παρακάτω εντολές:

```
>
>
>
> gapminder%>%filter(country == "Afghanistan") %>% ggplot(aes(x=year,y=li
feExp))+geom_line()
>
```

Εικόνα:Η συνάρτηση geom_line()



Εικόνα:Line plot

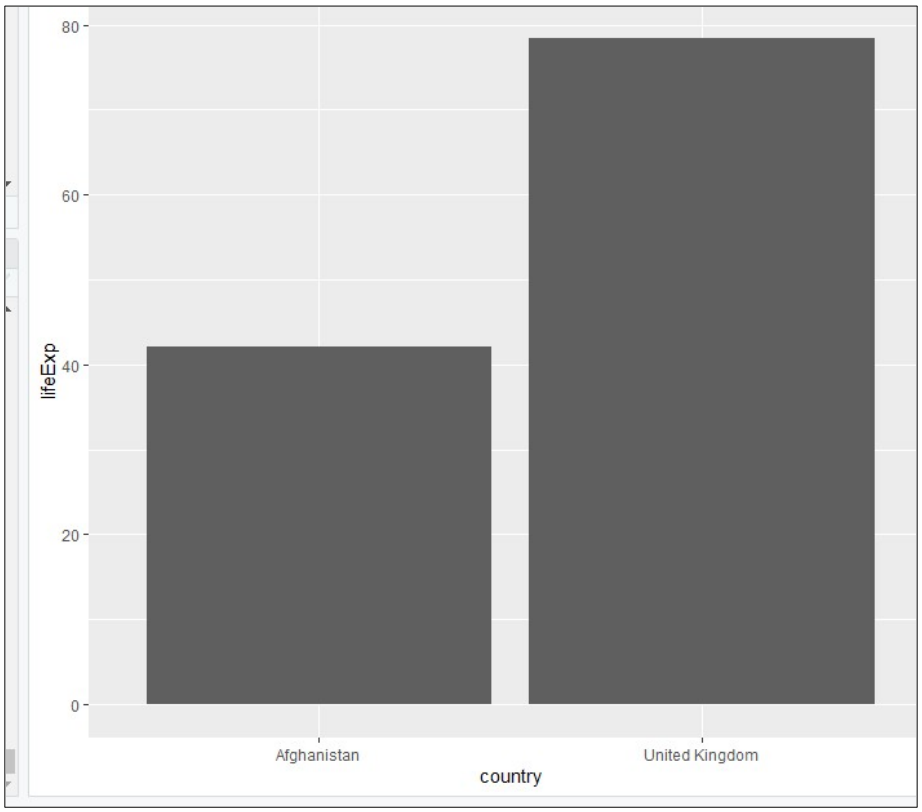
Διάγραμμα 4ο: Bar-plots

Πέρα από αυτά τα διαγράμματα υπάρχουν και τα bar-plots. Υπάρχουν δύο είδη bar-plots, το geom-col() και το geom-bar().Το geom-col() χρησιμοποιείται όταν θέλουμε να βρούμε την συσχέτιση δυο μεταβλητών ενώ το geom-bar() όταν θέλουμε να γίνει καταμέτρηση του πλήθους των γραμμών από το σύνολο των δεδομένων. Ας δούμε όμως πιο αναλυτικά τα δύο αυτά διαγράμματα.

Το geom-col() χρειάζεται δύο μεταβλητές την x που είναι κατηγορική μεταβλητή και την y που είναι συνεχής. Παρακάτω ακολουθεί ένα παράδειγμα για το Ηνωμένο Βασίλειο, το Αφγανιστάν και το προσδόκιμο ζωής τους:

```
>
>
>
>
>
> gapminder2002%>%filter(country %in% c("Afghanistan","United Kingdom"))%>%ggplot(aes(x=country,lifeExp))+geom_col()
>
```

Εικόνα:Η συνάρτηση geom_col()

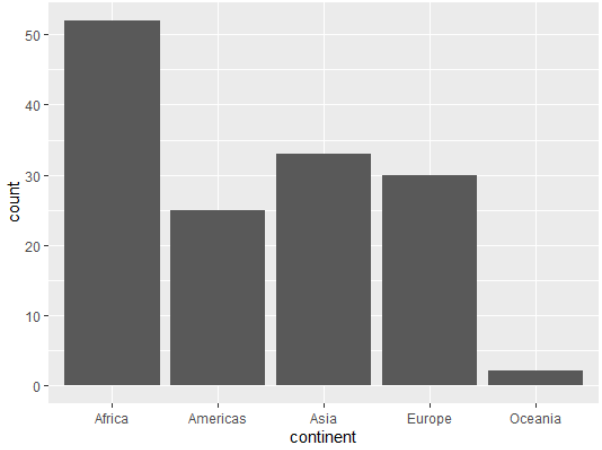


Εικόνα:Διάγραμμα geom-col()

Από την άλλη πλευρά, το geom-bar() χρειάζεται μία μόνο μεταβλητή την x που είναι κατηγορική μεταβλητή. Έστω ότι θέλουμε να συμπεριλάβουμε το άθροισμα όλων των ηπείρων.

```
>
> gapminder2002%>%ggplot(aes(x=continent))+geom_bar()
>
```

Εικόνα:Η συνάρτηση geom_bar()

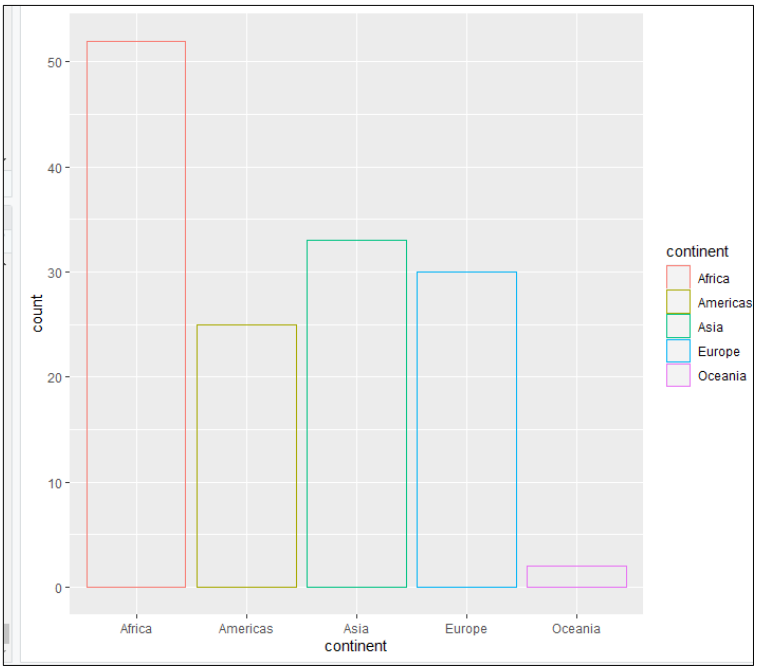


Εικόνα:Διάγραμμα geom-bar()

Στο παρακάτω διάγραμμα κάθε χώρα απεικονίζεται με διαφορετικό χρώμα:

```
>
>
>
> gapminder2002%>%ggplot(aes(x=continent,colour=continent))+geom_bar(fill=NA)
>
```

Εικόνα:Εμφάνιση αποτελεσμάτων με διαφορετικό χρώμα



Εικόνα:Τελικό αποτέλεσμα

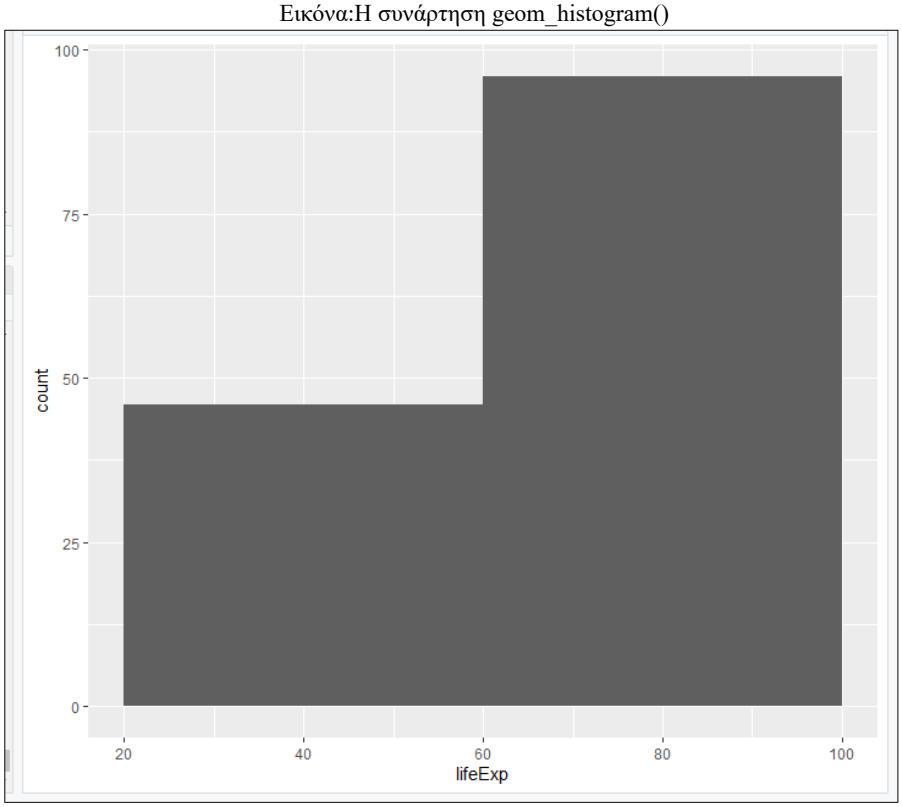
Στο σημείο αυτό πρέπει να τονισθεί η διαφορά ανάμεσα στη συνάρτηση fill() και στη συνάρτηση color(). Στην συνάρτηση color() τοποθετείται χρώμα στα όρια των διαγραμμάτων ενώ με τη συνάρτηση fill() τοποθετείται χρώμα στο εσωτερικό κάθε διαγράμματος.

Διάγραμμα 5ο:Histogram

Άλλο ένα διάγραμμα που υπάρχει στην R είναι το ιστόγραμμα στο οποίο απεικονίζεται η κατανομή των τιμών μιας συνεχής μεταβλητής. Στο παρακάτω παράδειγμα η μεταβλητή x είναι το προσδόκιμο ζωής και δημιουργείται ένα ιστόγραμμα ανά δέκα έτη:

```
>
> gapminder2002%>%ggplot(aes(x=lifeExp))+geom_histogram(binwidth=40)
>
```

Εικόνα:Η συνάρτηση geom_histogram()



Εικόνα:Ιστόγραμμα

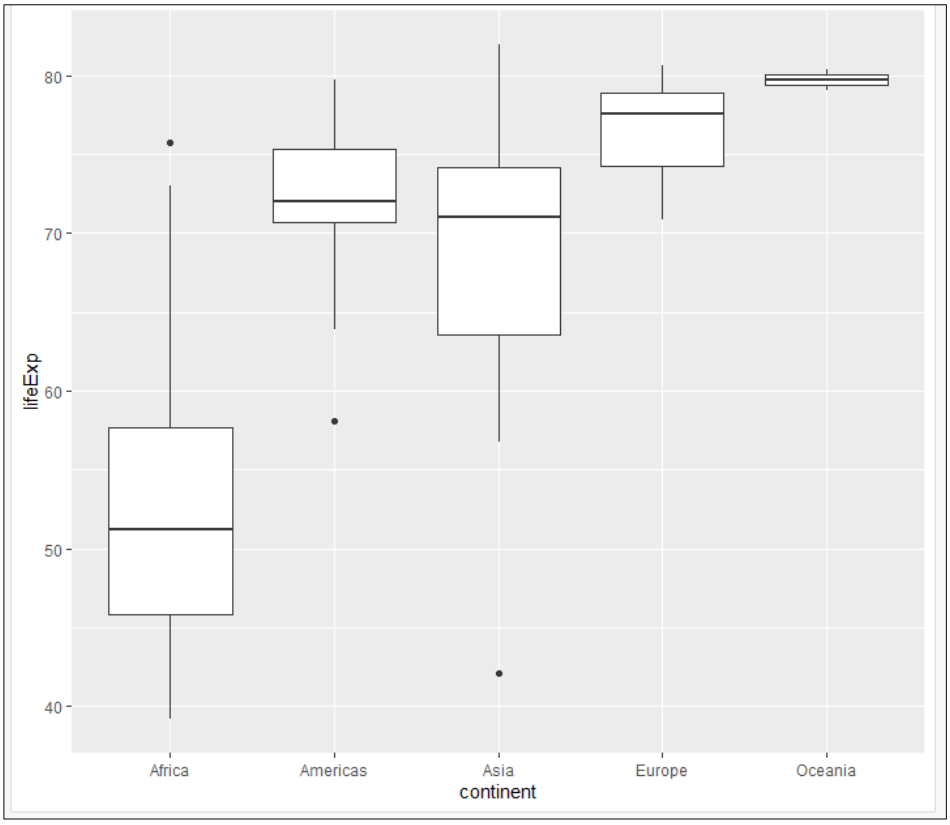
Διάγραμμα 6ο:Box plots

Το τελευταίο διάγραμμα που θα δούμε στην R είναι τα θηκογράμματα (box-plots) στα οποία η μεταβλητή εξόδου είναι συνεχής. Τα box-plots περιέχουν την μέση τιμή που είναι η μεσαία γραμμή σε κάθε θηκόγραμμα,την ελάχιστη τιμή, το 1ο και το 3ο τεταρτημόριο, τη μέγιστη

τιμή αλλά και πιθανές ακραίες τιμές (outliers). Ακολουθεί παρακάτω ένα παράδειγμα θηκογράμματος όπου το x είναι οι ήπειροι και το y είναι το προσδόκιμο ζωής:

```
>
>
>
> gapminder2002%>%ggplot(aes(x=continent,y=lifeExp))+geom_boxplot()
>
```

Εικόνα:Η συνάρτηση geom_boxplot()



Εικόνα:Το διάγραμμα box-plots

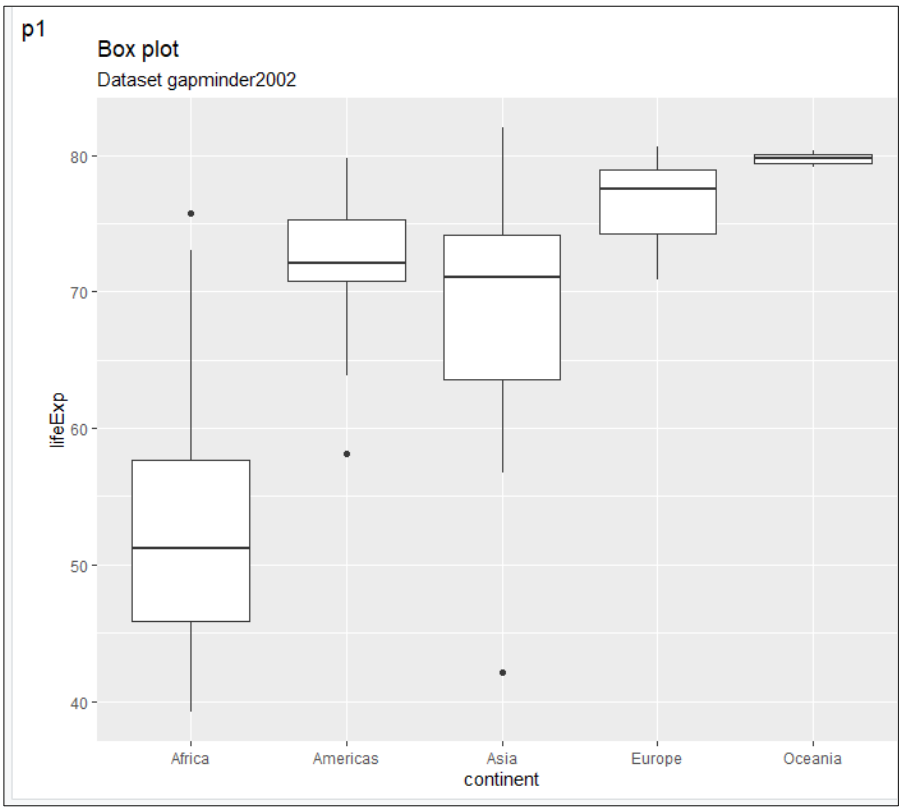
Όλα αυτά τα διαγράμματα που αναλύθηκαν παραπάνω μπορούν να χρησιμοποιηθούν το κάθε ένα μόνο του ή και σε συνδυασμό *(απαιτείται η βιβλιοθήκη patchwork)*. Επίσης, τα διαγράμματα που δημιουργούνται μπορούν να αποθηκευτούν σε μία μεταβλητή πχ. στο a και έτσι να εμφανίζονται πάνω αριστερά στο γραφικό περιβάλλον της R και να διατηρούνται.

Αφού παρουσιάστηκαν τα βασικά διαγράμματα της R παρακάτω θα αναλυθούν κάποια βασικά χαρακτηριστικά αυτών των διαγραμμάτων. Για να εκχωρηθούν τίτλος, υπότιτλος,

όνομα στους άξονες x και y αλλά και όνομα στο ολόκληρο το διάγραμμα χρησιμοποιείται η συνάρτηση labs().

```
>
>
>
>
> gapminder2002%>%ggplot(aes(x=continent,y=lifeExp))+geom_boxplot()+labs(title="Box plot",subtitle="Dataset gapminder2002",x="continent",y="life
Exp",tag="p1")
>
```

Εικόνα:Συνάρτηση labs()



Εικόνα:Τελικό αποτέλεσμα

Για να τοποθετήσουμε σχόλια μέσα στο διάγραμμα αλλά και στα όρια των αξόνων x και y χρησιμοποιείται η συνάρτηση annotate().

Τέλος, σε περίπτωση που ο χρήστης επιθυμεί να γίνει η αποθήκευση των διαγραμμάτων που κατασκεύασε σε μορφή pdf ή png πρέπει να χρησιμοποιήσει τη συνάρτηση ggsave() όπου μέσα στην παρένθεση τοποθετείται το διάγραμμα που θα αποθηκεύσει, το όνομα του αρχείου και τέλος οι διαστάσεις του αρχείου.

Κεφάλαιο 3ο-Ανάλυση Δεδομένων στην R

3.1:Tests για συνεχείς μεταβλητές

Στον τομέα της υγείας είναι πολύ χρήσιμη η ανάλυση των δεδομένων για την αξιολόγηση της πορείας των ασθενών αλλά και για τη σύγκριση διαφόρων τιμών της υγείας τους. Για τον σκοπό αυτό χρησιμοποιούνται διάφορα tests όπως είναι για παράδειγμα το t-test. Είναι σημαντικό κάθε φορά να χρησιμοποιείται το σωστό test για να εμφανίζονται ορθά αποτελέσματα αφού η R δεν προειδοποιεί τους χρήστες για τυχόν λάθος test. Στο κεφάλαιο αυτό όπως και στο προηγούμενο, θα χρησιμοποιηθεί το σύνολο δεδομένων gapminder (*R for Health Data ScienceEwen Harrison and Riinu Pius 2021-01-15*) το οποίο υπάρχει ήδη μέσα στην R για αυτό αρκεί μόνο να γίνει η χρήση της βιβλιοθήκης gapminder.

```
> library(gapminder)
> gapminder
# A tibble: 1,704 x 6
   country     continent  year lifeExp      pop gdpPercap
   <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
 1 Afghanist~ Asia       1952    28.8  8.43e6       779.
 2 Afghanist~ Asia       1957    30.3  9.24e6       821.
 3 Afghanist~ Asia       1962    32.0  1.03e7       853.
 4 Afghanist~ Asia       1967    34.0  1.15e7       836.
 5 Afghanist~ Asia       1972    36.1  1.31e7       740.
 6 Afghanist~ Asia       1977    38.4  1.49e7       786.
 7 Afghanist~ Asia       1982    39.9  1.29e7       978.
 8 Afghanist~ Asia       1987    40.8  1.39e7       852.
 9 Afghanist~ Asia       1992    41.7  1.63e7       649.
10 Afghanist~ Asia       1997    41.8  2.22e7       635.
# ... with 1,694 more rows
```

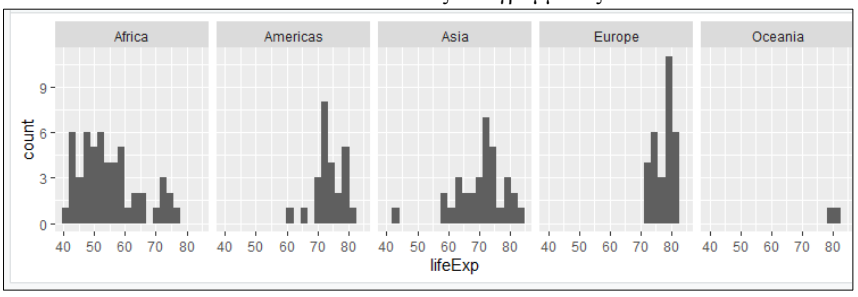
Εικόνα:Φόρτωση των δεδομένων

Πριν γίνει η ανάλυση δεδομένων πρώτα θα αναλυθούν διάφορα γραφήματα. Για την κατασκευή των διαγραμμάτων αυτών θα γίνει η σύγκριση των προσδόκιμων ζωής των 5 ηπείρων ξεχωριστά.

Ιστόγραμμα:

```
> gapminder%>%filter(year==2007)%>%ggplot(aes(x=lifeExp))+geom_histogram(bins=20)+facet_wrap(~continent)
> |
```

Εικόνα:Κώδικας Ιστογράμματος



Εικόνα:Τελικό αποτέλεσμα

**Με την εντολή geom_histogram (bins=20) τοποθετούνται μέχρι 20 στήλες σε κάθε ήπειρο.*

***Με την εντολή facet_wrap δημιουργούνται διαφορετικά πινακάκια για κάθε ήπειρο.*

Με βάση το παραπάνω διάγραμμα παρατηρείται ότι:

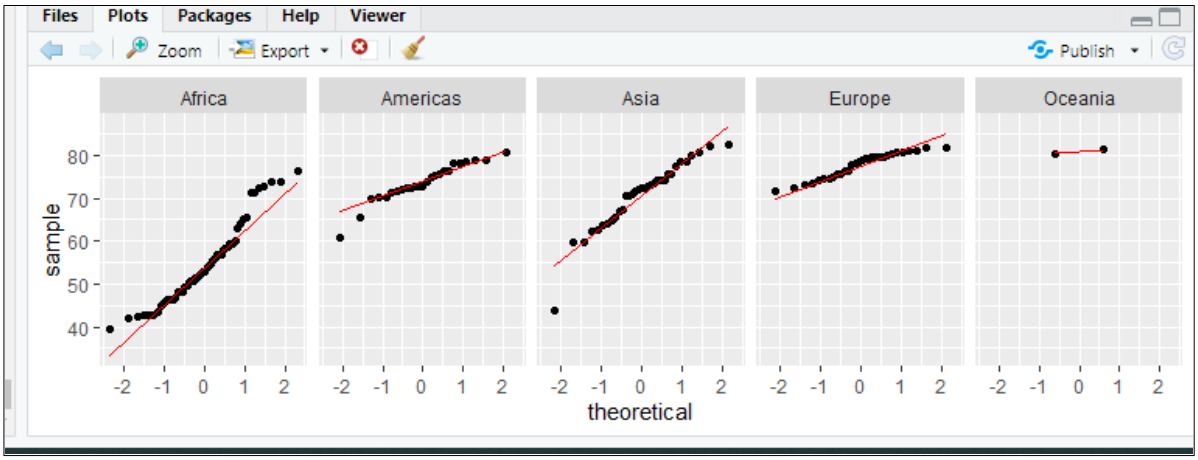
- η Αφρική έχει το μικρότερο προσδόκιμο ζωής διότι οι στήλες είναι πιο αριστερά,
- η Ωκεανία έχει λίγες γραμμές γιατί περιέχει μόνο δύο χώρες την Αυστραλία και την Νέα Ζηλανδία,
- το μεγαλύτερο προσδόκιμο ζωής το έχει η Ευρώπη διότι οι στήλες της είναι πιο δεξιά.

Q-Q Plots:

Τα Q-Q plots είναι γραφήματα τα οποία μας δείχνουν αν τα δεδομένα έχουν προκύψει από κάποια γνωστή κατανομή όπως για παράδειγμα την κανονική κατανομή.

```
>
>
>
> gapminder%>%filter(year==2007)%>%ggplot(aes(sample=lifeExp))+geom_qq()+geom_qq_line(colour="red")+facet_wrap(~co
ntinent)
>
```

Εικόνα:Κώδικας Q-Q plots



Εικόνα:Τελικό αποτέλεσμα

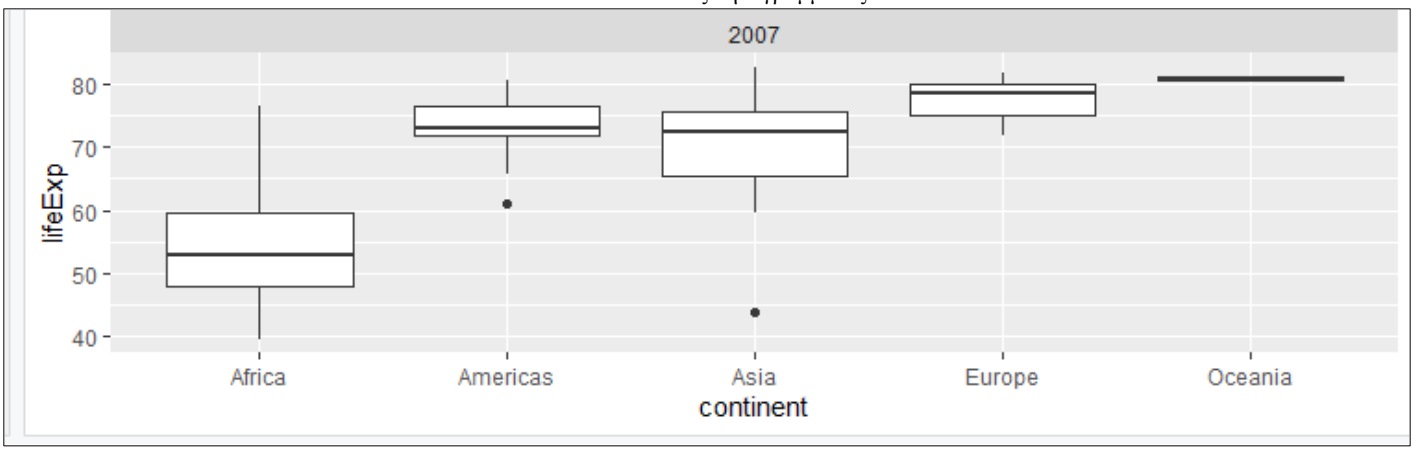
Παρατηρείται ότι όλοι οι ήπειροι ακολουθούν την γραμμή με μόνη εξαίρεση την Αφρική που δεν την ακολουθεί στο τέλος. Άρα η Αφρική είναι η μόνο ήπειρος που δεν ακολουθεί την κανονική κατανομή.

Θηκόγραμμα:

Τα θηκογράμματα είναι τα καλύτερα διαγράμματα για την εξέταση συνεχής μεταβλητής.

```
>
>
> gapminder%>%filter(year==2007)%>%ggplot(aes(x=continent,y=lifeExp))+geom_boxplot()+facet_wrap(~continent)
>
```

Εικόνα:Κώδικας θηκογράμματος



Εικόνα:Τελικό αποτέλεσμα

Από το διάγραμμα αυτό, φαίνεται ότι:

- η Αφρική έχει το μικρότερο μέσο προσδόκιμο ζωής,
- λόγω των λίγων παρατηρήσεων της Ωκεανίας το “πάχος του κουτιού” είναι πολύ λεπτό.

3.1.1:Tests για μία ομάδα (πληθυσμό):

T-Tests:

Στον έλεγχο των στατιστικών υποθέσεων πολύ σημαντικό ρόλο έχει η τιμή p-value, βάση της οποίας απορρίπτεται ή όχι η μηδενική υπόθεση H0. **Αν η p-value είναι μικρότερη από το επίπεδο σημαντικότητας (συνήθως είναι το 0.05) τότε η μηδενική υπόθεση απορρίπτεται και αποδεχόμαστε την εναλλακτική υπόθεση (H1).** Για τη μηδενική και την εναλλακτική υπόθεση ισχύει ο παρακάτω πίνακας:

Μηδενική Υπόθεση	*Εναλλακτική Υπόθεση*
H0: $\mu=\mu_o$	H1: $\mu<\mu_o$
H0: $\mu=\mu_o$	H1: $\mu>\mu_o$
H0: $\mu=\mu_o$	H1: $\mu\neq\mu_o$

Πίνακας:Μηδενική-Εναλλακτική Υπόθεση

Σε αυτό το t-test συγκρίνεται η μέση τιμή μίας ομάδας όπου περιλαμβάνει την εξαρτημένη μεταβλητή που είναι ποσοτική και την ανεξάρτητη που είναι ποιοτική. Η βασική σύνταξη της συνάρτησης t.test είναι η εξής:

t.test(x, mu=m_o)

όπου x=η ομάδα,mu=η μέση τιμή στην μηδενική υπόθεση H0.

