

# Practica 2 - Tipologia i cicle de vida de les dades

*Sara Quesada Gil*

*11 de junio, 2019*

## Contents

|           |  |           |
|-----------|--|-----------|
| <b>1</b>  | <b>Introducció</b>   | <b>1</b>  |
| <b>2</b>  | <b>Apartats Pràctica 2</b>   | <b>1</b>  |
| <b>3</b>  | <b>Apartat 1 - Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre?</b>   | <b>2</b>  |
| <b>4</b>  | <b>Apartat 2 - Integració i selecció de les dades d'interès a analitzar.</b>   | <b>2</b>  |
| <b>5</b>  | <b>Apartat 3 - Neteja de les dades.</b>  | <b>4</b>  |
| 5.1       | Apartat 3.1 - Les dades contenen zeros o elements buits? Com gestionaries aquests casos? . . . . .   | 7         |
| 5.2       | Apartat 3.2 - Identificació i tractament de valors extrems. . . . .  | 9         |
| <b>6</b>  | <b>Apartat 4 - Anàlisi de les dades.</b>   | <b>19</b> |
| 6.1       | Apartat 4.1 - Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar). . . . .   | 19        |
| 6.2       | Apartat 4.2 - Comprovació de la normalitat i homogeneïtat de la variància. . . . .   | 20        |
| 6.3       | Apartat 4.3 - Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d'anàlisi diferents. . . . . | 28        |
| <b>7</b>  | <b>Apartat 5 - Representació dels resultats a partir de taules i gràfiques.</b>  | <b>37</b> |
| <b>8</b>  | <b>Apartat 6 - Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?</b>  | <b>40</b> |
| <b>9</b>  | <b>Apartat 7 - Codi: Cal adjuntar el codi, preferiblement en R, amb el que s'ha realitzat la neteja, anàlisi i representació de les dades. Si ho preferiu, també podeu treballar en Python.</b>  | <b>41</b> |
| <b>10</b> | <b>Recursos</b>  | <b>41</b> |

## 1 Introducció

Aquesta és una proposta de resolució de la Pràctica 2 de l'assignatura *Tipologia i cicle de vida de les dades* de l'alumna *Sara Quesada*.

**Components del grup:** *Sara Quesada Gil*

## 2 Apartats Pràctica 2

La Pràctica 2 es composa de 7 apartats, dels quals es mostrarà la seva corresponent solució.

*Seguint les principals etapes d'un projecte analític, les diferents tasques a realitzar (i justificar) són les següents:*

### 3 Apartat 1 - Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre?

Després de realitzar una recerca exhaustiva entre els diferents repositoris de www.kaggle.com, per a trobar un dataset que respongués algun tema que considero interessant, finalment he decidit escollir aquest, degut no únicament a la pregunta que pretén respondre, sinó a l'estructura d'aquest, ja que és un conjunt de dades amb suficients observacions com per a aplicar diferents mètodes d'anàlisi, i aconseguir una resposta més acurada sobre la pregunta que pretén respondre. Es pretén saber quins factors influeixen en l'augment de les taxes de suïcidi, i si aquests es troben en les variables del dataset, és a dir, si aquestes variables són explicatives del nombre dels suïcidis.

El dataset s'ha extret de la següent url: (<https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016>)

Al següent apartat, realitzarem una descripció de cada variable del dataset, un cop hagim integrat les dades.

### 4 Apartat 2 - Integració i selecció de les dades d'interès a analitzar.

Per a descriure les variables del dataset, primer carregarem les dades:

```
suicideData<-read.csv("C:/PRAC2/suicide_rates.csv", header=T, sep = ",",na.strings = "NA")
```

Amb la funció dim(), podrem veure el nombre d'observacions del dataset i el nombre de variables. Aquest dataset té 27820 observacions i 12 variables.

```
dim(suicideData)
```

```
## [1] 27820    12
```

Mitjançant la funció str() mostrarem els tipus de dades de la variable, i alguns valors que prenen, però també podrem detectar a priori si hem de canviar el tipus de dades d'alguna variable.

```
str(suicideData)
```

```
## 'data.frame': 27820 obs. of 12 variables:
## $ .i..country : Factor w/ 101 levels "Albania","Antigua and Barbuda",...: 1 1 1 1 1 1 1 1 1 1 ...
## $ year       : int  1987 1987 1987 1987 1987 1987 1987 1987 1987 ...
## $ sex        : Factor w/ 2 levels "female","male": 2 2 1 2 2 1 1 1 2 1 ...
## $ age         : Factor w/ 6 levels "15-24 years",...: 1 3 1 6 2 6 3 2 5 4 ...
## $ suicides_no : int  21 16 14 1 9 1 6 4 1 0 ...
## $ population  : int  312900 308000 289700 21800 274300 35600 278800 257200 137500 311000 ...
## $ suicides.100k.pop: num  6.71 5.19 4.83 4.59 3.28 2.81 2.15 1.56 0.73 0 ...
## $ country.year : Factor w/ 2321 levels "Albania1987",...: 1 1 1 1 1 1 1 1 1 1 ...
## $ HDI.for.year : num  NA NA NA NA NA NA NA NA NA ...
## $ gdp_for_year....: Factor w/ 2321 levels "1,002,219,052,968",...: 727 727 727 727 727 727 727 727 727 727 ...
## $ gdp_per_capita....: int  796 796 796 796 796 796 796 796 796 796 ...
## $ generation     : Factor w/ 6 levels "Boomers","G.I. Generation",...: 3 6 3 2 1 2 6 1 2 3 ...
```

També mitjançant la funció summary() es mostren els valors que prenen les variables, amb una visió més amplia, i mostra alguns càculs estadístics de les variables quantitatives. Aquesta funció ens servirà per a comprovar si hi ha valors erronis, per exemple, veiem que la variable HDI.for.year conté 19456 NA's o elements buits.

```
summary(suicideData)
```

```
##          .i..country      year       sex           age
## Austria   : 382  Min.   :1985  female:13910  15-24 years:4642
## Iceland   : 382  1st Qu.:1995  male  :13910   25-34 years:4642
## Mauritius : 382  Median :2002                    35-54 years:4642
```

```

##  Netherlands: 382  Mean   :2001          5-14 years :4610
##  Argentina   : 372  3rd Qu.:2008          55-74 years:4642
##  Belgium     : 372  Max.    :2016          75+ years  :4642
##  (Other)      :25548
##  suicides_no   population    suicides.100k.pop
##  Min.       : 0.0  Min.       : 278  Min.       : 0.00
##  1st Qu.:  3.0  1st Qu.: 97498  1st Qu.:  0.92
##  Median   : 25.0  Median   :430150  Median   :  5.99
##  Mean     : 242.6  Mean     :1844794  Mean     : 12.82
##  3rd Qu.: 131.0  3rd Qu.:1486143  3rd Qu.: 16.62
##  Max.     :22338.0  Max.     :43805214  Max.     :224.97
##
##  country.year  HDI.for.year      gdp_for_year...
##  Albania1987: 12  Min.     :0.483  1,002,219,052,968: 12
##  Albania1988: 12  1st Qu.:0.713  1,011,797,457,139: 12
##  Albania1989: 12  Median   :0.779  1,016,418,229   : 12
##  Albania1992: 12  Mean     :0.777  1,018,847,043,277: 12
##  Albania1993: 12  3rd Qu.:0.855  1,022,191,296   : 12
##  Albania1994: 12  Max.     :0.944  1,023,196,003,075: 12
##  (Other)      :27748  NA's     :19456  (Other)      :27748
##  gdp_per_capita... generation
##  Min.     : 251  Boomers      :4990
##  1st Qu.: 3447  G.I. Generation:2744
##  Median   : 9372  Generation X  :6408
##  Mean     : 16866  Generation Z  :1470
##  3rd Qu.: 24874  Millenials   :5844
##  Max.     :126352  Silent       :6364
##

```

A continuació, farem una breu explicació de les variables del dataset:

- *country*: Variable de tipus factor que descriu el país d'estudi dels suïcidis. Pot prendre 101 valors, com hem vist en aplicar la funció str().

- *year*: Variable de tipus integer que indica l'any en el que s'han comès els suïcidis estudiats.

- *age*: Variable de tipus factor que indica el rang d'edat dels suïcidis comessos. Hi ha 6 valors diferents. Traurem la cadena "years" dels nivells del factor.

- *suicides\_no*: Variable de tipus integer que indica el nombre de suïcidis

- *population*: Variable de tipus integer que indica la població, segons un any, un rang d'edat i el sexe.

- *suicides.100k.pop*: \* Variable de tipus numeric que indica la taxa de suïcidis en relació a una població, en percentatge del 100000%. Aquest valor el podem obtenir aplicant un càlcul, pel que aquesta columna seria candidata a ser descartada. El càlcul seria el següent: (*suicides\_no*/*population*)100000

- *country.year*: Variable de tipus factor que concatena el país amb l'any d'ocurrència del suïcidi. Ja tenim la variable país i la variable any, llavors aquesta columna és una clara candidata a ser descartada del conjunt de dades.

- *HDI.for.year*: Variable de tipus numeric, que indica l'índex de desenvolupament humà. Un 70% de les observacions d'aquesta variable són valors buits, pel que aquesta variable és una clara candidata a ser descartada del dataset.

- *gdp\_for\_year*: Variable de tipus numeric que indica el PIB per any del país. Aquesta variable està erròniament tipada com a factor, ja que els separadors s'han aplicat amb el caràcter ','. A l'apartat de neteja de dades aplicarem la solució corresponent.

- *gdp\_per\_capita*: Variable de tipus integer que indica el PIB per capita, segons un país, un any, un rang d'edat i el sexe.

- *generation*: Variable de tipus factor amb 6 nivells, que indica la generació en la qual ha sigut classificada l'observació corresponent.

Com podem veure, el nom de les columnes no induceix a error i descriu correctament la variable, però alguns tenen punts al final, llavors en canviarem el nom.

A més, no farà falta discretitzar cap variable, el tipus de totes, a excepció de la variable gdp\_for\_year , són els adequats.

En resum, a l'apartat següent de neteja de dades, haurem de realitzar, entre d'altres, les següents tasques de neteja:

- a) Modificar el nom de les columnes.
- b) Traurem el caràcter ‘,’ a la variable gdp\_for\_year i canviar el tipus de dades a numeric. També traurem la cadena ” years” dels nivells del factor age.
- c) Mirar si no hi ha factors repetits, per exemple, que no hi hagi “Albania” i ” Albania”.
- d) Eliminar les variables: HDI for year, country-year, suicides.100k.pop.
- e) Buscarem valors perduts i outliers.

## 5 Apartat 3 - Neteja de les dades.

a) Començarem canviant el nom de les variables:

```
colnames(suicideData)[1] <- "country"
colnames(suicideData)[7] <- "suicides_100k"
colnames(suicideData)[8] <- "country_year"
colnames(suicideData)[9] <- "HDI_for_year"
colnames(suicideData)[10] <- "gdp_for_year"
colnames(suicideData)[11] <- "gdp_per_capita"
```

b) A continuació, substituirem el caràcter ‘,’ pel caràcter ‘?’ a la variable gdp\_for\_year i canviarem el tipus de dades a numeric.

Apliquem la funció options(scipen = 999) per a evitar el format amb la notació científica.

```
options(scipen = 999)
suicideData$gdp_for_year<-gsub(",","",suicideData$gdp_for_year)
suicideData$gdp_for_year<-as.numeric(suicideData$gdp_for_year)
```

I comprovarem mitjançant una taula, que el tipus de les variables és el correcte:

```
tipus <- sapply(suicideData,class)
kable(data.frame(Variables=names(tipus),Classe=as.vector(tipus)),
      caption = "Tipus de dades")
```

A continuació, traurem la cadena ” years” dels nivells del factor age:

```
suicideData$age<-as.factor(gsub(" years","",suicideData$age))
```

c) Ara, mirarem que no hi hagi factors a les variables de tipus factor:

No mirarem la variable country\_year, que era clara candidata a descartar.

Aplicarem la funció sapply(), concretament sapply(data1,levels), per a obtenir els nivells de les variables factors per a poder comprovar que aquestes són correctes, però al codi aplicarem la funció str() perquè en executar el codi i presentar-ho en pdf no surtin moltes pàgines generades pel gran volum de dades.

