M2.951 - TIPOLOGIA I CICLE DE VIDA DE LES DADES

Pràctica 2 - World Health Indicators Dataset

Anna de la Torre Suñe, Xavier Ventura de los Ojos

9/6/2020

Contents

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

El dataset que analitzarem és el **World Health Indicators** que vàrem obtenir de les Nacions Unides i del Banc Mundial a la pràctica 1 - web scraping. El WHI està format per 572 files i 22 columnes i conté dades relacionades amb l'àmbit de la salut pública i dades socioeconòmiques referents a 194 països i regions del món dels anys 2005, 2010 i 2019 (*).

El dataset té la següent estructura:

- country_code – Codi ISO3 de país
- country – Nom del país
- year – Any
- population – Població total
- population_grow – Creixement de la població (% anual)
- population_under_14 – Població d'edats entre els 0 – 14 anys (% de la població total)
- population_above_65 – Població d'edats superior als 65 anys (% de la població total)
- gdp - Producte Interior Brut (milions de dòlars actuals) – GDP per les seves sigles en anglès
- gdp_growth_rate – Taxa de creixement del Producte Interior Brut (% anual a preus constants de 2010)
- unemployment – Atur (% de població en edat laboral)
- education_gov_expenditure – Despesa governamental en educació (% GDP)
- health_expenditure – Despesa en salut (% GDP)
- life_expectancy_fem – Esperança de vida, dones (anys)
- life_expectancy_male – Esperança de vida, homes (anys)
- non_commun_disease_death – Morts per malalties no transmissibles (% del total)
- commun_disease_death – Morts per malalties transmissibles i condicions d'embaràs, prenatals i nutricionals (% del total)
- tuberculosis – Incidència de tuberculosi (per cada 100.000 habitants)
- hiv – Prevalença de VIH (en % de població d'edats entre els 15 i 49 anys)
- infant_mortality – Mortalitat infantil (per cada 1000 naixements amb vida)
- undernourishment – Prevalença de desnutrició (% població)
- hospital_beds – Llits d'hospital (per cada 1000 habitants)
- physicians – metges (per cada 1000 habitants)

(*) En el cas del Banc Mundial, s'ha hagut d'assimilar les dades de 2018 a les de 2019 a l'espera que es publiquin les dades de 2019.

Considerem que aquest és un dataset interessant que ens pot ajudar a entendre la situació sociosanitària de les diferents regions del món, determinar vulnerabilitats o zones de risc i per tant identificar on es fa més necessari actuar. S'analitzarà la correlació entre indicadors socioeconòmics i indicadors sanitaris, amb especial èmfasi en aquelles variables que afecten l'esperança de vida d'homes i dones. S'estudiarà amb detall les diferències en l'esperança de vida segons diferents rangs de despesa sanitària pública. Finalment s'implementarà un model de regressió que permeti predir l'esperança de vida en funció de diferents variables correlacionades.

2 Integració i selecció de les dades d'interès a analitzar.

En aquest cas no cal realitzar un procés d'integració de dades, ja que ja es va dur a terme en la Pràctica 1, de manera que partim d'un dataset que ja integra dades de diferents fonts.

Respecte a la selecció de dades, d'una banda s'eliminaran les columnes non_commun_disease_death, commun_disease_death, hospital_beds i undernourishment a causa de la important proporció de dades buides que contenen (veure secció 3.1), de manera que els tests i models que s'apliquin se sustentin en dades de major qualitat. I d'altra banda, gràcies als testos de correlació que s'aplicaran en la secció 4.3. observarem que necessitarem crear quatre noves variables calculades, que mostrin les variables gdp, gdp_growth_rate, education_gov_expenditure i health_expenditure per càpita. Si no es prenen els valors per càpita, s'observen correlacions molt febles entre les dades d'interès relacionades amb la salut dels habitants i les dades globals de gdp, gdp_growth_rate, education_gov_expenditure i health_expenditure pel conjunt del país.