Στον παρακάτω κώδικα αναλύεται αν η μέση τιμή του προσδόκιμου ζωής σε κάθε ήπειρο διαφέρει από την τιμή 78.

```
> gapminder %>%
+     filter(year == 2007) %>%
+     group_by(continent) %>%
+     do(
+         t.test(.$lifeExp, mu = 78) %>%
+             tidy()
+     )
# A tibble: 5 x 9
# Groups:   continent [5]
  continent estimate statistic  p.value parameter conf.low conf.high method           alternative
  <fct>        <dbl>     <dbl>    <dbl>     <dbl>    <dbl>     <dbl> <chr>            <chr>
1 Africa        54.8    -17.4   4.63e-23        51     52.1      57.5 One Sample t-test two.sided
2 Americas      73.6     -4.94  4.78e- 5        24     71.8      75.4 One Sample t-test two.sided
3 Asia          70.7     -5.25  9.74e- 6        32     67.9      73.6 One Sample t-test two.sided
4 Europe        77.6     -0.646 5.23e- 1        29     76.5      78.8 One Sample t-test two.sided
5 Oceania       80.7      5.28  1.19e- 1         1     74.2      87.3 One Sample t-test two.sided
>
```

Εικόνα:Έλεγχος μέσης τιμής

Με βάση τα αποτελέσματα συμπεραίνουμε ότι μόνο το προσδόκιμου όριο ζωής της Ευρώπης και της Ωκεανίας δεν διαφέρει από το 78 (αφού η Ευρώπη έχει διάστημα 76-78 και η Ωκεανία 74-87). Αυτό φαίνεται από το εύρος των στηλών conf.low και conf.high του πίνακα.

3.1.2:Tests για 2 ομάδες (πληθυσμούς):

Α)Τ-tests (Παραμετρική μέθοδος):

Σε αυτό το σημείο, θα αναλυθούν τα t-tests για δύο ομάδες. Τα t-tests είναι παραμετρικά (τα δεδομένα ακολουθούν γνωστή κατανομή για παράδειγμα κανονική κατανομή). Στην περίπτωση αυτή ισχύει ο παρακάτω πίνακας όπου μ_A είναι η μέση τιμή της μίας ομάδας και μ_B η μέση τιμή της άλλης ομάδας:

Μηδενική Υπόθεση	*Εναλλακτική Υπόθεση*
H0: $\mu_A=\mu_B$	H1: $\mu_A\neq\mu_B$
H0: $\mu_{A=}\mu_B$	H1: $\mu_A>\mu_B$
H0: $\mu_A=\mu_B$	H1: $\mu_A<\mu_B$

Πίνακας:Μηδενική-Εναλλακτική Υπόθεση

Έπειτα, θα συγκριθεί το προσδόκιμο ζωής της Αφρικής και της Ευρώπης το 2007. Πρέπει να σημειωθεί ότι θεωρείται κάθε μέτρηση ανεξάρτητη από τις άλλες έτσι ώστε να είναι η σύγκριση ορθή.

```
> ttest<-gapminder%>%filter(year==2007)%>% filter(continent %in% c
("Africa","Europe"))
> ttest%>%t.test(lifeExp~continent,data=.)

        Welch Two Sample t-test

data:  lifeExp by continent
t = -15.84, df = 66.128, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -25.72171 -19.96341
sample estimates:
mean in group Africa mean in group Europe
            54.80604             77.64860
```

Εικόνα:Σύγκριση προσδόκιμου ζωής

Από τον πίνακα αυτό προκύπτει ότι το p-value είναι μικρότερο από το 0.05 (δηλαδή οι μέσες τιμές διαφέρουν μεταξύ τους).

B) Wilcoxon Test (μη παραμετρική μέθοδος):

Η μέθοδος αυτή χαρακτηρίζεται ως μη παραμετρική διότι η κατανομή που ακολουθούν τα δεδομένα δεν είναι γνωστή Σε αυτά τα tests ισχύει το παρακάτω πινακάκι:

Μηδενική Υπόθεση	*Εναλλακτική Υπόθεση*
H0:οι δύο ομάδες δεν διαφέρουν	H1:οι δύο ομάδες διαφέρουν

Πίνακας:Μηδενική-Εναλλακτική Υπόθεση

Ακολουθεί παράδειγμα στο οποίο συγκρίνεται το προσδόκιμο ζωής στην Ευρώπη το 1982 και το 2002.

```
> europe<-gapminder%>%filter(year %in% c(1982,2002))%>%filter(cont
inent %in% c("Europe"))
> europe%>%wilcox.test(lifeExp~year,data=.)

        Wilcoxon rank sum test with continuity correction

data:  lifeExp by year
W = 161.5, p-value = 2.063e-05
alternative hypothesis: true location shift is not equal to 0
```

Εικόνα:Wilcoxon test

Το p-value είναι μικρότερο από 0.05 άρα απορρίπτεται η μηδενική υπόθεση που σημαίνει ότι υπάρχει σημαντική στατιστική διαφορά του προσδόκιμου ζωής στην Ευρώπη το 1982 και το 2002.

3.1.3:Tests για ζευγάρια:

A)T-tests για ζευγάρια

Μπορούμε επίσης, να βρούμε κατά πόσο μεταβλήθηκε το προσδόκιμο ζωής από την χρονιά που ξεκίνησε η μελέτη (1952) μέχρι το τέλος της (2007). Για να επιτευχθεί αυτό , απαιτείται η συνάρτηση summarise η οποία θα υπολογίσει την μεταβολή αυτή την συγκεκριμένη χρονική περίοδο.

```
> paired_data <- gapminder %>%
+     filter(year %in% c(1952, 2007)) %>%
+     filter(continent == "Europe")
> paired_data %>%
+     t.test(lifeExp ~ year, data = ., paired = TRUE)

        Paired t-test

data:  lifeExp by year
t = -16.132, df = 29, p-value = 5.066e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -14.91866 -11.56154
sample estimates:
mean of the differences
               -13.2401
```

Εικόνα:T-test για ζευγάρια

Όπως φαίνεται από τον προηγούμενο πίνακα το p-value είναι πολύ μικρό (πιο μικρό από 0.05) άρα απορρίπτεται η H0 και υπάρχει σημαντική στατιστική διαφορά στις χώρες της Ευρώπης το 1952 και το 2007.

3.1.4:Tests για περισσότερες από 2 ομάδες:

Α)ANOVA(παραμετρική μέθοδος)

Ένας τρόπος σύγκρισης των μέσων τιμών περισσότερων από 2 διαφορετικών ομάδων είναι ο ANOVA στον οποίο αναφέρονται αν υπάρχει διαφορά μεταξύ των μέσων των ομάδων που έχουν συμπεριληφθεί στην μελέτη. Στο σημείο αυτό θα εξεταστεί το προσδόκιμο όριο ζωής της Ευρώπης, της Αφρικής και της Αμερικής.

Σε αυτά τα tests ισχύει ο παρακάτω πίνακας:

Μηδενική Υπόθεση	*Εναλλακτική Υπόθεση*
H0:όλες οι μέσες τιμές είναι ίσες	H1:κάποιες μέσες τιμές δεν είναι ίσες

Πίνακας:Μηδενική-Εναλλακτική Υπόθεση

Με τις παρακάτω εντολές πραγματοποιείται το ANOVA test για το προσδόκιμο ζωής των τριών αυτών ηπείρων το 2007.

```
> gapminder%>%filter(year==2007)%>%filter(continent %in% c("America
s","Europe","Africa"))%>%aov(lifeExp~continent,data=.)%>%tidy()
# A tibble: 2 x 6
  term         df  sumsq meansq statistic   p.value
  <chr>     <dbl>  <dbl>  <dbl>     <dbl>     <dbl>
1 continent     2 12017. 6008.       114.  5.36e-27
2 Residuals   104  5461.   52.5       NA   NA
>
```

Εικόνα:Τελικό αποτέλεσμα

Όπως μπορείτε να παρατηρήσετε εύκολα στο προηγούμενο πινακάκι το p-value είναι πολύ μικρό για τον λόγο αυτό απορρίπτεται η H0 και έτσι καταλήγουμε ότι υπάρχει σημαντική στατιστική διαφορά μεταξύ των 3 ηπείρων.

B)Kruskal-Wallis test(μη παραμετρική μέθοδος):

Σε αυτό το σημείο πραγματοποιείται το test Kruskal-Wallis πάλι για την εξέταση του προσδόκιμου ζωής της Ευρώπης, της Αφρικής και της Αμερικής. Στο test αυτό ισχύει ο παρακάτω πίνακας:

Μηδενική υπόθεση	*Μη μηδενική υπόθεση*
Οι τιμές των ομάδων δεν διαφέρουν(είναι ίδιες).	Τουλάχιστον μία ομάδα εμφανίζει μεγαλύτερες τιμές από τις άλλες(υπάρχει στατιστική διαφορά μεταξύ των ομάδων).

Πίνακας:Μηδενική-Εναλλακτική Υπόθεση

```
> gapminder%>%filter(year==2007)%>%filter(continent %in% c("America
s","Europe","Africa"))%>%kruskal.test(lifeExp~continent,data=.)%>%tid
y()
# A tibble: 1 x 4
  statistic  p.value parameter method
      <dbl>    <dbl>     <int> <chr>
1      73.3 1.19e-16         2 Kruskal-Wallis rank sum test
>
```

Εικόνα:Test Kruskal-Wallis

Εύκολα αντιλαμβάνεστε ότι το p-value είναι μικρότερο από το επίπεδο σημαντικότητας (αν θεωρηθεί ότι είναι 0,05) άρα απορρίπτεται η H0 δηλαδή υπάρχει σημαντική στατιστική διαφορά μεταξύ των ομάδων.

3.2: Γραμμική Παλινδρόμηση

Σε αυτό το σημείο της εργασίας θα αναλυθεί η *γραμμική παλινδρόμηση στην R (Linear Regression, Selva Prabhakaran)* . Γενικά, η γραμμική παλινδρόμηση είναι μια μέθοδος η οποία δείχνει στους χρήστες της τη σχέση που υπάρχει ανάμεσα σε δύο μεταβλητές, όπου η μεταβλητή εξόδου είναι πάντοτε συνεχής. Ένα κλασικό παράδειγμα γραμμικής παλινδρόμησης είναι κατά πόσο η κατανάλωση του αλκοόλ επηρεάζει την υγεία των ατόμων. Στην γραμμική παλινδρόμηση στον άξονα x τοποθετείται η ανεξάρτητη μεταβλητή ενώ στον άξονα y η εξαρτημένη μεταβλητή. Ο γενικός τύπος της γραμμικής παλινδρόμησης είναι ο εξής:

$$\boldsymbol{Y=\beta_0+\beta_1\chi_1+\beta_2\chi_2+.....+\beta_v\chi_v+\varepsilon}$$

Ακόμα, στην γραμμική παλινδρόμηση ισχύει ο παρακάτω πίνακας :

Μηδενική Υπόθεση	***Μη μηδενική Υπόθεση***
Δεν υπάρχει σχέση μεταξύ των x και y	*Υπάρχει σχέση ανάμεσα στα x και y*

Πίνακας:Μηδενική-Εναλλακτική Υπόθεση

3.2.1:Απλή Γραμμική Παλινδρόμηση:

Στην απλή γραμμική παλινδρόμηση χρησιμοποιείται ο τύπος:

$Y=\beta_0+\beta1*X+\varepsilon$	Y=Εξαρτημένη μεταβλητή X=Ανεξάρτητη μεταβλητή β_0,β_1=Συντελεστές παλινδρόμησης ε=Σφάλμα

Εικόνα:Τύπος γραμμικής παλινδρόμησης

Για να είναι αποτελεσματικό το μοντέλο αυτό θα πρέπει οι διαφορές $\varepsilon=y_i-(\beta_o+\beta_1x_i)$ να είναι μικρές έτσι ώστε τα x_i,y_i να είναι κοντά σε μια ευθεία. Η ευθεία αυτή ονομάζεται βέλτιστη ευθεία και είναι η ευθεία που περνά πιο κοντά από τα δεδομένα μας.

Η συνάρτηση που χρησιμοποιείται στην R για να πραγματοποιηθεί γραμμική παλινδρόμηση είναι η *lm()*. Θα πρέπει να σημειωθεί ότι για να θεωρείται ότι η γραμμική παλινδρόμηση είναι ο ιδανικός τρόπος απεικόνισης των δεδομένων, θα πρέπει η γραμμή που προκύπτει να ταιριάζει στα δεδομένα. Σε αντίθετη περίπτωση θα πρέπει να εφαρμοστεί άλλος τρόπος απεικόνισης.

Ας δούμε ένα παράδειγμα γραμμικής παλινδρόμησης για τη χώρα μας όπου η ανεξάρτητη μεταβλητή(x) είναι το έτος και εξαρτημένη(y) το προσδόκιμο ζωής. Δηλαδή, με άλλα λόγια εξετάζεται κατά πόσο επηρεάζεται το προσδόκιμο ζωής ανά έτος:

```
>
>
> lm<-gapminder%>%filter(country=="Afghanistan")%>%lm(lifeExp~year,da
ta=.)
> lm%>%summary()

Call:
lm(formula = lifeExp ~ year, data = .)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5447 -0.9905 -0.2757  0.8847  1.6868

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -507.53427   40.48416  -12.54 1.93e-07 ***
year           0.27533    0.02045   13.46 9.84e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.223 on 10 degrees of freedom
Multiple R-squared:  0.9477,    Adjusted R-squared:  0.9425
F-statistic: 181.2 on 1 and 10 DF,  p-value: 9.835e-08

>
```

Εικόνα:Παράδειγμα γραμμικής παλινδρόμησης

i β_o:-507.53,

ii β_1: 0,2753,

iii άρα y=-507.53+0,2753*x.

iv T-value=Estimate/Std.Error.

v P-value= όπως φαίνεται από την τελευταία γραμμή το p-value είναι πολύ μικρό και πιο συγκεκριμένα πιο μικρό από το 0.05 που είναι το επίπεδο σημαντικότητας άρα καταλήγουμε στην απόρριψη της μηδενικής υπόθεσης H0 (έτσι υπάρχει σχέση μεταξύ του x και του y δηλαδή εξαρτάται το προσδόκιμο ζωής με το χρόνο).

vi Multiple R-squared (εκτίμηση του συντελεστή προσδιορισμού): Ο δείκτης αυτός εκφράζει το ποσοστό διακύμανσης της εξαρτημένης μεταβλητής το οποίο ερμηνεύεται από τη διακύμανση των τιμών της ανεξάρτητης μεταβλητής. Δηλαδή το 94.25% της

εξαρτημένης μεταβλητής (προσδόκιμο ζωής) ερμηνεύεται από την ανεξάρτητη μεταβλητή(έτος). Ο δείκτης αυτός ονομάζεται και συντελεστής προσδιορισμού.

vii Adjust R-squared: προσαρμοσμένη τιμή του R.

3.2.2: Πολλαπλή Γραμμική Παλινδρόμηση:

Για την πολλαπλή γραμμική παλινδρόμηση χρησιμοποιήθηκε το ίδιο dataset. Πιο συγκεκριμένα, για να γίνει πολλαπλή γραμμική παλινδρόμηση χρησιμοποιήθηκαν οι μεταβλητές gdpPercap,year και lifeExp.

```
> mlm<-gapminder%>%filter(country=="Afghanistan")%>%lm(lifeExp~year+gdpPercap,data=.)
> summary(mlm)

Call:
lm(formula = lifeExp ~ year + gdpPercap, data = .)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4699 -0.9875 -0.2011  0.9953  1.6675

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.094e+02  4.286e+01 -11.886 8.35e-07 ***
year         2.758e-01  2.150e-02  12.831 4.34e-07 ***
gdpPercap    1.133e-03  3.582e-03   0.316    0.759
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.282 on 9 degrees of freedom
Multiple R-squared:  0.9483,    Adjusted R-squared:  0.9368
F-statistic: 82.52 on 2 and 9 DF,  p-value: 1.626e-06

>
```

Εικόνα:Πολλαπλή γραμμική παλινδρόμηση

Ο τύπος της πολλαπλής γραμμικής παλινδρόμησης είναι ο εξής:

$$Y=\beta_0+\beta_1\chi_1+\beta_2\chi_2+\beta_3\chi_3+\varepsilon$$

Άρα στο παράδειγμα αυτό,

i Y=-0.05+0.27*year+0.00113*gdpPercap

ii R-squared=0,9368

iii P-value: είναι πολύ μικρό(1.626e-06) άρα και οι δύο μεταβλητές gdpPercap, year επηρεάζουν τη μεταβλητή lifeExp (προσδόκιμο ζωής) στο Αφγανιστάν.

3.3:Tests για κατηγορικές μεταβλητές

Όπως αναφέρθηκε και στην αρχή της εργασίας αυτής οι κατηγορικές μεταβλητές , που είναι πολύ συχνές στα δεδομένα υγείας, μπορεί να είναι:

a.Ομάδες (πχ.άντρες-γυναίκες),

b. Χαρακτήρες (γράμματα,προτάσεις),

c. Λογικές τελεστές (Ναι/Όχι).

Σε αυτό το σημείο θα χρησιμοποιηθεί το σύνολο δεδομένων melanoma (*R for Health Data ScienceEwen Harrison and Riinu Pius 2021-01-15*) το οποίο περιλαμβάνει ασθενείς που πάσχουν από αυτόν τον τύπο καρκίνου.

Τα συνεχή δεδομένα μπορούν να μετατραπούν σε κατηγορικά δεδομένα αρκεί να υπάρχει μεγάλη προσοχή έτσι ώστε να μην χαθούν πληροφορίες από τον πίνακα κατά τη μετατροπή.

```
>
>
>
>
> data<-meldata%>%mutate(ulcer.factor=factor(ulcer)%>%fct_recode("Present"="1","Absent"="0"))
>
```

Εικόνα:Μετατροπή συνεχών δεδομένων σε κατηγορικά

Στην προηγούμενη εικόνα δημιουργούμε μία νέα στήλη την ulcer.factor όπου τοποθετείται 1 όταν ο ασθενής έχει εμφανιστεί το μελάνωμα και 0 όταν δεν έχει.

3.3.1:Pearson's X^2 test:

Το συγκεκριμένο test χρησιμοποιείται για να εξεταστεί αν δύο κατηγορικές μεταβλητές είναι ανεξάρτητες ή όχι σε έναν πληθυσμό. Ισχύει ο παρακάτω πίνακας:

Μηδενική Υπόθεση	*Μη Μηδενική Υπόθεση*
Οι δύο μεταβλητές είναι ανεξάρτητες	Οι δύο μεταβλητές εξαρτώνται

Πίνακας:Μηδενική-Εναλλακτική Υπόθεση

Επίσης, δημιουργείται μία νέα στήλη Ages η οποία έχει young αν το άτομο είναι μικρότερο από 50 χρονών και old διαφορετικά.

```
>
>
> data<-meldata%>%mutate(ulcer.factor=factor(ulcer)%>%fct_recode("Present"="1","Absent"="0"))
> data1<-data%>%mutate(Ages=if_else(age<=50,"Young","Old"))
> table(data1$Ages,data1$ulcer.factor)

        Absent Present
  Old       60      58
  Young     55      32
> data1%$%table(data1$ulcer.factor,data1$Ages)%>%chisq.test()

        Pearson's Chi-squared test with Yates' continuity correction

data:  .
X-squared = 2.6298, df = 1, p-value = 0.1049

>
```

Εικόνα: Pearson's X^2 test

Όπως φαίνεται από την τελευταία γραμμή του προηγούμενου πίνακα το p-value είναι μεγαλύτερο από 0.05, άρα δεν απορρίπτεται η H0 με αποτέλεσμα η εμφάνιση του όγκου να μην εξαρτάται από την ηλικία του ατόμου.

3.3.2:Fisher's test:

Χρησιμοποιείται για μικρά δείγματα και είναι δύσκολο να υπολογιστεί με πράξεις στο χέρι.

```
> data1%$%table(data1$ulcer.factor,data1$Ages)%>%fisher.test()

        Fisher's Exact Test for Count Data

data:  .
p-value = 0.08846
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.3279618 1.0998505
sample estimates:
odds ratio
 0.6033942
```

Εικόνα:Fisher's test

Παρατηρούμε ότι το p-value είναι μεγαλύτερο από 0.05 άρα οι μεταβλητές δεν εξαρτώνται.

3.4:Λογιστική Παλινδρόμηση

Στο σημείο αυτό της εργασίας θα αναλυθεί η δυαδική λογιστική παλινδρόμηση όπου η μεταβλητή y (μεταβλητή εξόδου) είναι κατηγορική και μπορεί να έχει δύο τιμές (π.χ. 0 ή 1) ενώ η μεταβλητή x μπορεί να είναι είτε συνεχής μεταβλητή είτε κατηγορική. Για να γίνει κατανοητή η έννοια της δυαδικής λογιστικής παλινδρόμησης πρέπει πρώτα να αναλυθούν ορισμένες βασικές έννοιες.

1) Odds: Ο τύπος είναι p/p-1 όπου p είναι η πιθανότητα πραγματοποίησης ενός γεγονότος. Αν το odds είναι μεγαλύτερο από το 1 σημαίνει ότι η εμφάνιση της νόσου είναι μεγαλύτερη από την μη εμφάνιση της νόσου ενώ αν το odds είναι μικρότερο από 1 το αντίστροφο. Για παράδειγμα αν η πιθανότητα εμφάνισης της νόσου είναι ¼ τότε το odds θα είναι:

Odds=p/p-1=0.25/0.75=0.33

2) Πιθανότητα: Ο τύπος είναι odds/odds+1 και οι τιμές του κυμαίνεται από το 0 μέχρι το 1.

3)Odds Ratio: Είναι ο σχετικός λόγος συμπληρωματικών πιθανοτήτων να συμβεί ένα ενδεχόμενο υπό μία συνθήκη Α προς τον λόγο συμπληρωματικών πιθανοτήτων να συμβεί το ίδιο ενδεχόμενο υπό μία άλλη συνθήκη Β.

	Όχι Έκθεση	***Έκθεση***
Εμφάνιση Νόσου	*a*	*b*
Μη εμφάνιση Νόσου	*c*	*D*

Odds (εμφάνιση νόσου όταν καπνίζει): b/d

Odds (εμφάνιση νόσου όταν δεν καπνίζει): a/c

Odds Ratio: a*d/b*c

Για την παρουσίαση της λογιστικής παλινδρόμησης θα χρησιμοποιηθεί το dataset με τα δεδομένα για τους ασθενείς που έπασχαν από μελάνωμα που χρησιμοποιήθηκε και προηγουμένως.

Με τον παρακάτω κώδικα δημιουργούνται μία νέα στήλη:

- Η στήλη *status.factor* έχει *died melanoma* όπου η στήλη status έχει 1, *alive* όπου έχει 2 και *died* όπου έχει 3.

Επίσης με τον κώδικα αυτό αλλάζει το όνομα των στηλών *age.*

Στην συνέχεια γίνεται λογιστική παλινδρόμηση μεταξύ της στήλης *status.factor* (εξαρτημένη μεταβλητή) και της στήλης *ages* (ανεξάρτητη μεταβλητή).

```
>
> meldata <- boot::melanoma
> meldata1<-meldata%>%mutate(status.factor=factor(status)%>%fct_recode("Died Melanoma"="1","Ali
ve"="2","Died"="3"))
> meldata2<-meldata1%>%mutate(Ages=if_else(age<=50,"Young","Old"))
> fit<-glm(meldata2$status.factor~meldata2$Ages,family=binomial)
> summary(fit)

Call:
glm(formula = meldata2$status.factor ~ meldata2$Ages, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7147  -1.5230   0.7228   0.8675   0.8675

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)          0.7835     0.1984   3.949 7.86e-05 ***
meldata2$AgesYoung   0.4254     0.3230   1.317    0.188
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 242.35  on 204  degrees of freedom
Residual deviance: 240.58  on 203  degrees of freedom
AIC: 244.58

Number of Fisher Scoring iterations: 4

> |
```

Εικόνα:Λογιστική Παλινδρόμηση

Όπως μπορείτε να παρατηρήσετε το p-value είναι πιο μικρό από το 0.05 το οποίο μας οδηγεί στην απόρριψη της μηδενικής υπόθεσης δηλαδή μας οδηγεί στο συμπέρασμα ότι η εμφάνιση του όγκου διαδραματίζει σημαντικό ρόλο στον θάνατο των ατόμων.

3.5: Ανάλυση Επιβίωσης

Πέρα από την δυαδική λογιστική παλινδρόμηση που αναλύθηκε στο προηγούμενο κεφάλαιο πολλές φορές ο χρόνος που απαιτείται για να πραγματοποιηθεί ένα γεγονός είναι πολύ σημαντικός. Σε αυτήν την περίπτωση ακολουθείται η μέθοδος της ανάλυσης επιβίωσης η οποία μπορεί να πραγματοποιηθεί για μία αλλά και για περισσότερες μεταβλητές. Για τη μέθοδο αυτή χρησιμοποιήθηκε το σύνολο δεδομένων με τους ασθενείς που έπασχαν από

μελάνωμα οι οποίοι χειρουργήθηκαν. Με βάση αυτό το σύνολο δεδομένων εξετάζεται ο χρόνος επιβίωσης των ασθενών μετά την χειρουργική επέμβαση.

Στην στήλη status υπάρχει το 1 όταν ο ασθενής πέθανε από αυτόν τύπο καρκίνου (μελάνωμα), 2 όταν ο ασθενής είναι ακόμα ζωντανός και 3 όταν ο ασθενής πέθανε από άλλους λόγους.

Ακόμα, στην στήλη time μετρώνται οι μέρες από την χειρουργική επέμβαση μέχρι τον πιθανό θάνατο του ασθενή. Σε περίπτωση που έχουν χαθεί τα ίχνη του ασθενή μετά από κάποιο χρονικό διάστημα προσμετρώνται εκείνες οι ημέρες που υπήρχε επικοινωνία μεταξύ του ασθενή και του νοσοκομείου.

Ολοκληρώνοντας, στο σημείο αυτό θα μελετηθεί αν το φύλο του ατόμου διαδραματίζει καθοριστικό ρόλο στο χρόνο επιβίωσης του.

```
>
>
> meldata <- boot::melanoma
> view(meldata)
> meldata <- meldata %>%mutate(status_os = if_else(status == 2, 0, 1))
> meldata <- meldata %>%mutate(sex=factor(sex)%>%fct_recode("Male" = "1","Female" = "0"))
> view(meldata)
> dependent_os <- "Surv(time, status_os)"
> explanatory  <- c("sex")
> meldata%>%surv_plot(dependent_os, explanatory, pval = TRUE)
>
```

Εικόνα:Ανάλυση Επιβίωσης

meldata

Filter

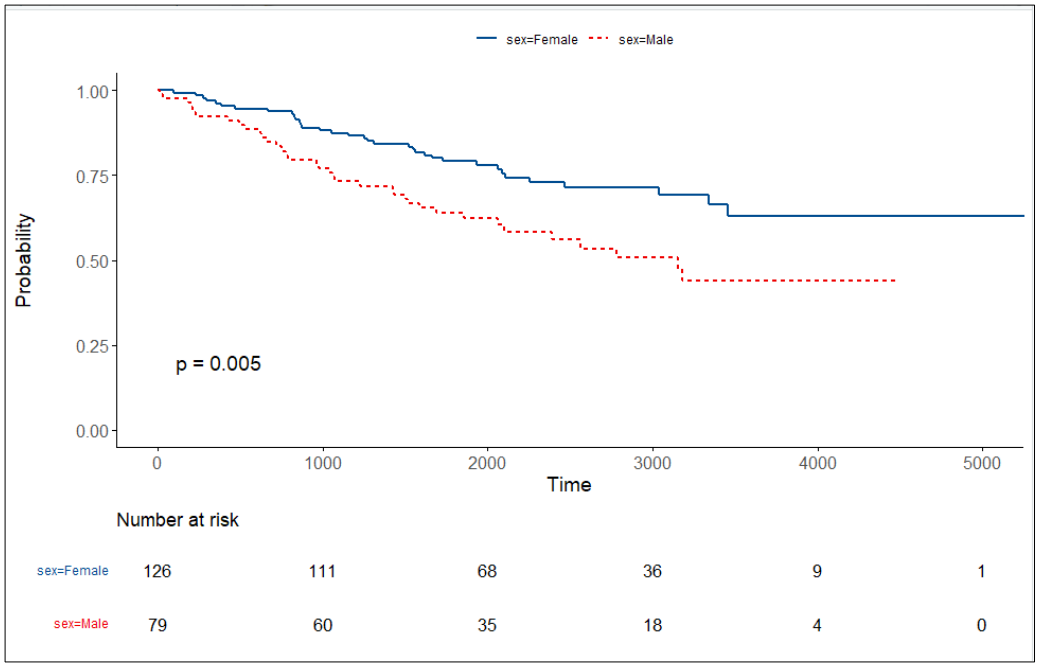
	time	status	sex	age	year	thickness	ulcer	status_os
1	10	3	Male	76	1972	6.76	1	1
2	30	3	Male	56	1968	0.65	0	1
3	35	2	Male	41	1977	1.34	0	0
4	99	3	Female	71	1968	2.90	0	1
5	185	1	Male	52	1965	12.08	1	1
6	204	1	Male	28	1971	4.84	1	1
7	210	1	Male	77	1972	5.16	1	1
8	232	3	Female	60	1974	3.22	1	1
9	232	1	Male	49	1968	12.88	1	1
10	279	1	Female	68	1971	7.41	1	1
11	295	1	Female	53	1969	4.19	1	1

Showing 1 to 11 of 205 entries, 8 total columns

Εικόνα:Ο πίνακας meldata

Με βάση τον παραπάνω κώδικα δημιουργούνται οι εξής νέες στήλες:

- status_os: η οποία έχει την τιμή 0 όταν ο ασθενής είναι ακόμα ζωντανός και 1 όταν ο ασθενής έχει πεθάνει είτε από το μελάνωμα είτε από άλλες αιτίες.
- στην στήλη sex όπου υπάρχει η τιμή 1 μπαίνει το male και όπου υπάρχει η τιμή 0 μπαίνει το female.



Εικόνα:Survival plot

Τα συμπεράσματα που εξάγονται είναι τα εξής:

- Την χρονική στιγμή 0 η πιθανότητα επιβίωσης και για τα δυο φύλα είναι 100%.