Table 1: Tipus de dades

| Variables      | Classe  |
|----------------|---------|
| country        | factor  |
| year           | integer |
| sex            | factor  |
| age            | factor  |
| suicides_no    | integer |
| population     | integer |
| suicides_100k  | numeric |
| country_year   | factor  |
| HDI_for_year   | numeric |
| gdp_for_year   | numeric |
| gdp_per_capita | integer |
| generation     | factor  |

```
qualitatives<-c(1,3,4,8,12)
data1<-suicideData[,qualitatives]
str(sapply(data1,levels),6)

## List of 5
## $ country      : chr [1:101] "Albania" "Antigua and Barbuda" "Argentina" "Armenia" ...
## $ sex          : chr [1:2] "female" "male"
## $ age          : chr [1:6] "15-24" "25-34" "35-54" "5-14" ...
## $ country_year: chr [1:2321] "Albania1987" "Albania1988" "Albania1989" "Albania1992" ...
## $ generation   : chr [1:6] "Boomers" "G.I. Generation" "Generation X" "Generation Z" ...
```

d) A continuació, procedim a eliminar les variables: HDI\_for\_year, country\_year, suicides\_100k\_pop.

Les columnes a eliminar són la 7, 8 i 9:

```
suicideData<-suicideData[,-(c(7,8,9))]
```

Com a tècnica de reducció de dimensió, pot ser interessant aplicar una Anàlisi de Components Principals (PCA).

L'objectiu de l'Anàlisi de Components Principals és sintetitzar el nombre d'atributs del dataset, és a dir, reduir el nombre d'atributs del dataset, el que resulta especialment beneficiós en datasets que tinguin un nombre molt elevat de variables, disminuint d'aquesta manera la dimensió de la problemàtica de l'anàlisi.

El procediment és transformar el conjunt de variables originals en un conjunt de noves variables que anomenarem components principals, i que la seva principal característica és que existeixi correlació entre alguna d'aquestes variables.

Per a aquest estudi únicament té sentit seleccionar les variables quantitatives, ja que per al càclul de les components necessitem estimar una matriu de correlació:

En el cas del nostre dataset, el qual únicament conté 6 variables quantitatives pot resultar no tan útil com si s'apliqués a un dataset de major dimensió.

```
quantitatives<-c(2,5,6,7,8)
df_quant<-suicideData[,quantitatives]

pca.quantitatives <- prcomp(df_quant,scale=T)
pca.quantitatives

## Standard deviations (1, ..., p=5):
```

```

## [1] 1.4990444 1.1506881 0.8256388 0.7245929 0.4712415
##
## Rotation (n x k) = (5 x 5):
##          PC1        PC2        PC3        PC4        PC5
## year      0.1093880  0.68794788 -0.69785754  0.1656590 -0.01771865
## suicides_no 0.5020203 -0.24185814 -0.32278450 -0.7157662 -0.27012806
## population  0.5919981 -0.20875303 -0.07438154  0.2414139  0.73630639
## gdp_for_year 0.5754973  0.03127484  0.25515971  0.5016230 -0.59253073
## gdp_per_capita 0.2331330  0.65090431  0.58152036 -0.3877235  0.18296750
summary(pca.quantitatives)

## Importance of components:
##          PC1        PC2        PC3        PC4        PC5
## Standard deviation 1.4990 1.1507 0.8256 0.7246 0.47124
## Proportion of Variance 0.4494 0.2648 0.1363 0.1050 0.04441
## Cumulative Proportion 0.4494 0.7142 0.8506 0.9556 1.00000

```

Podem veure que l'aplicació de la funció prcomp() ens proposa 5 components principals. La primera variable year, explica el 44,94% de la variància total, i els 4 primers components expliquen el 95,56%. Podríem prescindir del quint, però en aquest cas concret, com no tenim tantes variables, deixarem el dataset com ho tenim.

*e) Buscarem valors perduts i outliers.*

Això ho farem en els apartats 3.1 i 3.2 respectivament.

## 5.1 Apartat 3.1 - Les dades contenen zeros o elements buits? Com gestionaries aquests casos?

Els valors perduts o buits poden manifestar-se amb el nom “NA”, amb zeros o amb algun caràcter del tipus “”, “ ” o “?”.

A l'apartat 2, concretament al subapartat c), hem comprovat que els nivells de les variables qualitatives eren correctes i no contenien valors buits. A continuació, realitzarem la prova a tot el dataset, per analitzar especialment si les variables de tipus quantitatius contenen valors buits.

Podem aplicar la funció sapply() passant com a paràmetre la funció is.na() que s'aplicarà a tot el dataset:

```
sapply(suicideData, function(x) sum(is.na(x)))
```

```
##          country            year           sex         age suicides_no
##             0                  0                  0                  0                  0
##      population   gdp_for_year  gdp_per_capita generation
##             0                  0                  0                  0
```

Com podem comprovar, no es detecta cap valor perdut. Aplicarem una prova, indicant NA com a string:

```
grep("NA",suicideData)>0
```

```
## logical(0)
```

Comprovem que efectivament no es detecten valors perduts.

Com hem comentat, haguéssim detectat els valors perduts fàcilment amb l'aplicació de la funció summary() al dataset:

```
summary(suicideData)
```

```
##          country            year           sex         age
##    Austria     : 382  Min.   :1985  female:13910  15-24:4642
##    Iceland     : 382  1st Qu.:1995   male  :13910  25-34:4642
##    Mauritius   : 382  Median  :2002                35-54:4642
##    Netherlands: 382  Mean    :2001                5-14  :4610
##    Argentina   : 372  3rd Qu.:2008                55-74:4642
##    Belgium     : 372  Max.    :2016                75+   :4642
##    (Other)      :25548
##    suicides_no           population      gdp_for_year
##    Min.     : 0.0  Min.     : 278  Min.     : 46919625
##    1st Qu.: 3.0  1st Qu.: 97498  1st Qu.: 8985352832
##    Median   :25.0  Median   :430150  Median   :48114688201
##    Mean     :242.6  Mean     :1844794  Mean     :445580969026
##    3rd Qu.:131.0  3rd Qu.:1486143  3rd Qu.:260202429150
##    Max.    :22338.0  Max.    :43805214  Max.    :18120714000000
##
##          gdp_per_capita        generation
##    Min.     : 251  Boomers       :4990
##    1st Qu.: 3447 G.I. Generation:2744
##    Median   : 9372 Generation X  :6408
##    Mean     :16866 Generation Z   :1470
##    3rd Qu.:24874 Millennials   :5844
##    Max.    :126352 Silent       :6364
##
```

En cas de detectar valors perduts, hagués sortit la secció NA's. Si ens fixem, únicament la variable suicides\_no conté el valor zero, cas completament normal, ja que poden existir països, en relació a un any, a

un rang d'edat, a un sexe i a una població, que no hi hagi enregistrat cap suïcidi.

Per a respondre la pregunta de l'apartat, el nostre dataset no conté valors buits o zeros, és cert que la variable HDI.for.year sí contenia valors buits, però aquesta variable ha sigut descartada perquè un 70% de les dades era buida, a més, aquesta variable, encara que hi hagués continguts els valors correctes, no es considerava rellevant per al nostre estudi.

Si haguéssim hagut de tractar els elements buits o zeros, en primer lloc és primordial el volum de dades que engloba, és a dir, quants valors perduts s'han detectat:

- *Si tenim molts valors perduts:* Pot passar que aquests estiguin repartits en tot el dataset, en aquest cas tindríem un problema, ja que hauríem de predir les dades i llavors no estaríem treballant amb dades reals tretes d'una mostra, sinó d'una predicció.

En aquest cas hauríem de valorar si el dataset del que disposem és la millor opció.

Un altre cas pot ser el que ens ha passat al nostre dataset d'estudi, que els valors perduts estiguin concentrats en una mateixa variable, que aquests representin la major part de les observacions de la variable, i a més aquesta no es consideri rellevant per a l'estudi.

En aquest cas, el més encertat és discriminar la variable del dataset i realitzar l'estudi a partir de la resta de dades.

- *Si tenim pocs valors perduts:*

En aquest cas podem optar per eliminar les observacions, però estaríem perdent informació, i seria més aconsellable predir les dades i treballar amb aquestes, que prescindir completament d'aquesta informació. En aquest cas optaríem pel següent:

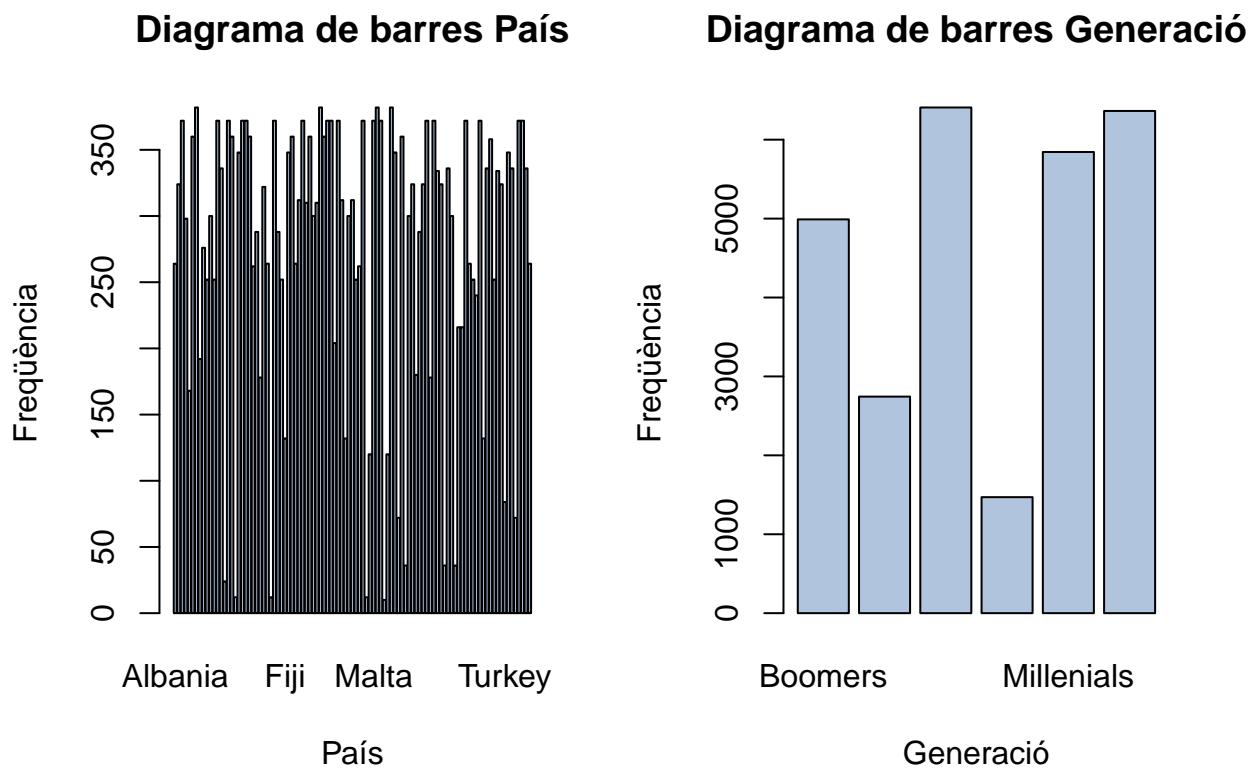
- a) Independentment del tipus de la variable: Podríem aplicar la funció knn() o k veïns més propers, que és de les eines d'aprenentatge automàtic supervisades més populars, basada en distàncies, i que el que fa és cercar en les observacions més pròximes a la qual s'està tractant de predir i classifica el punt d'interès basat en la majoria de dades que li envolten.
- b) Segons el tipus de la variable:
  - Si la variable és qualitativa: Podríem optar per emplenar el valor amb un nou nivell “Sense informació”, o emplenar-ho amb la moda, o valor més repetit, sempre tenint present els riscos que això comporta.
  - Si la variable és quantitativa: El més habitual és aplicar una mesura de tendència central com la mitjana o la mediana.

## 5.2 Apartat 3.2 - Identificació i tractament de valors extrems.

Els valors extrems o outliers, són observacions que semblen inconsistentes amb la resta dels valors de la mostra. Amb aquesta definició, podem intuir que una bona manera de detectar els outliers o valors extrems és mitjançant diagrames de caixa (boxplot) per a les variables quantitatives i diagrames circulars o diagrames de barres per a les quantitatives.

