METROPOLITAN COLLEGE

CENTRE OF EXCELLENCE

Spatial Data Analysis Laboratory Exercise

Student ID: 2018506

Μάθημα: Spatial Data Analysis

MSc Data Science
Σχολή Πληροφορικής
Μητροπολιτικό Κολλέγιο
Αθήνα, Ελλάδα

Ημερομηνία:
27/04/2021

Περιεχόμενα

Άσκηση 1

Τα δεδομένα με όνομα GR.Municipalities αφορούν μια γεωγραφική βάση δεδομένων με χωρικές οντότητες τους δήμους της Ελλάδας.

Ανάλυση του συνόλου δεδομένων GR.Municipalities

Το συγκεκριμένο σύνολο δεδομένων είναι ενσωματωμένο στο πακέτο lctools της γλώσσας R, το οποίο είναι μια βιβλιοθήκη που βοηθάει ένα Data Scientist με προβλήματα χωρικής ανάλυσης, μελέτη χωρικής αυτοσυσχέτισης και μετρήσεις χωρικών ανισοτήτων.

Τα δεδομένα GR.Municipalities αποτελούν μια γεωγραφική βάση δεδομένων με χωρικές οντότητες τους δήμους της Ελλάδας, η οποία βασίστηκε στη διοικητική διαίρεση του προγράμματος Καλλικράτης (2011). Τα δεδομένα περιέχουν επίσης αρκετά περιγραφικά δεδομένα από την Απογραφή Πληθυσμού του 2001, όπως πληθυσμός και ανεργία κατά φύλο. Επιπροσθέτως, περιέχονται και στοιχεία από τη Γενική Γραμματεία Πληροφοριακών Συστημάτων, όπως το μέσο ετήσιο δηλωθέν οικογενειακό εισόδημα.

Η παραπάνω συνένωση από 3 διαφορετικές πηγές, καθιστά το σύνολο δεδομένων GR.Municipalities ιδανικό για κάποιον που θέλει να εφαρμόσει χωρική ανάλυση στην Ελλάδα, και να προβεί σε κάποια χρήσιμα συμπεράσματα τα οποία βασίζονται τόσο στη γεωγραφία της χώρας όσο και σε δημογραφικά στοιχεία του ελληνικού πληθυσμού.

Ερώτημα α

Δημιουργείστε το διάγραμμα διασποράς των ζευγών κανονικοποιημένων τιμών εισοδήματος και κανονικοποιημένων σταθμισμένων αθροισμάτων των εισοδημάτων των έξι κοντινότερων γειτόνων για κάθε δήμο.

Απάντηση α

Αρχικά φορτώνουμε τα δεδομένα, και στη συνέχεια χρησιμοποιώντας την εντολή names βλέπουμε τα ονόματα των 14 μεταβλητών. Πριν από αυτό όμως για λόγους απόδοσης και αποδέσμευσης μνήμης αφαιρούμε το object GR.Municipalities καθώς ήδη έχουμε τα δεδομένα στη μεταβλητή gr_data:

```
library(lctools)

# Load data
data(GR.Municipalities)
# Assign data to local field
gr_data<-GR.Municipalities

# Remove GR.Municipalities for better performance
rm(GR.Municipalities)
names(gr_data)
```

Για το συγκεκριμένο ερώτημα, θα χρησιμοποιήσουμε τις μεταβλητές X, Y και Income01, που είναι οι καρτεσιανές συντεταγμένες (longitude, latitude) των γεωμετρικών κεντροειδών των δήμων και το μέσο ετήσιο δηλωθέν οικογενειακό εισόδημα, που αποκτήθηκε το 2001 και δηλώθηκε το 2002, σε επίπεδο δήμου Καλλικράτη. Οι συντεταγμένες είναι απαραίτητες για την δημιουργία των γεωγραφικών πολυγώνων των χωρικών οντοτήτων, έτσι ώστε να είναι δυνατός ο προσδιορισμός κάθε φορά των κοντινότερων γειτόνων.

Στη συνέχεια, φτιάχνουμε ένα dataframe με τις συντεταγμένες, και δημιουργούμε το object l.moran το οποίο είναι και αυτό ένα dataframe 9 μεταβλητών και 325 παρατηρήσεων. Σαν input στη συνάρτηση l.moransI θα δώσουμε τις συντεταγμένες, τον αριθμό 6 γιατί μας ενδιαφέρουν οι 6 κοντινότεροι γείτονες κάθε δήμου, και το income (εισόδημα):

```
# Create coordinates dataframe
coordinates<-cbind(gr_data$X, gr_data$Y)

# Create l.moran object
l.moran<-l.moransI(coordinates,6,gr_data$Income01)
```

	ID	Ii	Ei	Vi	Zi	p.value	Xi	wXj	Cluster
1	1	-1.09371445	-0.01851852	5.87648	-0.44353612	6.573780e-01	0.16776961	-1.2948251789	0
2	2	3.91436454	-0.01851852	5.87648	1.62237936	1.047221e-01	-1.38514004	-0.5912794880	0
3	3	4.47774922	-0.01851852	5.87648	1.85478487	6.362695e-02	-1.31370450	-0.7022489342	0
4	4	2.77004347	-0.01851852	5.87648	1.15032798	2.500088e-01	-1.28262442	-0.4643497329	0
5	5	-1.14026540	-0.01851852	5.87648	-0.46273916	6.435513e-01	0.27981786	-0.8292234414	0
6	6	2.48204490	-0.01851852	5.87648	1.03152380	3.022953e-01	-0.80242852	-0.6421800464	0
7	7	4.28161979	-0.01851852	5.87648	1.77387823	7.608329e-02	-0.99702585	-0.8710121823	0
8	8	1.20247834	-0.01851852	5.87648	0.50368142	6.144853e-01	-0.75450707	-0.3565385350	0
9	9	4.71974300	-0.01851852	5.87648	1.95461131	5.062897e-02	-0.95080013	-0.9985659894	0
10	10	-3.35411817	-0.01851852	5.87648	-1.37599007	1.688247e-01	0.79051938	-0.8612086843	0
11	11	1.67749892	-0.01851852	5.87648	0.69963527	4.841551e-01	-0.62466707	-0.5645064956	0
12	12	1.19707460	-0.01851852	5.87648	0.50145228	6.160529e-01	-0.41312250	-0.6049298798	0
13	13	0.48493873	-0.01851852	5.87648	0.20768445	8.354754e-01	-0.55064465	-0.2207028258	0
14	14	1.52824748	-0.01851852	5.87648	0.63806658	5.234303e-01	-0.55456797	-0.5779015109	0
15	15	1.54722740	-0.01851852	5.87648	0.64589611	5.183466e-01	-0.57590797	-0.5647273112	0