- Η πιθανότητα επιβίωσης είναι 75% την χρονική στιγμή 1000 (ημέρες) για τους άνδρες και περίπου 2000 (ημέρες) για τις γυναίκες (δηλαδή οι διπλάσιες).

- Στις 3000 ημέρες η επιβίωση των ανδρών είναι 50%, ενώ για τις γυναίκες είναι περίπου 75%. Άρα, οι γυναίκες εμφανίζουν μεγαλύτερη επιβίωση σε σύγκριση με τους άνδρες.

Κεφάλαιο 4ο-Δείκτες Νοσοκομείου

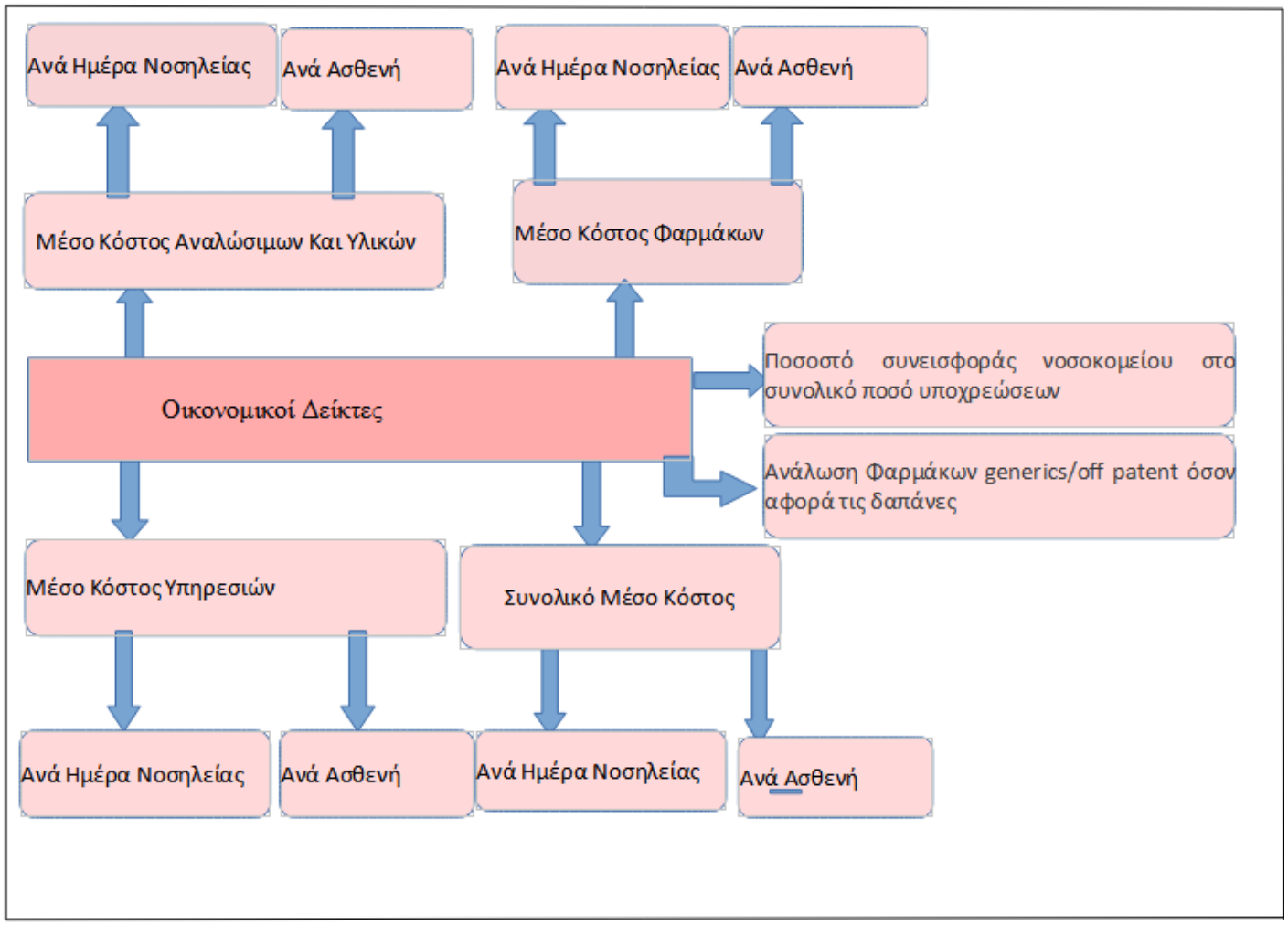
Τα δημόσια νοσοκομεία στη χώρας μας ανέρχονται στα 126 τα οποία μπορούν να ομαδοποιηθούν με βάση τον σκοπό της λειτουργίας τους, δηλαδή το είδος τους, ή με βάση την χωρητικότητα τους (πλήθος κλινών). Ως προς τον σκοπό λειτουργίας τους υπάρχουν 10 είδη νοσοκομείων στην Ελλάδα:

- *Γενικά*
- *Ογκολογικά*
- *Παιδιατρικά*
- *Δερματολογικά*
- *Γυναικολογικά-Μαιευτικά*
- *Ορθοπεδικά*
- *Αποκατάστασης*
- *Οφθαλμολογικά*
- *Ψυχιατρικά*
- *Θωρακικής Νόσου*

Έτσι με τον διαχωρισμό αυτό και παράλληλα με τη χρήση των οικονομικών και λειτουργικών δεικτών μπορεί να γίνει η αξιολόγηση των νοσοκομείων.

4.1.Οικονομικοί Δείκτες

Οι οικονομικοί δείκτες που θα εξετάσουμε στο υποκεφάλαιο αυτό είναι οι εξής:



Εικόνα:Οικονομικοί Δείκτες

4.1.1:Συνολικό Μέσο Κόστος Ανά Ασθενή/Ανά Ημέρα Νοσηλείας

Ο πρώτος δείκτης που θα αναλυθεί είναι το συνολικό μέσο κόστος ανά ασθενή και ανά ημέρα νοσηλείας. Το μέσο συνολικό κόστος ανά ασθενή δείχνει πώς κατανέμεται το συνολικό κόστος στον ετήσιο αριθμό ασθενών που επισκέπτονται το νοσοκομείο δηλαδή δείχνει τα συνολικά χρήματα που χρησιμοποιούνται για την θεραπεία του κάθε ασθενή. Το μέσο κόστος ανά ημέρα νοσηλείας δείχνει τον τρόπο που χρησιμοποιούνται τα χρήματα κάθε

ημέρα. Το κόστος αυτό περιλαμβάνει όλα τα έξοδα του νοσοκομείου (υπηρεσίες, φάρμακα, αντιδραστήρια, υλικά) εκτός από την μισθοδοσία των προσώπων του νοσοκομείου.

Το συνολικό κόστος ανά ασθενή περιγράφεται από τον τύπο που έχει στον αριθμητή το συνολικό κόστος και στον παρονομαστή το συνολικό αριθμό των ασθενών, ενώ το συνολικό κόστος ανά ημέρα νοσηλείας περιγράφεται από τον τύπο που έχει ως αριθμητή το συνολικό κόστος και ως παρονομαστή το συνολικό αριθμό ημερών νοσηλείας.

- ✓ ***Συνολικό Μέσο Κόστος Ανά Ασθενή = Συνολικό Κόστος / Αριθμός Ασθενών***
- ✓ ***Συνολικό Μέσο Κόστος Ανά Ημέρα Νοσηλείας = Συνολικό Κόστος/Αριθμός Ημερών Νοσηλειών***

4.1.2: Μέσο Κόστος Αναλώσιμων και Υλικών Ανά Ασθενή/Ανά Ημέρα Νοσηλείας

Το μέσο κόστος για αναλώσιμα και υλικά αναφέρεται στον τρόπο που δαπανήθηκε το ποσό χρημάτων για την θεραπεία του ασθενή εξαιρουμένου του κόστος των φαρμάκων. Το συνολικό κόστος αναλώσιμων και υλικών ανά ασθενή περιγράφεται από τον τύπο που έχει στον αριθμητή το συνολικό κόστος υλικών και αναλώσιμων και στον παρονομαστή το συνολικό αριθμό των ασθενών, ενώ το συνολικό κόστος ανά ημέρα νοσηλείας περιγράφεται από τον τύπο που έχει ως αριθμητή το συνολικό κόστος και ως παρονομαστή το συνολικό αριθμό ημερών νοσηλείας.

- ***Μέσο Κόστος Αναλώσιμων και Υλικών Ανά Ασθενή = Συνολικό Κόστος Αναλώσιμων και υλικών / Αριθμός Ασθενών***
- ***Μέσο Κόστος Αναλώσιμων και Υλικών Ανά Ημέρα Νοσηλείας = Συνολικό Κόστος Αναλώσιμων και υλικών/Αριθμός Ημερών Νοσηλειών***

4.1.3:Μέσο Κόστος Φαρμάκων Ανά Ασθενή/Ανά Ημέρα Νοσηλείας

Το μέσο κόστος φαρμάκων αναφέρεται στον τρόπο που δαπανήθηκε το ποσό των χρημάτων για την χρήση των κατάλληλων φαρμάκων για την θεραπεία των ασθενών. Το κόστος των φαρμάκων εξαρτάται άμεσα από το είδος του νοσοκομείου (για παράδειγμα τα ογκολογικά νοσοκομεία έχουν πολύ ακριβά φάρμακα για την θεραπεία των ασθενών). Όπως και οι προηγούμενοι δείκτες έτσι και αυτός διαχωρίζεται στο μέσο κόστος φαρμάκων ανά ασθενή (στον αριθμητή τοποθετείται το κόστος των φαρμάκων και στον παρονομαστή το πλήθος των νοσηλευομένων) αλλά και στο μέσο κόστος φαρμάκων ανά ημέρα νοσηλείας (στον αριθμητή τοποθετείται το κόστος των φαρμάκων και στον παρονομαστή το σύνολο των ημερών νοσηλείας).

- ***Μέσο Κόστος Φαρμάκων Ανά Ασθενή = Συνολικό Κόστος Φαρμάκων / Αριθμός Ασθενών***
- ***Μέσο Κόστος Φαρμάκων Ανά Ημέρα Νοσηλείας = Συνολικό Κόστος Φαρμάκων/Αριθμός Ημερών Νοσηλειών***

4.1.4:Μέσο Κόστος Υπηρεσιών Ανά Ασθενή/Ανά Ημέρα Νοσηλείας

Το μέσο κόστος υπηρεσιών περιλαμβάνει τα κόστη της εστίασης, καθαριότητας, ασφάλειας κτλπ αλλά και την ΔΕΚΟ (ΔΕΗ, ΟΤΕ κ.α.). Στον δείκτη αυτό περιλαμβάνεται και η μισθοδοσία. Ο δείκτης αυτός χωρίζεται στο μέσο κόστος υπηρεσιών ανά ασθενή (όπου στο αριθμητή είναι το σύνολο του κόστους των υπηρεσιών και στον παρονομαστή το πλήθος των ασθενών), αλλά και στο μέσο κόστος υπηρεσιών ανά ημέρα νοσηλείας (όπου στον αριθμητή έχουμε το συνολικό κόστος υπηρεσιών και στον παρονομαστή το πλήθος των ημερών νοσηλείας). Σε περίπτωση που το κόστος των υπηρεσιών είναι μεγάλο και το πλήθος των ασθενών είναι μικρό είναι απαραίτητο να γίνει επανεξέταση στο συγκεκριμένο νοσοκομείο.

- ***Μέσο Κόστος Υπηρεσιών Ανά Ασθενή = Συνολικό Κόστος Υπηρεσιών / Αριθμός Ασθενών***

- ***Μέσο Κόστος Υπηρεσιών Ανά Ημέρα Νοσηλείας = Συνολικό Κόστος Υπηρεσιών/Αριθμός Ημερών Νοσηλειών***

4.1.5: Ποσοστό συνεισφοράς νοσοκομείου στο συνολικό ποσό υποχρεώσεων

Η παρακολούθηση των υποχρεώσεων κάθε υγειονομικής μονάδας ως % των συνολικών υποχρεώσεων είναι πολύ σημαντική γιατί μας βοηθάει να εντοπίσουμε πόσο συνεισφέρει το νοσοκομείο στις συνολικές υποχρεώσεις. Ο δείκτης έχει στον αριθμητή το σύνολο των υποχρεώσεων του νοσοκομείου του τρέχοντος έτους και στον παρονομαστή το συνολικό κόστος.

- ***Ποσοστό συνεισφοράς νοσοκομείου στο συνολικό ποσό υποχρεώσεων = Υποχρεώσεις του νοσοκομείου /συνολικές υποχρεώσεις***

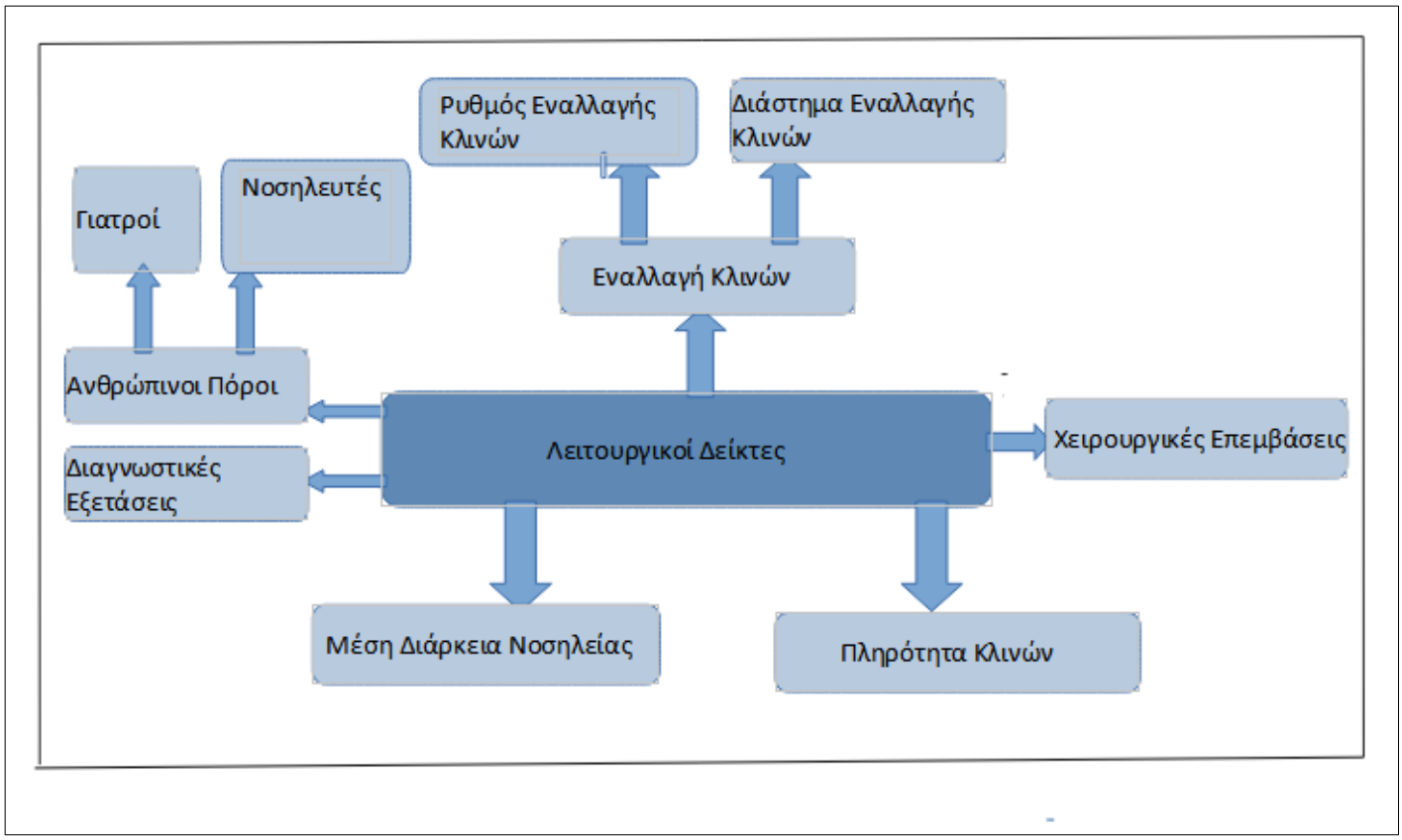
4.1.6:Ανάλωση Φαρμάκων generics/off patent όσον αφορά τις δαπάνες

Λόγω του μεγάλου κόστος των φαρμάκων πολλές χώρες χρησιμοποιούν τα γενόσημα φάρμακα δηλαδή τα φάρμακα που μοιάζουν πολύ με τα πρωτότυπα φάρμακα τα οποία όμως έχουν μικρότερο κόστος. Με τον τρόπο αυτό, αφού τα γενόσημα φάρμακα είναι πιο φτηνά μπορεί να μειωθεί το κόστος που σχετίζεται με τα φάρμακα χωρίς όμως να επηρεαστεί η αποτελεσματικότητα της θεραπείας των ασθενών. Ο τύπος του δείκτη αυτού είναι ο εξής:

- ***Ανάλωση Φαρμάκων generics/off patent = Δαπάνες Νοσοκομείου για χρήση Γενόσημων Φαρμάκων/Συνολικό κόστος για φάρμακα***

4.2.Λειτουργικοί Δείκτες

Οι λειτουργικοί δείκτες που θα αναλυθούν τώρα είναι οι εξής:



Εικόνα:Λειτουργικοί Δείκτες

4.2.1:Μέση Διάρκεια Νοσηλείας

Η μέση διάρκεια νοσηλείας αναφέρεται στον μέσο αριθμό ημερών που περνούν οι ασθενείς στο νοσοκομείο. Η μέση διάρκεια νοσηλείας περιγράφεται ως ένα κλάσμα που έχει ως αριθμητή το σύνολο των ημερών νοσηλείας και ως παρονομαστή το σύνολο των ασθενών. Αν η μέση διάρκεια νοσηλείας είναι μικρή αυτό σημαίνει ότι οι ασθενείς θα έχουν μικρό κόστος για τη νοσηλεία τους αφού θα πάρουν εξιτήριο πολύ γρήγορα. Ωστόσο μια πολύ μικρή μέση διάρκεια νοσηλείας μπορεί και να σημαίνει ότι το νοσοκομείο παρέχει υπηρεσίες χαμηλού επιπέδου με αποτέλεσμα να είναι πιθανή η επανεισαγωγή του ασθενή στο νοσοκομείο.

- ***Μέση Διάρκεια Νοσηλείας = Σύνολο ημερών νοσηλείας/Σύνολο ασθενών***

4.2.2:Πληρότητα Κλινών

Ο δείκτης της πληρότητας των κλινών είναι ένας δείκτης που αντικατοπτρίζει τον τρόπο που χρησιμοποιούνται οι πόροι του νοσοκομείου. Αν ο δείκτης είναι υψηλός τότε αυτό μπορεί να οφείλεται στην ορθή ιατρική πρακτική, ή στην μεγάλη διάρκεια παραμονής των ασθενών στο νοσοκομείο. Το ποσοστό της πληρότητας κλινών περιγράφεται από το κλάσμα που έχει στον αριθμητή το σύνολο των ημερών νοσηλείας και στον παρονομαστή το γινόμενο του πλήθους των ημερών νοσηλείας επί του συνόλου των κλινών.

- ***Πληρότητα Κλινών = (Ημέρες Νοσηλείας / (Ημέρες Νοσηλείας * Σύνολο Κλινών))*100***

4.2.3:Εναλλαγή Κλινών

4.2.3.1:Ρυθμός Εναλλαγής Κλινών

Ο δείκτης αυτός δείχνει πόσες φορές παρατηρήθηκε αλλαγή κλίνης από έναν ασθενή, δηλαδή δείχνει το ποσοστό των κλινών που επαναχρησιμοποιούνται. Ο δείκτης αυτός περιγράφεται από το κλάσμα όπου στον αριθμητή έχει το σύνολο των εξιτηρίων (πλήθος νοσηλευομένων) και στον παρονομαστή το σύνολο των κρεβατιών.

- ***Ρυθμός Εναλλαγής Κλινών = Σύνολο Εξιτηρίων / Σύνολο κρεβατιών***

4.2.3.2:Διάστημα Εναλλαγής Κλινών

Ο δείκτης αυτός περιγράφει τον χρόνο που είναι διαθέσιμες οι κλίνες του νοσοκομείου. Ο δείκτης αυτός περιγράφεται από το κλάσμα που έχει ως αριθμητή το σύνολο των κλινών επί τις ημέρες του χρόνου μείον τις συνολικές ημέρες νοσηλείας και ως παρονομαστή το σύνολο των εξιτηρίων. Όταν ο δείκτης αυτός είναι αρνητικός σημαίνει ότι το νοσοκομείο έχει έλλειψη κλινών, ενώ όταν είναι θετικός σημαίνει ότι το νοσοκομείο κάνει κακή διαχείριση των κλινών.

- ***Διάστημα Εναλλαγής Κλινών = ((Σύνολο Κλινών*ημέρες του χρόνου) - συνολικές ημέρες νοσηλείας) / Σύνολο Εξιτηρίων***

4.2.4:Χειρουργικές Επεμβάσεις

Ο δείκτης αυτός φανερώνει τον όγκο εργασίας των γιατρών σχετικά με τις χειρουργικές επεμβάσεις του νοσοκομείου. Σε περίπτωση που ο ο δείκτης αυτός είναι πολύ μεγάλος και ο αριθμός των γιατρών είναι πολύ μικρός είναι αναγκαίο να προσληφθούν νέα πρόσωπα με σκοπό την μείωση του φόρτου εργασίας των γιατρών. Ο τύπος του δείκτη είναι ο εξής :

- ***Χειρουργικές Επεμβάσεις = Σύνολο χειρουργείων/Σύνολο γιατρών***

4.2.5:Διαγνωστικές Εξετάσεις

Ο δείκτης αυτός μας δείχνει αν σε ένα νοσοκομείο γίνονται υπερβολικές εξετάσεις ανά γιατρό ή ανά κλίνη. Για τον λόγο αυτό είναι απαραίτητο να ελέγχονται αν οι εξετάσεις είναι απαραίτητες να γίνονται σε κάθε περίπτωση. Υπάρχουν δύο είδη του δείκτη αυτού που είναι οι εξής:

- ***Διαγνωστικές Εξετάσεις ανά Γιατρό = Εξετάσεις/Σύνολο γιατρών***
- ***Διαγνωστικές Εξετάσεις ανά Κλίνη = Εξετάσεις/Σύνολο κρεβατιών***

4.2.6:Ανθρώπινοι Πόροι

Ο δείκτης αυτός δείχνει αν έχουμε έλλειμμα ή πλεόνασμα ιατρικού προσωπικού(γιατρών,νοσηλευτών).Με την ανάλυση του δείκτη αυτού, τα νοσοκομεία μπορούν να λάβουν αποφάσεις έτσι ώστε να είναι καλύτερα οργανωμένο και άρα πιο αποτελεσματικό. Υπάρχουν δύο είδη αυτού του δείκτη:

- ***Γιατροί Ανά Κλίνη: Σύνολο Γιατρών/Σύνολο Κλινών***
- ***Νοσηλευτές Ανά Κλίνη: Σύνολο Νοσηλευτών/Σύνολο Κλινών***

Κεφάλαιο 5ο-Ανάλυση δεδομένων νοσοκομείων μέσω της R

Στο τελευταίο κεφάλαιο θα δούμε πώς μπορούν να προκύψουν χρήσιμα συμπεράσματα από την ανάλυση των δεδομένων των νοσοκομείων με την γλώσσα προγραμματισμού R. Όπως αναφέρθηκε και προηγουμένως, η ανάλυση των στοιχείων αυτών μπορεί να γίνει με βάση το μέγεθος των νοσοκομείων ή με βάση το είδος τους. Στην εργασία αυτή η ομαδοποίηση των νοσοκομείων θα πραγματοποιηθεί με βάση το είδος τους. Κάποιες φορές λοιπόν για τις ανάγκες της ανάλυσης των δεδομένων, η 3η και η 5η Υγειονομική Περιφέρεια θα αντιμετωπιστούν συνολικά και άλλες ξεχωριστά.

Γενικά Νοσοκομεία 3ης Υγειονομικής Περιφέρειας
Γ.Ν. "ΠΑΠΑΓΕΩΡΓΙΟΥ"
Γ.Ν. ΒΕΡΟΙΑΣ
Γ.Ν. ΓΙΑΝΝΙΤΣΩΝ
Γ.Ν. ΓΡΕΒΕΝΩΝ
Γ.Ν. ΕΔΕΣΣΑΣ
Γ.Ν. ΘΕΣ/ΝΙΚΗΣ "ΑΓ. ΔΗΜΗΤΡΙΟΣ"
Γ.Ν. ΘΕΣ/ΝΙΚΗΣ "Γ. ΓΕΝΝΗΜΑΤΑΣ"
Γ.Ν. ΘΕΣ/ΝΙΚΗΣ "Γ. ΠΑΠΑΝΙΚΟΛΑΟΥ"
Γ.Ν. ΚΑΣΤΟΡΙΑΣ
Γ.Ν. ΚΑΤΕΡΙΝΗΣ
Γ.Ν. ΚΟΖΑΝΗΣ "ΜΑΜΑΤΣΕΙΟ"
Γ.Ν. ΝΑΟΥΣΑΣ
Γ.Ν. ΠΤΟΛΕΜΑΪΔΑΣ "ΜΠΟΔΟΣΑΚΕΙΟ"
Γ.Ν. ΦΛΩΡΙΝΑΣ "ΕΛΕΝΗ Θ. ΔΗΜΗΤΡΙΟΥ"

Πίνακας:Γενικά Νοσοκομεία 3ης Υγειονομικής Περιφέρειας

Ψυχιατρικό Νοσοκομείο 3ης Υγειονομικής Περιφέρειας
ΨΥΧΙΑΤΡΙΚΟ ΝΟΣΟΚΟΜΕΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ

Πίνακας:Ψυχιατρικό Νοσοκομείο 3ης Υγειονομικής Περιφέρειας

Γενικά Νοσοκομεία 5ης Υγειονομικής Περιφέρειας
Γ.Ν. ΑΜΦΙΣΣΑΣ
Γ.Ν. ΒΟΛΟΥ "ΑΧΙΛΛΟΠΟΥΛΕΙΟ"
Γ.Ν. ΘΗΒΩΝ
Γ.Ν. ΚΑΡΔΙΤΣΑΣ
Γ.Ν. ΚΑΡΠΕΝΗΣΙΟΥ
Γ.Ν. ΛΑΜΙΑΣ
Γ.Ν. ΛΑΡΙΣΑΣ "ΚΟΥΤΛΙΜΠΑΝΕΙΟ"
Γ.Ν. ΛΙΒΑΔΕΙΑΣ
Γ.Ν. ΤΡΙΚΑΛΩΝ
Γ.Ν. ΧΑΛΚΙΔΑΣ

Πίνακας:Γενικά Νοσοκομεία 5ης Υγειονομικής Περιφέρειας

Κέντρα Υγείας-Γενικά Νοσοκομεία
Γ.Ν.- Κ.Υ. ΚΑΡΥΣΤΟΥ
Γ.Ν.- Κ.Υ. ΚΥΜΗΣ

Πίνακας:Κέντρα Υγείας-Γενικά Νοσοκομεία 5ης Υγειονομικής Περιφέρειας

Πανεπιστημιακό Νοσοκομείο
ΠΑΝΕΠΙΣΤΗΜΙΑΚΟ Γ.Ν. ΛΑΡΙΣΑΣ

Πίνακας:Πανεπιστημιακό Νοσοκομείο 5ης Υγειονομικής Περιφέρειας

Πιο συγκεκριμένα, στο πρώτο μέρος του κεφαλαίου αυτού θα αναλυθούν και θα αναπαρασταθούν τα δεδομένα των νοσοκομείων στην R με τη βοήθεια των οικονομικών δεικτών ενώ στο δεύτερο μέρος με τη βοήθεια των λειτουργικών δεικτών.