- *Diagrama de barres per a les qualitatives:*

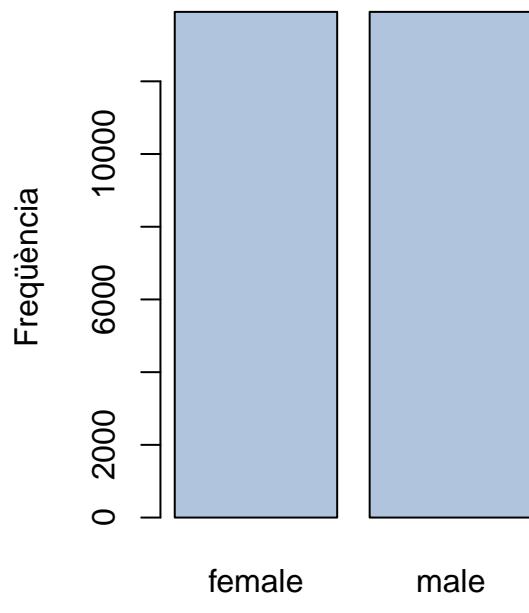
```
par(mfrow = c(1:2))
barplot(table(suicideData$country), xlab="País", ylab="Freqüència",
        main="Diagrama de barres País", col="lightsteelblue")
barplot(table(suicideData$generation), xlab="Generació", ylab="Freqüència",
        main="Diagrama de barres Generació", col="lightsteelblue")
```



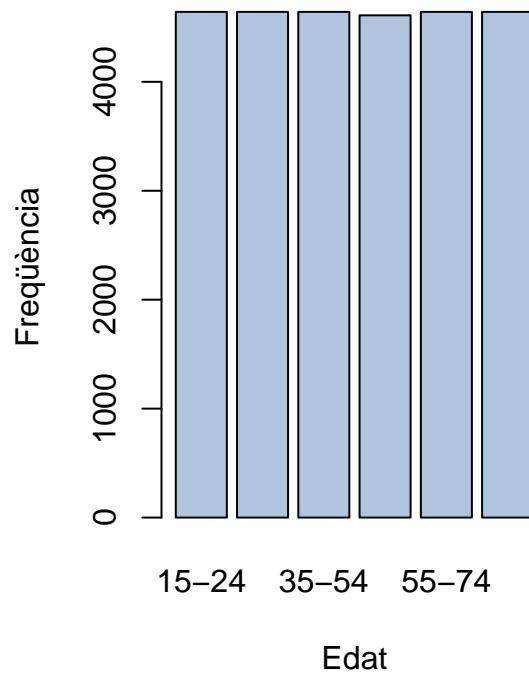
No observem cap valor extrem, les dades estan repartides de manera uniforme, és clar que per exemple, en relació als països, hi haurà més observacions per a uns que per a altres, però podem considerar els valors com normals.

```
par(mfrow = c(1:2))
barplot(table(suicideData$sex), xlab="Sexe", ylab="Freqüència",
        main="Diagrama de barres Sexe", col="lightsteelblue")
barplot(table(suicideData$age), xlab="Edat", ylab="Freqüència",
        main="Diagrama de barres Edat", col="lightsteelblue")
```

**Diagrama de barres Sexe**



**Diagrama de barres Edat**



Si observem els diagrames, veiem que el nombre d'observacions per a cada nivell de cada factor és molt similar. Ho comparem amb taules per a confirmar-ho:

```
table(suicideData$sex)
```

```
##  
## female male  
## 13910 13910
```

```
table(suicideData$age)
```

```
##  
## 15-24 25-34 35-54 5-14 55-74 75+  
## 4642 4642 4642 4610 4642 4642
```

-Boxplot per a les quantitatives:

A continuació, analitzarem si existeixen outliers en les variables quantitatives. En apartats anteriors, mitjançant l'aplicació de la funció `summary()`, vam veure que les variables quantitatives, a excepció de la variable `year`, prenien valors molt diferents, llavors els diagrames de caixa sortiran esbiaixats, però els resultats són correctes, no hi ha presència d'outliers, simplement hi ha valors molt baixos i molts alts en la mateixa variable, això afecta a totes les variables quantitatives, a excepció de la variable `year`.

Per a mostrar que els resultats del boxplot no són deguts a presència d'outliers, sinó a la diferència exagerada en els valors d'una mateixa variable, possarem d'exemple la variable `suicides_no`, i primer distribuirem les observacions en rangs per poder representar de millor manera les dades, mitjançant l'algorisme d'Sturges, i després ho mostrarem gràficament.

Ens ajudarem de les següents taules:

Table 2: Variables qualitatives

| country          | sex          | age        | generation           |
|------------------|--------------|------------|----------------------|
| Austria : 382    | female:13910 | 15-24:4642 | Boomers :4990        |
| Iceland : 382    | male :13910  | 25-34:4642 | G.I. Generation:2744 |
| Mauritius : 382  |              | 35-54:4642 | Generation X :6408   |
| Netherlands: 382 |              | 5-14 :4610 | Generation Z :1470   |
| Argentina : 372  |              | 55-74:4642 | Millenials :5844     |
| Belgium : 372    |              | 75+ :4642  | Silent :6364         |
| (Other) :25548   |              |            |                      |

Table 3: Variables quantitatives discretes

| year         | suicides_no    | population       | gdp_per_capita |
|--------------|----------------|------------------|----------------|
| Min. :1985   | Min. : 0.0     | Min. : 278       | Min. : 251     |
| 1st Qu.:1995 | 1st Qu.: 3.0   | 1st Qu.: 97498   | 1st Qu.: 3447  |
| Median :2002 | Median : 25.0  | Median : 430150  | Median : 9372  |
| Mean :2001   | Mean : 242.6   | Mean : 1844794   | Mean : 16866   |
| 3rd Qu.:2008 | 3rd Qu.: 131.0 | 3rd Qu.: 1486143 | 3rd Qu.: 24874 |
| Max. :2016   | Max. :22338.0  | Max. :43805214   | Max. :126352   |
|              |                |                  |                |

```
options(knitr.kable.NA = ' ')
qualitatives<-c(1,3,4,9)
discretes<-c(2,5,6,8)
continues<-c(7)
kable(summary(suicideData)[,qualitatives],
      digits=2, align='l', caption="Variables qualitatives")
```

- Estudi de variables quantitatives discretes:

```
kable(summary(suicideData)[,discretes],
      digits=2, align='l', caption="Variables quantitatives discretes")
```

- Estudi de variables quantitatives contínues:

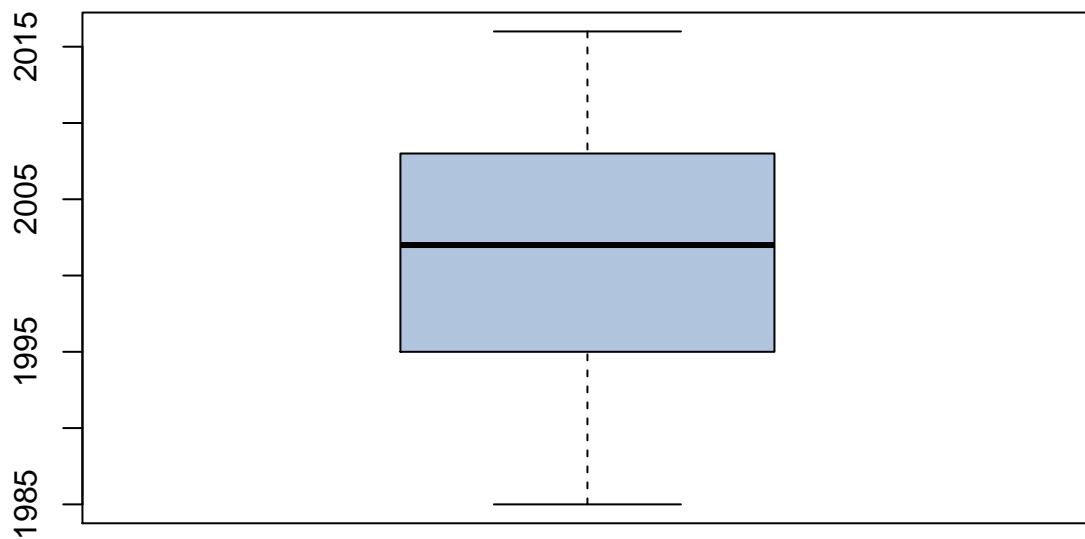
```
kable(summary(suicideData)[,continues],
      digits=2, align='l', caption="Variables quantitatives contínues")
```

```
boxplot(suicideData$year,main="Box plot: Any", col="lightsteelblue",na.rm=TRUE)
```

Table 4: Variables quantitatives continues

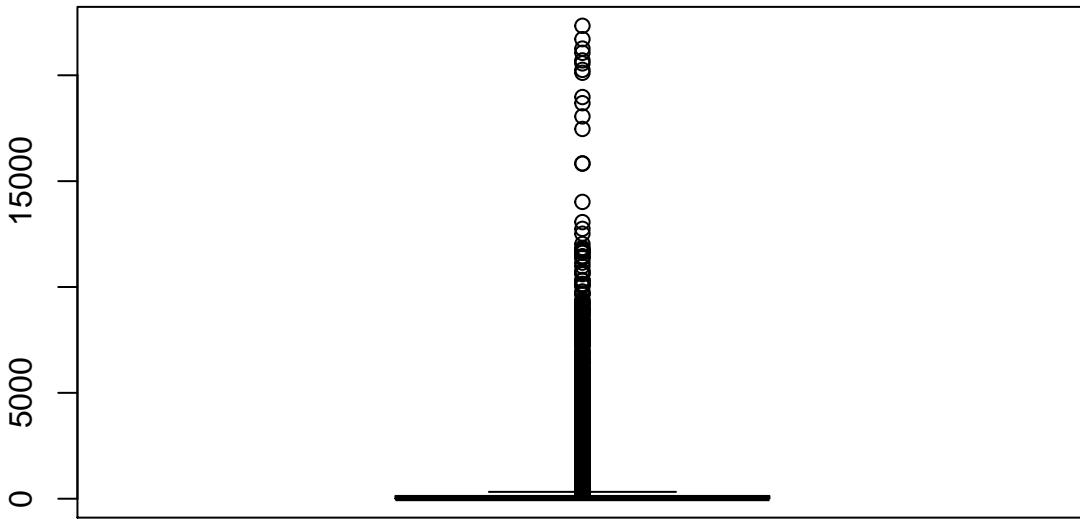
| x                     |
|-----------------------|
| Min. : 46919625       |
| 1st Qu.: 8985352832   |
| Median : 48114688201  |
| Mean : 445580969026   |
| 3rd Qu.: 260202429150 |
| Max. : 18120714000000 |

### Box plot: Any



```
boxplot(suicideData$suicides_no, main="Box plot: Nombre de suicidis",
        col="lightsteelblue", na.rm=TRUE)
```

## Box plot: Nombre de suïcidis



Apliquem Sturges:

```
rang.h<-range(suicideData$suicides_no,na.rm=TRUE)
nclass.Sturges(suicideData$suicides_no)

## [1] 16
seq(rang.h[1],rang.h[2],length=nclass.Sturges(suicideData$suicides_no))

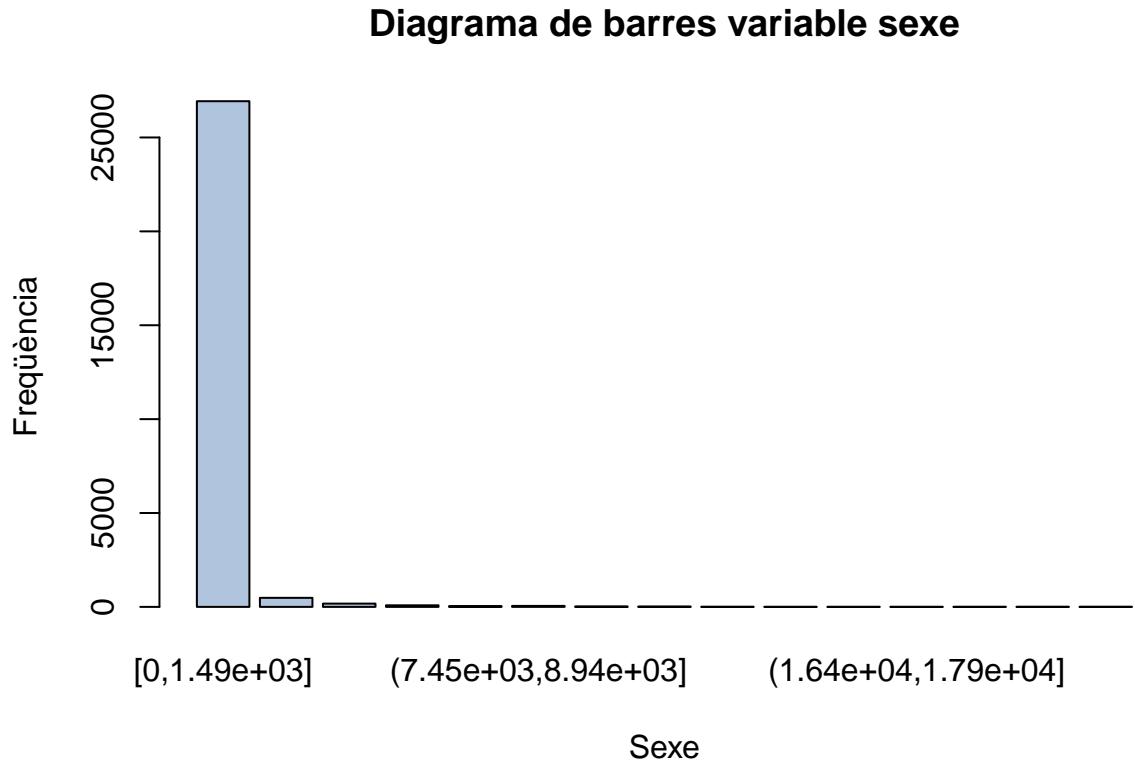
## [1] 0.0 1489.2 2978.4 4467.6 5956.8 7446.0 8935.2 10424.4
## [9] 11913.6 13402.8 14892.0 16381.2 17870.4 19359.6 20848.8 22338.0
intervalshs=cut(suicideData$suicides_no,breaks=seq(rang.h[1],rang.h[2],length=
nclass.Sturges(suicideData$suicides_no)),include.lowest=TRUE)
table(intervalshs)

## intervalshs
## [0,1.49e+03] (1.49e+03,2.98e+03] (2.98e+03,4.47e+03]
## 26932 481 174
## (4.47e+03,5.96e+03] (5.96e+03,7.45e+03] (7.45e+03,8.94e+03]
## 84 40 52
## (8.94e+03,1.04e+04] (1.04e+04,1.19e+04] (1.19e+04,1.34e+04]
## 21 17 4
## (1.34e+04,1.49e+04] (1.49e+04,1.64e+04] (1.64e+04,1.79e+04]
## 1 2 1
## (1.79e+04,1.94e+04] (1.94e+04,2.08e+04] (2.08e+04,2.23e+04]
## 3 4 4
```