Procedim ara a llegir les dades del fitxer "world_health_indicators.csv" i fem una anàlisi descriptiva de les variables d'interès.

```
# Carreguem el fitxer de dades world_health_indicators.csv
health_data <- read.csv('world_health_indicators.csv', stringsAsFactors = F)
health_rows=dim(health_data)[1]
```

Visualitzem les primeres files del dataset

country_code	country	year	population	population_grow	population_under_14	population_above_65
ARE	United Arab Emirates (the)	2005	4588225	12.02	18.34	0.89
ARE	United Arab Emirates (the)	2010	8549988	7.69	13.16	0.69
ARE	United Arab Emirates (the)	2019	9630959	1.50	14.60	1.09
AFG	Afghanistan	2005	25654277	3.68	47.86	2.23
AFG	Afghanistan	2010	29185507	2.75	48.18	2.33
AFG	Afghanistan	2019	37172386	2.38	43.09	2.58
ATG	Antigua and Barbuda	2005	81465	1.40	26.51	6.98
ATG	Antigua and Barbuda	2010	88028	1.47	24.34	7.36
ATG	Antigua and Barbuda	2019	96286	0.90	22.08	8.80

country_code	year	gdp	gdp_growth_rate	unemployment	education_gov_expenditure
ARE	2005	182978	4.9	3.1	NA
ARE	2010	289787	1.6	3.7	NA
ARE	2019	382575	0.8	1.7	NA
AFG	2005	6622	9.9	8.5	NA
AFG	2010	16078	3.2	7.8	3.5
AFG	2019	21993	2.5	8.8	3.9
ATG	2005	1022	6.4	8.4	3.4
ATG	2010	1152	-7.2	NA	2.5
ATG	2019	1510	3.0	NA	NA

country_code	year	health_expenditure	life_expectancy_fem	life_expectancy_male	non_commun_disease_death	commun_disease_death
ARE	2005	2.3	76.82	74.60	NA	NA
ARE	2010	3.9	77.77	75.62	70.4	7.5
ARE	2019	3.5	79.16	77.13	NA	NA
AFG	2005	9.9	59.63	57.04	NA	NA
AFG	2010	8.6	62.46	59.68	39.2	50.2
AFG	2019	10.2	66.03	63.05	NA	NA
ATG	2005	4.5	76.34	73.58	NA	NA
ATG	2010	5.4	77.12	74.44	79.1	13.2
ATG	2019	4.3	77.98	75.72	NA	NA