5.1.ΜΕΣΟ ΚΟΣΤΟΣ ΦΑΡΜΑΚΩΝ

5.1.1.ΜΕΣΟ ΚΟΣΤΟΣ ΦΑΡΜΑΚΩΝ ΑΝΑ ΑΣΘΕΝΗ ΚΑΙ ΑΝΑ ΗΜΕΡΑ ΝΟΣΗΛΕΙΑΣ 5ΗΣ ΥΓΕΙΟΝΟΜΙΚΗΣ ΠΕΡΙΦΕΡΕΙΑΣ ΜΕ GEOM-COL

- *Βήμα 1ο*: Στο βήμα αυτό φορτώνονται τα Νοσοκομεία με τα δεδομένα τους (σύνολο ασθενών, σύνολο ημερών νοσηλείας και συνολικό κόστος φαρμάκων) της 5ης Υγειονομικής Περιφέρειας από το αρχείο Excel στο RStudio:

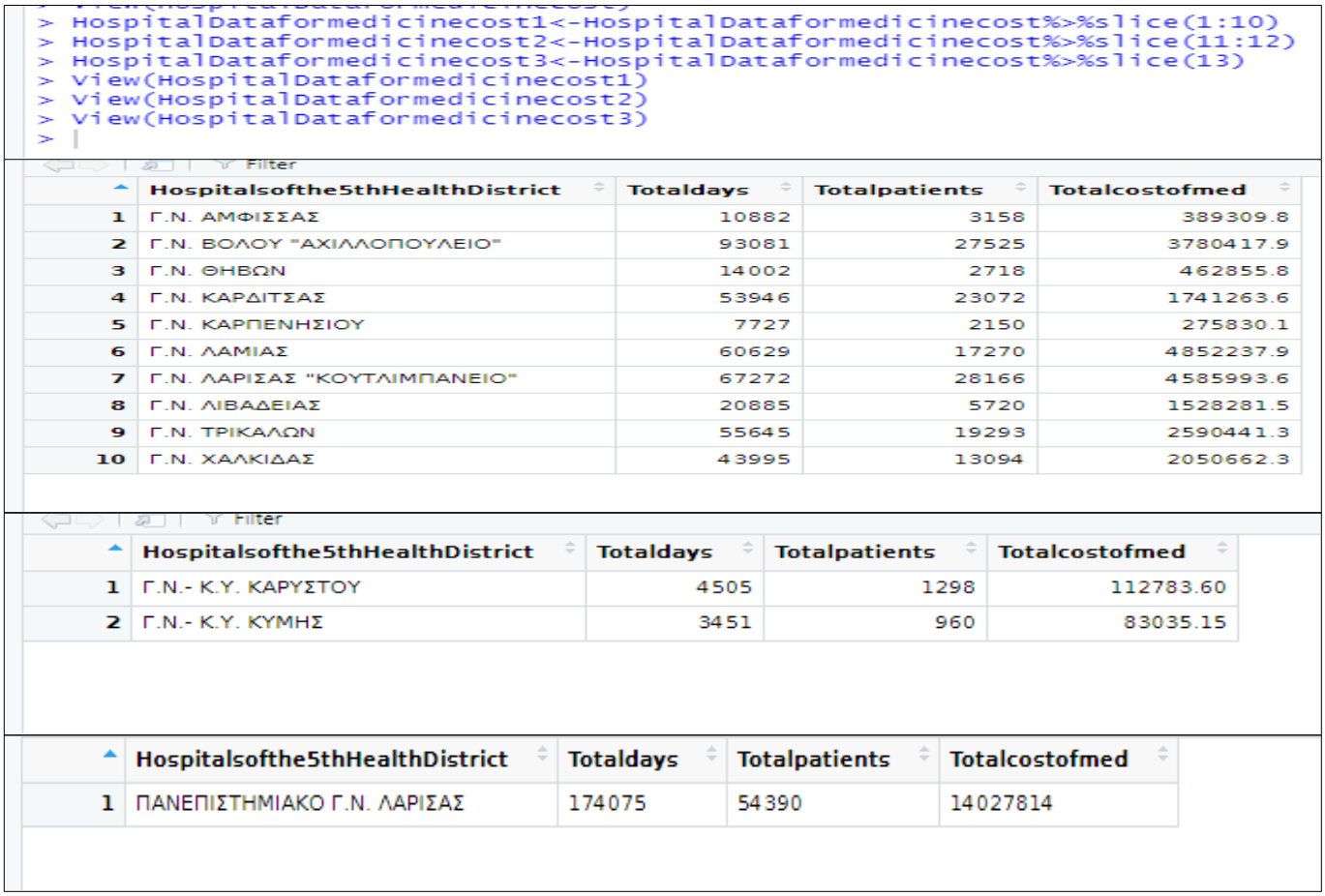
```
> library(readxl)
> HospitalDataformedicinecost <- read_excel("HospitalDataformedicinecost.xlsx")
> View(HospitalDataformedicinecost)
>
```

Filter

	HospitalsoftheSthHealthDistrict	Totaldays	Totalpatients	Totalcostofmed
1	Γ.Ν. ΑΜΦΙΣΣΑΣ	10882	3158	389309.81
2	Γ.Ν. ΒΟΛΟΥ "ΑΧΙΛΛΟΠΟΥΛΕΙΟ"	93081	27525	3780417.85
3	Γ.Ν. ΘΗΒΩΝ	14002	2718	462855.83
4	Γ.Ν. ΚΑΡΔΙΤΣΑΣ	53946	23072	1741263.64
5	Γ.Ν. ΚΑΡΠΕΝΗΣΙΟΥ	7727	2150	275830.14
6	Γ.Ν. ΛΑΜΙΑΣ	60629	17270	4852237.91
7	Γ.Ν. ΛΑΡΙΣΑΣ "ΚΟΥΤΛΙΜΠΑΝΕΙΟ"	67272	28166	4585993.60
8	Γ.Ν. ΛΙΒΑΔΕΙΑΣ	20885	5720	1528281.51
9	Γ.Ν. ΤΡΙΚΑΛΩΝ	55645	19293	2590441.31
10	Γ.Ν. ΧΑΛΚΙΔΑΣ	43995	13094	2050662.28
11	Γ.Ν.- Κ.Υ. ΚΑΡΥΣΤΟΥ	4505	1298	112783.60
12	Γ.Ν.- Κ.Υ. ΚΥΜΗΣ	3451	960	83035.15
13	ΠΑΝΕΠΙΣΤΗΜΙΑΚΟ Γ.Ν. ΛΑΡΙΣΑΣ	174075	54390	14027813.85

Εικόνα:Εμφάνιση των νοσοκομείων με τα δεδομένα τους

- *Βήμα 2ο*: Στην συνέχεια πραγματοποιείται ο διαχωρισμός των Νοσοκομείων με βάση το είδος τους με την βοήθεια της εντολής slice. Τα πρώτα 10 Νοσοκομεία είναι τα Γενικά Νοσοκομεία τα επόμενα 2 είναι τα Κέντρα Υγείας – Γενικά Νοσοκομεία και το τελευταίο είναι το Πανεπιστημιακό Νοσοκομείο:



```
> HospitalDataformedicinecost1<-HospitalDataformedicinecost%>%slice(1:10)
> HospitalDataformedicinecost2<-HospitalDataformedicinecost%>%slice(11:12)
> HospitalDataformedicinecost3<-HospitalDataformedicinecost%>%slice(13)
> View(HospitalDataformedicinecost1)
> View(HospitalDataformedicinecost2)
> View(HospitalDataformedicinecost3)
>
```

Filter

	Hospitalsofthe5thHealthDistrict	Totaldays	Totalpatients	Totalcostofmed
1	Γ.Ν. ΑΜΦΙΣΣΑΣ	10882	3158	389309.8
2	Γ.Ν. ΒΟΛΟΥ "ΑΧΙΛΛΟΠΟΥΛΕΙΟ"	93081	27525	3780417.9
3	Γ.Ν. ΘΗΒΩΝ	14002	2718	462855.8
4	Γ.Ν. ΚΑΡΔΙΤΣΑΣ	53946	23072	1741263.6
5	Γ.Ν. ΚΑΡΠΕΝΗΣΙΟΥ	7727	2150	275830.1
6	Γ.Ν. ΛΑΜΙΑΣ	60629	17270	4852237.9
7	Γ.Ν. ΛΑΡΙΣΑΣ "ΚΟΥΤΛΙΜΠΑΝΕΙΟ"	67272	28166	4585993.6
8	Γ.Ν. ΛΙΒΑΔΕΙΑΣ	20885	5720	1528281.5
9	Γ.Ν. ΤΡΙΚΑΛΩΝ	55645	19293	2590441.3
10	Γ.Ν. ΧΑΛΚΙΔΑΣ	43995	13094	2050662.3

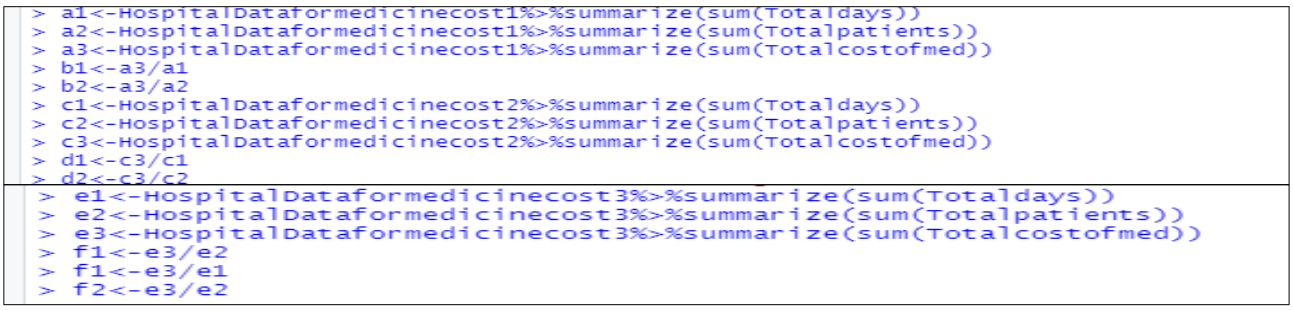
Filter

	Hospitalsofthe5thHealthDistrict	Totaldays	Totalpatients	Totalcostofmed
1	Γ.Ν.- Κ.Υ. ΚΑΡΥΣΤΟΥ	4505	1298	112783.60
2	Γ.Ν.- Κ.Υ. ΚΥΜΗΣ	3451	960	83035.15

	Hospitalsofthe5thHealthDistrict	Totaldays	Totalpatients	Totalcostofmed
1	ΠΑΝΕΠΙΣΤΗΜΙΑΚΟ Γ.Ν. ΛΑΡΙΣΑΣ	174075	54390	14027814

Εικόνα:Διαχωρισμός των νοσοκομείων ανάλογα με το είδος τους

- *Βήμα 3ο:*Στην συνέχεια υπολογίζεται ο συνολικός αριθμός των ασθενών, το συνολικό κόστος για τα φάρμακα και οι συνολικές ημέρες νοσηλείας για κάθε είδος Νοσοκομείου όπου τα αποτελέσματα που προκύπτουν τοποθετούνται σε μεταβλητές:



```
> a1<-HospitalDataformedicinecost1%>%summarize(sum(Totaldays))
> a2<-HospitalDataformedicinecost1%>%summarize(sum(Totalpatients))
> a3<-HospitalDataformedicinecost1%>%summarize(sum(Totalcostofmed))
> b1<-a3/a1
> b2<-a3/a2
> c1<-HospitalDataformedicinecost2%>%summarize(sum(Totaldays))
> c2<-HospitalDataformedicinecost2%>%summarize(sum(Totalpatients))
> c3<-HospitalDataformedicinecost2%>%summarize(sum(Totalcostofmed))
> d1<-c3/c1
> d2<-c3/c2
> e1<-HospitalDataformedicinecost3%>%summarize(sum(Totaldays))
> e2<-HospitalDataformedicinecost3%>%summarize(sum(Totalpatients))
> e3<-HospitalDataformedicinecost3%>%summarize(sum(Totalcostofmed))
> f1<-e3/e2
> f1<-e3/e1
> f2<-e3/e2
```

Εικόνα:Εύρεση αθροισμάτων

Το a1 είναι οι συνολικές ημέρες, το a2 είναι το σύνολο των ασθενών και το a3 είναι το κόστος φαρμάκων των Γενικών Νοσοκομείων. Το b1 και το b2 είναι το συνολικό κόστος φαρμάκων ανά ημέρα νοσηλείας και ανά ασθενή αντίστοιχα. Το c1, c2 και c3 είναι το σύνολο των ημερών,ασθενών και του κόστους των φαρμάκων των Κ.Υ.-Γ.Ν αντίστοιχα ενώ το d1 είναι το μέσο κόστος φαρμάκων ανά ημέρα νοσηλείας ενώ το d2 είναι το μέσο κόστος φαρμάκων ανά ασθενή. Τέλος, το e1, e2, e3 είναι το σύνολο των ημερών, ασθενών και του κόστους των φαρμάκων του Πανεπιστημιακού Νοσοκομείου και το f1 και το f2 είναι το συνολικό κόστος φαρμάκων ανά ημέρα νοσηλείας και ανά ασθενή αντίστοιχα.

- *Βήμα 4ο:* Έπειτα, κατασκευάζουμε έναν πίνακα με 3 μόνο παρατηρήσεις (γραμμές) με την εντολή tibble για τα Γενικά Νοσοκομεία, τα Κέντρα Υγείας και το Πανεπιστημιακό Νοσοκομείο. Ο πίνακας αυτός περιλαμβάνει τα αποτελέσματα του 3ου βήματος:

```
>
>
>
>
> mydata<-tibble(hospitals=c("General Hospitals","Health Center","University Hospital"),numberofpatients=c(a2,c2,e2),numberofdays=c(a1,c1,e1),costofm
ed=c(a3,c3,e3),costperday=c(b1,d1,f1),costperpatient=c(b2,d2,f2))
>
```

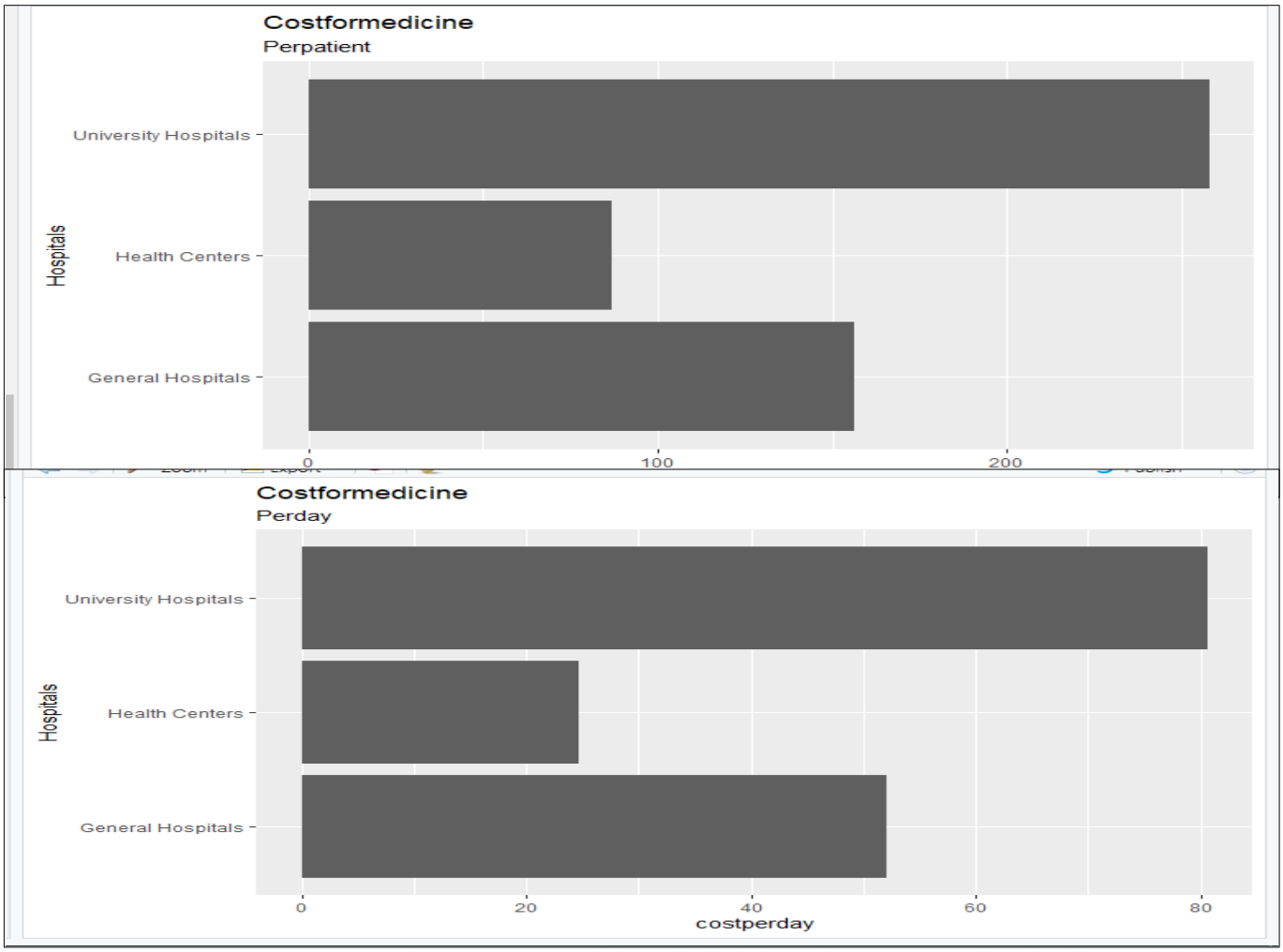
Filter

	hospitals	numberofpatients	numberofdays	costofmed	costperday	costperpatient
1	General Hospitals	142166	428064	22257293.88	51.995248093743	156.55848712069
2	Health Center	2258	7956	195818.75	24.6127136752137	86.7222099202834
3	University Hospital	54390	174075	14027813.85	80.5848849633778	257.911635410921

Εικόνα:Πίνακας για τις παρατηρήσεις με τα συνολικά στοιχεία

Με βάση τα αποτελέσματα αυτά μπορούμε να εμφανίσουμε τα αντίστοιχα διαγράμματα (bar plots):

```
>
>
> mydata%>%ggplot(aes(x=costperpatient,y=Hospitals))+geom_col()+labs(title="Costformedicine",subtitle="Perpatient")
> mydata%>%ggplot(aes(x=costperday,y=Hospitals))+geom_col()+labs(title="Costformedicine",subtitle="Perday")
>
```



Εικόνα:Εμφάνιση αποτελεσμάτων

Ακολουθώντας τα βήματα αυτά, καταλήξαμε σε ένα διάγραμμα το οποίο μας εμφανίζει το κόστος φαρμάκων ανά ασθενή και ανά ημέρα νοσηλείας στις ομάδες των νοσοκομείων της 5ης ΥΠΕ. Όπως φαίνεται το Πανεπιστημιακό Νοσοκομείο της Λάρισας έχει το μέγιστο κόστος φαρμάκων ανά ασθενή και ανά ημέρα νοσηλείας και ακολουθούν στην συνέχεια τα Γενικά Νοσοκομεία και τέλος τα Κέντρα Υγείας-Γενικά Νοσοκομεία. Πιθανώς, τα Κέντρα Υγείας-Γενικά Νοσοκομεία έχουν το πιο μικρό κόστος φαρμάκων και ανά ασθενή και ανά ημέρα νοσηλείας διότι είναι Νοσοκομεία μικρότερης χωρητικότητας από τα υπόλοιπα και επίσης συνήθως οι ασθενείς μεταφέρονται σε μεγαλύτερα Νοσοκομεία και δεν νοσηλεύονται σε αυτά, με αποτέλεσμα να μην γίνεται μεγάλη χρήση φαρμάκων.

5.1.2.ΘΗΚΟΓΡΑΜΜΑ ΓΙΑ ΤΟ ΜΕΣΟ ΚΟΣΤΟΣ ΦΑΡΜΑΚΩΝ ΑΝΑ ΑΣΘΕΝΗ ΤΗΣ 3ΗΣ ΚΑΙ ΤΗΣ 5ΗΣ ΥΓΕΙΟΝΟΜΙΚΗΣ ΠΕΡΙΦΕΡΕΙΑΣ ΣΥΝΟΛΙΚΑ

- *Βήμα 1ο:*Αρχικά, πραγματοποιείται η φόρτωση των δεδομένων από το Excel στο RStudio που περιλαμβάνουν το συνολικό μέσο κόστος για τα φάρμακα.

	Hospitals	Health District	Totalcostformedperpatient
1	Γ.Ν. "ΠΑΠΑΓΕΩΡΓΙΟΥ"	3	344.44520
2	Γ.Ν. ΒΕΡΟΙΑΣ	3	144.25558
3	Γ.Ν. ΓΙΑΝΝΙΤΣΩΝ	3	143.72806
4	Γ.Ν. ΓΡΕΒΕΝΩΝ	3	71.64021
5	Γ.Ν. ΕΔΕΣΣΑΣ	3	111.96839
6	Γ.Ν. ΘΕΣ/ΝΙΚΗΣ "ΑΓ. ΔΗΜΗΤΡΙΟΣ"	3	96.83919
7	Γ.Ν. ΘΕΣ/ΝΙΚΗΣ "Γ. ΓΕΝΝΗΜΑΤΑΣ"	3	109.97213
8	Γ.Ν. ΘΕΣ/ΝΙΚΗΣ "Γ. ΠΑΠΑΝΙΚΟΛΑΟΥ"	3	237.45468
9	Γ.Ν. ΚΑΣΤΟΡΙΑΣ	3	69.08837
10	Γ.Ν. ΚΑΤΕΡΙΝΗΣ	3	97.01601
11	Γ.Ν. ΚΟΖΑΝΗΣ "ΜΑΜΑΤΣΕΙΟ"	3	128.08967
12	Γ.Ν. ΝΑΟΥΣΑΣ	3	60.41141
13	Γ.Ν. ΠΤΟΛΕΜΑΪΔΑΣ "ΜΠΟΔΟΣΑΚΕΙΟ"	3	162.48182
14	Γ.Ν. ΦΛΩΡΙΝΑΣ "ΕΛΕΝΗ Θ. ΔΗΜΗΤΡΙΟΥ"	3	87.04233
15	ΨΥΧΙΑΤΡΙΚΟ ΝΟΣΟΚΟΜΕΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ	3	53.41106
16	Γ.Ν. ΑΜΦΙΣΣΑΣ	5	123.27733
17	Γ.Ν. ΒΟΛΟΥ "ΑΧΙΛΛΟΠΟΥΛΕΙΟ"	5	137.34488
18	Γ.Ν. ΘΗΒΩΝ	5	170.29280
19	Γ.Ν. ΚΑΡΔΙΤΣΑΣ	5	75.47086
20	Γ.Ν. ΚΑΡΠΕΝΗΣΙΟΥ	5	128.29309
21	Γ.Ν. ΛΑΜΙΑΣ	5	280.96340
22	Γ.Ν. ΛΑΡΙΣΑΣ "ΚΟΥΤΛΙΜΠΑΝΕΙΟ"	5	162.82019
23	Γ.Ν. ΛΙΒΑΔΕΙΑΣ	5	267.18208
24	Γ.Ν. ΤΡΙΚΑΛΩΝ	5	134.26846
25	Γ.Ν. ΧΑΛΚΙΔΑΣ	5	156.61084
26	Γ.Ν.- Κ.Υ. ΚΑΡΥΣΤΟΥ	5	86.89029
27	Γ.Ν.- Κ.Υ. ΚΥΜΗΣ	5	86.49495
28	ΠΑΝΕΠΙΣΤΗΜΙΑΚΟ Γ.Ν. ΛΑΡΙΣΑΣ	5	257.91164

Εικόνα:Φόρτωση δεδομένων

- *Βήμα 2ο:* Στην συνέχεια, δημιουργείται το θηκόγραμμα το οποίο έχει σαν άξονα x τις Υγειονομικές Περιφέρειες και σαν άξονα y το συνολικό μέσο κόστος φαρμάκων.

```
>
>
>
> HD%>%filter(`Health District` %in% c("3","5"))%>%ggplot(aes(x=`Health District`,y=Totalcostformedperpatient))+geom_boxplot()+labs(title
="Boxplot")
>
```

Εικόνα:Δημιουργία θηκογράμματος

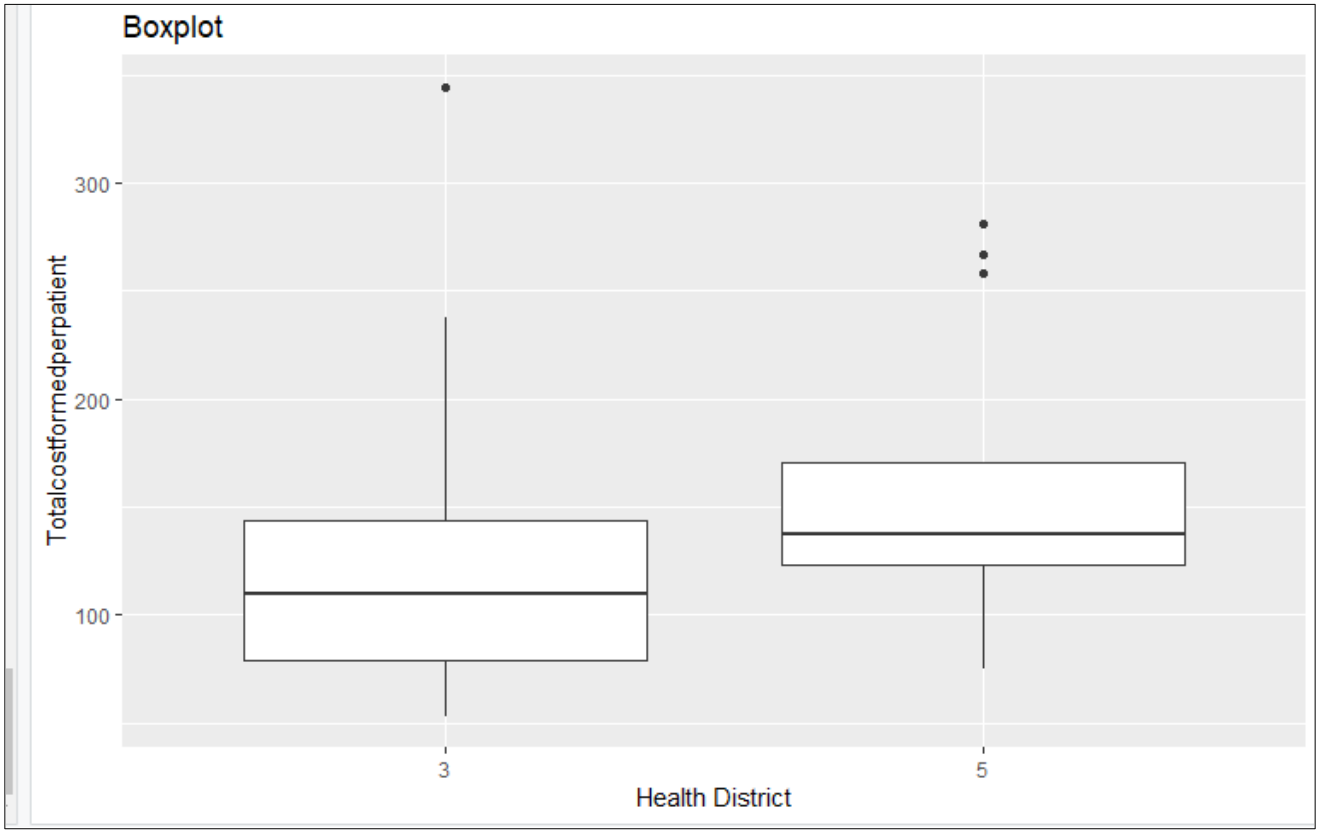
- *Βήμα 3ο:* Στο τελευταίο αυτό βήμα δημιουργείται το θηκόγραμμα που θέλουμε από το οποίο προκύπτουν τα εξής συμπεράσματα:

1 Στην 3η Υγειονομική Περιφέρεια υπάρχει μία ακραία μεγάλη τιμή που είναι μεγαλύτερη από την τιμή 300 (έκτροπη παρατήρηση), ενώ στην 5η Υγειονομική Περιφέρεια οι ακραίες τιμές είναι τρεις.

2 Η διάμεσος της 3ης Υγειονομικής Περιφέρειας είναι σχεδόν 100(χωρίς ασυμμετρία) ενώ της 5ης Υγειονομικής Περιφέρειας είναι λίγο μεγαλύτερη(θετική ασυμμετρία=τα δεδομένα που είναι μεγαλύτερα από τη διάμεσο έχουν μεγαλύτερη διακύμανση).

3 Αφού υπάρχουν ακραίες τιμές και οι απολήξεις είναι ασύμμετρες συμπεραίνουμε ότι οι τιμές δεν ακολουθούν κανονική κατανομή.

4 Η 3η Υγειονομική Περιφέρεια έχει τη μεγαλύτερη διακύμανση (απόσταση των παρατηρήσεων από την μέση τιμή) του δείκτη αυτού λόγω του μεγάλου "πάχους" του κουτιού.