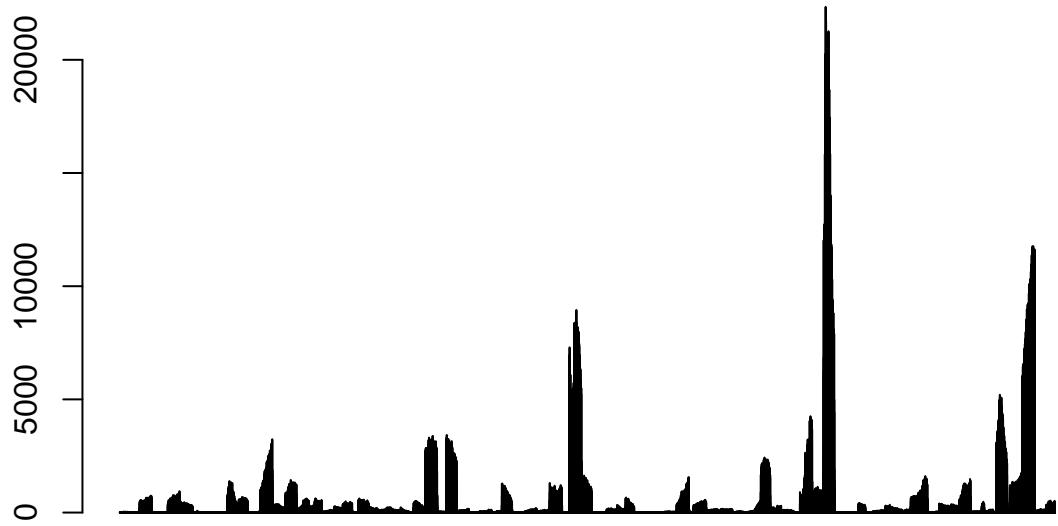
Efectivament veiem que no són outliers, sinó que la major part de les observacions es concentra en el primer

valor, i la resta estan repartits, però són correctes. La mateixa situació s'aplica als gràfics de la resta de variables quantitatives:

```
barplot(table(intervalshs), xlab="Sexe", ylab="Freqüència",
        main="Diagrama de barres variable sexe", col="lightsteelblue")
```

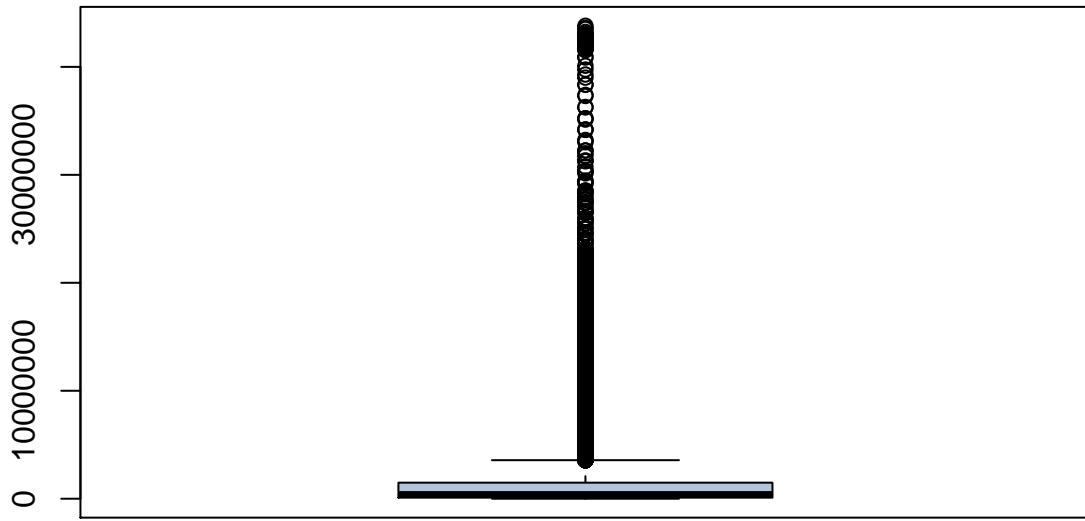


```
barplot(suicideData$suicides_no)
```



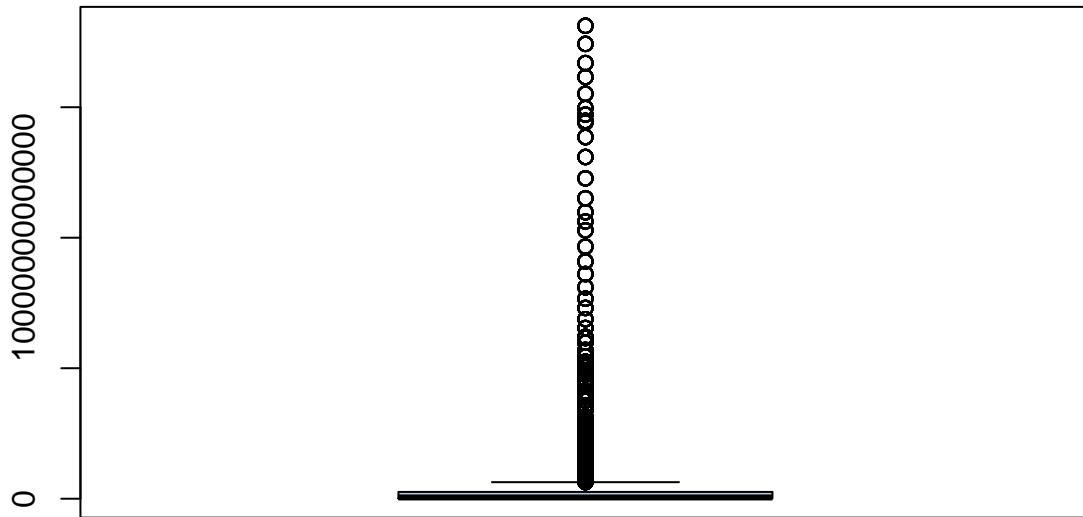
```
boxplot(suicideData$population, main="Box plot: Població", col="lightsteelblue", na.rm=TRUE)
```

## Box plot: Població



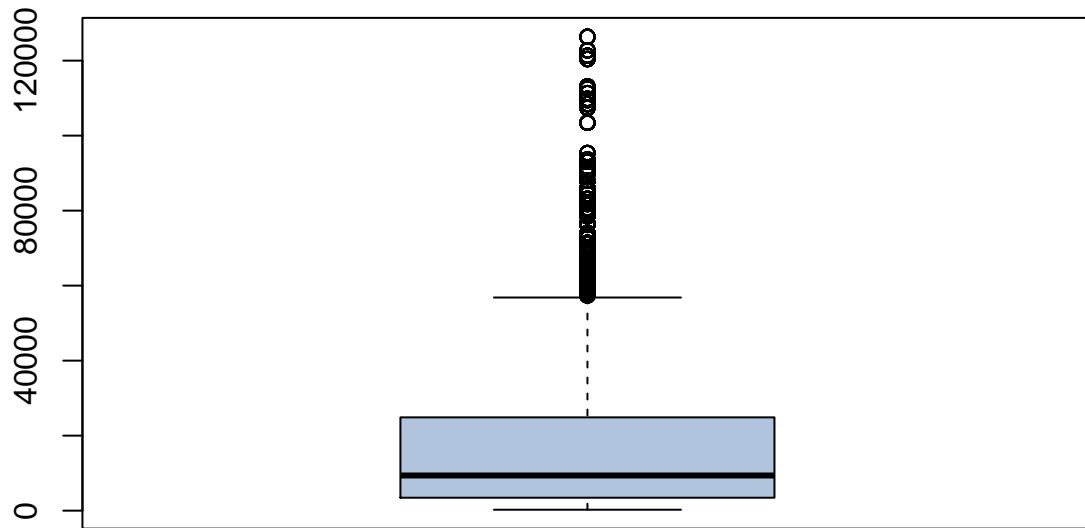
```
boxplot(suicideData$gdp_for_year, main="Box plot: PIB per any", col="lightsteelblue", na.rm=TRUE)
```

### Box plot: PIB per any



```
boxplot(suicideData$gdp_per_capita, main="Box plot: PIB per capita", col="lightsteelblue", na.rm=TRUE)
```

## Box plot: PIB per capita



Guardarem el conjunt de dades net:

```
ruta <- "C:/PRAC2/suicide_rates_Solution.csv"  
write.csv(suicideData,file=ruta, row.names=F)
```

## 6 Apartat 4 - Anàlisi de les dades.

### 6.1 Apartat 4.1 - Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar).

Per tal d'assolir l'objectiu d'aquesta anàlisi, concretament, quins factors influeixen en l'augment de les taxes de suïcidi, seria interessant comparar els següents grups de dades:

- suicides\_no, sex, age
- suicides\_no, generation
- suicides\_no, gdp\_per\_capita

Les preguntes que ens farem per tal d'assolir el nostre objectiu principal són les següents:

- a) Quines variables quantitatives tenen més influència en el nombre de suïcidis?
- b) El nombre de suïcidis està influït pel PIB per capita, la població, el rang d'edat i la generació?
- c) Existeixen diferències significatives en el nombre de suïcidis dels homes en relació a les dones?

Per a dur-lo a terme, construirem un nou dataset a l'apartat b, on incorporarem dues noves variables qualitatives o dummy, que ens permetran assolir el nostre objectiu.

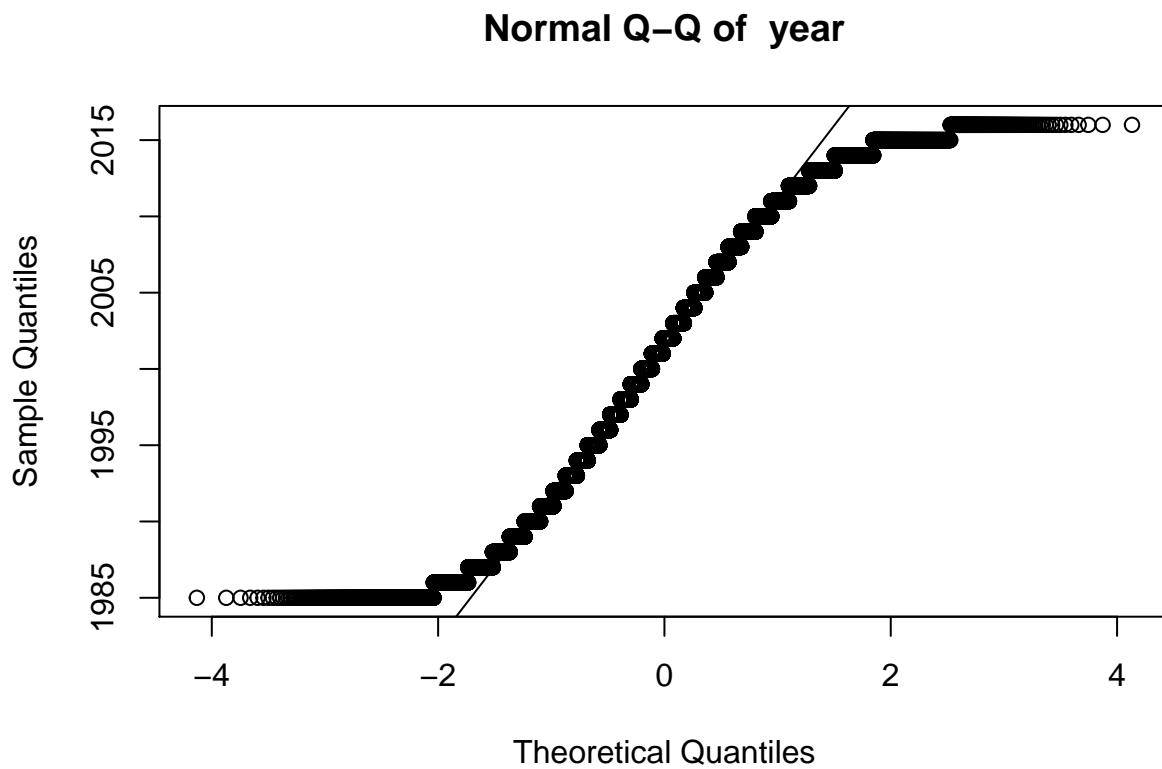
## 6.2 Apartat 4.2 - Comprovació de la normalitat i homogeneïtat de la variància.