Εικόνα 1 - Αντικείμενο l.moran

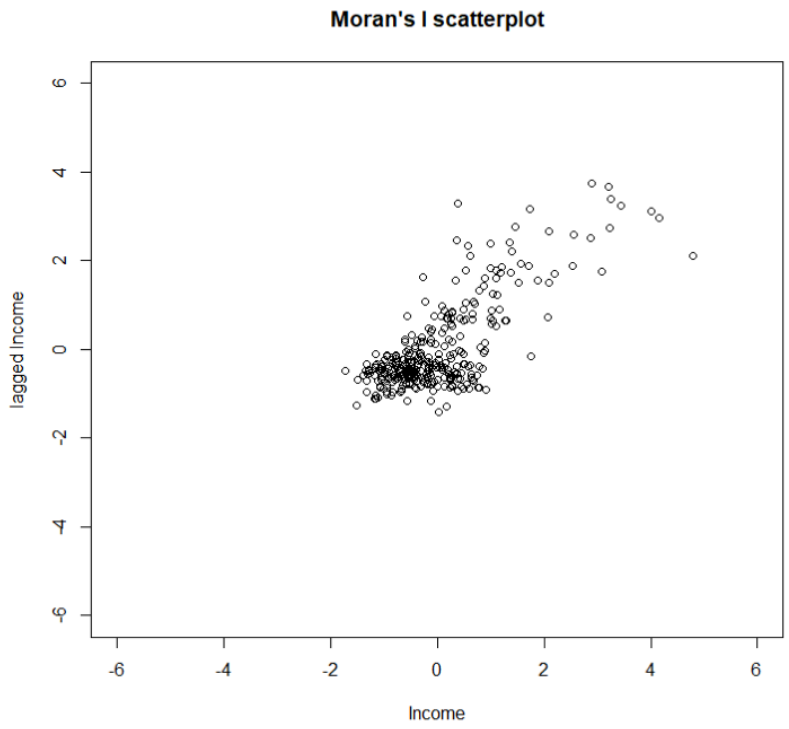
Στη συνέχεια, με τον παρακάτω κώδικα δημιουργούμε τις ελάχιστες και μέγιστες τιμές για τους άξονες x & y:

```r
# calculate range values in order to create scatter plot
xmin<-round(ifelse(abs(min(l.moran[,7])) > abs(min(l.moran[,8])),
                   abs(min(l.moran[,7])), abs(min(l.moran[,8]))))
xmax<-round(ifelse(abs(max(l.moran[,7])) > abs(max(l.moran[,8])),
                   abs(max(l.moran[,7])), abs(max(l.moran[,8]))))
xmax<-ifelse(xmin>xmax,xmin,xmax)+1
ymax<-xmax
xmin<- -xmax
ymin<- -ymax
```

Χρησιμοποιήσαμε τις τιμές l.moran[,7] και l.moran[,8], γιατί από την παραπάνω εικόνα (Εικόνα 1) βλέπουμε ότι αυτές είναι οι στήλες που μας ενδιαφέρουν, καθώς η στήλη 7 αντιστοιχεί στο Xi και η στήλη 8 στο wXj.

Εν συνεχεία, δημιουργούμε την γραμμή παλινδρόμησης που θα χρησιμοποιηθεί στο διάγραμμα, και ζωγραφίζουμε με την εντολή plot το αρχικό διάγραμμα που περιέχει τις παρατηρήσεις μας:

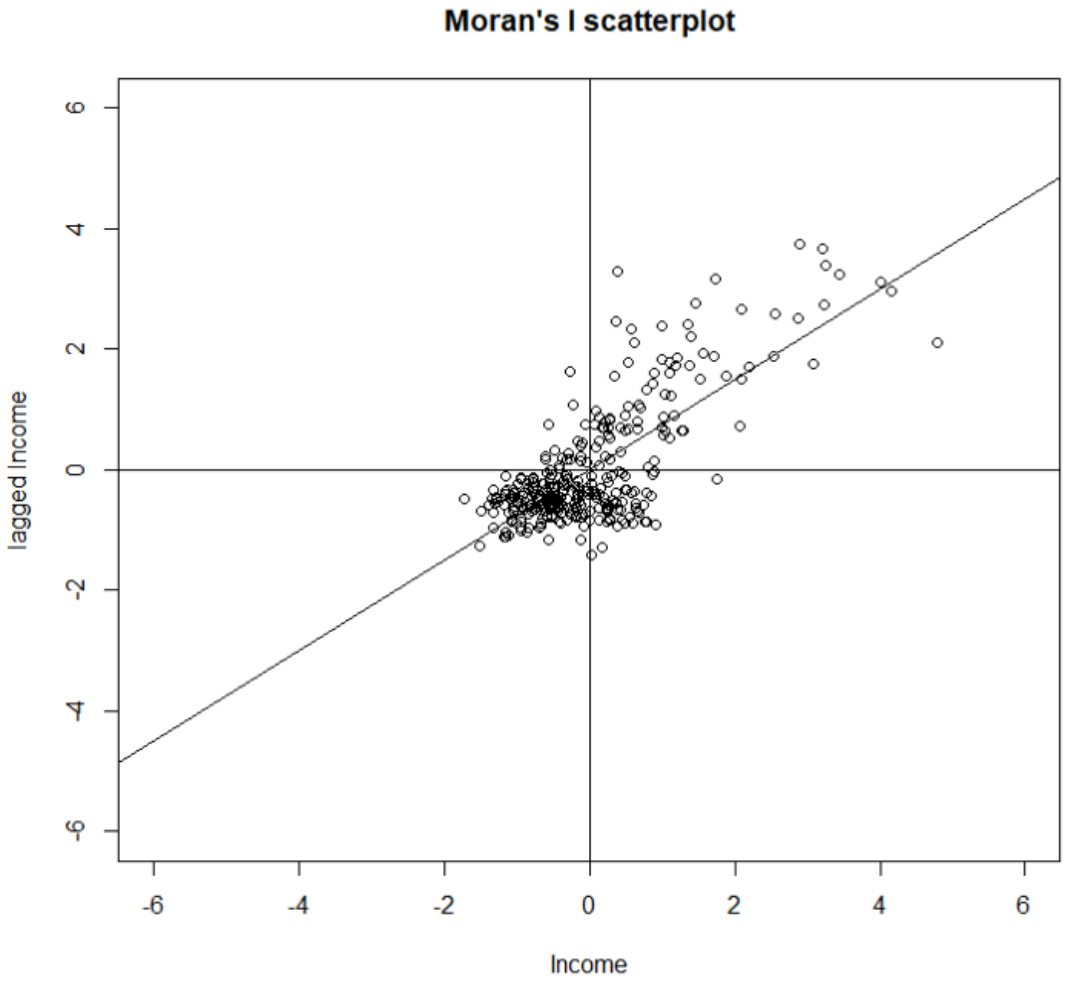
```r
# create regression line
regressionLine <- lm(l.moran[,8]~l.moran[,7])

# draw initial plot
plot(l.moran[,7], l.moran[,8], main="Moran's I scatterplot", sub="",
     xlab="Income", ylab="lagged Income", xlim=c(xmin, xmax),
     ylim=c(ymin, ymax))
```