Εικόνα:Εμφάνιση θηκογράμματος

5.2.ΜΕΣΟ ΚΟΣΤΟΣ ΥΛΙΚΩΝ ΚΑΙ ΑΝΑΛΩΣΙΜΩΝ

5.2.1.ΜΕΣΟ ΚΟΣΤΟΣ ΥΛΙΚΩΝ ΚΑΙ ΑΝΑΛΩΣΙΜΩΝ ΑΝΑ ΑΣΘΕΝΗ ΚΑΙ ΑΝΑ ΗΜΕΡΑ ΝΟΣΗΛΕΙΑΣ 5ης ΥΓΕΙΟΝΟΜΙΚΗΣ ΠΕΡΙΦΕΡΕΙΑΣ ΜΕ GEOM-COL

- *Βήμα 1ο*:

```
> library(readxl)
> HospitalDataformaterialscost <- read_excel("HospitalDataformaterialscost.xlsx")
> View(HospitalDataformaterialscost)
```

	HospitalsofSthHealthDistrict	Totaldays	Totalpatients	HygMaterial	OrthopMaterial	Reag.	OtherMat	Gases	Fuels	Other
1	Γ.Ν. ΑΜΦΙΣΣΑΣ	10882	3158	498360.89	207861.57	227871.58	132718.94	42024.19	230940.74	89072.91
2	Γ.Ν. ΒΟΛΟΥ "ΑΧΙΛΛΟΠΟΥΛΕΙΟ"	93081	27525	2843298.51	678146.82	1813931.49	497902.99	170639.34	456436.37	633035.23
3	Γ.Ν. ΘΗΒΩΝ	14002	2718	217608.09	133430.26	136253.63	109171.75	0.00	362493.57	66648.78
4	Γ.Ν. ΚΑΡΔΙΤΣΑΣ	53946	23072	1202215.35	503627.49	678962.69	300285.62	247791.53	474464.90	217312.55
5	Γ.Ν. ΚΑΡΠΕΝΗΣΙΟΥ	7727	2150	247475.47	67910.06	99903.56	75448.28	3181.26	164752.16	56202.78
6	Γ.Ν. ΛΑΜΙΑΣ	60629	17270	2458330.28	562382.85	1330661.72	784263.83	682101.28	168168.06	366042.75
7	Γ.Ν. ΛΑΡΙΣΑΣ "ΚΟΥΤΛΙΜΠΑΝΕΙΟ"	67272	28166	2691647.59	409754.03	1717072.09	440729.82	155833.80	557394.24	679314.47
8	Γ.Ν. ΛΙΒΑΔΕΙΑΣ	20885	5720	928701.87	231433.70	392590.99	263969.90	51736.62	335272.52	202819.52
9	Γ.Ν. ΤΡΙΚΑΛΩΝ	55645	19293	1619235.20	362935.88	762901.09	204943.97	NA	439394.30	241112.44
10	Γ.Ν. ΧΑΛΚΙΔΑΣ	43995	13094	1006175.24	408326.61	818743.27	372410.74	97154.73	203482.79	315167.33
11	Γ.Ν.- Κ.Υ. ΚΑΡΥΣΤΟΥ	4505	1298	42557.78	69679.44	122938.35	43529.85	14609.78	13675.68	35984.45
12	Γ.Ν.- Κ.Υ. ΚΥΜΗΣ	3451	960	85455.79	40599.77	94556.49	41808.61	20174.88	35796.83	42348.10
13	ΠΑΝΕΠΙΣΤΗΜΙΑΚΟ Γ.Ν. ΛΑΡΙΣΑΣ	174075	54390	10709958.30	1577620.71	3999582.72	1140145.36	NA	1137481.29	1032110.34

Εικόνα:Νοσοκομεία της 5ης ΥΠΕ με τα στοιχεία τους

- *Βήμα 2ο*:

```
>
>
> HospitalDataformaterialscost1<-HospitalDataformaterialscost%>%slice(1:10)
> HospitalDataformaterialscost2<-HospitalDataformaterialscost%>%slice(11:12)
> HospitalDataformaterialscost3<-HospitalDataformaterialscost%>%slice(13)
>
```

HospitalDataformaterialscost1 | HospitalDataformaterialscost2 | HospitalDataformaterialscost3

Filter

	HospitalsofSthHealthDistrict	Totaldays	Totalpatients	HygMaterial	OrthopMaterial	Reag.	OtherMat	Gases	Fuels	Other
1	Γ.Ν. ΑΜΦΙΣΣΑΣ	10882	3158	498360.9	207861.57	227871.58	132718.94	42024.19	230940.7	89072.91
2	Γ.Ν. ΒΟΛΟΥ "ΑΧΙΛΛΟΠΟΥΛΕΙΟ"	93081	27525	2843298.5	678146.82	1813931.49	497902.99	170639.34	456436.4	633035.23
3	Γ.Ν. ΘΗΒΩΝ	14002	2718	217608.1	133430.26	136253.63	109171.75	0.00	362493.6	66648.78
4	Γ.Ν. ΚΑΡΔΙΤΣΑΣ	53946	23072	1202215.4	503627.49	678962.69	300285.62	247791.53	474464.9	217312.55
5	Γ.Ν. ΚΑΡΠΕΝΗΣΙΟΥ	7727	2150	247475.5	67910.06	99903.56	75448.28	3181.26	164752.2	56202.78
6	Γ.Ν. ΛΑΜΙΑΣ	60629	17270	2458330.3	562382.85	1330661.72	784263.83	682101.28	168168.1	366042.75
7	Γ.Ν. ΛΑΡΙΣΑΣ "ΚΟΥΤΛΙΜΠΑΝΕΙΟ"	67272	28166	2691647.6	409754.03	1717072.09	440729.82	155833.80	557394.2	679314.47
8	Γ.Ν. ΛΙΒΑΔΕΙΑΣ	20885	5720	928701.9	231433.70	392590.99	263969.90	51736.62	335272.5	202819.52
9	Γ.Ν. ΤΡΙΚΑΛΩΝ	55645	19293	1619235.2	362935.88	762901.09	204943.97	NA	439394.3	241112.44
10	Γ.Ν. ΧΑΛΚΙΔΑΣ	43995	13094	1006175.2	408326.61	818743.27	372410.74	97154.73	203482.8	315167.33

HospitalDataformaterialscost1 | HospitalDataformaterialscost2 | HospitalDataformaterialscost3

Filter

	HospitalsofSthHealthDistrict	Totaldays	Totalpatients	HygMaterial	OrthopMaterial	Reag.	OtherMat	Gases	Fuels	Other
1	Γ.Ν.- Κ.Υ. ΚΑΡΥΣΤΟΥ	4505	1298	42557.78	69679.44	122938.35	43529.85	14609.78	13675.68	35984.45
2	Γ.Ν.- Κ.Υ. ΚΥΜΗΣ	3451	960	85455.79	40599.77	94556.49	41808.61	20174.88	35796.83	42348.10

HospitalDataformaterialscost1 | HospitalDataformaterialscost2 | HospitalDataformaterialscost3

Filter

	HospitalsofSthHealthDistrict	Totaldays	Totalpatients	HygMaterial	OrthopMaterial	Reag.	OtherMat	Gases	Fuels	Other
1	ΠΑΝΕΠΙΣΤΗΜΙΑΚΟ Γ.Ν. ΛΑΡΙΣΑΣ	174075	54390	10709958	1577621	3999583	1140145	NA	1137481	1032110

Εικόνα:Χρήση της εντολής slice

- *Βήμα 3ο:*Στο βήμα αυτό ακολουθείται η ίδια μεθοδολογία με την διαφορά ότι χρησιμοποιείται η εντολή mutate για να δημιουργήσουμε μια νέα στήλη η οποία θα περιλαμβάνει το συνολικό κόστος για τα υλικά και τα αναλώσιμα των νοσοκομείων:

```
> HospitalDataformaterialscost1%>%mutate(summarterials=HygMaterial+OrthopMaterial+Reag.+OtherMat+Fuels+Gases+Other)
# A tibble: 10 x 11
   HospitalsofSthH~ Totaldays Totalpatients HygMaterial
   <chr>               <dbl>         <dbl>       <dbl>
 1 "Γ.Ν. ΑΜΦΙΣΣΑΣ"      10882          3158     498361.
 2 "Γ.Ν. ΒΟΛΟΥ \"Α~      93081         27525    2843299.
 3 "Γ.Ν. ΘΗΒΩΝ"          14002          2718     217608.
 4 "Γ.Ν. ΚΑΡΔΙΤΣΑΣ"      53946         23072    1202215.
 5 "Γ.Ν. ΚΑΡΠΕΝΗΣΙ~       7727          2150     247475.
 6 "Γ.Ν. ΛΑΜΙΑΣ"         60629         17270    2458330.
 7 "Γ.Ν. ΛΑΡΙΣΑΣ \~      67272         28166    2691648.
 8 "Γ.Ν. ΛΙΒΑΔΕΙΑΣ"      20885          5720     928702.
 9 "Γ.Ν. ΤΡΙΚΑΛΩΝ"       55645         19293    1619235.
10 "Γ.Ν. ΧΑΛΚΙΔΑΣ"       43995         13094    1006175.
# ... with 7 more variables: OrthopMaterial <dbl>, Reag. <dbl>,
#   OtherMat <dbl>, Gases <dbl>, Fuels <dbl>, Other <dbl>,
#   summarterials <dbl>
> HospitalDataformaterialscost2%>%mutate(summarterials=HygMaterial+OrthopMaterial+Reag.+OtherMat+Fuels+Gases+Other)
# A tibble: 2 x 11
  HospitalsofSthH~ Totaldays Totalpatients HygMaterial
  <chr>               <dbl>         <dbl>       <dbl>
1 Γ.Ν.- Κ.Υ. ΚΑΡΥ~       4505          1298       42558.
2 Γ.Ν.- Κ.Υ. ΚΥΜΗΣ        3451           960       85456.
# ... with 7 more variables: OrthopMaterial <dbl>, Reag. <dbl>,
#   OtherMat <dbl>, Gases <dbl>, Fuels <dbl>, Other <dbl>,
#   summarterials <dbl>
> HospitalDataformaterialscost3%>%mutate(summarterials=HygMaterial+OrthopMaterial+Reag.+OtherMat+Fuels+Gases+Other)
# A tibble: 1 x 11
  HospitalsofSthH~ Totaldays Totalpatients HygMaterial
  <chr>               <dbl>         <dbl>       <dbl>
1 ΠΑΝΕΠΙΣΤΗΜΙΑΚΟ ~    174075         54390    10709958.
# ... with 7 more variables: OrthopMaterial <dbl>, Reag. <dbl>,
#   OtherMat <dbl>, Gases <dbl>, Fuels <dbl>, Other <dbl>,
#   summarterials <dbl>
```

```
> a1<-HospitalDataformaterialscost1%>%summarize(sum(Totaldays))
> a2<-HospitalDataformaterialscost1%>%summarize(sum(Totalpatients))
> a3<-HospitalDataformaterialscost1%>%summarize(sum(HygMaterial))
> a4<-HospitalDataformaterialscost1%>%summarize(sum(OrthopMaterial))
> a5<-HospitalDataformaterialscost1%>%summarize(sum(Reag.))
> a6<-HospitalDataformaterialscost1%>%summarize(sum(OtherMat))
```

```
> a7<-HospitalDataformaterialscost1%>%summarize(sum(Gases,na.rm=TRUE))
> a7
# A tibble: 1 x 1
  `sum(Gases, na.rm = TRUE)`
                       <dbl>
1                    1450463.
```

```
> a10<-sum(a3,a4,a5,a6,a7,a8,a9)
> b10<-sum(b3,b4,b5,b6,b7,b8,b9)
> c10<-sum(c3,c4,c5,c6,c8,c9)
```

```
> mpp1<-a10/a2
> view(mpp1)
> mpp2<-b10/b2
> mpp3<-c10/c2
```

```
> mpd1<-a10/a1
> mpd2<-b10/b1
> mpd3<-c10/c1
>
```

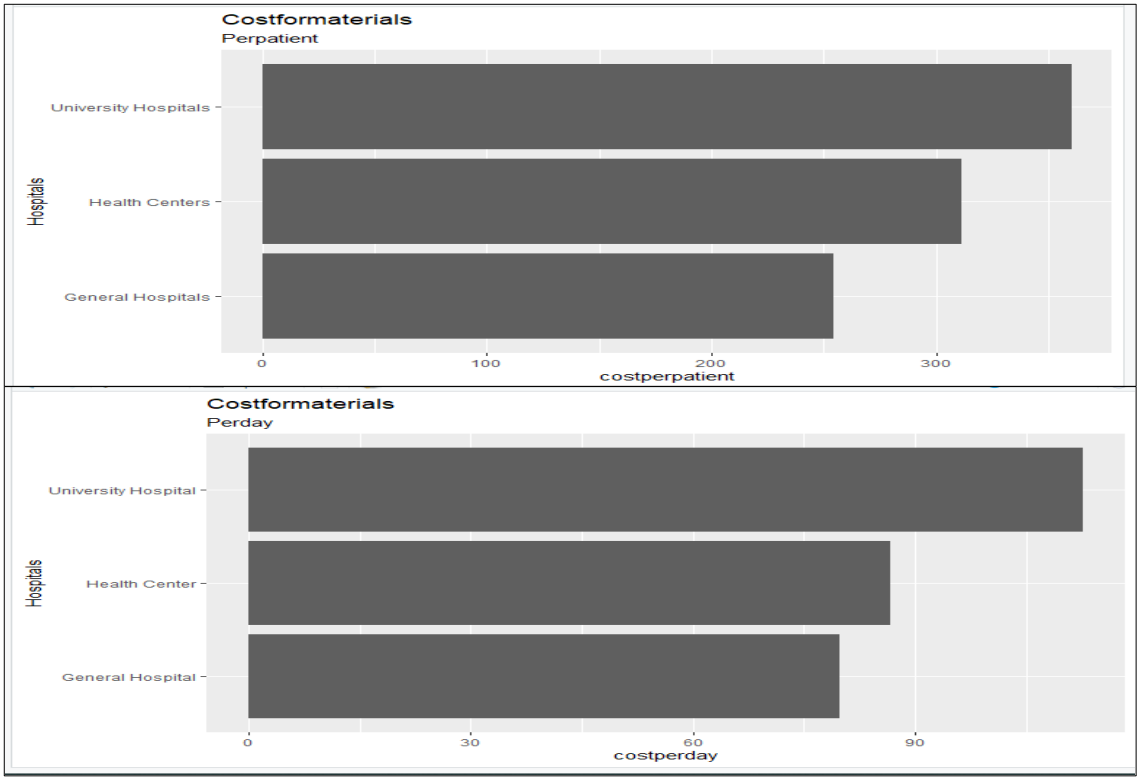
```
> a8<-HospitalDataformaterialscost1%>%summarize(sum(Fuels))
> a9<-HospitalDataformaterialscost1%>%summarize(sum(Other))
> b1<-HospitalDataformaterialscost2%>%summarize(sum(Totaldays))
> b2<-HospitalDataformaterialscost2%>%summarize(sum(Totalpatients))
> b3<-HospitalDataformaterialscost2%>%summarize(sum(HygMaterial))
> b4<-HospitalDataformaterialscost2%>%summarize(sum(OrthopMaterial))
> b5<-HospitalDataformaterialscost2%>%summarize(sum(Reag.))
> b6<-HospitalDataformaterialscost2%>%summarize(sum(OtherMat))
> b7<-HospitalDataformaterialscost2%>%summarize(sum(Gases))
> b8<-HospitalDataformaterialscost2%>%summarize(sum(Fuels))
> b9<-HospitalDataformaterialscost2%>%summarize(sum(Other))
> c1<-HospitalDataformaterialscost3%>%summarize(sum(Totaldays))
> c2<-HospitalDataformaterialscost3%>%summarize(sum(Totalpatients))
> c3<-HospitalDataformaterialscost3%>%summarize(sum(HygMaterial))
> c4<-HospitalDataformaterialscost3%>%summarize(sum(OrthopMaterial))
> c5<-HospitalDataformaterialscost3%>%summarize(sum(Reag.))
> c6<-HospitalDataformaterialscost3%>%summarize(sum(OtherMat))
```

Εικόνα:Εύρεση αθροισμάτων

Βήμα 4ο:

```
> mydata<-tibble(Hospitals=c("General Hospital","Health Center","University Hospital"),numberofpatients=c(a2,b2,c2),numberofdays=c(a1,b
1,c1),costperday=c(mpd1,mpd2,mpd3),costperpatient=c(mpp1,mpp2,mpp3))
> mydata%>%ggplot(aes(x=costperpatient,y=Hospitals))+geom_col()+labs(title="Costformaterials",subtitle="Perpatient")
> mydata%>%ggplot(aes(x=costperday,y=Hospitals))+geom_col()+labs(title="Costformaterials",subtitle="Perday")
>
```

Εικόνα: Κώδικας για την δημιουργία διαγραμμάτων



Εικόνα:Διαγράμματα κόστους αναλώσιμων και υλικών ανά ασθενή και ανά ημέρα νοσηλείας

Με βάση τα διαγράμματα αυτά, *τα Γενικά Νοσοκομεία έχουν το μικρότερο κόστος αναλώσιμων και υλικών ανά ασθενή και ανά ημέρα νοσηλείας, ενώ το μεγαλύτερο κόστος το έχει το Πανεπιστημιακό Νοσοκομείο της Λάρισας.*

5.3.ΜΕΣΟ ΚΟΣΤΟΣ ΥΠΗΡΕΣΙΩΝ

5.3.1.ΜΕΣΟ ΚΟΣΤΟΣ ΥΠΗΡΕΣΙΩΝ ΑΝΑ ΑΣΘΕΝΗ ΚΑΙ ΑΝΑ ΗΜΕΡΑ ΝΟΣΗΛΕΙΑΣ 5ης ΥΓΕΙΟΝΟΜΙΚΗΣ ΠΕΡΙΦΕΡΕΙΑΣ ΜΕ GEOM-COL

- *Βήμα 1ο*:

```
> library(readxl)
> HospitalDataforservicescost <- read_excel("HospitalDataforservicescost.xlsx")
> View(HospitalDataforservicescost)
>
```

	HospitalsoftheSthHealthDistrict	Totaldays	Totalpatients	Supplpayment	Additpayment	Publicservices	Security	Costforcleaning	Catering	Other
1	Γ.Ν. ΑΜΦΙΣΣΑΣ	10882	3158	90531.22	79517.80	137283.26	1551.27	158399.60	NA	226434.0
2	Γ.Ν. ΒΟΛΟΥ "ΑΧΙΛΛΟΠΟΥΛΕΙΟ"	93081	27525	800563.72	414653.08	948206.02	185186.26	599071.50	NA	837525.8
3	Γ.Ν. ΘΗΒΩΝ	14002	2718	95137.68	22322.50	327542.46	1415.03	318924.28	0.00	362114.0
4	Γ.Ν. ΚΑΡΔΙΤΣΑΣ	53946	23072	123374.56	152890.37	508934.17	179823.14	562958.70	NA	754102.3
5	Γ.Ν. ΚΑΡΠΕΝΗΣΙΟΥ	7727	2150	26697.34	20922.88	104923.51	21996.91	78601.70	NA	226962.8
6	Γ.Ν. ΛΑΜΙΑΣ	60629	17270	452080.84	274435.24	1012230.86	331752.22	862189.10	NA	1617364.2
7	Γ.Ν. ΛΑΡΙΣΑΣ "ΚΟΥΤΛΙΜΠΑΝΕΙΟ"	67272	28166	150437.27	157659.78	708903.86	107390.81	554892.29	62922.83	1048961.9
8	Γ.Ν. ΛΙΒΑΔΕΙΑΣ	20885	5720	148069.35	67844.80	436133.37	1502.86	20000.00	NA	915363.6
9	Γ.Ν. ΤΡΙΚΑΛΩΝ	55645	19293	57188.23	166079.49	789479.57	87285.61	736460.39	NA	1368018.9
10	Γ.Ν. ΧΑΛΚΙΔΑΣ	43995	13094	200398.48	114499.55	438770.05	NA	NA	NA	1285275.5
11	Γ.Ν.- Κ.Υ. ΚΑΡΥΣΤΟΥ	4505	1298	36905.87	0.00	63451.45	NA	34112.82	NA	145556.5
12	Γ.Ν.- Κ.Υ. ΚΥΜΗΣ	3451	960	10020.62	0.00	63853.27	3629.65	57392.15	NA	230983.1
13	ΠΑΝΕΠΙΣΤΗΜΙΑΚΟ Γ.Ν. ΛΑΡΙΣΑΣ	174075	54390	387236.73	228047.79	2201596.58	200196.76	1489884.33	NA	5825799.0

Εικόνα:Νοσοκομεία της 5ης ΥΠΕ με τα στοιχεία τους

- *Βήμα 2ο*:

```
>
> HospitalDataforservicescost1<-HospitalDataforservicescost%>%slice(1:10)
> HospitalDataforservicescost2<-HospitalDataforservicescost%>%slice(11:12)
> HospitalDataforservicescost3<-HospitalDataforservicescost%>%slice(13)
>
```

	HospitalsoftheSthHealthDistrict	Totaldays	Totalpatients	Supplpayment	Additpayment	Publicservices	Security	Costforcleaning	Catering	Other
1	Γ.Ν. ΑΜΦΙΣΣΑΣ	10882	3158	90531.22	79517.80	137283.3	1551.27	158399.6	NA	226434.0
2	Γ.Ν. ΒΟΛΟΥ "ΑΧΙΛΛΟΠΟΥΛΕΙΟ"	93081	27525	800563.72	414653.08	948206.0	185186.26	599071.5	NA	837525.8
3	Γ.Ν. ΘΗΒΩΝ	14002	2718	95137.68	22322.50	327542.5	1415.03	318924.3	0.00	362114.0
4	Γ.Ν. ΚΑΡΔΙΤΣΑΣ	53946	23072	123374.56	152890.37	508934.2	179823.14	562958.7	NA	754102.3
5	Γ.Ν. ΚΑΡΠΕΝΗΣΙΟΥ	7727	2150	26697.34	20922.88	104923.5	21996.91	78601.7	NA	226962.8
6	Γ.Ν. ΛΑΜΙΑΣ	60629	17270	452080.84	274435.24	1012230.9	331752.22	862189.1	NA	1617364.2
7	Γ.Ν. ΛΑΡΙΣΑΣ "ΚΟΥΤΛΙΜΠΑΝΕΙΟ"	67272	28166	150437.27	157659.78	708903.9	107390.81	554892.3	62922.83	1048961.9
8	Γ.Ν. ΛΙΒΑΔΕΙΑΣ	20885	5720	148069.35	67844.80	436133.4	1502.86	20000.0	NA	915363.6
9	Γ.Ν. ΤΡΙΚΑΛΩΝ	55645	19293	57188.23	166079.49	789479.6	87285.61	736460.4	NA	1368018.9
10	Γ.Ν. ΧΑΛΚΙΔΑΣ	43995	13094	200398.48	114499.55	438770.0	NA	NA	NA	1285275.5

HospitalDataforservicescost | HospitalDataforservicescost1 | HospitalDataforservicescost2 | HospitalDataforservicescost3

Filter

	HospitalsoftheSthHealthDistrict	Totaldays	Totalpatients	Supplpayment	Additpayment	Publicservices	Security	Costforcleaning	Catering	Other
1	Γ.Ν.- Κ.Υ. ΚΑΡΥΣΤΟΥ	4505	1298	36905.87	0	63451.45	NA	34112.82	NA	145556.5
2	Γ.Ν.- Κ.Υ. ΚΥΜΗΣ	3451	960	10020.62	0	63853.27	3629.65	57392.15	NA	230983.1

HospitalDataforservicescost | HospitalDataforservicescost1 | HospitalDataforservicescost2 | HospitalDataforservicescost3

Filter

	HospitalsoftheSthHealthDistrict	Totaldays	Totalpatients	Supplpayment	Additpayment	Publicservices	Security	Costforcleaning	Catering	Other
1	ΠΑΝΕΠΙΣΤΗΜΙΑΚΟ Γ.Ν. ΛΑΡΙΣΑΣ	174075	54390	387236.7	228047.8	2201597	200196.8	1489884	NA	5825799

Εικόνα:Χρήση της εντολής slice

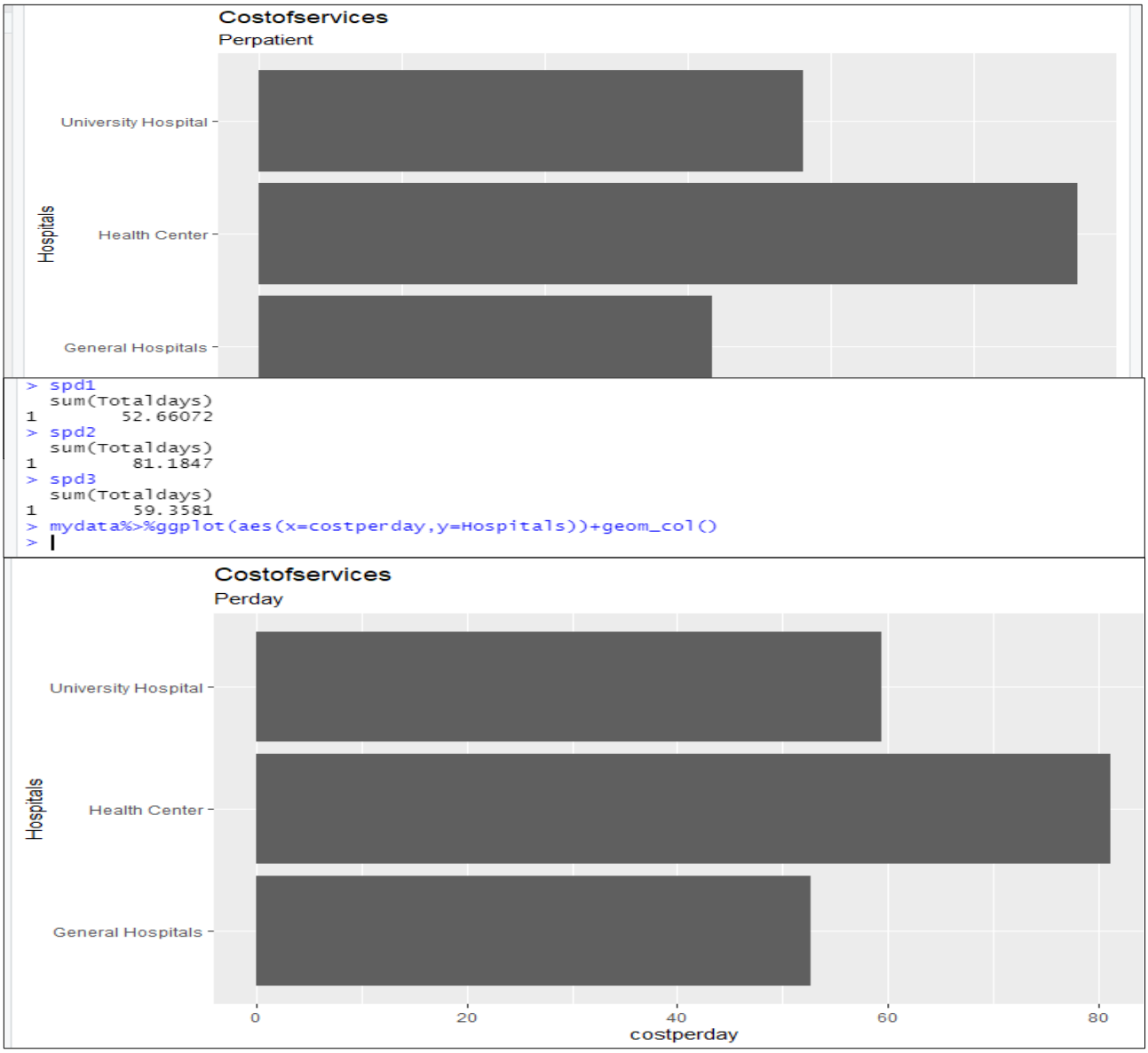
■ *Βήμα 3ο:*

```
> a1<-HospitalDataforservicescost1%>%summarise(sum(Totaldays))
> a2<-HospitalDataforservicescost1%>%summarise(sum(Totalpatients))
> a3<-HospitalDataforservicescost1%>%summarise(sum(Supplpayment))
> a4<-HospitalDataforservicescost1%>%summarise(sum(Additpayment))
> a5<-HospitalDataforservicescost1%>%summarise(sum(Publicservices))
> a6<-HospitalDataforservicescost1%>%summarise(sum(Security,na.rm=TRUE))
> a7<-HospitalDataforservicescost1%>%summarise(sum(Costforcleaning,na.rm=TRUE))
> a8<-HospitalDataforservicescost1%>%summarise(sum(Catering,na.rm=TRUE))
> a9<-HospitalDataforservicescost1%>%summarise(sum(Other))
> a10<-sum(a3,a4,a5,a6,a7,a8,a9)
> b1<-HospitalDataforservicescost2%>%summarise(sum(Totaldays))
> b2<-HospitalDataforservicescost2%>%summarise(sum(Totalpatients))
> b3<-HospitalDataforservicescost2%>%summarise(sum(Supplpayment))
> c1<-HospitalDataforservicescost3%>%summarise(sum(Totaldays))
> c2<-HospitalDataforservicescost3%>%summarise(sum(Totalpatients)
> c3<-HospitalDataforservicescost3%>%summarise(sum(Supplpayment))
> c4<-HospitalDataforservicescost3%>%summarise(sum(Additpayment))
> c5<-HospitalDataforservicescost3%>%summarise(sum(Publicservices
> c6<-HospitalDataforservicescost3%>%summarise(sum(Security))
> c7<-HospitalDataforservicescost3%>%summarise(sum(Costforcleanin
> c9<-HospitalDataforservicescost3%>%summarise(sum(Other))
> c10<-sum(c3,c4,c5,c6,c7,c9)
> spp1<-a10/a2
> spp2<-b10/b2
> spp3<-c10/c2
> spd1<-a10/a1
> spd2<-b10/b1
> spd3<-c10/c1
>
```

Εικόνα:Εύρεση αθροισμάτων

Βήμα 4ο:

```
> mydata<-tibble(Hospitals=c("General Hospital","Health Centers","University Hospital"),numberofpatients=c(a2,b2,c
2),numberofdays=c(a1,b1,c1),costperday=c(spd1,spd2,spd3),costperpatient=c(spp1,spp2,spp3))
> mydata%>%ggplot(aes(x=costperpatient,y=Hospitals))+geom_col()
> spp1
  sum(Totalpatients)
1          158.5622
> spp2
  sum(Totalpatients)
1           286.052
> spp3
  sum(Totalpatients)
1          189.9754
```



Εικόνα:Διαγράμματα κόστους υπηρεσιών ανά ασθενή και ανά ημέρα νοσηλείας

Όπως παρατηρούμε το Κέντρα Υγείας-Γενικά Νοσοκομεία έχουν το μεγαλύτερο κόστος υπηρεσιών ανά ασθενή και ανά ημέρα νοσηλείας συγκριτικά με τα άλλα είδη Νοσοκομείων. Λόγω της μικρής χωρητικότητας συγκριτικά με τα άλλα νοσοκομεία, θα πρέπει να γίνει επανεξέταση στα νοσοκομεία αυτά.