### a) Estudi de la normalitat

A continuació, estudiarem si les variables qualitatives segueixen una distribució normal. Ho farem mitjançant la representació gràfica de la corba de la normalitat i mitjançant el test de normalitat Lilliefors de Komolgorov-Smirnov, ja que és el test que més s'adqua al nostre dataset.

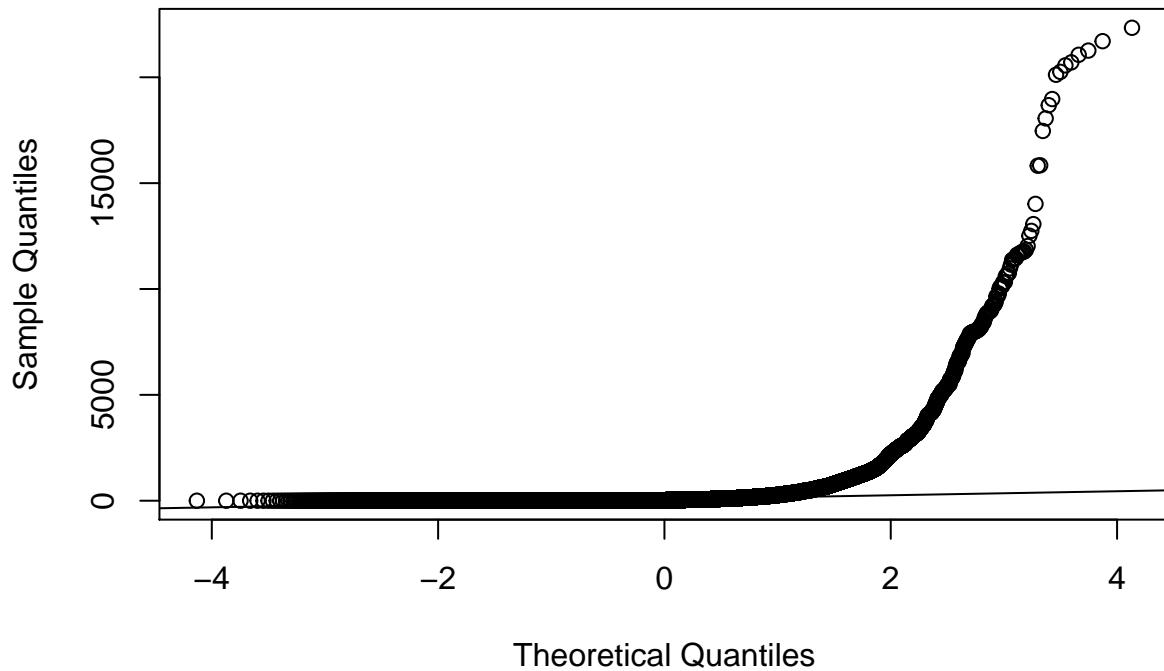
El test de Shapiro Wilk és adequat quant el nombre d'observacions és petit. Es recomana per a un nombre d'observacions inferior o igual a 30, encara que es pot aplicar fins a 5000 observacions, com veurem en un exemple:

```
quantit<-colnames(suicideData[,quantitatives])
compt<-0
for(i in quantitatives){
  compt<-compt+1
  qqnorm(suicideData[,i],main=paste ("Normal Q-Q of ", quantit[compt]))
  qqline(suicideData[,i])
  print(lillie.test(suicideData[,i]))
}
```



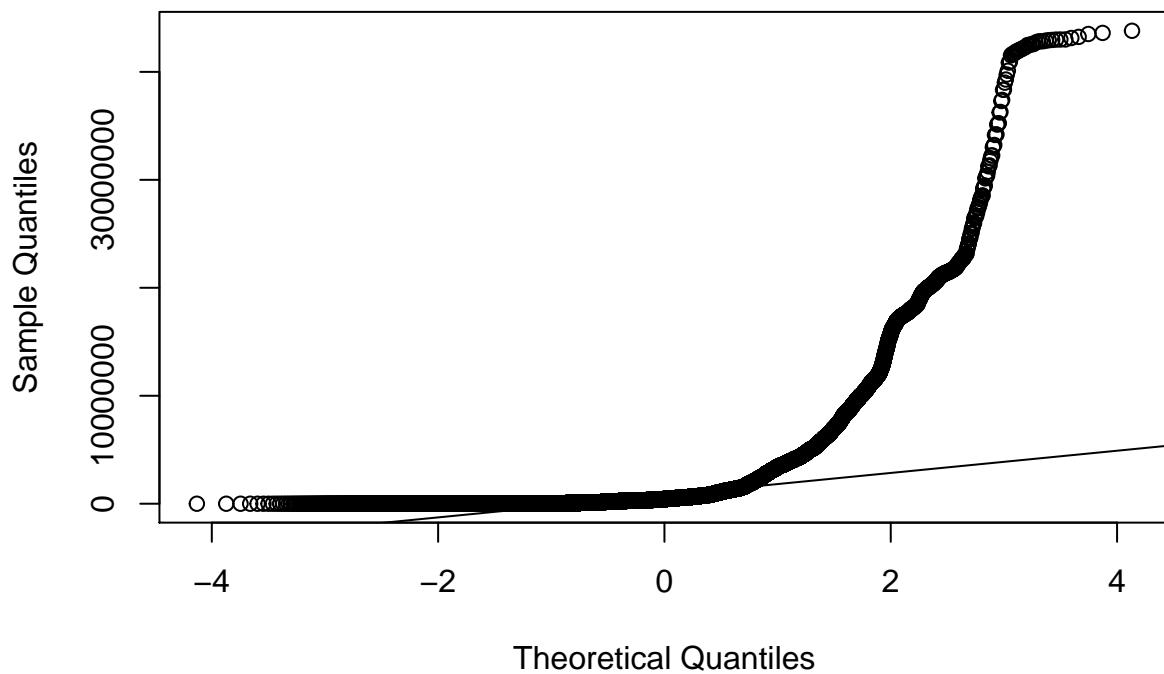
```
##  
##  Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data: suicideData[, i]  
## D = 0.073889, p-value < 0.0000000000000022
```

### Normal Q–Q of suicides\_no

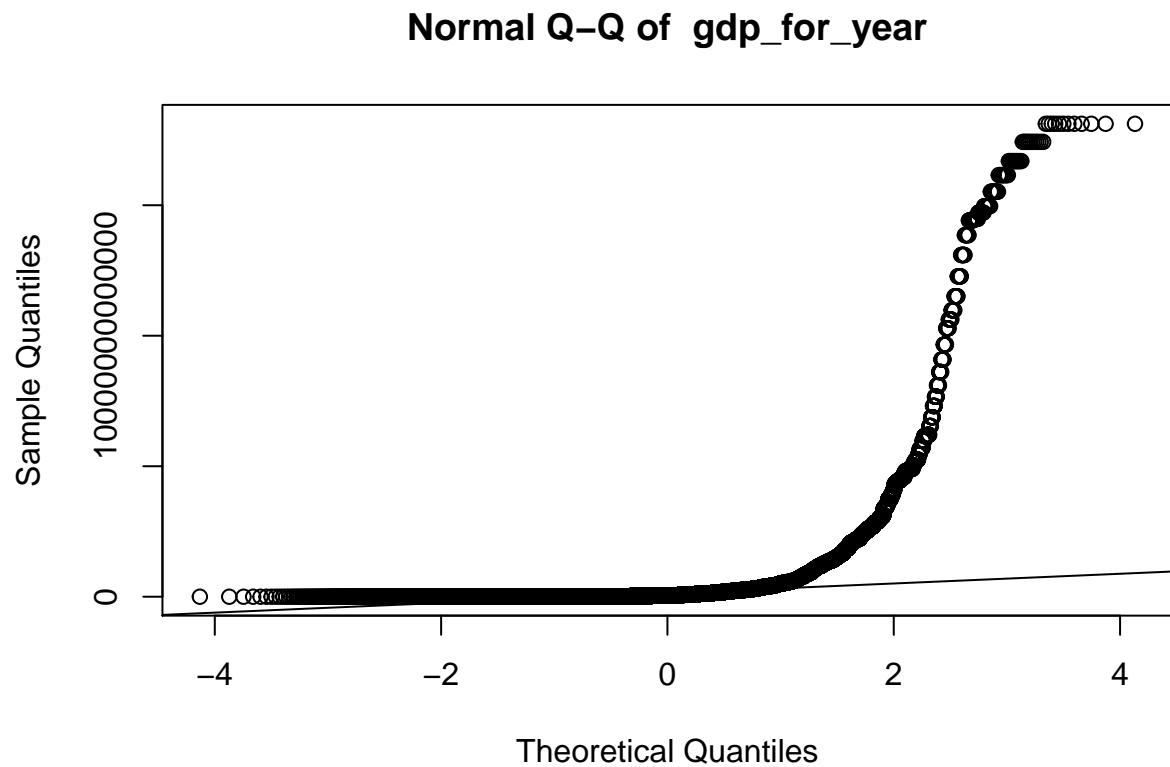


```
##  
## Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data: suicideData[, i]  
## D = 0.394, p-value < 0.0000000000000022
```

## Normal Q–Q of population

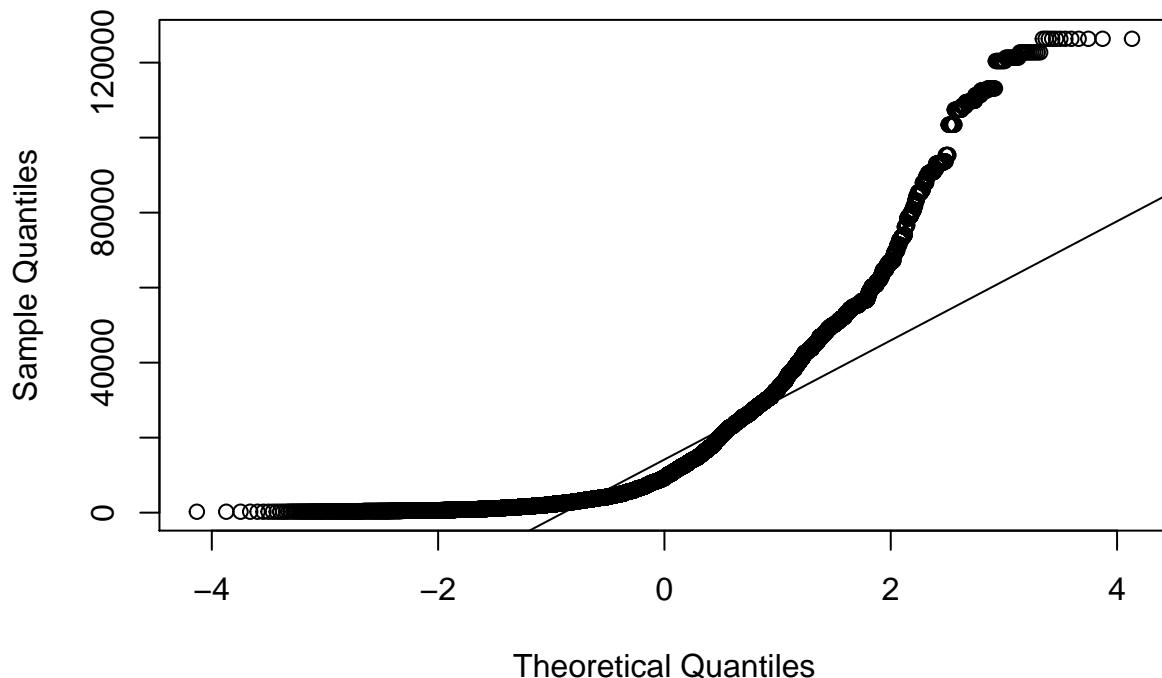


```
##  
## Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data: suicideData[, i]  
## D = 0.31863, p-value < 0.0000000000000022
```



```
##  
## Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data: suicideData[, i]  
## D = 0.37961, p-value < 0.00000000000000022
```

## Normal Q–Q of gdp\_per\_capita



```
##  
## Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data: suicideData[, i]  
## D = 0.18965, p-value < 0.0000000000000022
```

El gràfic “Normal Q–Q” mesura si es compleix amb una distribució normal.