Εικόνα 2 - Αρχικό διάγραμμα χωρίς άξονες

Τέλος, για να ολοκληρωθεί το διάγραμμα διασποράς, χρησιμοποιούμε την εντολή abline με τις κατάλληλες παραμέτρους, για να σχηματιστούν οι άξονες x, y καθώς και η γραμμή παλινδρόμησης:

```
# draw horizontal axis
abline(h=0)

# draw vertical axis
abline(v=0)

# γραμμή παλινδρόμησης
abline(regressionLine)
```



Εικόνα 3- Τελικό διάγραμμα διασποράς

Στο συνημμένο αρχείο υπάρχει και ολόκληρος ο κώδικας R του ερωτήματος 1α για ευκολία ανάγνωσης:

exercise_1a.r

Ερώτημα β

Να παρουσιαστούν και να αναλυθούν τα βήματα υπολογισμού των ολικών και τοπικών δεικτών Moran's I για το ποσοστό ανεργίας με το πακέτο lctools της R με εναλλακτικούς αριθμούς κοντινότερων γειτόνων π.χ. k = {3, 5, 9, 12, 15, 18, 20, 24, 30} και να σχολιαστούν τα αποτελέσματα.

Απάντηση β

Το ποσοστό ανεργίας που αναφέρεται στην παραπάνω εκφώνηση είναι η μεταβλητή UnemrT01 από το dataset. Όπως και στο ερώτημα α, αρχικά θα πρέπει με την χρήση κατάλληλου κώδικα να φορτωθούν τα δεδομένα, και να φτιαχτεί το vector coordinates που περιέχει πληροφορίες σχετικές με τις συντεταγμένες:

```
# Load libraries
library(lctools)
library(ggplot2)
library(rgeos)
# Load data and create initial object
data(GR.Municipalities)
gr_data<-GR.Municipalities
coordinates<-cbind(gr_data$X, gr_data$Y)
```

Στη συνέχεια, θα φτιάξουμε ένα dataframe moran, το οποίο θα περιέχει 9 γραμμές (1 γραμμή για κάθε k, δηλαδή για κάθε εναλλακτικό αριθμό κοντινότερων γειτόνων) και 7 στήλες. Θα αλλάξουμε το όνομα της κάθε στήλης και θα εκτυπώσουμε το αντικείμενο για να δούμε τα περιεχόμενα του:

```
# Create moran dataframe for multiple k values. k stands for
# municipality neighbour.
bw<-c(3, 5, 9, 12, 15, 18, 20, 24, 30)
moran<-matrix(data=NA, nrow=9, ncol=7)
counter<-1
for(b in bw){
    moranI<-moransI(coordinates,b,gr_data$UnemrT01)
    moran[counter,1]<-counter
    moran[counter,2]<-b
    moran[counter,3]<-moranI$Morans.I
    moran[counter,4]<-moranI$z.resampling
    moran[counter,5]<-moranI$p.value.resampling
    moran[counter,6]<-moranI$z.randomization
```

```
    moran[counter,7]<-moranI$p.value.randomization
    counter<-counter+1
}
# Change column names in moran object
colnames(moran)<-c("ID","k", "Moran's I", "Z resampling", "P-value
resampling", "Z randomization", "P-value
randomization")

# View moran object
moran
```