5.4.ΜΕΣΟ ΣΥΝΟΛΙΚΟ ΚΟΣΤΟΣ

5.4.1.ΜΕΣΟ ΣΥΝΟΛΙΚΟ ΚΟΣΤΟΣ ΑΝΑ ΑΣΘΕΝΗ ΚΑΙ ΑΝΑ ΗΜΕΡΑ ΝΟΣΗΛΕΙΑΣ 5ης ΥΓΕΙΟΝΟΜΙΚΗΣ ΠΕΡΙΦΕΡΕΙΑΣ ΜΕ GEOM-COL

- *Βήμα 1ο*:

```
> library(readxl)
> HospitalDatafortotalCost <- read_excel("HospitalDatafortotalcost.xlsx")
> View(HospitalDatafortotalCost)
```

	Hospitalsofthe5thHealthDistrict	Totaldays	Totalpatients	Costformaterial	Costforconsum	Costforservices	Costformedic	Total
1	Γ.Ν. ΑΜΦΙΣΣΑΣ	10882	3158	1456122.8	362037.84	523668.2	389309.81	2731138.6
2	Γ.Ν. ΒΟΛΟΥ "ΑΧΙΛΛΟΠΟΥΛΕΙΟ"	93081	27525	9613697.7	1260110.94	2569989.6	3780417.85	17224216.1
3	Γ.Ν. ΘΗΒΩΝ	14002	2718	1059319.6	429142.35	1009995.8	462855.83	2961313.5
4	Γ.Ν. ΚΑΡΔΙΤΣΑΣ	53946	23072	4426354.8	939568.98	2005818.3	1741263.64	9113005.7
5	Γ.Ν. ΚΑΡΠΕΝΗΣΙΟΥ	7727	2150	766567.5	224136.20	432484.9	275830.14	1699018.8
6	Γ.Ν. ΛΑΜΙΑΣ	60629	17270	9987876.6	1216312.09	3823536.4	4852237.91	19879962.9
7	Γ.Ν. ΛΑΡΙΣΑΣ "ΚΟΥΤΛΙΜΠΑΝΕΙΟ"	67272	28166	9845197.1	1392542.51	2483071.6	4585993.60	18306804.9
8	Γ.Ν. ΛΙΒΑΔΕΙΑΣ	20885	5720	3344978.0	589828.66	1372999.8	1528281.51	6836087.9
9	Γ.Ν. ΤΡΙΚΑΛΩΝ	55645	19293	5540457.5	680506.74	2981244.5	2590441.31	11792650.0
10	Γ.Ν. ΧΑΛΚΙΔΑΣ	43995	13094	4656318.1	615804.85	1724045.6	2050662.28	9046830.8
11	Γ.Ν.- Κ.Υ. ΚΑΡΥΣΤΟΥ	4505	1298	391489.0	64269.91	243120.8	112783.60	811663.3
12	Γ.Ν.- Κ.Υ. ΚΥΜΗΣ	3451	960	345455.8	98319.81	355858.1	83035.15	882668.9
13	ΠΑΝΕΠΙΣΤΗΜΙΑΚΟ Γ.Ν. ΛΑΡΙΣΑΣ	174075	54390	31455120.9	2169591.63	9717476.6	14027813.85	57370003.1

- *Βήμα 2ο*:

```
> HospitalDataformaterialsCost2<-HospitalDatafortotalCost%>%slice(11:12)
> HospitalDataformaterialsCost1<-HospitalDatafortotalCost%>%slice(1:10)
> HospitalDataformaterialsCost2<-HospitalDatafortotalCost%>%slice(11:12)
> HospitalDataformaterialsCost3<-HospitalDatafortotalCost%>%slice(13)
>
```

	Hospitalsofthe5thHealthDistrict	Totaldays	Totalpatients	Costformaterial	Costforconsum	Costforservices	Costformedic	Total
1	Γ.Ν. ΑΜΦΙΣΣΑΣ	10882	3158	1456122.8	362037.8	523668.2	389309.8	2731139
2	Γ.Ν. ΒΟΛΟΥ "ΑΧΙΛΛΟΠΟΥΛΕΙΟ"	93081	27525	9613697.7	1260110.9	2569989.6	3780417.9	17224216
3	Γ.Ν. ΘΗΒΩΝ	14002	2718	1059319.6	429142.3	1009995.8	462855.8	2961314
4	Γ.Ν. ΚΑΡΔΙΤΣΑΣ	53946	23072	4426354.8	939569.0	2005818.3	1741263.6	9113006
5	Γ.Ν. ΚΑΡΠΕΝΗΣΙΟΥ	7727	2150	766567.5	224136.2	432484.9	275830.1	1699019
6	Γ.Ν. ΛΑΜΙΑΣ	60629	17270	9987876.6	1216312.1	3823536.4	4852237.9	19879963
7	Γ.Ν. ΛΑΡΙΣΑΣ "ΚΟΥΤΛΙΜΠΑΝΕΙΟ"	67272	28166	9845197.1	1392542.5	2483071.6	4585993.6	18306805
8	Γ.Ν. ΛΙΒΑΔΕΙΑΣ	20885	5720	3344978.0	589828.7	1372999.8	1528281.5	6836088
9	Γ.Ν. ΤΡΙΚΑΛΩΝ	55645	19293	5540457.5	680506.7	2981244.5	2590441.3	11792650
10	Γ.Ν. ΧΑΛΚΙΔΑΣ	43995	13094	4656318.1	615804.8	1724045.6	2050662.3	9046831

	HospitalsoftheSthHealthDistrict	Totaldays	Totalpatients	Costformaterial	Costforconsum	Costforservices	Costformedic	Total
1	Γ.Ν.- Κ.Υ. ΚΑΡΥΣΤΟΥ	4505	1298	391489.0	64269.91	243120.8	112783.60	811663.3
2	Γ.Ν.- Κ.Υ. ΚΥΜΗΣ	3451	960	345455.8	98319.81	355858.1	83035.15	882668.9

Filter

	HospitalsoftheSthHealthDistrict	Totaldays	Totalpatients	Costformaterial	Costforconsum	Costforservices	Costformedic	Total
1	ΠΑΝΕΠΙΣΤΗΜΙΑΚΟ Γ.Ν. ΛΑΡΙΣΑΣ	174075	54390	31455121	2169592	9717477	14027814	57370003

Εικόνα:Χρήση της εντολής slice

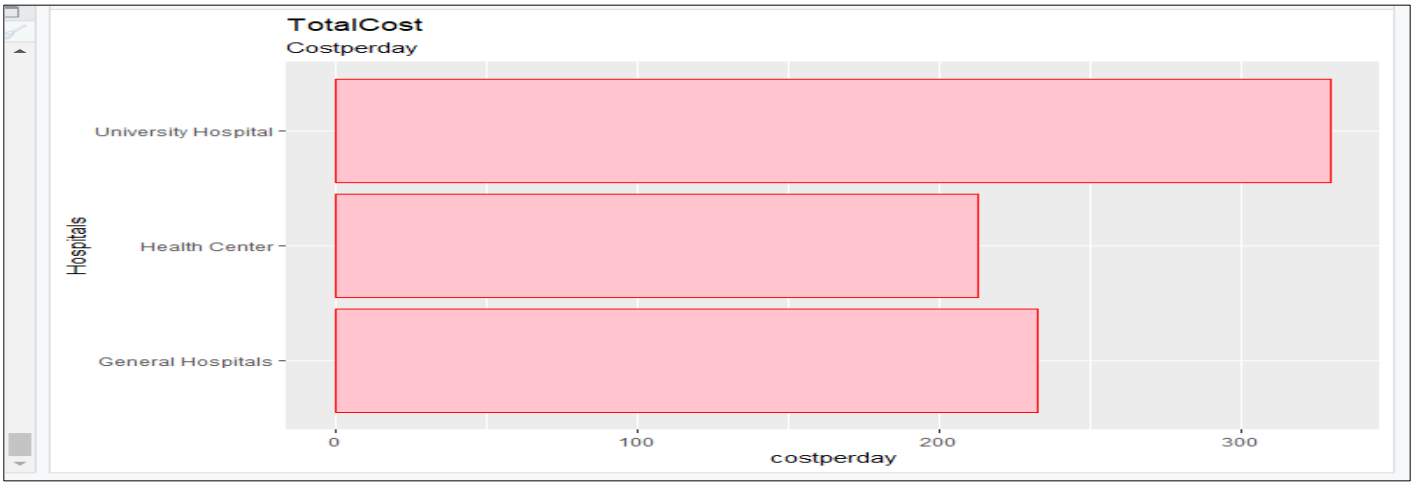
- ***Βήμα 3ο:***

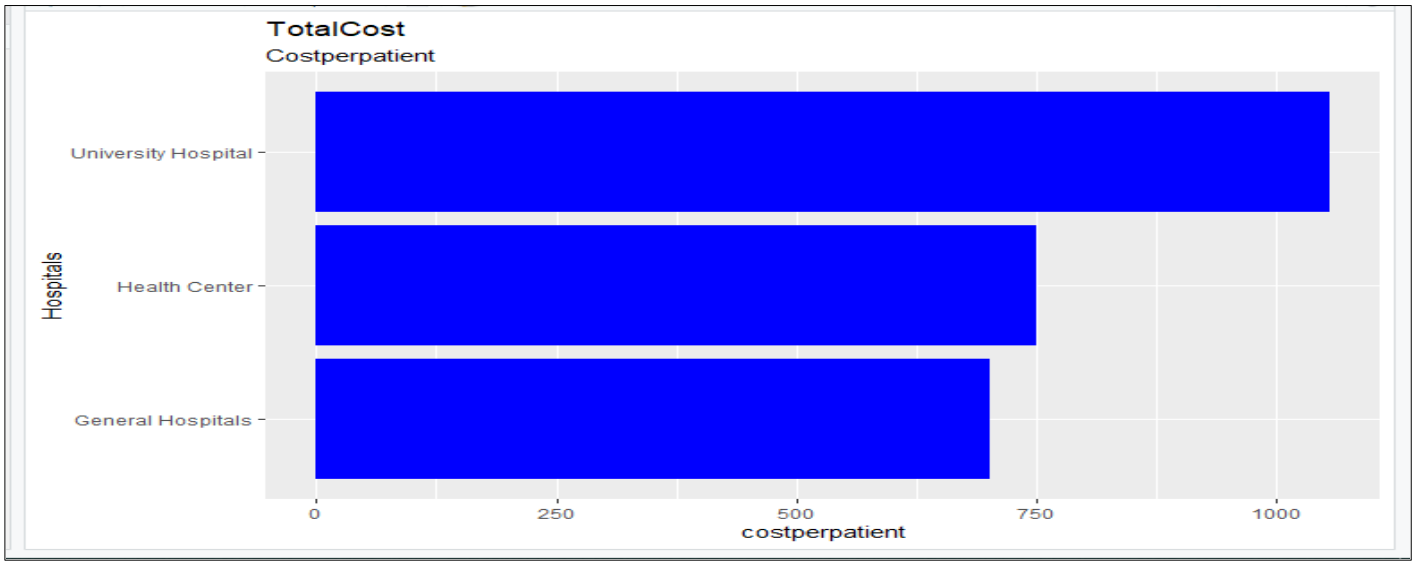
```
> a1<-HospitalDatafortotalcost1%>%summarise(sum(Totaldays))
> a2<-HospitalDatafortotalcost1%>%summarise(sum(Totalpatients))
> a3<-HospitalDatafortotalcost1%>%summarise(sum(Costformaterial))
> a4<-HospitalDatafortotalcost1%>%summarise(sum(Costforconsum))
> a5<-HospitalDatafortotalcost1%>%summarise(sum(Costforservices))
> a6<-HospitalDatafortotalcost1%>%summarise(sum(Costformedic))
> b1<-HospitalDatafortotalcost2%>%summarise(sum(Totaldays))
> b2<-HospitalDatafortotalcost2%>%summarise(sum(Totalpatients))
> b3<-HospitalDatafortotalcost2%>%summarise(sum(Costformaterial))
> b4<-HospitalDatafortotalcost2%>%summarise(sum(Costforconsum))
> b5<-HospitalDatafortotalcost2%>%summarise(sum(Costforservices))
> b6<-HospitalDatafortotalcost2%>%summarise(sum(Costformedic))
> c1<-HospitalDatafortotalcost3%>%summarise(sum(Totaldays))
> c2<-HospitalDatafortotalcost3%>%summarise(sum(Totalpatients))
> c3<-HospitalDatafortotalcost3%>%summarise(sum(Costformaterial))
> c4<-HospitalDatafortotalcost3%>%summarise(sum(Costforconsum))
> c5<-HospitalDatafortotalcost3%>%summarise(sum(Costforservices))
> c6<-HospitalDatafortotalcost3%>%summarise(sum(Costformedic))
> a7<-sum(a3,a4,a5,a6)
> b7<-sum(b3,b4,b5,b6)
> c7<-sum(c3,c4,c5,c6)
> spp1<-a7/a2
> spp2<-b7/b2
> spp3<-c7/c2
> spd1<-a7/a1
> spd2<-b7/b1
> spd3<-c7/c1
>
```

Εικόνα:Μεταβλητές

- ***Βήμα 4ο:***

```
>
>
>
>
>
>
> mydata<-tibble(Hospitals=c("General Hospitals","Health Center","University Hospital"),numberofpatients=c(a2,b2,c2),numberofdays=c(a1,b1,c1),costperday=c(spd1,spd2,spd3),costperpatient=c(spp1,spp2,spp3))
> mydata%>%ggplot(aes(x=costperday,y=Hospitals))+geom_col(fill="pink",color="red")+labs(title="TotalCost",subtitle="Costperday")
> mydata%>%ggplot(aes(x=costperpatient,y=Hospitals))+geom_col(fill="blue")+labs(title="TotalCost",subtitle="Costperpatient")
>
```





Εικόνα:Εμφάνιση αποτελεσμάτων

Όπως φαίνεται από τα διαγράμματα αυτά το Πανεπιστημιακό Νοσοκομείο έχει το μεγαλύτερο μέσο κόστος και στις δύο περιπτώσεις.

5.4.2.ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ ΓΙΑ ΤΟΝ ΔΕΙΚΤΗ ΤΟΥ ΜΕΣΟΥ ΣΥΝΟΛΙΚΟΥ ΚΟΣΤΟΥΣ ΑΝΑ ΑΣΘΕΝΗ ΚΑΙ ΑΝΑ ΗΜΕΡΑ ΝΟΣΗΛΕΙΑΣ ΤΗΣ 5ΗΣ ΥΓΕΙΟΝΟΜΙΚΗΣ ΠΕΡΙΦΕΡΕΙΑΣ

Η Γραμμική Παλινδρόμηση αποτελεί ένα σπουδαίο εργαλείο με το οποίο μπορούμε να εξάγουμε σημαντικά αποτελέσματα σε κάθε έρευνα. Στην εικόνα που ακολουθεί, στο πρώτο μέρος με την εντολή lm πραγματοποιήθηκε Γραμμική Παλινδρόμηση ανάμεσα στο συνολικό κόστος και στο πλήθος των ασθενών ενώ στο δεύτερο μέρος ανάμεσα στο συνολικό κόστος και στο πλήθος των ημερών νοσηλείας.

Ένα στοιχείο της Γραμμικής Παλινδρόμησης που επιφέρει σημαντικές πληροφορίες είναι το p-value. Όσο πιο μικρή είναι η τιμή του p-value τόσο πιο στατιστικά σημαντική είναι η μελέτη της γραμμικής παλινδρόμησης. Συνήθως, το p-value συγκρίνεται με το 0.05 έτσι εάν είναι μικρότερο από το 0.05 τότε η μελέτη είναι στατιστικά σημαντική. Όπως φαίνεται, και στις 2 περιπτώσεις το p-value είναι μικρότερο από 0.05 οπότε συμπεραίνουμε ότι η γραμμική παλινδρόμηση και στις 2 περιπτώσεις είναι στατιστικά σημαντική. Επίσης, εφόσον το p-value είναι μικρότερο από 0.05 συμπεραίνουμε ότι το συνολικό μέσο κόστος εξαρτάται και από το πλήθος των ασθενών και από τις ημέρες νοσηλείας.

Ένα ακόμα σημαντικό στοιχείο της Γραμμικής Παλινδρόμησης είναι ο δείκτης R^2 ο οποίος όσο μεγαλύτερος τόσο πιο συνεπής είναι ο δείκτης που μελετάται. Στην περίπτωση αυτή, παρατηρείται ότι το R^2 του μέσου κόστους ανά ημέρα νοσηλείας είναι πιο μεγάλος από το R^2 του μέσου κόστους ανά ασθενή. Αυτό μας οδηγεί στο συμπέρασμα, ότι ο δείκτης του μέσου κόστους ανά ημέρα νοσηλείας είναι πιο συνεπής. Ακόμα, το 86% του μέσου συνολικού κόστους ερμηνεύεται από το πλήθος των ασθενών και το 92% από τις συνολικές ημέρες νοσηλείας.

```
> Linearfortotalcost5<-HospitalDatafortotalcost%>%mutate(sum=Costformaterial+Costforconsum+Costforservices)
> View(Linearfortotalcost5)
> g<-Linearfortotalcost5%>%lm(sum~Totalpatients, data = .)
> g%>%summary()

Call:
lm(formula = sum ~ Totalpatients, data = .)

Residuals:
     Min       1Q   Median       3Q      Max
-7372880 -2944939  1059188  1744046  7071311

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.114e+06  1.632e+06  -0.682    0.509
Totalpatients  6.873e+02  7.636e+01   9.002 2.09e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4112000 on 11 degrees of freedom
Multiple R-squared:  0.8805,    Adjusted R-squared:  0.8696
F-statistic: 81.03 on 1 and 11 DF,  p-value: 2.092e-06

> f<-Linearfortotalcost5%>%lm(sum~Totaldays, data = .)
> f%>%summary()

Call:
lm(formula = sum ~ Totaldays, data = .)

Residuals:
     Min       1Q   Median       3Q      Max
-6600467 -1724717  1068768  1431620  4613863

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.428e+06  1.211e+06   -1.18    0.263
Totaldays    2.307e+02  1.845e+01   12.50 7.63e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3050000 on 11 degrees of freedom
Multiple R-squared:  0.9342,    Adjusted R-squared:  0.9283
F-statistic: 156.3 on 1 and 11 DF,  p-value: 7.634e-08
```

Εικόνα: Γραμμική Παλινδρόμηση ανάμεσα στο συνολικό μέσο κόστος και στο σύνολο των ασθενών και στο σύνολο των ημερών νοσηλείας.

5.5.ΜΕΣΗ ΔΙΑΡΚΕΙΑ ΝΟΣΗΛΕΙΑΣ

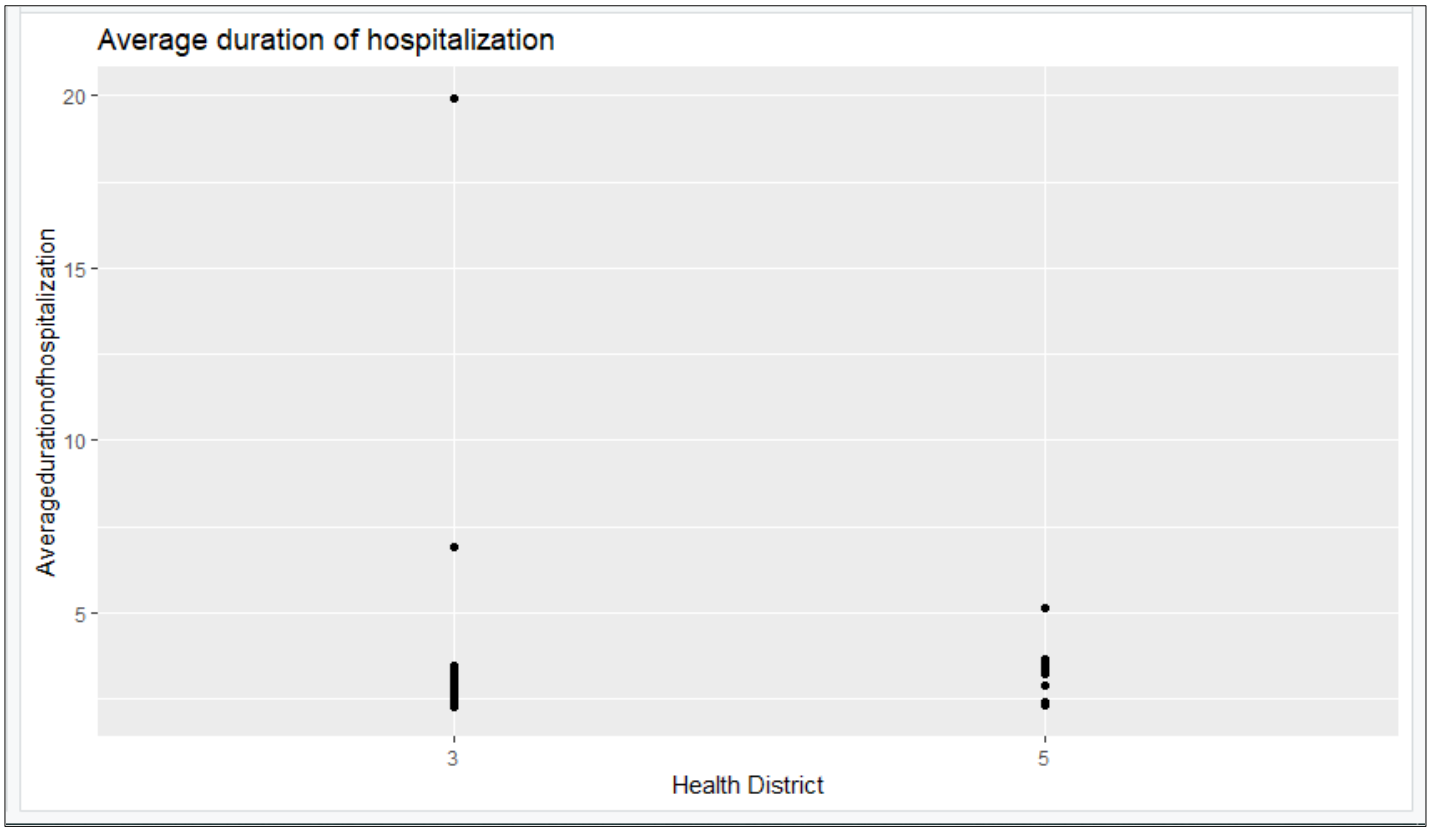
5.5.1.ΜΕΣΗ ΔΙΑΡΚΕΙΑ ΝΟΣΗΛΕΙΑΣ- GEOM_POINT

Με την μέθοδο αυτή γίνεται μια απλή αναπαράσταση της μέσης διάρκειας νοσηλείας των νοσοκομείων της 3ης και της 5ης Υγειονομικής Περιφέρειας. Κάθε κουκίδα απεικονίζει και ένα από τα νοσοκομεία. Η 3η Υγειονομική Περιφέρεια έχει την μεγαλύτερη διάρκεια νοσηλείας σε νοσοκομείο που είναι του Ψυχιατρικού Νοσοκομείου το οποίο απαιτεί περισσότερες μέρες για την θεραπεία των ασθενών από ότι τα υπόλοιπα Νοσοκομεία.

```
> View(MesiDiarkeiaNosileias)
> library(readxl)
> average_duration_of_hospitalization <- read_excel("average duration of hospitalization.xlsx")
> view(average_duration_of_hospitalization)
> average_duration_of_hospitalization<-average_duration_of_hospitalization%>%mutate(Averagedurationofhospitalization=Sumdays/Sump
at)
> average_duration_of_hospitalization%>%ggplot(aes(x=`Health District`,y=Averagedurationofhospitalization))+geom_point()+labs(tit
le="Average duration of hospitalization")
>
```

	Hospitals	Health District	Sumpat	Sumdays	Averagedurationofhospitalization
1	Γ.Ν. "ΠΑΠΑΓΕΩΡΓΙΟΥ"	3	63434	200000	3.152883
2	Γ.Ν. ΒΕΡΟΙΑΣ	3	12137	37649	3.102002
3	Γ.Ν. ΓΙΑΝΝΙΤΣΩΝ	3	11126	37056	3.330577
4	Γ.Ν. ΓΡΕΒΕΝΩΝ	3	6138	15247	2.484034
5	Γ.Ν. ΕΔΕΣΣΑΣ	3	11323	31681	2.797933
6	Γ.Ν. ΘΕΣ/ΝΙΚΗΣ "ΑΓ. ΔΗΜΗΤΡΙΟΣ"	3	12833	31181	2.429751
7	Γ.Ν. ΘΕΣ/ΝΙΚΗΣ "Γ. ΓΕΝΝΗΜΑΤΑΣ"	3	14339	49501	3.452193
8	Γ.Ν. ΘΕΣ/ΝΙΚΗΣ "Γ. ΠΑΠΑΝΙΚΟΛΑΟΥ"	3	61504	165038	2.683370
9	Γ.Ν. ΚΑΣΤΟΡΙΑΣ	3	7164	16404	2.289782
10	Γ.Ν. ΚΑΤΕΡΙΝΗΣ	3	14518	100110	6.895578
11	Γ.Ν. ΚΟΖΑΝΗΣ "ΜΑΜΑΤΣΕΙΟ"	3	10856	34003	3.132185
12	Γ.Ν. ΝΑΟΥΣΑΣ	3	6558	19821	3.022415
13	Γ.Ν. ΠΤΟΛΕΜΑΪΔΑΣ "ΜΠΟΔΟΣΑΚΕΙΟ"	3	12289	35857	2.917813
14	Γ.Ν. ΦΛΩΡΙΝΑΣ "ΕΛΕΝΗ Θ. ΔΗΜΗΤΡΙΟΥ"	3	6488	16697	2.573520
15	ΨΥΧΙΑΤΡΙΚΟ ΝΟΣΟΚΟΜΕΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ	3	11862	236627	19.948322
16	Γ.Ν. ΑΜΦΙΣΣΑΣ	5	3158	10882	3.445852
17	Γ.Ν. ΒΟΛΟΥ "ΑΧΙΛΛΟΠΟΥΛΕΙΟ"	5	27525	93081	3.381689
18	Γ.Ν. ΘΗΒΩΝ	5	2718	14002	5.151582
19	Γ.Ν. ΚΑΡΔΙΤΣΑΣ	5	23072	53946	2.338159
20	Γ.Ν. ΚΑΡΠΕΝΗΣΙΟΥ	5	2150	7727	3.593953
21	Γ.Ν. ΛΑΜΙΑΣ	5	17270	60629	3.510654
22	Γ.Ν. ΛΑΡΙΣΑΣ "ΚΟΥΤΛΙΜΠΑΝΕΙΟ"	5	28166	67272	2.388412
23	Γ.Ν. ΛΙΒΑΔΕΙΑΣ	5	5720	20885	3.651224
24	Γ.Ν. ΤΡΙΚΑΛΩΝ	5	19293	55645	2.884207
25	Γ.Ν. ΧΑΛΚΙΔΑΣ	5	13094	43995	3.359936
26	Γ.Ν.- Κ.Υ. ΚΑΡΥΣΤΟΥ	5	1298	4505	3.470724
27	Γ.Ν.- Κ.Υ. ΚΥΜΗΣ	5	960	3451	3.594792
28	ΠΑΝΕΠΙΣΤΗΜΙΑΚΟ Γ.Ν. ΛΑΡΙΣΑΣ	5	54390	174075	3.200496

Εικόνα:Φόρτωση δεδομένων

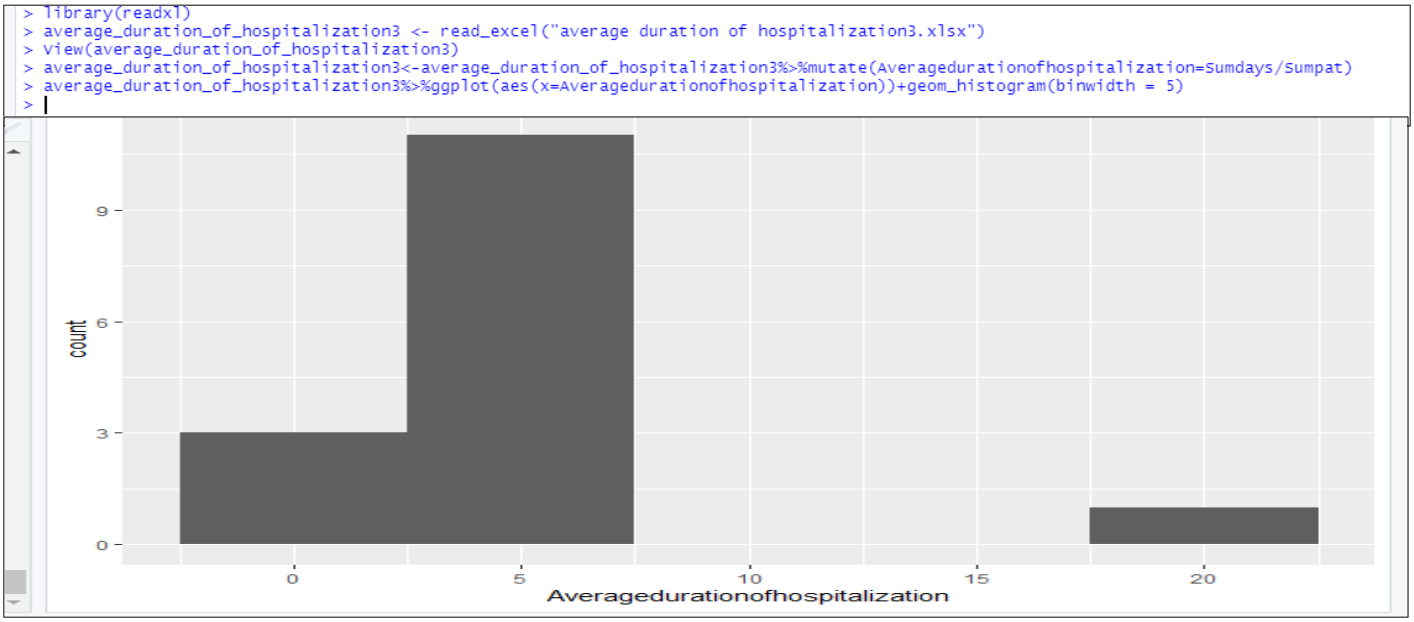


Εικόνα:Εμφάνιση αποτελεσμάτων

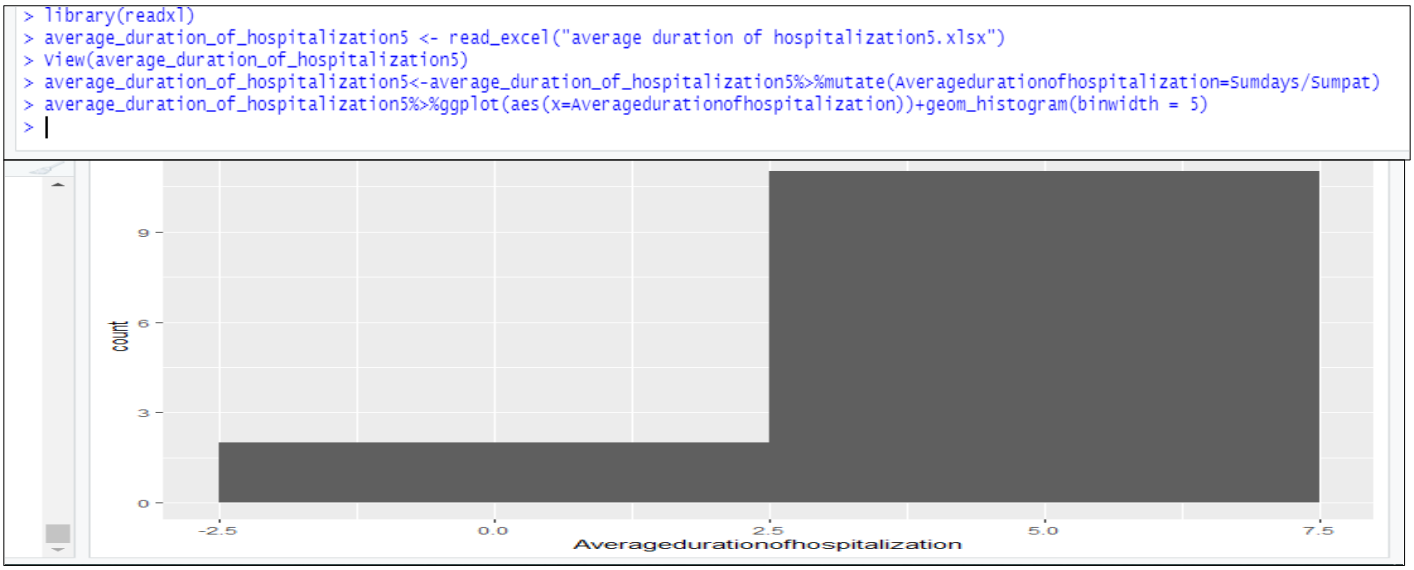
5.5.2.ΣΥΓΚΡΙΣΗ ΜΕΣΗΣ ΔΙΑΡΚΕΙΑΣ ΝΟΣΗΛΕΙΑΣ ΤΩΝ ΔΥΟ ΥΓΕΙΟΝΟΜΙΚΩΝ ΠΕΡΙΦΕΡΕΙΩΝ

Για να συγκρίνουμε τη μέση διάρκεια νοσηλείας των 2 Υγειονομικών Περιφερειών πρέπει να αποφασίσουμε ποιο test θα χρησιμοποιήσουμε. Στην απόφαση αυτή, διαδραματίζει μεγάλο ρόλο η κατανομή που ακολουθούν τα δεδομένα μας. Για να ελέγξουμε ποια κατανομή ακολουθούν τα δεδομένα μας πρέπει να εξετάσουμε το ιστόγραμμα τους και τα Q-Q plots.