Aquest gràfic ens indica en quina mesura les dades s'allunyen de la distribució normal que representa la línia recta.

Si el supòsit de normalitat és cert, llavors s'esperaria que els punts del diagrama es distribueixin a l'atzar a la proximitat de la línia recta de referència.

A primera vista, el gràfic en relació a year, sembla que segueixen una distribució normal, ja que els punts de la gràfica segueixen la línia de la distribució normal i únicament uns pocs dels extrems es desvien una mica.

A priori, aquest és un cas clar de distribució normal, ja que les observacions es troben a prop de la línia, l'allunyament és molt petit.

La resta de gràfics no són del tot clars, perquè si hi ha moltes observacions properes a la línia, però hi ha d'altres que estan més lluny.

Necessitarem comprovar mitjançant el p-value resultant de l'aplicació del test Lilliefors de Komolgorov-Smirnov, si aquest és inferior a 0.05.

En cas que així sigui, la variable seguirà una distribució normal. En el nostre cas d'estudi, les 5 variables quantitatives testejades tenen un p-valor molt petit, molt proper a 0, molt inferior a 0.05, pel que podem afirmar que les variables segueixen una distribució normal.

Si hi apliquem el shapiro test, a les primeres 5000 observacions (és el límit) ens surt el mateix valor del p-value.

```
for(i in quantitatives){  
  print(shapiro.test(suicideData[, i][0:5000]))  
}  
  
##  
## Shapiro-Wilk normality test  
##  
## data: suicideData[, i][0:5000]  
## W = 0.96211, p-value < 0.0000000000000022  
##  
##  
## Shapiro-Wilk normality test  
##  
## data: suicideData[, i][0:5000]  
## W = 0.49326, p-value < 0.0000000000000022  
##  
##  
## Shapiro-Wilk normality test  
##  
## data: suicideData[, i][0:5000]  
## W = 0.42422, p-value < 0.0000000000000022  
##  
##  
## Shapiro-Wilk normality test  
##  
## data: suicideData[, i][0:5000]  
## W = 0.55824, p-value < 0.0000000000000022  
##  
##  
## Shapiro-Wilk normality test  
##  
## data: suicideData[, i][0:5000]  
## W = 0.8366, p-value < 0.0000000000000022
```

### b) Estudi de la homogeneïtat de la variància

Per a realitzar l'estudi de la homocedasticitat o igualtat de variàncies dels grups a comparar, aplicarem un test de Levene, prova paramètrica que s'aplica quan les dades segueixen una distribució normal, com és el cas del dataset en estudi.

En cas que el p-value sigui superior o igual a 0.05, no rebutjarem la hipòtesi nul · la que les variàncies són iguals.

En cas contrari, rebutjarem la hipòtesi nul · la en favor de l'alternativa que indica que les variàncies dels nivells del factor analitzat són diferents.

Realitzarem l'estudi amb totes les variables categòriques, en funció del nombre de suïcidis. Les variables categòriques seran les explicatives de la variable explicada suicides\_no:

```
leveneTest(suicides_no~country, data = suicideData)  
  
## Levene's Test for Homogeneity of Variance (center = median)  
##              Df F value          Pr(>F)  
## group     100 159.04 < 0.0000000000000022 ***  
##             27719
```

```

## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
leveneTest(suicides_no~sex, data = suicideData)

## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value           Pr(>F)
## group      1  569.3 < 0.0000000000000022 ***
## 27818
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
leveneTest(suicides_no~age, data = suicideData)

## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value           Pr(>F)
## group      5  181.87 < 0.0000000000000022 ***
## 27814
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
leveneTest(suicides_no~generation, data = suicideData)

## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value           Pr(>F)
## group      5  101.88 < 0.0000000000000022 ***
## 27814
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Com podem veure, l'estadístic de contrast F ens surt molt alt, és molt superior a 2 en tots els tests. A més, podem comprovar que totes les variables són significatives, totes tenen els tres \*\*\*, i el seu p-value és molt inferior a 0.05, proper a 0, pel que rebutjarem la hipòtesi nul·la que les variàncies de les variables del grup són iguals.

També podríem haver aplicat el test de Kruskal Wallis, per a determinar si qualsevol de les diferències entre les mitjanes (no variàncies) són estadísticament significatives, i el test pairwise.wilcox.test per a determinar quins són els nivells dels factors amb mitjanes que són iguals o diferents. Ho aplicarem d'exemple a la variable generation:

```

kruskal.test(suicides_no~generation,data=suicideData)

##
##  Kruskal-Wallis rank sum test
##
## data: suicides_no by generation
## Kruskal-Wallis chi-squared = 2388.4, df = 5, p-value <
## 0.0000000000000022

pairwise.wilcox.test(suicideData$suicides_no,suicideData$generation,p.adj='bonferroni',exact=F)

##
##  Pairwise comparisons using Wilcoxon rank sum test
##
## data: suicideData$suicides_no and suicideData$generation
##
##          Boomers          G.I. Generation
## G.I. Generation < 0.0000000000000002 -
## Generation X    < 0.0000000000000002 0.000000000000089

```

```

## Generation Z      < 0.0000000000000002 < 0.0000000000000002
## Millenials       < 0.0000000000000002 < 0.0000000000000002
## Silent          < 0.0000000000000002 0.000000004279
##                  Generation X           Generation Z
## G.I. Generation -                   -
## Generation X     -                   -
## Generation Z     < 0.0000000000000002 -
## Millenials        < 0.0000000000000002 < 0.0000000000000002
## Silent           1                   < 0.0000000000000002
##                  Millenials
## G.I. Generation -
## Generation X     -
## Generation Z     -
## Millenials        -
## Silent           < 0.0000000000000002
##
## P value adjustment method: bonferroni

```

Efectivament, ens surt que hi ha una diferència significativa entre els diferents nivells del factor (p-value molt petit per a tots els nivells de la matriu).

### 6.3 Apartat 4.3 - Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d'anàlisi diferents.

a) Quines variables quantitatives tenen més influència en el nombre de suïcidis?

En primer lloc procedirem a realitzar una anàlisi de correlació per a tractar d'assolir el nostre objectiu d'estudi, on volem saber quines variables tenen més influència en el nombre de suïcidis, concretament, ho aplicarem a les variables de tipus quantitatius. El mètode que aplicarem serà el Coeficient de correlació de Pearson, ja que les dades segueixen una distribució normal, i sempre que sigui possible hem d'utilitzar proves paramètriques, ja que són més robustes.

Aquest mètode avalua el grau de relació entre dues variables quantitatives.

Hem de tenir en compte, que els coeficients de la correlació que obtindrem mitjançant l'aplicació del mètode del Coeficient de correlació de Pearson, prendran valors entre -1 i 1, on -1 i 1 indiquen que existeix una correlació perfecta entre les variables estudiades, i 0 indica cap relació. Hem de tenir en compte que existirà una relació positiva quan el signe del coeficient sigui positiu, és a dir, les dues variables es mouen en el mateix sentit, si una augmenta de valor, l'altra també, no necessàriament en la mateixa escala, i existirà una relació negativa o inversa quan el signe del coeficient sigui negatiu, és a dir, les dues variables es mouen en sentits opositos, si una augmenta, l'altra disminueix, no necessàriament en la mateixa escala.

En definitiva, quan els valors absoluts del coeficient siguin alts (propers a |1|), amb independència del seu signe, existirà correlació entre les variables.

La funció que ens permet calcular la correlació entre dues variables de tipus quantitatius és cor():

```
cor(x=suicideData$suicides_no,y=suicideData$year,method="pearson")
```

```
## [1] -0.004545958
```

Aplicarem el test, perquè ens indiqui si la correlació entre variables és diferent de zero i també l'interval de confiança al 95%.

```
for(i in quantit[-2]){
  print(cor.test(suicideData$suicides_no,suicideData[,i]))
}

##
## Pearson's product-moment correlation
##
## data: suicideData$suicides_no and suicideData[, i]
## t = -0.75822, df = 27818, p-value = 0.4483
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## -0.016296036 0.007205377
## sample estimates:
##           cor
## -0.004545958
##
##
## Pearson's product-moment correlation
##
## data: suicideData$suicides_no and suicideData[, i]
## t = 130.48, df = 27818, p-value < 0.0000000000000022
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
```

```

##  0.6088195 0.6233995
## sample estimates:
##      cor
##  0.6161623
##
##
## Pearson's product-moment correlation
##
## data: suicideData$suicides_no and suicideData[, i]
## t = 79.459, df = 27818, p-value < 0.0000000000000022
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4204700 0.4396249
## sample estimates:
##      cor
##  0.4300959
##
##
## Pearson's product-moment correlation
##
## data: suicideData$suicides_no and suicideData[, i]
## t = 10.248, df = 27818, p-value < 0.0000000000000022
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.04961456 0.07302807
## sample estimates:
##      cor
##  0.06132975

```

També podem obtenir directament una matriu de correlacions:

```

quantitatives_data<-suicideData[,quantitatives]
cor(quantitatives_data)

```

|                   | year         | suicides_no  | population | gdp_for_year |
|-------------------|--------------|--------------|------------|--------------|
| ## year           | 1.000000000  | -0.004545958 | 0.00885017 | 0.09452857   |
| ## suicides_no    | -0.004545958 | 1.000000000  | 0.61616227 | 0.43009585   |
| ## population     | 0.008850170  | 0.616162268  | 1.00000000 | 0.71069732   |
| ## gdp_for_year   | 0.094528572  | 0.430095852  | 0.71069732 | 1.00000000   |
| ## gdp_per_capita | 0.339134280  | 0.061329749  | 0.08150986 | 0.30340454   |
| ## gdp_per_capita | 0.33913428   |              |            |              |
| ## year           | 0.33913428   |              |            |              |
| ## suicides_no    | 0.06132975   |              |            |              |
| ## population     | 0.08150986   |              |            |              |
| ## gdp_for_year   | 0.30340454   |              |            |              |
| ## gdp_per_capita | 1.00000000   |              |            |              |

Per a respondre a la pregunta feta a aquest apartat, podem determinar que existeix una forta correlació positiva entre la variable suicide\_no i la variable population.

D'altra banda, si analitzem la correlació entre la resta de variables podem determinar que existeix una forta correlació positiva entre les variables population i gdp\_for\_year.

b) El nombre de suïcidis està influït pel PIB per capita, la població, el rang d'edat i la generació?

En segon lloc, per a poder respondre a la pregunta realitzada, aplicarem un model de regressió lineal múltiple, tenint en compte que ho aplicarem a un conjunt de dades amb regressors quantitatius i qualitatius.

L'acceptació de la hipòtesi nul · la implicaria que cap de les variables explicatives PIB per capita, poblacio, rang d'edat i generació expliquen la variable Y nombre de suïcidis, llavors, la situació ideal seria poder rebutjar la hipòtesi nul · la H0.

Primer definirem les categories de referència per a les noves variables ageR i generacioR i les incorporarem al conjunt de dades. Necessitem fer això per incloure els nivells dels factors a l'estudi, ja que estem treballant amb variables de tipus qualitatiu.

Establirem “25-34” com a categoria de referència per a la variable ageR i “Boomers” per a la variable generation, i afegirem les noves variables al conjunt de dades:

```
ageR<-relevel(suicideData$age,ref="25-34")
generationR<-relevel(suicideData$generation,ref="Boomers")
suicideData_regres<-data.frame(suicideData,ageR,generationR)
```

Comprovem que ho hem fet correctament:

També comprovem si s'han creat correctament els subnivells. Per exemple, el valor “Boomers” de la variable generation quedarà incorporat a l'intercepte B0, ja que es donarà quan a totes les Bi el valor per a aquesta variable generation sigui 0.