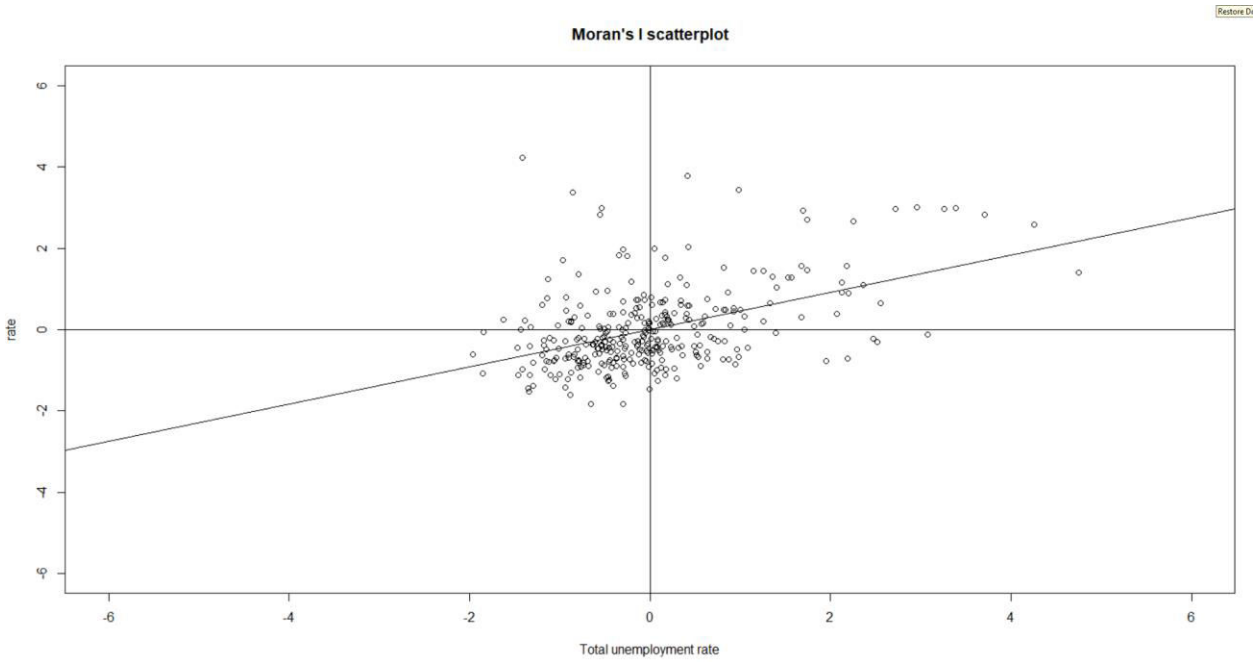
	ID	k	Moran's I	Z resampling	P-value resampling	Z randomization	P-value randomization
1	1	3	0.3216971	7.797530	6.313046e-15	7.838659	4.553815e-15
2	2	5	0.3099953	9.666900	4.168143e-22	9.717886	2.529801e-22
3	3	9	0.2412746	10.138538	3.726352e-24	10.191979	2.153345e-24
4	4	12	0.2163795	10.556241	4.753186e-26	10.611852	2.624950e-26
5	5	15	0.1901998	10.461191	1.302082e-25	10.516273	7.269233e-26
6	6	18	0.1687433	10.255480	1.118189e-24	10.309443	6.387017e-25
7	7	20	0.1618714	10.414360	2.132278e-25	10.469142	1.197238e-25
8	8	24	0.1359871	9.718285	2.519908e-22	9.769353	1.524230e-22
9	9	30	0.1047954	8.583662	9.190002e-18	8.628666	6.207168e-18

Εικόνα 4- moran dataframe

Ακολούθως, θα φτιάξουμε το αντίστοιχο διάγραμμα διασποράς με βάση την συνάρτηση l.moran, αυτή τη φορά βασιζόμενοι στο ποσοστό ανεργίας και όχι στο εισόδημα:

```
# Create l.moran object based on Unemployment
l.moran<-l.moransI(coordinates,6,gr_data$UnemrT01)
# calculate range values in order to create scatter plot
xmin<-round(ifelse(abs(min(l.moran[,7])) > abs(min(l.moran[,8])),
                   abs(min(l.moran[,7])), abs(min(l.moran[,8]))))
xmax<-round(ifelse(abs(max(l.moran[,7])) > abs(max(l.moran[,8])),
                   abs(max(l.moran[,7])), abs(max(l.moran[,8]))))
xmax<-ifelse(xmin>xmax,xmin,xmax)+1
ymax<-xmax
xmin<- -xmax
ymin<- -ymax
# create regression line
reg1 <- lm(l.moran[,8]~l.moran[,7])
# create initial plot
plot(l.moran[,7], l.moran[,8], main="Moran's I scatterplot", sub="",
```

```
     xlab="Total unemployment rate", ylab="lagged Total unemployment
rate", xlim=c(xmin, xmax), ylim=c(ymin, ymax))
# draw lines
abline(h=0)
abline(v=0)
abline(reg1)
```