country_code	year	tuberculosis	hiv	infant_mortality	undernourishment	hospital_beds	physicians
ARE	2005	2.6	NA	9.0	4.1	2.2	1.5
ARE	2010	1.8	NA	7.7	5.9	1.9	1.5
ARE	2019	1.0	NA	6.5	NA	NA	2.4
AFG	2005	189.0	0.1	84.6	33.2	0.4	0.2
AFG	2010	189.0	0.1	72.2	22.1	0.4	0.2
AFG	2019	189.0	0.1	60.1	NA	NA	0.3
ATG	2005	8.5	NA	11.5	NA	2.4	NA
ATG	2010	9.1	NA	9.0	NA	2.2	1.6
ATG	2019	6.0	NA	6.8	NA	NA	2.8

```r
# Verifiquem l'estructura del joc de dades
str(health_data)
```

```
## 'data.frame':    572 obs. of  22 variables:
##  $ country_code             : chr  "ARE" "ARE" "ARE" "AFG" ...
##  $ country                  : chr  "United Arab Emirates (the)" "United Arab Emirates (the)" "United
##  $ year                     : int  2005 2010 2019 2005 2010 2019 2005 2010 2019 2005 ...
##  $ population               : num  4588225 8549988 9630959 25654277 29185507 ...
##  $ population_grow          : num  12.02 7.69 1.5 3.68 2.75 ...
##  $ population_under_14      : num  18.3 13.2 14.6 47.9 48.2 ...
##  $ population_above_65      : num  0.89 0.69 1.09 2.23 2.33 2.58 6.98 7.36 8.8 8.5 ...
##  $ gdp                      : int  182978 289787 382575 6622 16078 21993 1022 1152 1510 8052 ...
##  $ gdp_growth_rate          : num  4.9 1.6 0.8 9.9 3.2 2.5 6.4 -7.2 3 5.5 ...
##  $ unemployment             : num  3.1 3.7 1.7 8.5 7.8 8.8 8.4 NA NA 17.5 ...
##  $ education_gov_expenditure: num  NA NA NA NA 3.5 3.9 3.4 2.5 NA 3.2 ...
##  $ health_expenditure       : num  2.3 3.9 3.5 9.9 8.6 10.2 4.5 5.4 4.3 6.3 ...
##  $ life_expectancy_fem      : num  76.8 77.8 79.2 59.6 62.5 ...
##  $ life_expectancy_male     : num  74.6 75.6 77.1 57 59.7 ...
##  $ non_commun_disease_death : num  NA 70.4 NA NA 39.2 NA NA 79.1 NA NA ...
##  $ commun_disease_death     : num  NA 7.5 NA NA 50.2 NA NA 13.2 NA NA ...
##  $ tuberculosis             : num  2.6 1.8 1 189 189 189 8.5 9.1 6 19 ...
##  $ hiv                      : num  NA NA NA 0.1 0.1 0.1 NA NA NA NA ...
##  $ infant_mortality         : num  9 7.7 6.5 84.6 72.2 60.1 11.5 9 6.8 21.2 ...
##  $ undernourishment         : num  4.1 5.9 NA 33.2 22.1 NA NA NA NA 10.9 ...
##  $ hospital_beds            : num  2.2 1.9 NA 0.4 0.4 NA 2.4 2.2 NA 3.1 ...
##  $ physicians               : num  1.5 1.5 2.4 0.2 0.2 0.3 NA 1.6 2.8 NA ...
```

```r
# Convertim la variable gdp de "integer" a numèrica:
health_data$gdp <- as.numeric(health_data$gdp)

# Anàlisi estadístic descriptiu de health_data
summary(health_data)
```

```
##  country_code         country                year        population
##  Length:572         Length:572         Min.   :2005   Min.   :1.978e+04
##  Class :character   Class :character   1st Qu.:2005   1st Qu.:2.073e+06
##  Mode  :character   Mode  :character   Median :2010   Median :7.655e+06
##                                        Mean   :2011   Mean   :3.660e+07
##                                        3rd Qu.:2019   3rd Qu.:2.530e+07
##                                        Max.   :2019   Max.   :1.393e+09
##
```

```
##  population_grow  population_under_14 population_above_65      gdp
##  Min.   :-3.980   Min.   :11.90       Min.   : 0.690      Min.   :     112
##  1st Qu.: 0.490   1st Qu.:18.88       1st Qu.: 3.310      1st Qu.:    6296
##  Median : 1.350   Median :28.16       Median : 5.720      Median :   22979
##  Mean   : 1.462   Mean   :29.09       Mean   : 7.863      Mean   :  337383
##  3rd Qu.: 2.370   3rd Qu.:38.85       3rd Qu.:11.980      3rd Qu.:  147788
##  Max.   :13.870   Max.   :50.04       Max.   :27.580      Max.   :19485394
##  NA's   :3        NA's   :3           NA's   :3           NA's   :1
##  gdp_growth_rate    unemployment   education_gov_expenditure health_expenditure
##  Min.   :-14.000   Min.   : 0.10   Min.   : 1.000            Min.   : 1.300
##  1st Qu.:  2.100   1st Qu.: 4.20   1st Qu.: 3.300            1st Qu.: 4.325
##  Median :  3.900   Median : 6.65   Median : 4.400            Median : 6.000
##  Mean   :  4.365   Mean   : 8.24   Mean   : 4.602            Mean   : 6.285
##  3rd Qu.:  6.450   3rd Qu.:10.70   3rd Qu.: 5.500            3rd Qu.: 8.100
##  Max.   : 64.000   Max.   :37.20   Max.   :12.800            Max.   :17.100
##  NA's   :1         NA's   :10      NA's   :113               NA's   :34
##  life_expectancy_fem life_expectancy_male non_commun_disease_death
##  Min.   :44.60       Min.   :40.57        Min.   :14.90
##  1st Qu.:66.82       1st Qu.:62.72        1st Qu.:42.35
##  Median :75.46       Median :69.32        Median :73.00
##  Mean   :72.68       Mean   :67.83        Mean   :66.07
##  3rd Qu.:79.40       3rd Qu.:74.33        3rd Qu.:87.05
##  Max.   :87.70       Max.   :82.30        Max.   :95.10
##  NA's   :3           NA's   :3            NA's   :389
##  commun_disease_death  tuberculosis         hiv         infant_mortality
##  Min.   : 1.50        Min.   :   0.0   Min.   : 0.100   Min.   :  1.60
##  1st Qu.: 6.60        1st Qu.:  15.0   1st Qu.: 0.100   1st Qu.:  8.30
##  Median :13.50        Median :  51.0   Median : 0.400   Median : 19.10
##  Mean   :24.79        Mean   : 131.1   Mean   : 1.985   Mean   : 30.59
##  3rd Qu.:46.60        3rd Qu.: 176.0   3rd Qu.: 1.500   3rd Qu.: 51.10
##  Max.   :71.50        Max.   :1280.0   Max.   :27.400   Max.   :141.90
##  NA's   :389          NA's   :3        NA's   :157      NA's   :3
##  undernourishment hospital_beds      physicians
##  Min.   : 2.50    Min.   : 0.100   Min.   :0.000
##  1st Qu.: 2.50    1st Qu.: 1.600   1st Qu.:0.200
##  Median : 7.40    Median : 3.000   Median :1.100
##  Mean   :12.48    Mean   : 3.502   Mean   :1.491
##  3rd Qu.:18.20    3rd Qu.: 5.100   3rd Qu.:2.400
##  Max.   :57.10    Max.   :14.300   Max.   :8.200
##  NA's   :247      NA's   :319      NA's   :102
```