```
head(model.matrix(suicides_no~gdp_per_capita+ageR+generationR+population,data=suicideData_regres),6)

##   (Intercept) gdp_per_capita ageR15-24 ageR35-54 ageR5-14 ageR55-74
## 1           1          796       1       0       0       0
## 2           1          796       0       1       0       0
## 3           1          796       1       0       0       0
## 4           1          796       0       0       0       0
## 5           1          796       0       0       0       0
## 6           1          796       0       0       0       0
##   ageR75+ generationRG.I. Generation generationRGeneration X
## 1       0             0           0           1
## 2       0             0           0           0
## 3       0             0           0           1
## 4       1             1           0           0
## 5       0             0           0           0
## 6       1             1           0           0
##   generationRGeneration Z generationRMillenials generationRSilent
## 1                   0           0           0
## 2                   0           0           1
## 3                   0           0           0
## 4                   0           0           0
## 5                   0           0           0
## 6                   0           0           0
##   population
## 1     312900
## 2     308000
## 3     289700
## 4     21800
## 5     274300
## 6     35600
```

Ara apliquem la funció lm() per a obtenir el model de regressió:

```

model<-lm(suicides_no~gdp_per_capita+ageR+generationR+population,data=suicideData_regres)
summary(model)

##
## Call:
## lm(formula = suicides_no ~ gdp_per_capita + ageR + generationR +
##     population, data = suicideData_regres)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -3300.6  -106.9   -38.2    86.5 19554.5 
##
## Coefficients:
##                               Estimate Std. Error t value
## (Intercept)           15.880968782 16.860492743  0.942
## gdp_per_capita        0.000680139  0.000233101  2.918
## ageR15-24            -58.501186188 16.011108453 -3.654
## ageR35-54             73.448712248 17.587074667  4.176
## ageR5-14              -219.060178132 19.082751543 -11.479
## ageR55-74             51.061368339 25.264079269  2.021
## ageR75+               12.626965936 28.445847861  0.444
## generationRG.I. Generation 31.924192245 25.853570164  1.235
## generationRGeneration X -54.918147484 17.577399046 -3.124
## generationRGeneration Z -44.192110717 30.585174121 -1.445
## generationRMillenials  -60.370986464 21.988730376 -2.746
## generationRSilent      9.302509748 20.387428687  0.456
## population            0.000141366 0.000001097 128.900
##                               Pr(>|t|) 
## (Intercept)           0.346250
## gdp_per_capita        0.003528 ** 
## ageR15-24            0.000259 *** 
## ageR35-54             0.0000297 *** 
## ageR5-14              < 0.0000000000000002 ***
## ageR55-74             0.043278 *  
## ageR75+               0.657122
## generationRG.I. Generation 0.216912
## generationRGeneration X 0.001784 ** 
## generationRGeneration Z 0.148501
## generationRMillenials 0.006045 ** 
## generationRSilent     0.648187
## population            < 0.0000000000000002 ***
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 700.2 on 27807 degrees of freedom
## Multiple R-squared:  0.3977, Adjusted R-squared:  0.3974 
## F-statistic:  1530 on 12 and 27807 DF,  p-value: < 0.0000000000000022

```

Veiem que el coeficient de determinació és 0.3477, és un valor baix, únicament explica el 34,77% de la variabilitat total del nivell del nombre de suïcidis (suicide\_no) a partir de les variables explicatives.

D'altra banda, tenim que de les noves variables, la variable age\_R explica molt bé el nombre de suïcidis (mirar Estimate), únicament la categoria “ageR75”, la de més elevada edat, és la que no és significativa (p-valor superior a 0.05 i estadístic t inferior a 2).

També expliquen bé la variable Y tots els nivells de generació, encara que 3 d'aquests nivells o categories no són significatius, ja que si mirem el seu p-value, aquest és molt alt, a excepció de les categories “generationRGeneration X” i “generationRMillenials”, on el seu p-value sí és inferior a 0.05. analitzem el p-valor, veiem que és molt.

Les variables quantitatives gdp\_per\_capita i population, són significants pel model, encara que no expliquen massa bé la variable suicide\_no.

Podem dir que aquesta variable generation no és significativa pel model.

Respecte al coeficient R2 ajustat, el qual és similar a R2, amb la diferència de què l'ajustat penalitza la introducció de variables independents poc rellevants en el model, a l'hora d'explicar la variable dependent Y.

Així, sempre es complirà que R2ajustat<=R2.

Com hem mencionat, la variable generation no és significativa, a continuació provarem de fer el mateix estudi treient-la de l'anàlisi:

```
model<-lm(suicides_no~gdp_per_capita+ageR+population,data=suicideData_regres)
summary(model)
```

```
##
## Call:
## lm(formula = suicides_no ~ gdp_per_capita + ageR + population,
##      data = suicideData_regres)
##
## Residuals:
##    Min      1Q  Median      3Q     Max
## -3287.1  -103.7   -37.1    86.1 19562.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -23.901862830 11.070320362 -2.159 0.0309
## gdp_per_capita 0.000536112 0.000223072  2.403 0.0163
## ageR15-24    -74.176642412 14.536416187 -5.103 0.000000336837819
## ageR35-54     105.304912033 14.603820434  7.211 0.000000000000571
## ageR5-14      -231.323483193 14.561546969 -15.886 < 0.000000000000002
## ageR55-74     104.031064597 14.536591475  7.156 0.000000000000848
## ageR75+       74.444680791 14.600175021  5.099 0.000000343890831
## population    0.000141366 0.000001096 128.949 < 0.000000000000002
##
## (Intercept) *
## gdp_per_capita *
## ageR15-24 ***
## ageR35-54 ***
## ageR5-14 ***
## ageR55-74 ***
## ageR75+ ***
## population ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 700.3 on 27812 degrees of freedom
## Multiple R-squared: 0.3974, Adjusted R-squared: 0.3973
## F-statistic: 2620 on 7 and 27812 DF, p-value: < 0.0000000000000022
```

Com podem veure, el model ha millorat notablement, el coeficient de determinació ha varian mínimament,

però el p-value de les categories ha millorat, ara totes les categories de la variable age són significatives i expliquen molt bé la variable dependent suicide\_no.

Podem conculoure que la incorporació de la variable explicativa generation no s'ajusta de manera acceptable al nostre model de dades, però podem rebutjar la hipòtesi nul · la H0, ja que la resta de variables expliquen bé la variable explicada i són significatives, en definitiva, el nombre de suïcidis està influït pel PIB per capita, la població i el rang d'edat.

c) Existeixen diferències significatives en el nombre de suïcidis dels homes en relació a les dones?

Per a trobar una resposta a aquesta pregunta, aplicarem un mètode sobre la diferència de mitjanes en el cas de variàncies poblacionals desconegudes, però iguals.

Per interpretar de millor manera el diagrama de caixa, recuperarem la taula d'informació conjunta per intervals de valors del nombre de suïcidis, variable suicide\_no (perquè sigui més llegible).

```
table(intervalshs,suicideData$sex)
```

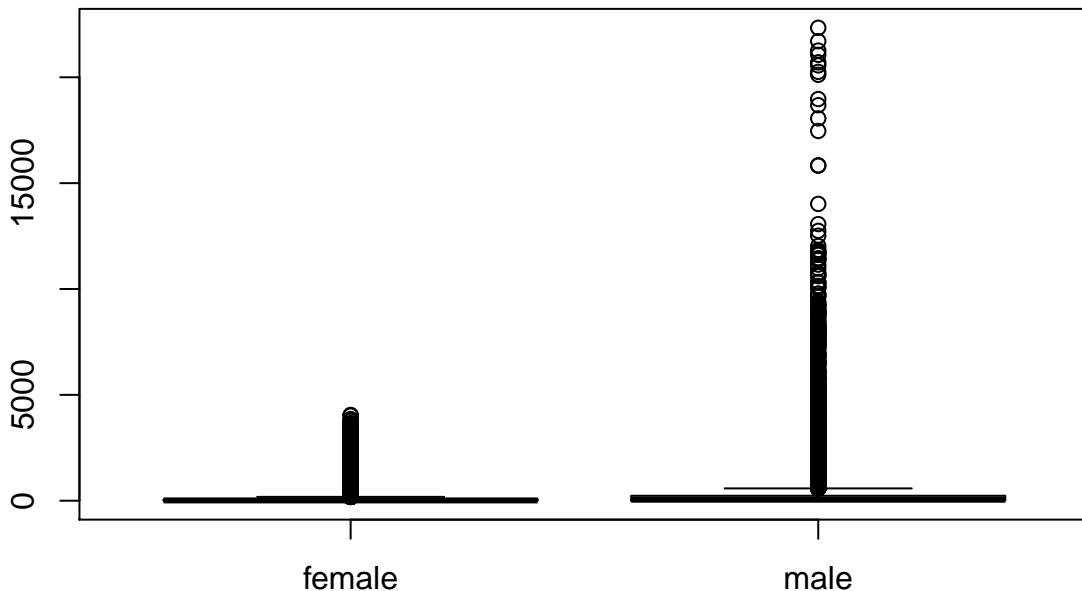
```
##  
## intervalshs      female   male  
## [0,1.49e+03]      13714 13218  
## (1.49e+03,2.98e+03]    164   317  
## (2.98e+03,4.47e+03]    32    142  
## (4.47e+03,5.96e+03]    0     84  
## (5.96e+03,7.45e+03]    0     40  
## (7.45e+03,8.94e+03]    0     52  
## (8.94e+03,1.04e+04]    0     21  
## (1.04e+04,1.19e+04]    0     17  
## (1.19e+04,1.34e+04]    0     4  
## (1.34e+04,1.49e+04]    0     1  
## (1.49e+04,1.64e+04]    0     2  
## (1.64e+04,1.79e+04]    0     1  
## (1.79e+04,1.94e+04]    0     3  
## (1.94e+04,2.08e+04]    0     4  
## (2.08e+04,2.23e+04]    0     4
```

Perquè encara es vegi millor, calcularem la mitjana de la variable suicide\_no per cada categoria de la variable sex:

```
mean.suicide_noVSsex<-aggregate(suicideData$suicides_no,by=list(suicideData$sex),mean,na.rm=TRUE)  
mean.suicide_noVSsex
```

```
##   Group.1      x  
## 1   female 112.1143  
## 2     male 373.0345  
boxplot(suicides_no~sex, data=suicideData,main="Box plot", col="grey",na.rm=TRUE)
```

## Box plot



A continuació expressem les hipòtesis:

- *Hipòtesi nul · la:*  $H_0: \mu_1 - \mu_2 = 0$  - *Hipòtesis alternativa:*  $H_1: \mu_1 - \mu_2 \neq 0$  (Bilateral) (on  $\mu_1$  i  $\mu_2$  són dues poblacions diferents que indiquen el valor de la mitjana d'homes i dones respectivament, poblacions on contrastarem la hipòtesi nul · la expressada).

Com ja tenim expressades les hipòtesis, determinarem el nivell de significació: alfa=0.05, ja que el nivell de confiança és del 95%.

Estem en un cas de contrast sobre la diferència de mitjanes en el cas de variàncies poblacionals desconegudes, però iguals. Hem d'analitzar si el nivell de satisfacció és igual en homes que en dones. Suposarem que totes dues poblacions es distribueixen normalment amb variàncies iguals i desconegudes. Haurem de contrastar la diferència de mitjanes per a saber si hi ha una diferència significativa o podem considerar que aquestes són iguals.