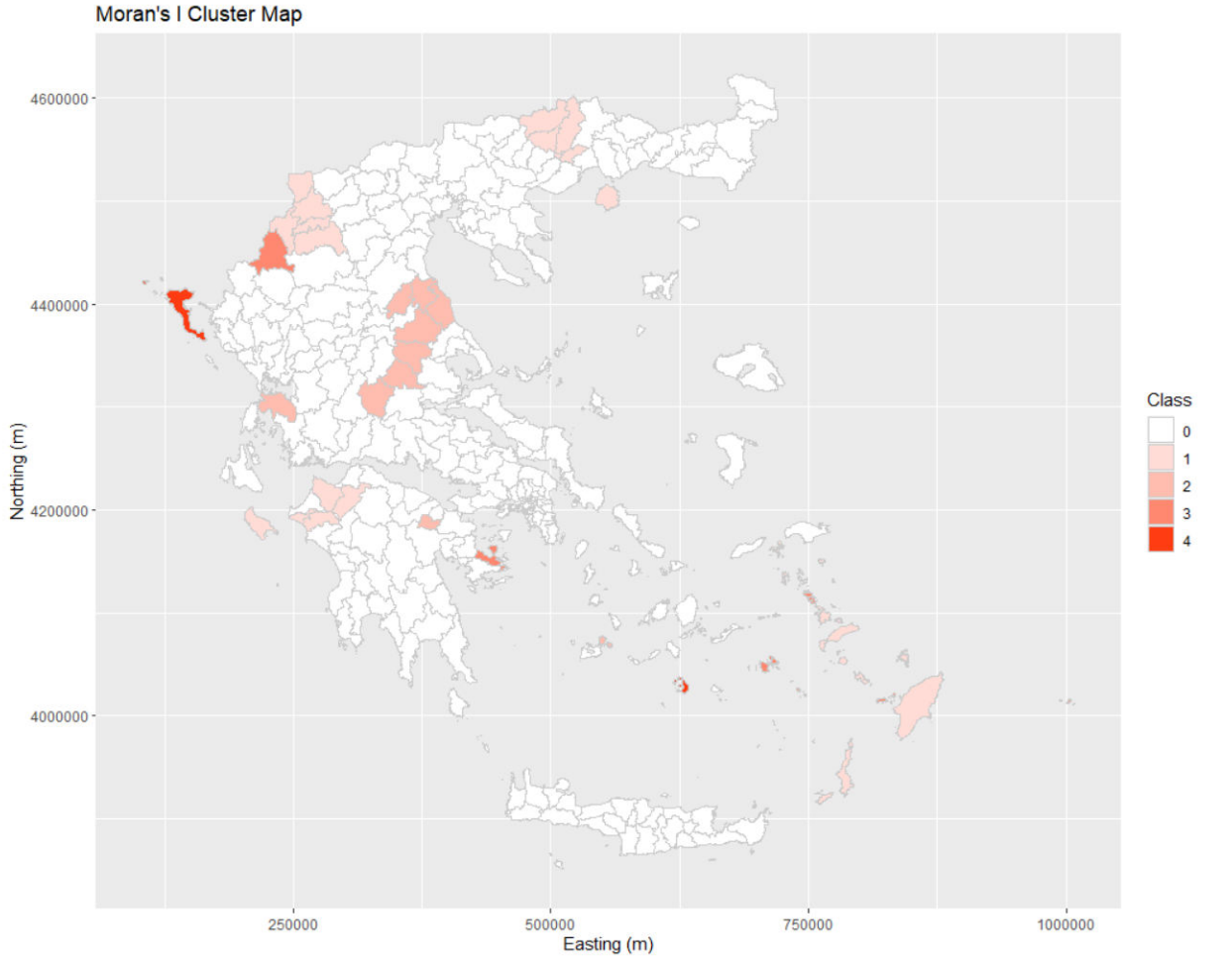


Εικόνα 5- διάγραμμα διασποράς με βάση το ποσοστό ανεργίας

Προκειμένου να προβούμε σε χρήσιμα συμπεράσματα, θα φτιάξουμε και τον αντίστοιχο χάρτη ο οποίος θα μας δείξει με χρώματα τους δήμους με τα υψηλά ποσοστά ανεργίας:

```
# prepare data for map creation
gr_data@data$Idx <- seq_len(nrow(gr_data))
gr_data_tmp <- merge(gr_data@data, l.moran, by.x="OBJECTID",
                     by.y="ID", sort=FALSE, all=TRUE)
gr_data@data<-gr_data_tmp[order(gr_data_tmp$Idx),]

# prepare map
map.f <- fortify(gr_data, region = "OBJECTID")
map.f <- merge(map.f, gr_data@data, by.x = "id", by.y = "OBJECTID")

# create map showing municipalities based on unemployment rate
# when an area is more red, it means the unemployment rate is higher
map <- ggplot(map.f, aes(long, lat, group = group)) +
    geom_polygon(colour="gray80",aes(fill=as.factor(Cluster))) +
```

```
    scale_fill_manual(values=c("white", "#FFDDD6", "#FFBDAF",
                               "#FF8970", "#FF3C12")) +
    coord_equal() +
    labs(x= "Easting (m)", y= "Northing (m)", fill= "Class") +
    ggtitle("Moran's I Cluster Map")
map
```



Εικόνα 6- Χάρτης Ελλάδας με ποσοστά ανεργίας

Συμπεράσματα:

Παρατηρώντας τον ολικό δείκτη Moran's I, δηλαδή το αντικείμενο moran που φτιάξαμε στην R, βλέπουμε ότι για k=5 το moran's I είναι 0.31, που σημαίνει ότι υπάρχει θετική και στατιστικά σημαντική χωρική αυτοσυσχέτιση στο ποσοστό ανεργίας στους δήμους της Ελλάδας. Όσο αυξάνεται το k μειώνεται η αυτοσυσχέτιση (δηλαδή τα ποσά είναι αντιστρόφως ανάλογα), κάτι το οποίο είναι λογικό. Επίσης, το συγκεκριμένο εύρημα παραμένει σταθερό μετά από ανάλυση ευαισθησίας (διαφορετικά βάρη).

Από την εικόνα 6 μπορούμε εύκολα να παρατηρήσουμε ότι υπάρχουν χωρικές εστίες γειτονικών δήμων με υψηλό ποσοστό ανεργίας, στη Μακεδονία και Θράκη, καθώς και στη Δυτική Ελλάδα, με την Κέρκυρα να βρίσκεται στο κόκκινο. Παρατηρούμε επίσης πως και στα

περισσότερα νησιά από τα Δωδεκάνησα, παρατηρείται σχετικά υψηλή ανεργία. Κρήτη, Πελοπόννησος, Θεσσαλία, Στερεά Ελλάδα και δη η Αττική, δεν αντιμετωπίζουν πρόβλημα ανεργίας, και αυτό μπορεί να οφείλεται στο γεγονός ότι οι συγκεκριμένες περιοχές βασίζονται αρκετά στον πρωτογενή τομέα και στη βιομηχανία για να κινείται η αγορά η εργασίας.

Συνημμένος ολόκληρος ο κώδικας του ερωτήματος 1β:

exercise_1b.r

Άσκηση 2

Δημιουργείστε δύο αντικριστά ιστογράμματα των ποσοστών ανεργίας ανδρών και γυναικών για κάποιο έτος της επιλογής σας και συγκρίνετε την κατανομή τους.

Απάντηση 2

Ένα ιστόγραμμα είναι μια προσεγγιστική απεικόνιση της κατανομής που ακολουθούν αριθμητικά δεδομένα (Pearson, 1895). Για να δημιουργηθεί ένα ιστόγραμμα, πρέπει πρώτα να υπολογιστεί ο λεγόμενος κουβάς, δηλαδή το εύρος (μέγιστο - ελάχιστο) που έχουν τα υπό απεικόνιση δεδομένα. Στη συνέχεια το εύρος χωρίζεται σε n ίσα μέρη, και προσμετράται ο αριθμός των στοιχείων που αναλογούν σε κάθε μεσοδιάστημα. Αυτό το βήμα καθορίζει και το ύψος που θα έχει το κάθε μεσοδιάστημα (interval). Ο κουβάς αποτελείται από πολλά μεσοδιαστήματα, τα οποία συνήθως είναι το ένα δίπλα στο άλλο και ίδιου μήκους/μεγέθους.