3. Neteja de les dades.

3.1. Les dades contenen zeros o elements buits? Com gestionaries aquests casos?

Analitzem els casos de zeros

Mostrem les files amb alguna columna amb valor 0.

```
columnes <- colSums(health_data == 0,na.rm=TRUE) > 0
columnes[c('country','year')] <- TRUE

zeros <- health_data[rowSums(health_data == 0,na.rm = TRUE) > 0,columnes]
```

Table 1: Valors zero

country	year	population_grow	gdp_growth_rate	tuberculosis	physicians
Burkina Faso	2010	3.02	8.4	58	**0**
Burundi	2005	3.22	0.9	195	**0**
Burundi	2010	3.26	5.1	144	**0**
Burundi	2019	3.17	**0**	111	**0**
Benin	2005	2.95	1.7	73	**0**
Central African Republic (the)	2010	1.13	3.6	540	**0**
Spain	2010	0.46	**0**	17	3.7
Ethiopia	2005	2.8	11.8	341	**0**
Ethiopia	2010	2.78	12.6	268	**0**
Saint Kitts and Nevis	2005	1.14	8.8	**0**	1.1
Liberia	2005	2.6	5.3	267	**0**
Liberia	2010	3.59	7.3	293	**0**
Liberia	2019	2.45	2.5	308	**0**
Lesotho	2005	-0.49	2.7	1280	**0**
Malawi	2005	2.6	3.3	398	**0**
Malawi	2010	2.87	5.5	310	**0**
Malawi	2019	2.64	5.1	181	**0**
Mozambique	2005	2.89	8.7	523	**0**
Mozambique	2010	2.74	6.7	545	**0**
Niger (the)	2005	3.73	7.4	143	**0**
Niger (the)	2010	3.84	8.4	113	**0**
Niger (the)	2019	3.82	4.9	87	**0**
Papua New Guinea	2005	2.19	3.9	432	**0**
Poland	2019	**0**	4.8	16	2.4
Qatar	2019	2.07	1.6	31	**0**
Sierra Leone	2005	3.82	4.5	316	**0**
Sierra Leone	2010	2.25	5.3	317	**0**
Sierra Leone	2019	2.14	3.8	298	**0**
Somalia	2010	2.75	2.6	286	**0**
Somalia	2019	2.83	1.8	262	**0**
Chad	2005	3.65	7.9	151	**0**
Chad	2010	3.33	15	147	**0**
Chad	2019	3.02	-3.1	142	**0**
Togo	2005	2.6	-4.7	71	**0**
Togo	2019	2.45	4.4	36	**0**
Tanzania, United Republic of	2005	2.82	7.4	510	**0**
Tanzania, United Republic of	2019	2.98	7	253	**0**

Hi ha quatre casuístiques diferents:

- Un valor 0 al creixement de la població de Polònia l'any 2019.

```
health_data[health_data$country=="Poland",c('country','year',"population","population_grow")]
```

	country	year	population	population_grow
417	Poland	2005	38165445	-0.04
418	Poland	2010	38042794	-0.29
419	Poland	2019	37974750	0.00

Observem que podria ser un valor correcte tenint en compte que la variació de la població entre els anys 2010 i 2019 a Polonia ha estat del -0.1788617%.