```
t.test(suicideData$suicides_no[suicideData$sex=="male"],
        suicideData$suicides_no[suicideData$sex=="female"], alternative = "two.sided")

##
##  Welch Two Sample t-test
##
## data:  suicideData$suicides_no[suicideData$sex == "male"] and suicideData$suicides_no[suicideData$se...
```

```
## 373.0345 112.1143
```

Com podem comprovar, hem obtingut com a resultat un estadístic de contrast prou alt i un p-value molt petit, el qual és inferior a 0.05 (nivell de significància fixat 5% amb una confiança del 95%), en definitiva, podem rebutjar la hipòtesi nul·la  $H_0$ , que les mitjanes poblacionals eren iguals pels homes que per a les dones, existeixen diferències significatives en el nombre de suïcidis dels homes en relació a les dones.

## 7 Apartat 5 - Representació dels resultats a partir de taules i gràfiques.

**NOTA IMPORTAT:** Aquest apartat ha sigut resolt durant l'elaboració de la pràctica, ja que s'han mostrat gràfiques i taules en cada exercici. Igualment, per a complementar l'apartat, a continuació mostrarem altres gràfiques amb taules amb informació de suport addicional, però cal indicar que el pes més gran de les gràfiques es trova a la resta d'apartats, i per no repetir informació, no seran afegits en aquest apartat.

Realitzarem l'anàlisi visual de la variable suicide\_no en funció del sexe i en funció del rang d'edat. El gràfic ha de permetre avaluar si hi ha interacció entre factors.

- 1: Agrupació de dades:

```
taula1<-suicideData%>%
  group_by(sex,age)%>%
  summarise(Mitjana=mean(suicides_no))
```

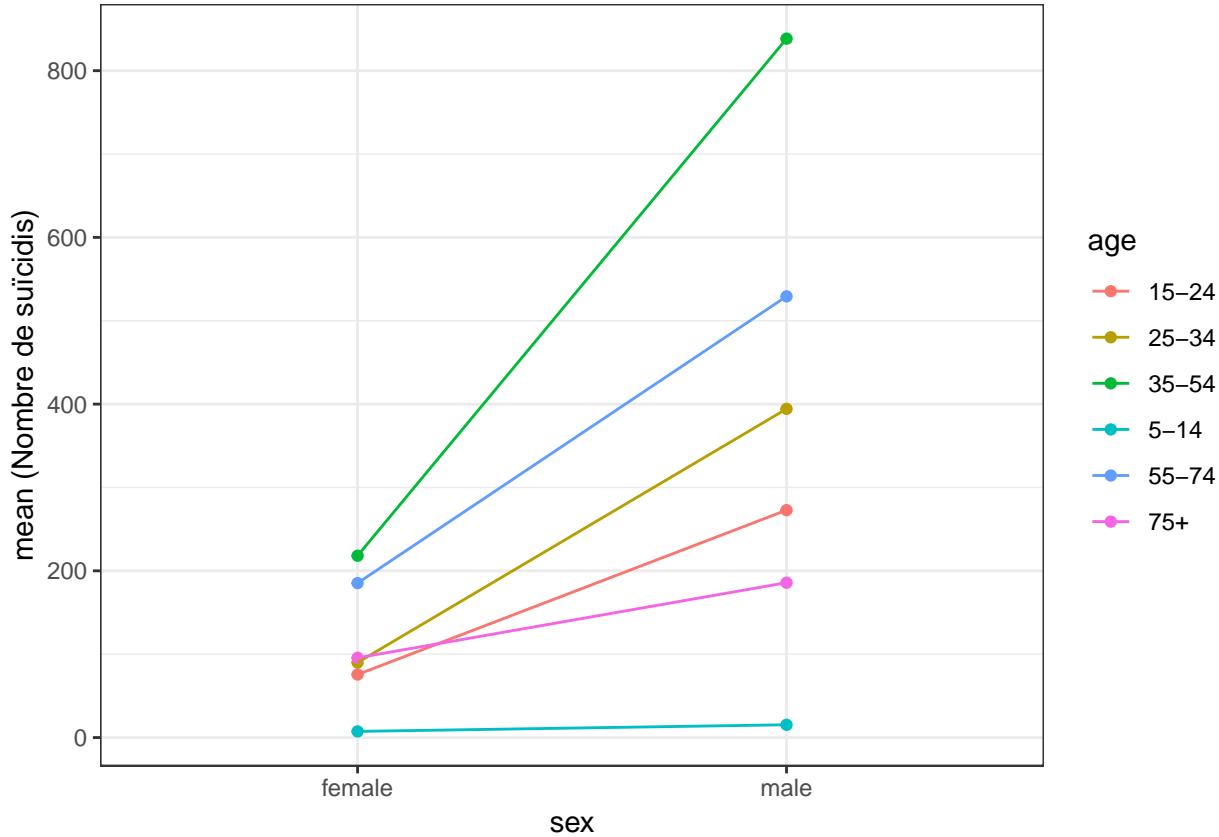
- 2: Mostrar conjunt de dades en format taula:

```
taula1
```

```
## # A tibble: 12 x 3
## # Groups:   sex [2]
##   sex     age   Mitjana
##   <fct>   <fct>   <dbl>
## 1 female  15-24    75.6
## 2 female  25-34    90.0
## 3 female  35-54   218.
## 4 female  5-14     7.37
## 5 female  55-74   185.
## 6 female  75+     95.6
## 7 male    15-24   273.
## 8 male    25-34   394.
## 9 male    35-54   838.
## 10 male   5-14    15.3
## 11 male   55-74   529.
## 12 male   75+     186.
```

- 3: Mostrar gràfic:

```
ggplot(data = taula1, aes(x=sex, y=Mitjana, colour=age,
                           group = age)) +
  stat_summary(fun.y = mean, geom = "point") +
  stat_summary(fun.y = mean, geom = "line") +
  labs(y = 'mean (Nombre de suicidis)') +
  theme_bw()
```



#### - 4: Conclusions gràfic 1:

En el gràfic podem veure que l'increment del nombre de suïcidis entre els 6 tipus rangs d'edat, no és proporcional pels dos tipus de sexe, sembla que hi ha interacció entre les variables sexe i edat.

Podem veure que entre les categories “35-54” i “55-74” per a les dones no hi molta diferència, en canvi sí hi ha una diferència molt potent en els homes quan el rang d'edat és igual a “35-54”.

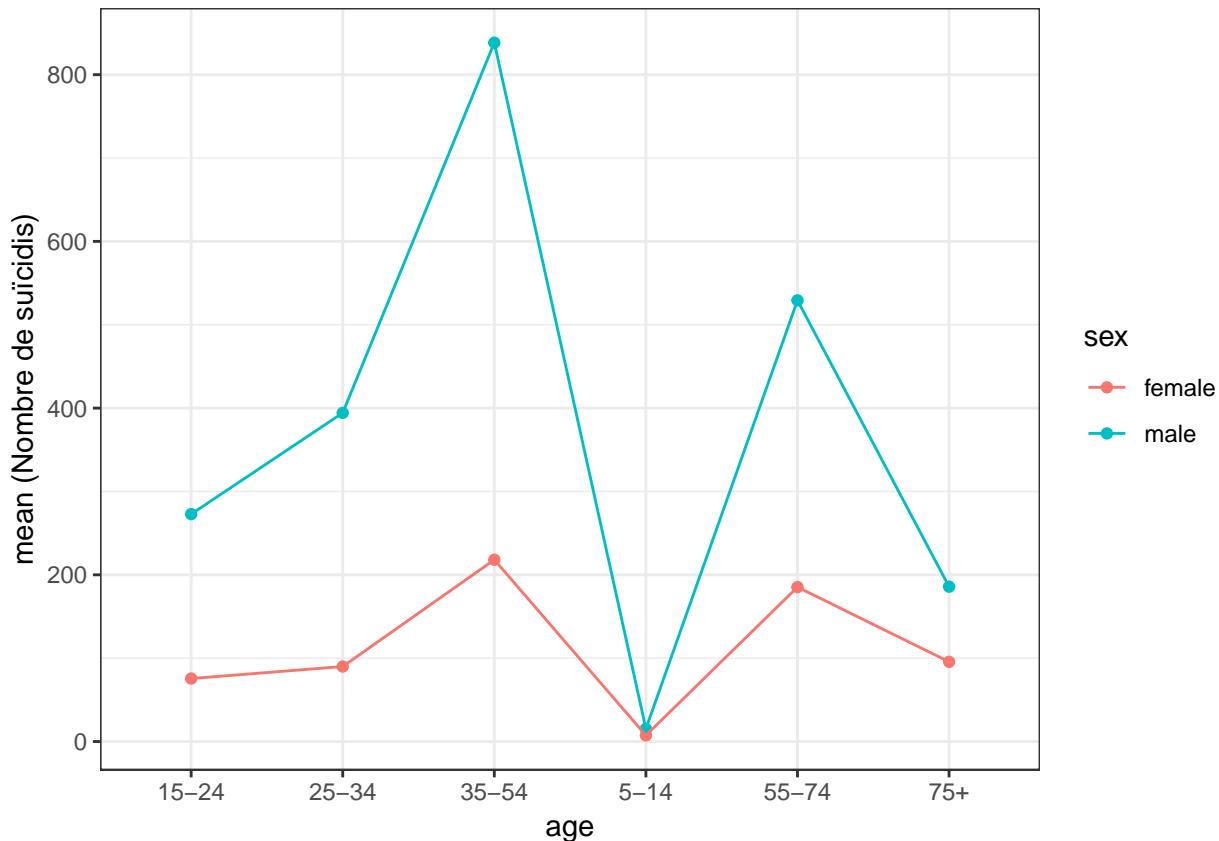
D'altra banda, en relació a la categoria de “5-14”, no hi ha molta diferència, en relació als altres nivells d'edat.

Per tot això, tenim que hi ha diferències significatives i hi ha interacció entre variables.

Totes aquestes conclusions, es veuen reforçades amb la taula calculada.

#### - 5: Ara construirem el gràfic agrupant per sexe:

```
ggplot(data = taula1, aes(x=age, y=Mitjana, colour=sex,
                           group = sex)) +
  stat_summary(fun.y = mean, geom = "point") +
  stat_summary(fun.y = mean, geom = "line") +
  labs(y = 'mean (Nombre de suïcidis)') +
  theme_bw()
```



- 6: *Conclusions gràfic 2:*

Com es pot observar, es veu una clara interacció, ja que les línies no són rectes, es veu fàcilment que hi ha desviació en el sexe, especialment notable al factor “Inc4”5-14“, pel que podem esperar al fet que hi hagi una interacció estadísticament significativa.

## **8 Apartat 6 - Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?**

Amb la finalitat de donar resposta a la pregunta plantejada a l'inici del desenvolupament d'aquesta pràctica, on volíem conèixer quins factors influïen en l'augment de les taxes de suïcidi i si aquests es trobaven en les variables del dataset, si aquestes variables són explicatives del nombre dels suïcidis, s'han realitzat tres tipus de proves estadístiques sobre el conjunt de dades escollit.

Per a poder realitzar una anàlisi exhaustiu del problema, s'han obtingut gràfics i taules com a suport addicional amb l'objectiu de facilitar l'anàlisi.

S'ha desenvolupat un *estudi de la normalitat* i de l'*homogeneïtat* de la variància, aplicant el *test de normalitat Lilliefors de Komolgorov-Smirnov*, obtenint com a resultat que les variables del conjunt de dades seguia una distribució normal, i hem empleat un *test de Levene* per a estudiar la homogeneïtat o igualtat de variàncies dels grups a comparar, obtenint com a resultat que les variàncies del grup eren significativament diferents.

Posteriorment, s'han realitzat 3 proves estadístiques més d'anàlisi de dades.

Les proves realitzades pretenien donar resposta a les següents preguntes:

- a) Quines variables quantitatives tenen més influència en el nombre de suïcidis?
- b) El nombre de suïcidis està influït pel PIB per capita, la població, el rang d'edat i la generació?
- c) Existeixen diferències significatives en el nombre de suïcidis dels homes en relació a les dones?

Per a donar resposta a la primera pregunta, hem decidit realitzar una *anàlisi de correlació*, utilitzant el *Coeficient de correlació de Pearson*, mitjançant el qual hem determinat que existia una forta correlació positiva entre la variable *suicide\_no* i la variable *population*.

Per a donar resposta a la segona pregunta, hem aplicat un *model de regressió lineal múltiple*, mitjançant el qual hem arribat a la conclusió de què el nombre de suïcidis està influït pel PIB per capita, la població i el rang d'edat, ja que aquestes explicaven bé la variable *suicides\_no* i han resultat ser significatives.

Per finalitzar, i per a donar resposta a la tercera pregunta, s'ha decidit aplicar un *mètode sobre la diferència de mitjanes en el cas de variàncies poblacionals desconegudes*, però iguals, obtenint com a conclusió de què existeixen diferències significatives en el nombre de suïcidis dels homes en relació a les dones.

Finalment, s'ha realitzat un petit anàlisi visual de la variable *suicide\_no* en funció del sexe i en funció del rang d'edat, on s'ha pogut veure que hi havia interacció entre les variables.

**9 Apartat 7 - Codi: Cal adjuntar el codi, preferiblement en R, amb el que s'ha realitzat la neteja, anàlisi i representació de les dades. Si ho preferiu, també podeu treballar en Python.**

Facilitat amb el lliurament de la pràctica.

## 10 Recursos

Calvo M., Subirats L., Pérez D. (2019). Introducción a la limpieza y análisis de los datos. Editorial UOC.

Megan Squire (2015). Clean Data. Packt Publishing Ltd.

Jiawei Han, Micheine Kamber, Jian Pei (2012). Data mining: concepts and techniques. Morgan Kaufmann.

Jason W. Osborne (2010). Data Cleaning Basics: Best Practices in Dealing with Extreme Scores. Newborn and Infant Nursing Reviews; 10 (1): pp. 1527-3369.

Peter Dalgaard (2008). Introductory statistics with R. Springer Science & Business Media.

Wes McKinney (2012). Python for Data Analysis. O'Reilley Media, Inc.

Tutorial de Github <https://guides.github.com/activities/hello-world>.

| Contribucions             | Signa  |
|---------------------------|--------|
| Recerca prèvia            | S.Q.G. |
| Redacció de les respostes | S.Q.G. |
| Desenvolupament codi      | S.Q.G. |