Σύμφωνα με το επίσημο documentation της R, και όπως έχει αναφερθεί και στη ενότητα «Ανάλυση του συνόλου δεδομένων GR.Municipalities», το πακέτο GR.Municipalities περιέχει δημογραφικά στοιχεία (όπως είναι και το ποσοστό ανεργίας) από την απογραφή πληθυσμού του 2001, συνεπώς τα ιστογράμματα θα δημιουργηθούν βάσει της συγκεκριμένης χρονιάς. (rdocumentation, 2011)

`UnemrM01`

a numeric vector of unemployment rate for males in 2001 (Census)

`UnemrF01`

a numeric vector of unemployment rate for females in 2001 (Census)

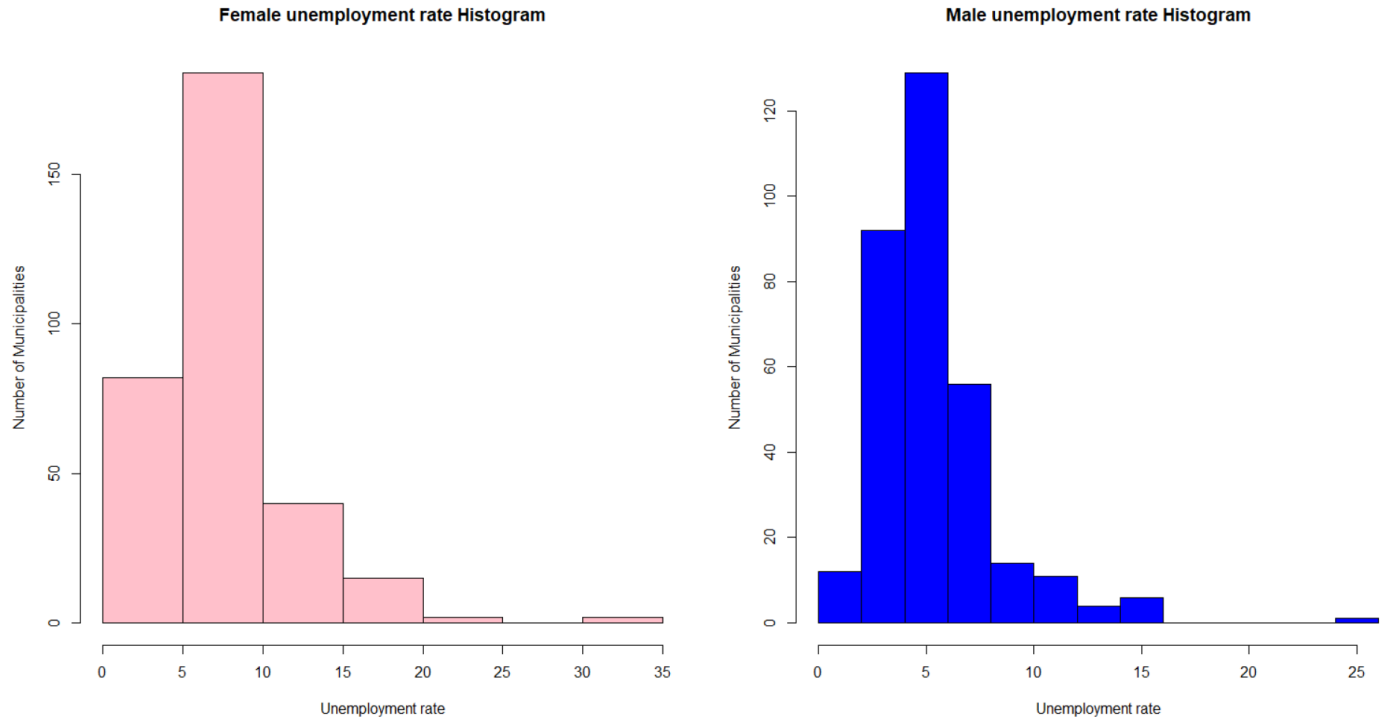
`UnemrT01`

a numeric vector of total unemployment rate in 2001 (Census)

Εικόνα 7-Μεταβλητές για ποσοστά ανεργιάς ανδρών, γυναικών και συνολικά

Κώδικας R για να φτιάξουμε τα αντικριστά ιστογράμματα:

```r
# load library
library(lctools)

# load data
data(GR.Municipalities)
gr_data <- GR.Municipalities

# get data for male and female unemployment
male_unem <- gr_data$UnemrM01
female_unem <- gr_data$UnemrF01

# create histograms
par(mfrow=c(1,2))

female.hist <- hist(female_unem, breaks=10, col="pink",
                    main="Female unemployment rate Histogram",
                    xlab="Unemployment rate",
                    ylab="Number of Municipalities")

male.hist <- hist(male_unem, breaks=10, col="blue",
                  main="Male unemployment rate Histogram",
                  xlab="Unemployment rate",
                  ylab="Number of Municipalities")
```



Εικόνα 8- Ιστογράμματα ανεργίας για άντρες και γυναίκες

Συνημμένο το αρχείο με κώδικα R για την λύση της άσκησης 2:

exercise_2.r

Παρατηρούμε ότι τα διαγράμματα έχουν παρόμοια κατανομή ωστόσο φαίνεται ότι το ποσοστό ανεργίας των ανδρών είναι κάτω από 10% για τους περισσότερους δήμους ενώ παρατηρούνται αρκετοί δήμοι με ποσοστό ανεργίας γυναικών μεγαλύτερο από 10%.

Από την εικόνα 8, βλέπουμε επίσης πως για τις γυναίκες, παρατηρείται ποσοστό ανεργίας της τάξης 5-10% σε πάνω από 150 δήμους, ενώ αντίστοιχα για τους άντρες η μεγαλύτερη συγκέντρωση είναι στο γύρω από το ποσοστό 5% (4-6%) και παρατηρείται σε λίγο περισσότερους από 120 δήμους. Στους άντρες βλέπουμε επίσης πως μετά το 7.5%, σε ελάχιστους δήμους (κάτω από 20) παρατηρούνται μεγαλύτερα ποσοστά ανεργίας. Στις γυναίκες, ακόμα και στο range ανεργίας 10-15%, βλέπουμε πως συναντάται σε ένα μέγεθος της τάξης κοντά στους 40 δήμους.