Ιστόγραμμα:



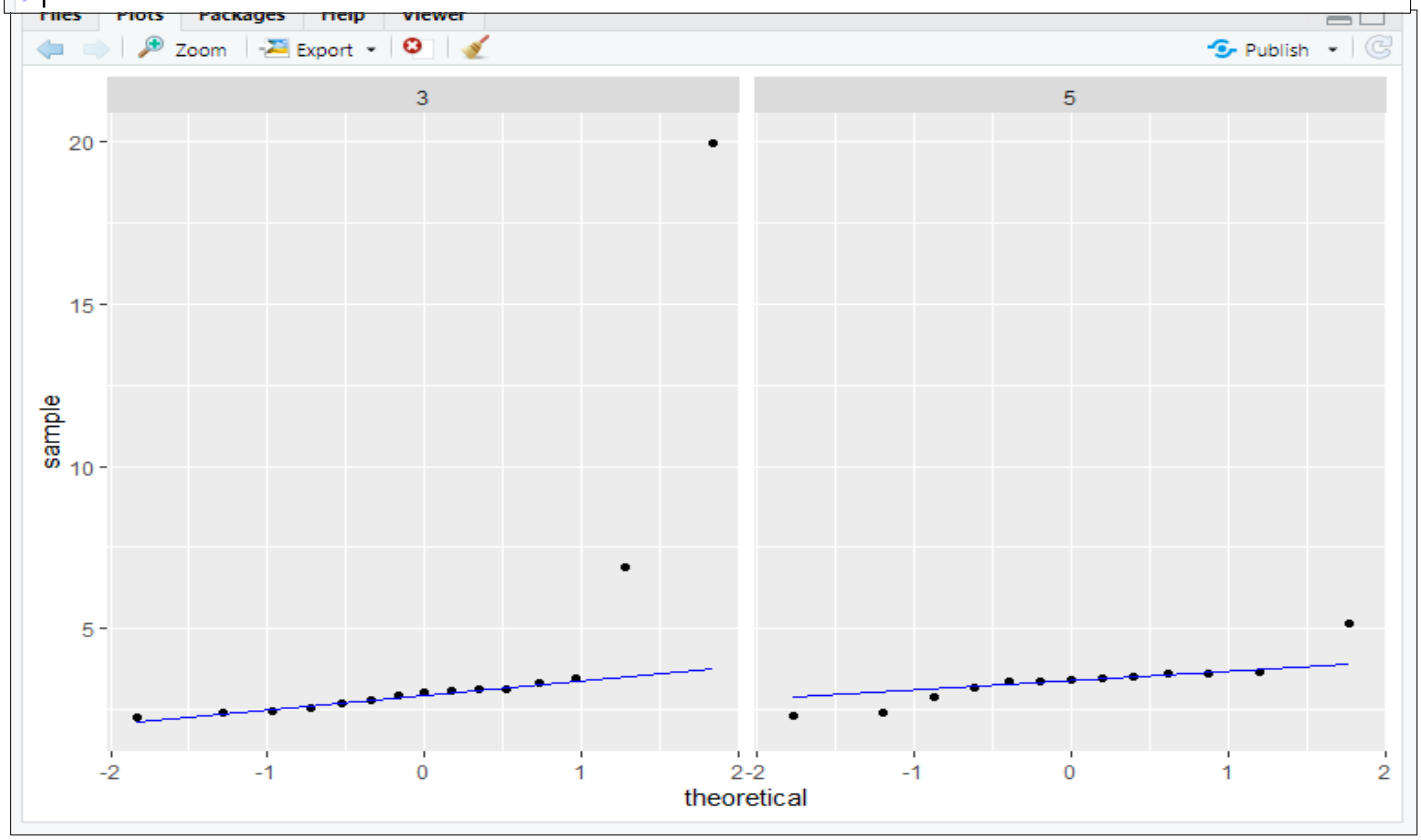
Εικόνα:Εμφάνιση Ιστογράμματος



Εικόνα:Εμφάνιση Ιστογράμματος

Q-Q Plots:

```
> library(readxl)
> average_duration_of_hospitalization <- read_excel("average duration of hospitalization.xlsx")
> View(average_duration_of_hospitalization)
> average_duration_of_hospitalization<-average_duration_of_hospitalization%>%mutate(Averagedurationofhospitalization=Sumdays/Sumpat)
> average_duration_of_hospitalization%>%ggplot(aes(sample=Averagedurationofhospitalization))+geom_qq()+geom_qq_line(colour="blue")+facet_grid(~`H
ealth District`)
>
```



Εικόνα:Εμφάνιση Q-Q plots

Παρατηρούμε, ότι η 5η Υγειονομική Περιφέρεια δεν ακολουθεί την γραμμή στο q-q plot άρα δεν ακολουθούν τα δεδομένα της την κανονική κατανομή. Για τον λόγο αυτό θα χρησιμοποιηθεί το test Wilcoxon για να εξεταστεί αν διαφέρουν οι μέσες διάρκειες νοσηλείες στην 3η και στην 5η Υγειονομική Περιφέρεια.

```
> average_duration_of_hospitalization%>%wilcox.test(Averagedurationofhospitalization~`Health District`,data=.)

        Wilcoxon rank sum exact test

data:  Averagedurationofhospitalization by Health District
W = 62, p-value = 0.1077
alternative hypothesis: true location shift is not equal to 0

>
```

Εικόνα:To Test Wilcoxon

Όπως, παρατηρείτε το p-value του test είναι μεγαλύτερο από 0.05 με αποτέλεσμα να δεχόμαστε την μηδενική υπόθεση δηλαδή η μέση διάρκεια νοσηλείας δεν διαφέρει στατιστικά σημαντικά ανάμεσα στην 3η και στην 5η Υγειονομική Περιφέρεια (στην 3η Υγειονομική Περιφέρεια η ΜΔΝ είναι ίση με 3.9 ενώ στην 5η είναι ίση με 3.06).

5.6.ΠΛΗΡΟΤΗΤΑ ΚΛΙΝΩΝ

5.6.1.ΠΛΗΡΟΤΗΤΑ ΚΛΙΝΩΝ-GEOM_COL ΓΙΑ ΤΗΝ 3Η ΥΓΕΙΟΝΟΜΙΚΗ ΠΕΡΙΦΕΡΕΙΑ

- *Βήμα 1ο*:

Filter

	Hospitals	Health District	Sumpat	Sumdays	Sumbeds
1	Γ.Ν. "ΠΑΠΑΓΕΩΡΓΙΟΥ"	3	63434	200000	709
2	Γ.Ν. ΒΕΡΟΙΑΣ	3	12137	37649	186
3	Γ.Ν. ΓΙΑΝΝΙΤΣΩΝ	3	11126	37056	175
4	Γ.Ν. ΓΡΕΒΕΝΩΝ	3	6138	15247	99
5	Γ.Ν. ΕΔΕΣΣΑΣ	3	11323	31681	161
6	Γ.Ν. ΘΕΣ/ΝΙΚΗΣ "ΑΓ. ΔΗΜΗΤΡΙΟΣ"	3	12833	31181	151
7	Γ.Ν. ΘΕΣ/ΝΙΚΗΣ "Γ. ΓΕΝΝΗΜΑΤΑΣ"	3	14339	49501	279
8	Γ.Ν. ΘΕΣ/ΝΙΚΗΣ "Γ. ΠΑΠΑΝΙΚΟΛΑΟΥ"	3	61504	165038	630
9	Γ.Ν. ΚΑΣΤΟΡΙΑΣ	3	7164	16404	107
10	Γ.Ν. ΚΑΤΕΡΙΝΗΣ	3	14518	100110	375
11	Γ.Ν. ΚΟΖΑΝΗΣ "ΜΑΜΑΤΣΕΙΟ"	3	10856	34003	194
12	Γ.Ν. ΝΑΟΥΣΑΣ	3	6558	19821	125
13	Γ.Ν. ΠΤΟΛΕΜΑΪΔΑΣ "ΜΠΟΔΟΣΑΚΕΙΟ"	3	12289	35857	200
14	Γ.Ν. ΦΛΩΡΙΝΑΣ "ΕΛΕΝΗ Θ. ΔΗΜΗΤΡΙΟΥ"	3	6488	16697	102
15	ΨΥΧΙΑΤΡΙΚΟ ΝΟΣΟΚΟΜΕΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ	3	11862	236627	604

Εικόνα:Φόρτωση Δεδομένων

- *Βήμα 2ο*:

```
>
>
> Fullnessofbeds1<-Fullnessofbeds%>%slice(1:14)
> Fullnessofbeds2<-Fullnessofbeds%>%slice(15)
```

	Hospitals	Health District	Sumpat	Sumdays	Sumbeds
1	Γ.Ν. "ΠΑΠΑΓΕΩΡΓΙΟΥ"	3	63434	200000	709
2	Γ.Ν. ΒΕΡΟΙΑΣ	3	12137	37649	186
3	Γ.Ν. ΓΙΑΝΝΙΤΣΩΝ	3	11126	37056	175
4	Γ.Ν. ΓΡΕΒΕΝΩΝ	3	6138	15247	99
5	Γ.Ν. ΕΔΕΣΣΑΣ	3	11323	31681	161
6	Γ.Ν. ΘΕΣ/ΝΙΚΗΣ "ΑΓ. ΔΗΜΗΤΡΙΟΣ"	3	12833	31181	151
7	Γ.Ν. ΘΕΣ/ΝΙΚΗΣ "Γ. ΓΕΝΝΗΜΑΤΑΣ"	3	14339	49501	279
8	Γ.Ν. ΘΕΣ/ΝΙΚΗΣ "Γ. ΠΑΠΑΝΙΚΟΛΑΟΥ"	3	61504	165038	630
9	Γ.Ν. ΚΑΣΤΟΡΙΑΣ	3	7164	16404	107
10	Γ.Ν. ΚΑΤΕΡΙΝΗΣ	3	14518	100110	375
11	Γ.Ν. ΚΟΖΑΝΗΣ "ΜΑΜΑΤΣΕΙΟ"	3	10856	34003	194
12	Γ.Ν. ΝΑΟΥΣΑΣ	3	6558	19821	125
13	Γ.Ν. ΠΤΟΛΕΜΑΪΔΑΣ "ΜΠΟΔΟΣΑΚΕΙΟ"	3	12289	35857	200
14	Γ.Ν. ΦΛΩΡΙΝΑΣ "ΕΛΕΝΗ Θ. ΔΗΜΗΤΡΙΟΥ"	3	6488	16697	102

	Hospitals	Health District	Sumpat	Sumdays	Sumbeds
1	ΨΥΧΙΑΤΡΙΚΟ ΝΟΣΟΚΟΜΕΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ	3	11862	236627	604

Εικόνα:Χρήση της εντολής slice

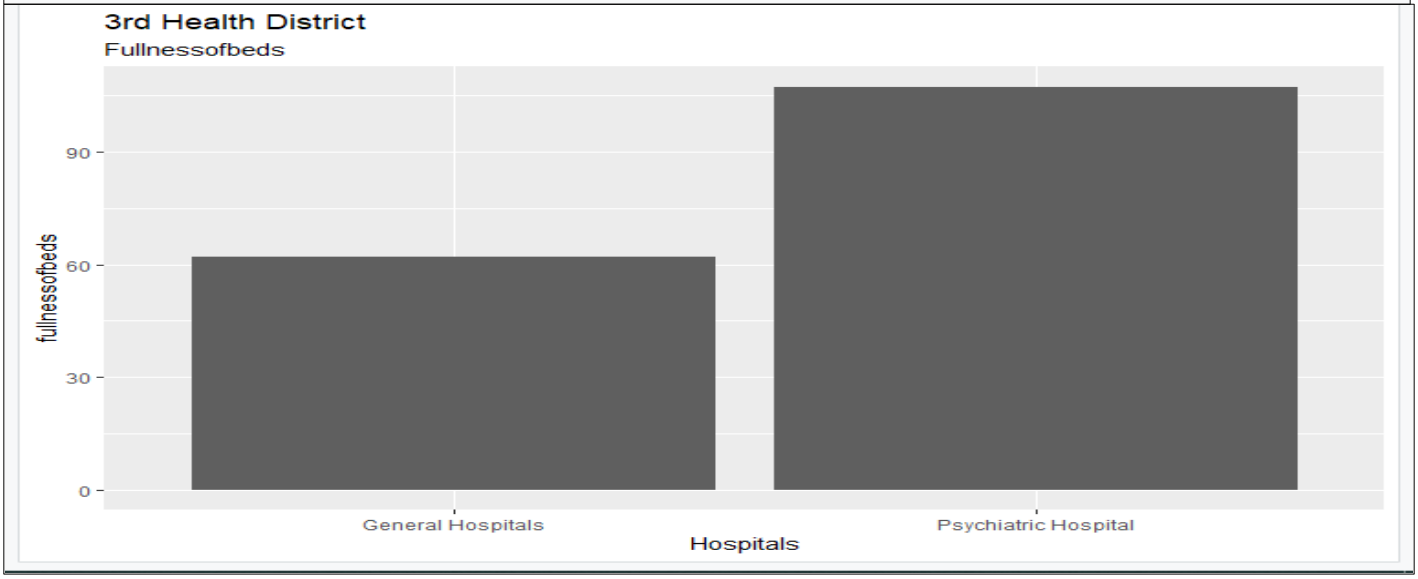
- *Βήμα 3ο*:

```
> a1<-Fullnessofbeds1%>%summarise(sum(Sumbeds))
> a1<-a1*365
> a2<-Fullnessofbeds1%>%summarise(sum(Sumdays))
> a3<-a2/a1
> a3<-a3*100
> a4<-Fullnessofbeds2%>%summarise(sum(Sumbeds))
> a4<-a4*365
> a5<-Fullnessofbeds2%>%summarise(sum(Sumdays))
> a6<-a5/a4
> a6<-a6*100
```

Εικόνα:Εύρεση αθροισμάτων

- *Βήμα 4ο*:

```
> mydata<-tibble(Hospitals=c("General Hospitals","Psychiatric Hospital"),fullnessofbeds=c(a3,a6))
> mydata%>%ggplot(aes(y=fullnessofbeds,x=Hospitals))+geom_col()+labs(title="3rd Health District",subtitle="Fullnessofbeds")
```



Εικόνα:Εμφάνιση αποτελεσμάτων

Στην 3η Υγειονομική Περιφέρεια το Ψυχιατρικό Νοσοκομείο έχει τη μεγαλύτερη πληρότητα κλινών πιθανώς λόγω της μεγάλης μέσης διάρκειας νοσηλείας.

5.6.2.ΠΛΗΡΟΤΗΤΑ ΚΛΙΝΩΝ- ΓΙΑ ΤΙΣ 2 ΥΓΕΙΟΝΟΜΙΚΕΣ ΠΕΡΙΦΕΡΕΙΕΣ ΜΑΖΙ

```
> library(readxl)
> FullnessofBeds35 <- read_excel("FullnessofBeds35.xlsx",
+     col_types = c("text", "text", "numeric",
+         "numeric", "numeric"))
> View(FullnessofBeds35)
> FullnessofBeds3<-FullnessofBeds35%>%slice(1:15)
> FullnessofBeds5<-FullnessofBeds35%>%slice(16:28)
> a1<-FullnessofBeds3%>%summarise(sum(Sumbeds))
> a1<-a1*365
> a2<-FullnessofBeds3%>%summarise(sum(Sumdays))
> a3<-a2/a1
> a3<-a3*100
> a3
  sum(Sumdays)
1     68.66849
```

```
> a5<-FullnessofBeds5%>%summarise(sum(Sumbeds))
> a5<-a5*365
> a6<-FullnessofBeds5%>%summarise(sum(Sumdays))
> a7<-a6/a5
> a7<-a7*100
> a7
  sum(Sumdays)
1     63.99285
```

Εικόνα:Σύγκριση της πληρότητας κλινών των 2 Υγειονομικών Περιφερειών

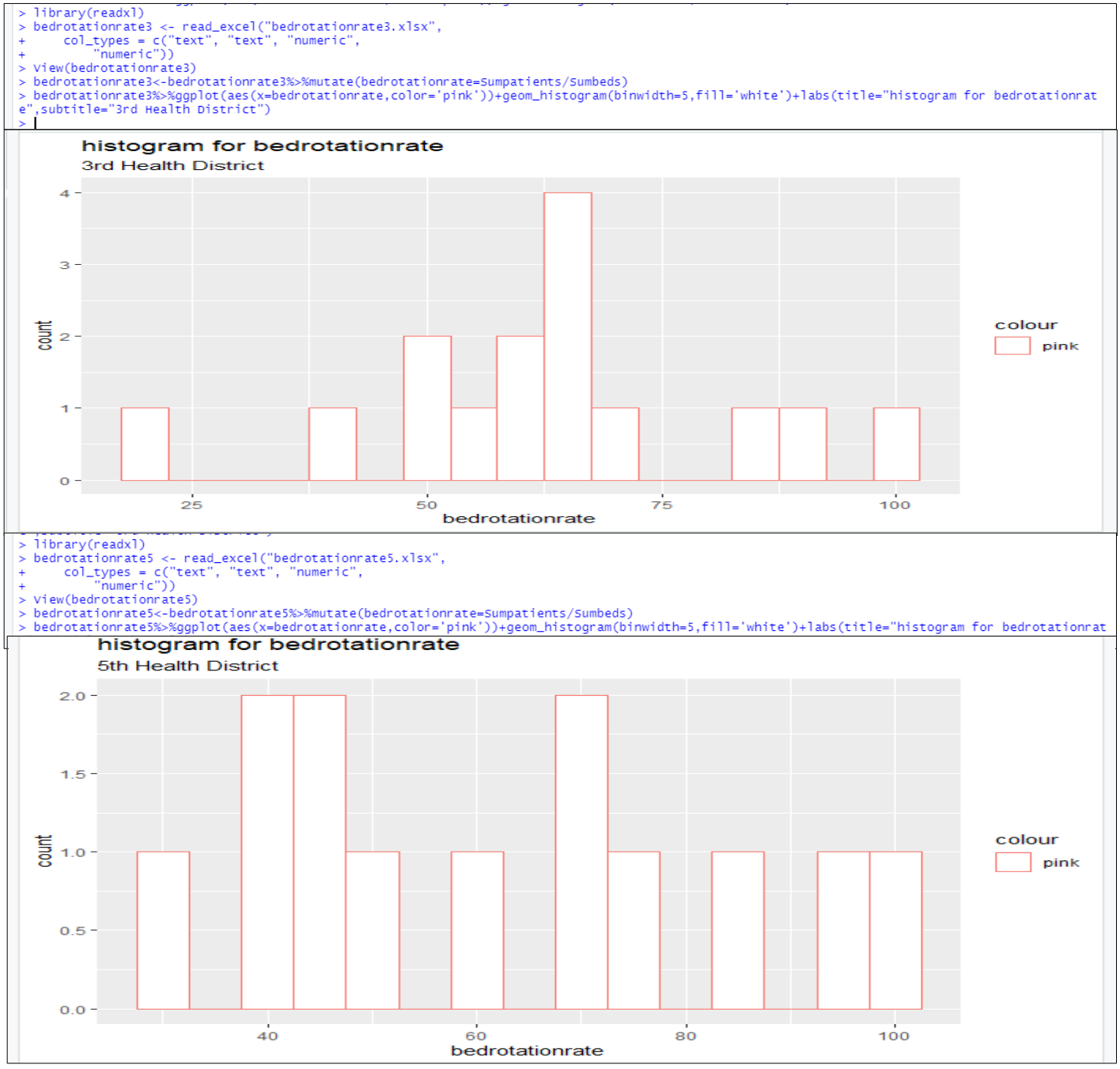
Όπως εύκολα παρατηρείτε η 3η Υγειονομική Περιφέρεια έχει μεγαλύτερη πληρότητα κλινών από την 5η Υγειονομική Περιφέρεια (η 3η έχει 68% πληρότητα ενώ η 5η 63%). Συμπεριλαμβάνοντας και τα αποτελέσματα της μέσης διάρκειας νοσηλείας μπορούμε να καταλήξουμε στο συμπέρασμα ότι πιθανώς, η αυξημένη πληρότητα οφείλεται στη μεγάλη μέση διάρκεια νοσηλείας και όχι στην ορθή ιατρική πρακτική.

5.7.ΕΝΑΛΛΑΓΗ ΚΛΙΝΩΝ

5.7.1.ΣΥΓΚΡΙΣΗ ΤΟΥ ΡΥΘΜΟΥ ΕΝΑΛΛΑΓΗΣ ΚΛΙΝΩΝ ΤΩΝ 2 ΥΓΕΙΟΝΟΜΙΚΩΝ ΠΕΡΙΦΕΡΕΙΩΝ

**Ακολουθείται η διαδικασία που πραγματοποιήθηκε και στη μέση διάρκεια νοσηλείας!!

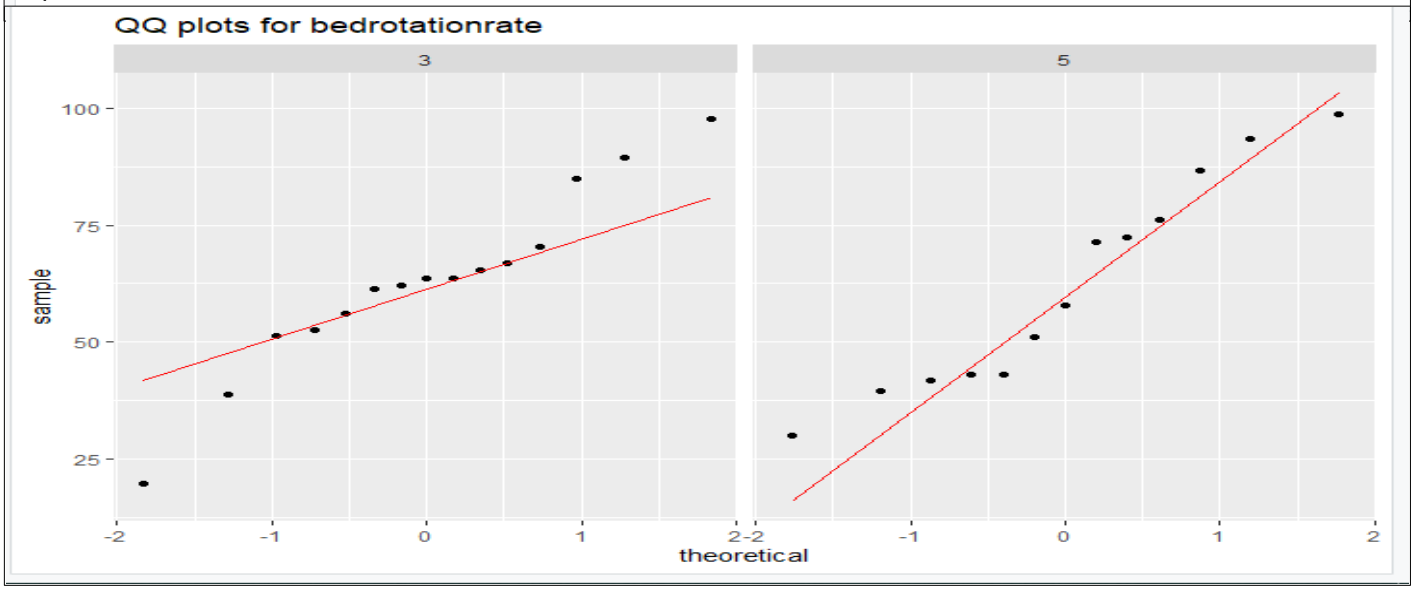
Ιστόγραμμα:



Εικόνα:Ιστόγραμμα

Q-Q plots:

```
> library(readxl)
> bedrotationrate <- read_excel("bedrotationrate.xlsx",
+     col_types = c("text", "text", "numeric",
+         "numeric"))
> View(bedrotationrate)
> bedrotationrate<-bedrotationrate%>%mutate(bedrotationrate=Sumpatients/Sumbeds)
> bedrotationrate%>%ggplot(aes(sample=bedrotationrate))+geom_qq()+geom_qq_line(colour="red")+facet_grid(~`Health District`)+labs(title="QQ plot
s for bedrotationrate")
>
```



Εικόνα:Q-Q plots

Από τα διαγράμματα αυτά συμπεραίνουμε ότι τα δεδομένα δεν ακολουθούν κανονική κατανομή για αυτόν τον λόγο θα χρησιμοποιήσουμε το test Wilcoxon για να συγκρίνουμε τον ρυθμό εναλλαγής κλινών των 2 Υγειονομικών Περιφερειών.

```
> bedrotationrate%>%wilcox.test(bedrotationrate~`Health District`,data=.)

        Wilcoxon rank sum exact test

data:  bedrotationrate by Health District
W = 101, p-value = 0.8919
alternative hypothesis: true location shift is not equal to 0
```

Εικόνα:Wilcoxon Test

Όπως, παρατηρείτε το p-value του test είναι μεγαλύτερο από 0.05 με αποτέλεσμα να δεχόμαστε την μηδενική υπόθεση δηλαδή ο ρυθμός εναλλαγής κλινών δεν διαφέρει στατιστικά σημαντικά ανάμεσα στην 3η και στην 5η Υγειονομική Περιφέρεια (η 3η έχει 64% ρυθμό εναλλαγής κλινών και η 5η 76%).

```
> bedrotationrate1<-bedrotationrate%>%slice(1:15)
> bedrotationrate2<-bedrotationrate%>%slice(16:28)
> a1<-bedrotationrate1%>%summarise(sum(Sumpatients))
```

Εικόνα:Η εντολή slice

```
> a1<-bedrotationrate1%>%summarise(sum(Sumpatients))
> a2<-bedrotationrate1%>%summarise(sum(Sumbeds))
> a3<-bedrotationrate2%>%summarise(sum(Sumpatients))
> a4<-bedrotationrate2%>%summarise(sum(Sumbeds))
> a5<-a1/a2
> a5
  sum(Sumpatients)
1         64.08811
> a6<-a3/a4
> a6
  sum(Sumpatients)
1         76.11562
```