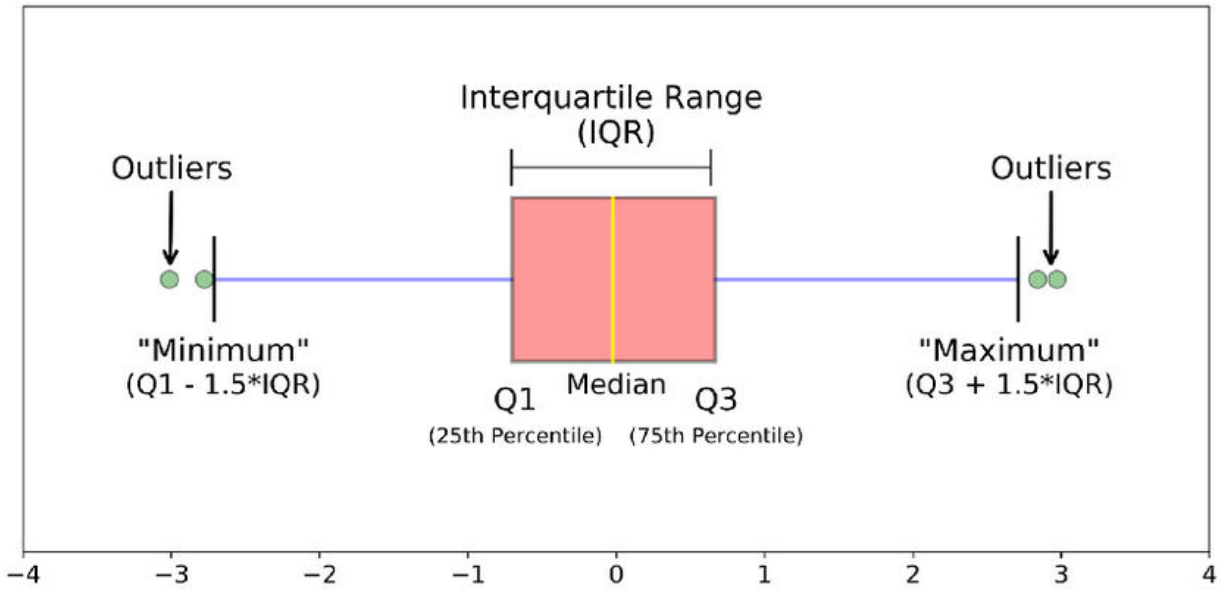
Από τις παραπάνω παρατηρήσεις, το τελικό συμπέρασμα είναι πως για την χρονιά 2001 υπήρχαν σημαντικά περισσότερες άνεργες γυναίκες στη χώρα μας, απ’ ότι άντρες.

Άσκηση 3

Δημιουργείστε ένα θηκόγραμμα με τα ποσοστά ανεργίας ανδρών και γυναικών το 2001 και συγκρίνετε τις τυχούσες έκτροπες παρατηρήσεις τους.

Απάντηση 3

Στην περιγραφική στατιστική, ένα θηκόγραμμα είναι μια μέθοδος γραφιστικής απεικόνισης συνόλων δεδομένων, μέσω της χρήσης των τεταρτημορίων τους. Τα θηκογράμματα μπορεί επίσης να περιέχουν γραμμές που επεκτείνονται εκτός των κύριων κουτιών, υποδεικνύοντας έτσι κατανομή των δεδομένων εκτός του μέγιστου και ελάχιστου τεταρτημορίου. Ανάλογα με το dataset, μπορεί να απεικονίζονται και τυχούσες έκτροπες παρατηρήσεις, συνήθως με την μορφή μικρών κύκλων ή αστερίσκων (Galarnyk, 2018).



Εικόνα 9 - Διαφορετικά μέρη ενός θηκογράμματος (οριζόντια απεικόνιση)

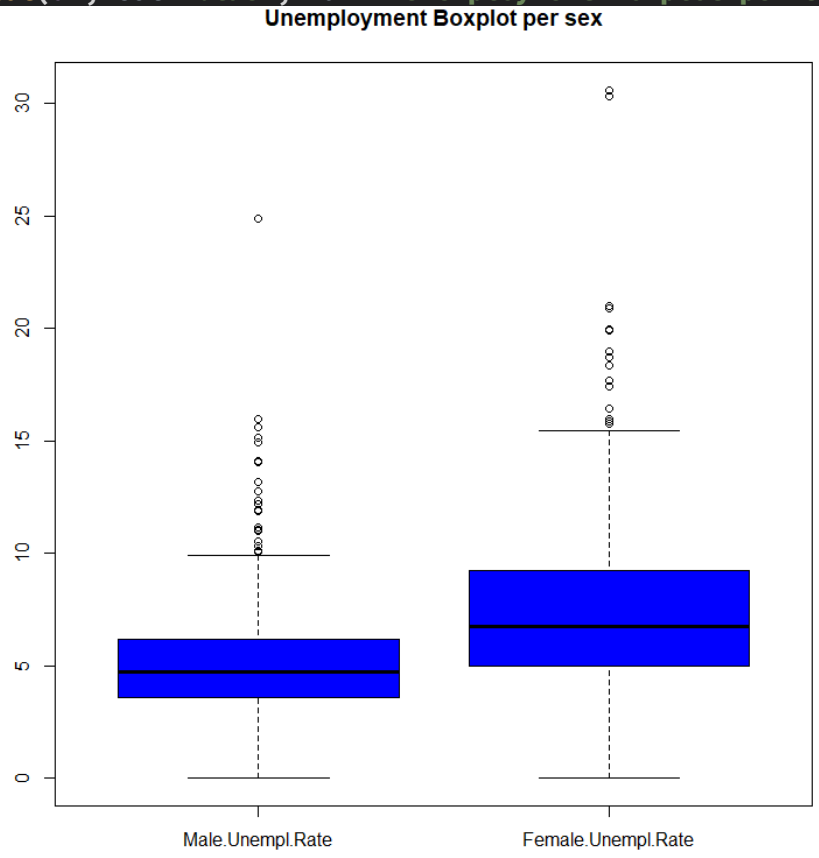
Ένα θηκόγραμμα, που μπορεί να απεικονιστεί είτε οριζόντια είτε κάθετα, αποτελείται από τα εξής μέρη:

- Το ελάχιστο (τεταρτημόριο 0, Q_0), το οποίο είναι το ελάχιστο σημείο στα δεδομένα, εξαιρώντας τυχούσες έκτροπες τιμές.
- Το μέγιστο (τεταρτημόριο 4, Q_4), το οποίο είναι το μέγιστο σημείο στα δεδομένα, εξαιρώντας τυχούσες έκτροπες τιμές.
- Το πρώτο τεταρτημόριο (Q_1 ή 25%) που είναι η μέση τιμή του κάτω μισού μέρους των δεδομένων.
- Το τρίτο τεταρτημόριο (Q_3 ή 75%) που είναι η μέρη τιμή του άνω μισού μέρους των δεδομένων.
- Το μέσο (Q_2 ή 50%) που είναι η μέση τιμή του συνόλου δεδομένων.
- Τυχούσες έκτροπες τιμές, οι οποίες είναι δεδομένα τα οποία βρίσκονται εκτός του Q_0 και του Q_4.

Όπως αναλύθηκε και στην άσκηση 2, έτσι και εδώ θα χρησιμοποιήσουμε τις ίδιες μεταβλητές. Προκειμένου να δημιουργήσουμε το ζητούμενο θηκόγραμμα, θα φτιάξουμε ένα αρχικό dataframe object το οποίο θα περιέχει τα ποσοστά ανεργείας ανά φύλο, τόσο για τους άντρες όσο και για τις γυναίκες. Στη συνέχεια θα φτιάξουμε το θηκόγραμμα δίνοντας το dataframe που δημιουργήσαμε ως input:

```
# Load library and data
library(lctools)
data(GR.Municipalities)
```

```r
gr_data<-GR.Municipalities

# create dataframe that includes both male and female unempl. rates
df<-data.frame(Male.Unempl.Rate= gr_data$UnemrM01,
               Female.Unempl.Rate= gr_data$UnemrF01)

# create boxplot based on above dataframe, which shows unemployment per sex
b.plot<-boxplot(df, col="blue", main="Unemployment Boxplot per sex")
```



Εικόνα 10 - θηκόγραμμα ανεργίας ανά φύλο

Στη συνέχεια, βασιζόμενοι στο παραπάνω boxplot θα εξάγουμε τις έκτροπες τιμές ανά φύλο, και θα υπολογίσουμε τα όρια έξω από τα οποία μια τιμή του συνόλου δεδομένων θα θεωρείται έκτροπη.

```r
# extract outliers per sex
outliers<-cbind(out=b.plot$out,sex=b.plot$group)

# calculate outlier limit per sex
out.limits <- aggregate(out ~ sex, data=outliers, min)
```

Έχοντας κάνει όλη την προεργασία, θα εξάγουμε από το αρχικό dataset τις έκτροπες τιμές για κάθε φύλο, και θα τις κάνουμε print.

```r
# extract outliers from initial data, for males
male.outliers<-gr_data@data[which(gr_data$UnemrM01>=out.limits$out[1]),]

#view male outliers
male.outliers[,c(4,9)]

# extract outliers from initial data, for females
female.outliers<-gr_data@data[which(gr_data$UnemrF01>=out.limits$out[2]),]

#view female outliers
female.outliers[,c(4,10)]
```

```
> male.outliers[,c(4,9)]
                                 Name UnemrM01
5                             DOXATOU 10.09678
8                          PROSOTSANIS 10.03695
14                              THASOU 11.03057
59                           SITHONIAS 11.93038
61                              VOIOU 15.13032
63                  SERVION - VELVENTOU 10.29525
66                           KASTORIAS 14.04237
67                           NESTORIOU 13.16568
68                           ORESTIDOS 14.93939
134 MANTOUDIOU - LIMNIS - AGIAS ANNAS 15.99265
140                           KERKYRAS 10.31348
149                     DYTIKIS ACHAIAS 10.33253
255                              POROU 11.89050
260                     FOURNON KORSEON 14.07942
265                           OINOUSSON 24.86773
274                           KALYMNION 10.51821
275                       AGATHONISIOU 12.76596
281                              KASOU 11.17021
285                           NISYROU 15.58442
295                             RODOU 12.18173
297                             SYMIS 12.33886
308                       CHERSONISOU 11.00833
>
```

Εικόνα 11 - Έκτροπες τιμές για άντρες

```
> female.outliers[,c(4,10)]
            Name UnemrF01
5        DOXATOU 15.85945
14       THASOU  18.37838
59     SITHONIAS 15.94635
66     KASTORIAS 18.95463
67     NESTORIOU 30.60498
68     ORESTIDOS 19.95478
266       PSARON 15.78947
269       THIRAS 18.70048
279       PATMOU 19.89101
280    KARPATHOU 20.87227
281        KASOU 15.87302
284           KO 16.44737
285      NISYROU 30.33708
295        RODOU 20.99539
297        SYMIS 17.67442
308 CHERSONISOU 17.43067
```

Εικόνα 12 - Έκτροπες τιμές για γυναίκες

Συνημμένος κώδικας R για την άσκηση 3:

exercise_3.r

Παρατηρώντας τα θηκογράμματα της εικόνας 9, προκύπτει ότι σε περισσότερους από τους μισούς (50%) δήμους της Ελλάδας το ποσοστό ανεργίας των γυναικών είναι μεγαλύτερο από αυτό των γυναικών. Ωστόσο, σε σχέση με την κατανομή των ποσοστών, στα δεδομένα για την ανεργία των ανδρών υπάρχουν περισσότερες έκτροπες παρατηρήσεις.

References

Galarnyk, M. (2018). *towardsdatascience*. Retrieved from towardsdatascience: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

Pearson, K. (1895). Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material. In K. Pearson.

rdocumentation. (2011). *rdocumentation*. Retrieved from rdocumentation: https://www.rdocumentation.org/packages/lctools/versions/0.2-8/topics/GR.Municipalities