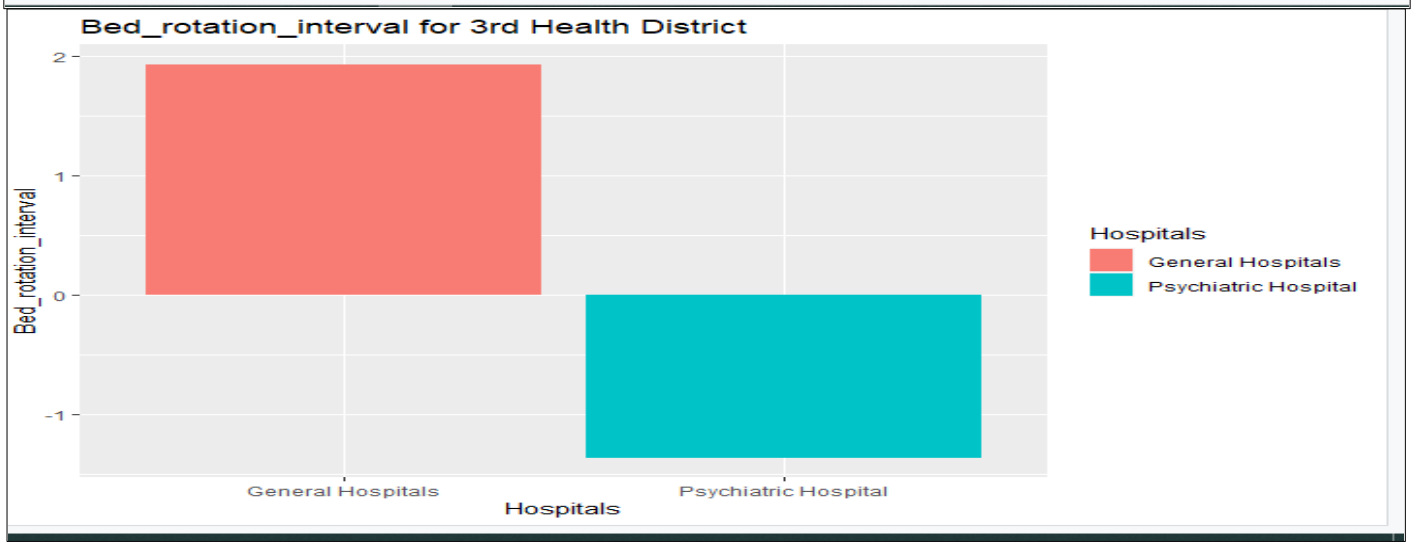
Εικόνα:Αποτελέσματα

5.7.2.ΔΙΑΣΤΗΜΑ ΕΝΑΛΛΑΓΗΣ ΚΛΙΝΩΝ -GEOM_COL ΓΙΑ ΤΗΝ 3Η ΥΓΕΙΟΝΟΜΙΚΗ ΠΕΡΙΦΕΡΕΙΑ

```
> library(readxl)
> Bed_rotation_interval <- read_excel("Bed rotation interval.xlsx")
> View(Bed_rotation_interval)
> Bed_rotation_interval1<-Bed_rotation_interval%>%slice(1:14)
> Bed_rotation_interval2<-Bed_rotation_interval%>%slice(15)
> a1<-Bed_rotation_interval1%>%summarise(sum(Sumbeds))
> a2<-Bed_rotation_interval1%>%summarise(sum(Sumdays))
> a3<-Bed_rotation_interval1%>%summarise(sum(Sumpat))
> a4<-((a1*365)-a2)/a3
> a4
  sum(Sumbeds)
1     1.933333
> b1<-Bed_rotation_interval2%>%summarise(sum(Sumbeds))
> b2<-Bed_rotation_interval2%>%summarise(sum(Sumdays))
> b3<-Bed_rotation_interval2%>%summarise(sum(Sumpat))
> b4<-((b1*365)-b2)/b3
> b4
  sum(Sumbeds)
1    -1.362924
```

Εικόνα:Κώδικας για το διάστημα εναλλαγής κλινών

```
>
>
>
>
> mydata%>%ggplot(aes(x=Hospitals,y=Bed_rotation_interval,fill=Hospitals))+geom_col()+labs(title="Bed_rotation_interval for 3rd Health Distric
t")
>
```



Εικόνα:Εμφάνιση της εναλλαγής κλινών με geom-col

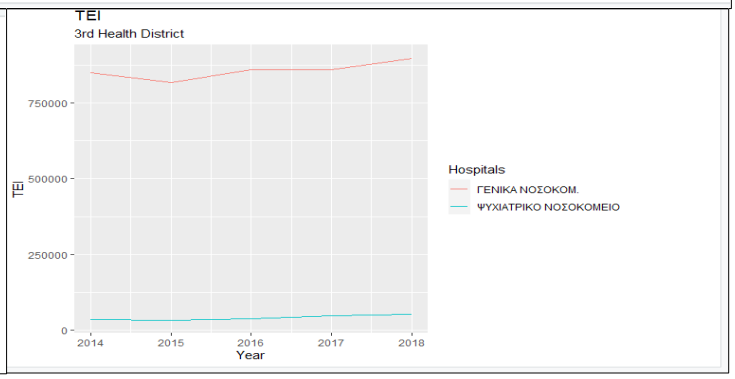
Όπως παρατηρείτε τα Γενικά Νοσοκομεία έχουν θετικό διάστημα εναλλαγής κλινών ενώ το Ψυχιατρικό Νοσοκομείο έχει αρνητικό. Αυτό πρακτικά σημαίνει ότι, τα Γενικά Νοσοκομεία κάνουν κακή διαχείριση των κλινών τους ενώ το Ψυχιατρικό Νοσοκομείο έχει έλλειψη κλινών.

ΕΠΙΠΡΟΣΘΕΤΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

ΤΕΙ 3η Υγειονομική Περιφέρεια

```
> TEI%>%ggplot(aes(x=Year,y=TEI,group=Hospitals,color=Hospitals))+geom_line()+labs(title="TEI",subtitle="3rd Health District")
> library(readxl)
> TEI <- read_excel("TEI.xlsx")
> View(TEI)
> TEI%>%ggplot(aes(x=Year,y=TEI,group=Hospitals,color=Hospitals))+geom_line()+labs(title="TEI",subtitle="3rd Health District")
>
```

	Year	Hospitals	TEI
1	2014	ΓΕΝΙΚΑ ΝΟΣΟΚΟΜ.	849499
2	2015	ΓΕΝΙΚΑ ΝΟΣΟΚΟΜ.	816364
3	2016	ΓΕΝΙΚΑ ΝΟΣΟΚΟΜ.	861069
4	2017	ΓΕΝΙΚΑ ΝΟΣΟΚΟΜ.	861069
5	2018	ΓΕΝΙΚΑ ΝΟΣΟΚΟΜ.	897926
6	2014	ΨΥΧΙΑΤΡΙΚΟ ΝΟΣΟΚΟΜΕΙΟ	36231
7	2015	ΨΥΧΙΑΤΡΙΚΟ ΝΟΣΟΚΟΜΕΙΟ	33754
8	2016	ΨΥΧΙΑΤΡΙΚΟ ΝΟΣΟΚΟΜΕΙΟ	38393
9	2017	ΨΥΧΙΑΤΡΙΚΟ ΝΟΣΟΚΟΜΕΙΟ	47072
10	2018	ΨΥΧΙΑΤΡΙΚΟ ΝΟΣΟΚΟΜΕΙΟ	53139



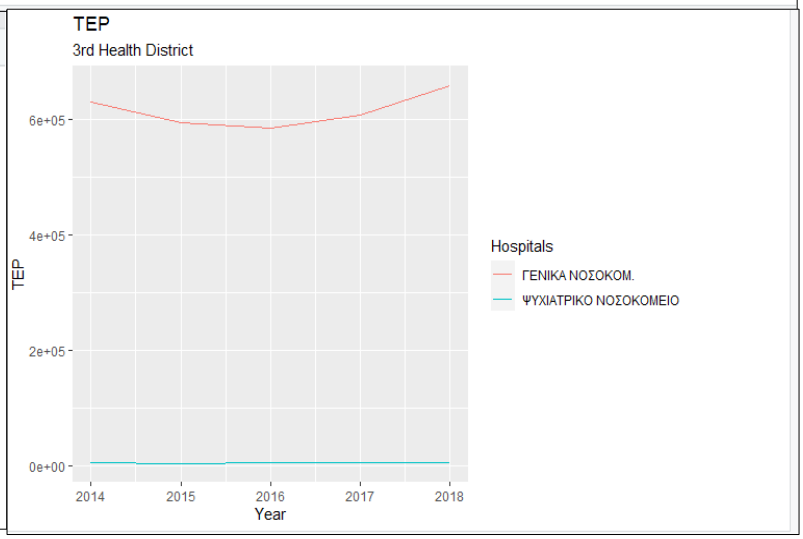
Εικόνα:Εμφάνιση των ΤΕΙ της 3ης Υγειονομικής Περιφέρειας με την συνάρτηση geom-line()

ΤΕΠ 3η Υγειονομική Περιφέρεια

```
>
> library(readxl)
> TEP <- read_excel("TEP.xlsx", col_types = c("numeric",
+     "text", "numeric"))
> View(TEP)
> TEP%>%ggplot(aes(x=Year,y=TEP,group=Hospitals,color=Hospitals))+geom_line()+labs(title="TEP",subtitle="3rd Health District")
>
```

Filter

	Year	Hospitals	TEP
1	2014	ΓΕΝΙΚΑ ΝΟΣΟΚΟΜ.	631553
2	2015	ΓΕΝΙΚΑ ΝΟΣΟΚΟΜ.	594638
3	2016	ΓΕΝΙΚΑ ΝΟΣΟΚΟΜ.	585474
4	2017	ΓΕΝΙΚΑ ΝΟΣΟΚΟΜ.	608769
5	2018	ΓΕΝΙΚΑ ΝΟΣΟΚΟΜ.	659678
6	2014	ΨΥΧΙΑΤΡΙΚΟ ΝΟΣΟΚΟΜΕΙΟ	5659
7	2015	ΨΥΧΙΑΤΡΙΚΟ ΝΟΣΟΚΟΜΕΙΟ	4278
8	2016	ΨΥΧΙΑΤΡΙΚΟ ΝΟΣΟΚΟΜΕΙΟ	5630
9	2017	ΨΥΧΙΑΤΡΙΚΟ ΝΟΣΟΚΟΜΕΙΟ	6281
10	2018	ΨΥΧΙΑΤΡΙΚΟ ΝΟΣΟΚΟΜΕΙΟ	4842



Εικόνα:Εμφάνιση των ΤΕΙ της 3ης Υγειονομικής Περιφέρειας με την συνάρτηση geom-line()

5.9.CLUSTERING- ΔΙΑΦΟΡΕΤΙΚΗ ΟΜΑΔΟΠΟΙΗΣΗ ΤΩΝ ΝΟΣΟΚΟΜΕΙΩΝ

Στην ανάλυση και στην απεικόνιση των αποτελεσμάτων στο προηγούμενο κεφάλαιο στηριχθήκαμε κυρίως στην ομαδοποίηση νοσοκομείων με βάση την Υγειονομική Περιφέρεια στην οποία ανήκαν αλλά και στο είδος τους. Ωστόσο στην R και γενικότερα στον προγραμματισμό μπορεί να γίνει ομαδοποίηση των δεδομένων με έναν αλγόριθμο που ονομάζεται ***K-means*** *(Partinional Clustering in R: The Essentials).* Ο αλγόριθμος αυτός έχει ως κεντρική ιδέα την ομαδοποίηση δεδομένων με κοινά χαρακτηριστικά.

Τα βασικά πλεονεκτήματα και μειονεκτήματα του αλγορίθμου αυτού είναι τα παρακάτω:

Πλεονεκτήματα	***Μειονεκτήματα***
1.Εύκολος και κατανοητός αλγόριθμος	1.Δεν υπάρχει αυτοματοποιημένος τρόπος για να βρούμε το πλήθος των συστάδων που απαιτούνται.
2.Πιο γρήγορος από άλλους αλγορίθμους	

Πίνακας:Πλεονεκτήματα και μειονεκτήματα του αλγορίθμου k-means

Στην συνέχεια θα αναφερθούμε στα βήματα του αλγορίθμου αυτού:

1 **Επιλέγεται ο αριθμός των συστάδων-ομάδων.**

2 **Στην συνέχεια, τοποθετούνται τυχαία παρατηρήσεις ως κέντρα των συστάδων.**

3 **Έπειτα κάθε παρατήρηση τοποθετείται στην κοντινότερη συστάδα.**

4 **Τέλος, υπολογίζουμε τα νέα κέντρα κάθε συστάδας ως μέσος όρο όλων των στοιχείων τους. Τα βήματα επαναλαμβάνονται μέχρι να μην υπάρχουν μεταβολές στις συστάδες.**

Στην εργασία αυτή, για την πραγματοποίηση του clustering με τον αλγόριθμο k-means χρησιμοποιήθηκαν τα νοσοκομεία της 3ης και της 5ης Υγειονομικής Περιφέρειας και όλα τα στοιχεία τους.

Παρακάτω αναφέρονται τα βήματα του αλγορίθμου k-means στο RStudio έτσι ώστε να προκύψουν ομάδες νοσοκομείων *(How to Use and Visualize K-Means Clustering in R,Jan 19) :*

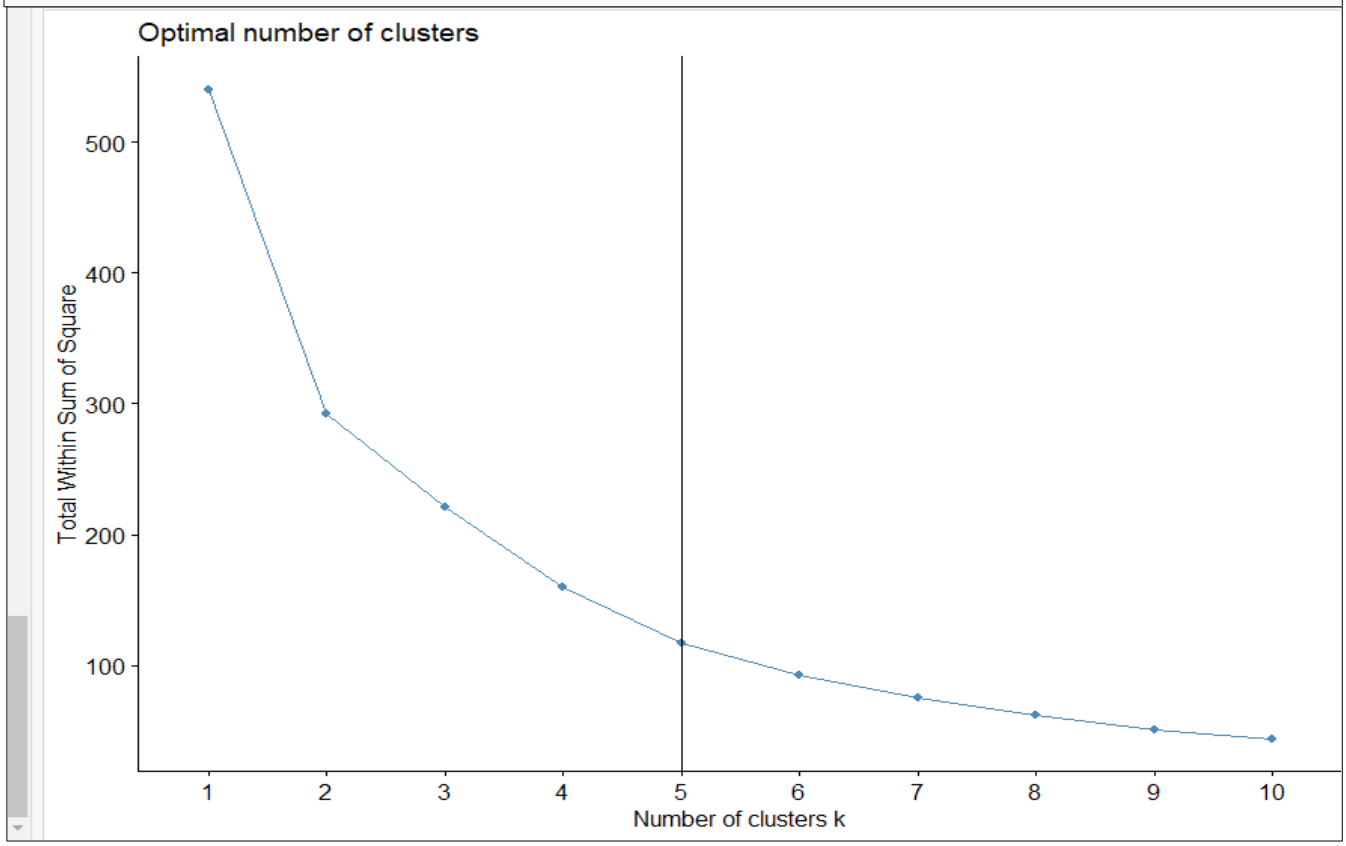
1) Ελέγχουμε αν υπάρχουν **ΝΑ τιμές** στον πίνακα δεδομένων μας. Στην περίπτωση μας δεν έχουμε οπότε το βήμα αυτό παραλήφθηκε.

2)Επιλέγουμε **τον αριθμό των ομάδων-συστάδων**. Η συνάρτηση **fviz_nbclust** χρησιμοποιείται για να υπολογίσει το κατάλληλο πλήθος ομάδων-συστάδων που πρέπει να χρησιμοποιηθούν. Η συνάρτηση αυτή έχει 4 ορίσματα τα οποία είναι τα εξής:

- *scale(Clustering[,2:22]):κανονικοποίηση των δεδομένων από την στήλη 2 μέχρι την στήλη 22.*
- *kmeans*
- *nstart=100:παράγονται 100 διαφορετικά κεντροειδή και χρησιμοποιούνται εκείνα που είναι καλύτερα για τον αλγόριθμο.*
- *method="wss":* μέθοδος Elbow

Επίσης η συνάρτηση geom_vline χρησιμοποιείται για να δημιουργηθεί η κάθετη γραμμή στο κατάλληλο πλήθος των συστάδων.(στην περίπτωση αυτή στο νούμερο 5)

```
>
> fviz_nbclust(scale(clustering[,2:22]), kmeans, nstart=100, method = "wss") +
+ geom_vline(xintercept = 5, linetype = 1)
>
```



Εικόνα: Επιλογή του κατάλληλου αριθμού ομάδων- clusters

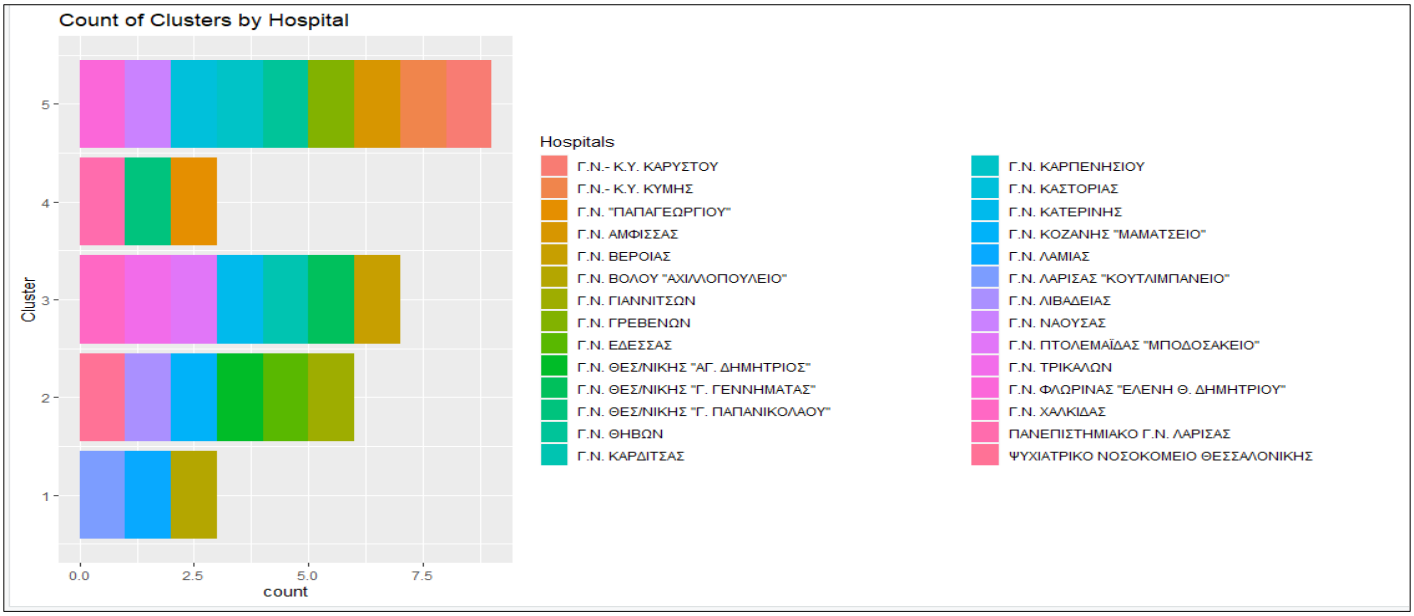
3)Στην συνέχεια χρησιμοποιείται η συνάρτηση **Set.seed** με έναν οποιοδήποτε ακέραιο αριθμό σαν όρισμα έτσι ώστε ο κώδικας να είναι αναπαραγωγίσιμος δηλαδή να παράγει σε κάθε επανάληψη τα ίδια αποτελέσματα σε όλους τους χρήστες.**(set.seed(1234))**

4)Με την εντολή ***kmeans_basic ←kmeans(Clustering[,2:22], centers = 5)*** δημιουργούνται 5 συστάδες με βάση τα δεδομένα που βρίσκονται από την στήλη 2 μέχρι την στήλη 21.

5)Με την εντολή ***kmeans_basic_df ←data.frame (Cluster =kmeans_basic$cluster, Clustering)*** δημιουργείται ένας πίνακας με τα νοσοκομεία και την ομάδα (cluster) στην οποία ανήκουν.

6)Επίσης. με την εντολή ***ggplot(data=kmeans_basic_df,aes(y=Cluster)) +geom_bar(aes(fill=Hospitals))+ggtitle("Count of Clusters by Hospital")*** δημιουργούνται οι συστάδες με τα στοιχεία τους.

7)Με τις εντολές αυτές προκύπτει το παρακάτω διάγραμμα στο οποίο μπορείτε να παρατηρήσετε εύκολα ότι τα νοσοκομεία ομαδοποιήθηκαν με διαφορετικό τρόπο σε σχέση με την υπόλοιπη εργασία. (στην υπόλοιπη εργασία ομαδοποιήθηκαν με βάση την Υγειονομική Περιφέρεια στην οποία ανήκαν, αλλά και με βάση το είδος τους)



Εικόνα:Εμφάνιση ομαδοποίησης δεδομένων

Clusters:

1ο:Γ.Ν.Λάρισας,Γ.Ν.Λαμίας,Γ.Ν. Βόλου.

2ο:Γ.Ν.Γιαννιτσά, Γ.Ν.Έδεσσας, Γ.Ν.Θεσσαλονίκης Αγ. Δημήτριος,, Γ.Ν.Κοζάνης Μαμάτσειο, Ψυχιατρικό Νοσοκομείο Θεσσαλονίκης, Γ.Ν.Τρικάλων.

3ο:Γ.Ν. Γεννηματάς, Γ.Ν. Κατερίνης, Γ.Ν. Καρδίτσας, Γ.Ν.Τρίκαλα, Γ.Ν.Χαλκίδας, Γ.Ν. Πτολεμαΐδας, Γ.Ν.Βέροιας

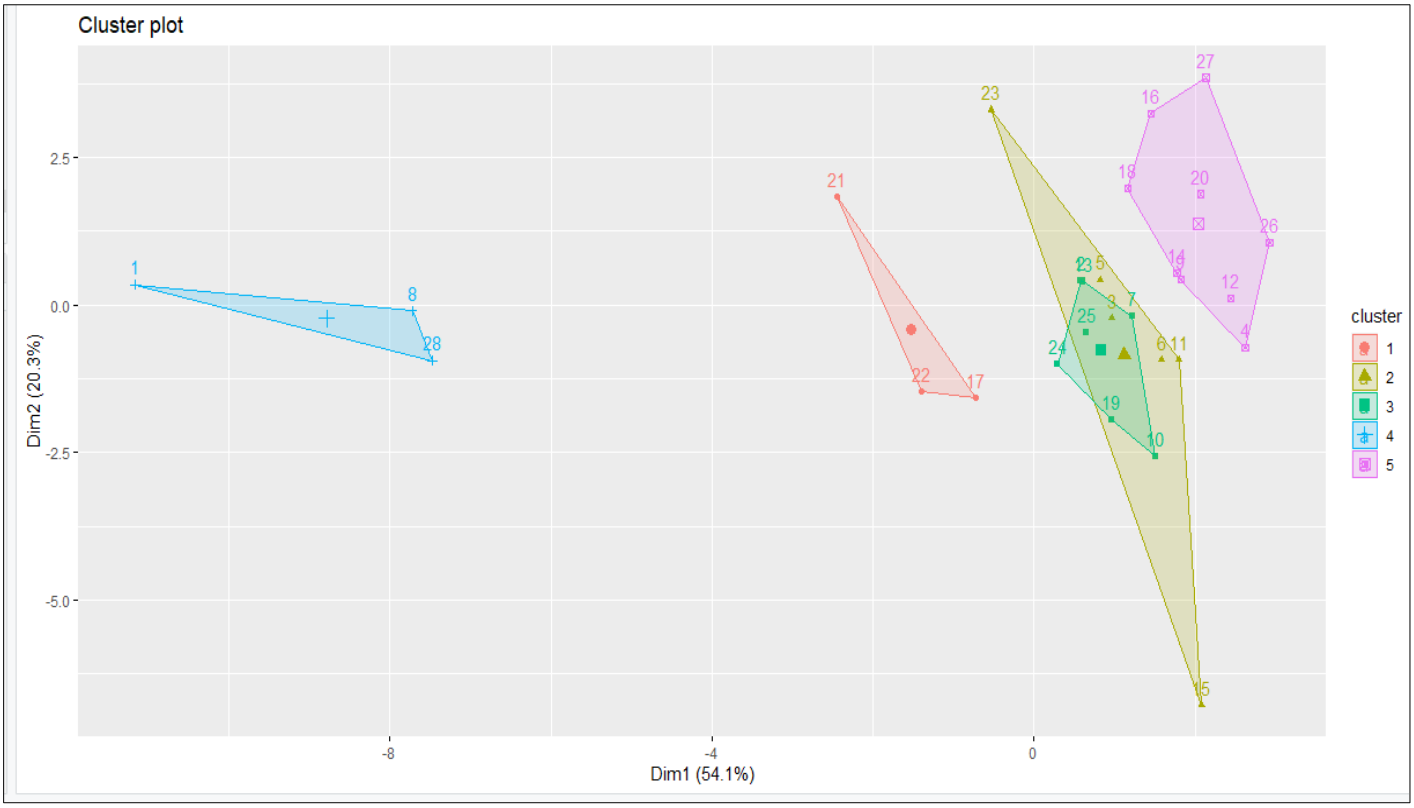
4ο:Γ.Ν.Παπαγεωργίου, Γ.Ν.Παπανικολάου,Πανεπιστημιακό Νοσοκομείο Λάρισας.

5ο:Γ.Ν.Γρεβενών, Γ.Ν.Νάουσας, Γ.Ν.Άμφισσας, Γ.Ν.Θηβών, Γ.Ν.Καρπενήσι, Γ.Ν-Κ.Υ.Κύμης και Καρύστου, Γ.Ν. Φλώρινας "Ελένη Δημητρίου", Γ.Ν. Καστοριάς.

Ένας άλλος τρόπος απεικόνισης των δεδομένων ομαδοποιημένα είναι μέσω της εντολής:

- ***fviz_cluster(kmeans_basic, data = Clustering[,2:22]).***

- Όπως παρατηρείτε, ο άξονας x έχει Dim (54.1%) και ο άξονας y έχει Dim (20.3%). Η ομαδοποίηση αυτή έγινε με βάση τα δεδομένα που υπάρχουν στις στήλες 2 μέχρι 22 του πίνακα Clustering. Όταν λοιπόν τα δεδομένα τα οποία συμμετέχουν στην ομαδοποίηση των δεδομένων είναι παραπάνω από 2 όπως σε αυτήν την περίπτωση τότε επιλέγονται εκείνα τα δύο, τα οποία συμμετέχουν περισσότερο στην διακύμανση (διακύμανση γενικά είναι η απόσταση των σημείων από την μέση τιμή δηλαδή στον kmeans από το κέντρο των συστάδων). Άρα, εδώ τα δύο αυτά στοιχεία αντιπροσωπεύουν το 74.4% της διακύμανσης των δεδομένων.



Εικόνα:Διαφορετική ομαδοποίηση δεδομένων

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1]Πληροφοριακά Συστήματα Υγείας (2020), Χαράλαμπος Καρανίκας, Πλατφόρμα Τηλεκπαίδευσης Πανεπιστημίου Θεσσαλίας: https://eclass.uth.gr/index.php?logout=yes

[2]Αξιολόγηση Συστημάτων Υγείας και Κατασκευή Βάσης Δεδομένων δεικτών απόδοσης του ΓΝ Λαμίας (2011), Ανεστίδης Δημήτριος.

[3]R For Health Data Ewen Harrison and Riinu Pius (2020-09-16): https://argoshare.is.ed.ac.uk/healthyr_book/

[4]Cluster Analysis Lecture (2018-04-25): https://lukedaniels1.github.io/Bio381_2018/Daniels_Cluster_Analysis_Lecture.html

[5]K-means Cluster Analysis: https://uc-r.github.io/kmeans_clustering

[6]How to Use and Visualize K-Means Clustering in R**:**https://towardsdatascience.com/how-to-use-and-visualize-k-means-clustering-in-r-19264374a53c

[7]Partinional Clustering in R:The Essentials:https://www.datanovia.com/en/lessons/k-means-clustering-in-r-algorith-and-practical-examples/

[8]Big Data: Νέοι ρόλοι και ευκαιρίες για τους επαγγελματίες πληροφόρησης Πτυχιακή Εργασία Θάνος Ευάγγελος:http://hypatia.teiath.gr/xmlui/bitstream/handle/11400/20185/lb_04174_thanos_thesis.pdf?sequence=1

[9]Statistical Tests:http://r-statistics.co/Statistical-Tests-in-R.html

[10]Linear Regression:http://r-statistics.co/Linear-Regression.html