- Paisos amb GDP Growth Rate igual a 0

```
cgdp0 <- health_data[health_data$gdp_growth_rate == 0,"country"]
health_data[health_data$country %in% cgdp0,c('country','year',"gdp","gdp_growth_rate")]
```

	country	year	gdp	gdp_growth_rate
55	Burundi	2005	1117	0.9
56	Burundi	2010	2032	5.1
57	Burundi	2019	3155	0.0
158	Spain	2005	1157248	3.7
159	Spain	2010	1431617	0.0
160	Spain	2019	1314314	3.0

Consultant altres fonts (tradingeconomics.com), corregim el valor del gdp_growth de Burundi per a l'any 2019 que va ser de 3.3.

```
health_data[health_data$country == "Burundi" & health_data$year == 2019,]$gdp_growth_rate <- 3.3
```

En el cas d'Espanya, el valor 0 per l'any 2010 és correcte.

- Paisos amb 0 casos de tuberculosi per 100.000 habitants.

```
ctry <- health_data[which(health_data$tuberculosis == 0),"country"]
health_data[health_data$country %in% ctry,c("country","year","population","tuberculosis")]
```

	country	year	population	tuberculosis
275	Saint Kitts and Nevis	2005	46857	0

Es tracta d'un terrori tant petit que és possible que aquest valor sigui correcte.

- Països amb 0.0 metges per cada 1000 habitants

```
ctry <- health_data[which(health_data$physicians == 0),"country"]
health_data[health_data$country %in%
        ctry,c('country','year',"health_expenditure","population","physicians")]
```

	country	year	health_expenditure	population	physicians
46	Burkina Faso	2005	4.4	13421930	0.1
47	Burkina Faso	2010	5.9	15605217	0.0
48	Burkina Faso	2019	6.8	19751535	0.1
55	Burundi	2005	8.4	7364862	0.0
56	Burundi	2010	11.3	8675602	0.0
57	Burundi	2019	7.7	11175378	0.0
58	Benin	2005	4.0	7982225	0.0
59	Benin	2010	4.1	9199259	0.1
60	Benin	2019	3.9	11485048	0.2
91	Central African Republic (the)	2005	4.8	4038382	0.1
92	Central African Republic (the)	2010	3.7	4386768	0.0
93	Central African Republic (the)	2019	4.3	4666377	0.1
161	Ethiopia	2005	4.1	76346311	0.0
162	Ethiopia	2010	5.5	87639964	0.0
163	Ethiopia	2019	4.0	109224559	0.1
300	Liberia	2005	8.8	3218116	0.0
301	Liberia	2010	8.8	3891356	0.0
302	Liberia	2019	9.6	4818977	0.0
303	Lesotho	2005	5.1	1996114	0.0
304	Lesotho	2010	7.7	1995581	0.1
305	Lesotho	2019	8.1	2108132	NA
357	Malawi	2005	6.1	12625952	0.0
358	Malawi	2010	7.2	14539612	0.0
359	Malawi	2019	9.8	18143315	0.0
366	Mozambique	2005	6.4	20493925	0.0
367	Mozambique	2010	5.1	23531574	0.0
368	Mozambique	2019	5.1	29495962	0.1
375	Niger (the)	2005	7.5	13624467	0.0
376	Niger (the)	2010	6.5	16464025	0.0
377	Niger (the)	2019	6.2	22442948	0.0
408	Papua New Guinea	2005	2.5	6494903	0.0
409	Papua New Guinea	2010	2.1	7310507	0.1
410	Papua New Guinea	2019	2.0	8606316	NA
433	Qatar	2005	2.6	865416	2.5
434	Qatar	2010	1.8	1856327	3.9
435	Qatar	2019	3.1	2781677	0.0
472	Sierra Leone	2005	11.0	5645624	0.0
473	Sierra Leone	2010	10.9	6415634	0.0
474	Sierra Leone	2019	16.5	7650154	0.0
478	Somalia	2005	NA	10446863	NA
479	Somalia	2010	NA	12043883	0.0
480	Somalia	2019	NA	15008154	0.0
498	Chad	2005	4.8	10096633	0.0
499	Chad	2010	4.1	11952136	0.0
500	Chad	2019	4.5	15477751	0.0
501	Togo	2005	3.9	5611640	0.0
502	Togo	2010	5.9	6421679	0.2
503	Togo	2019	6.6	7889094	0.0
528	Tanzania, United Republic of	2005	6.4	38450320	0.0
529	Tanzania, United Republic of	2010	5.3	44346525	NA
530	Tanzania, United Republic of	2019	4.1	56318348	0.0

Segons la WHO, *Available statistics show that over 45% of WHO Member States report to have less than 1 physician per 1000 population* això vol dir que aquests valors 0 poden ser correctes.

Analitzem els casos amb elements buits.

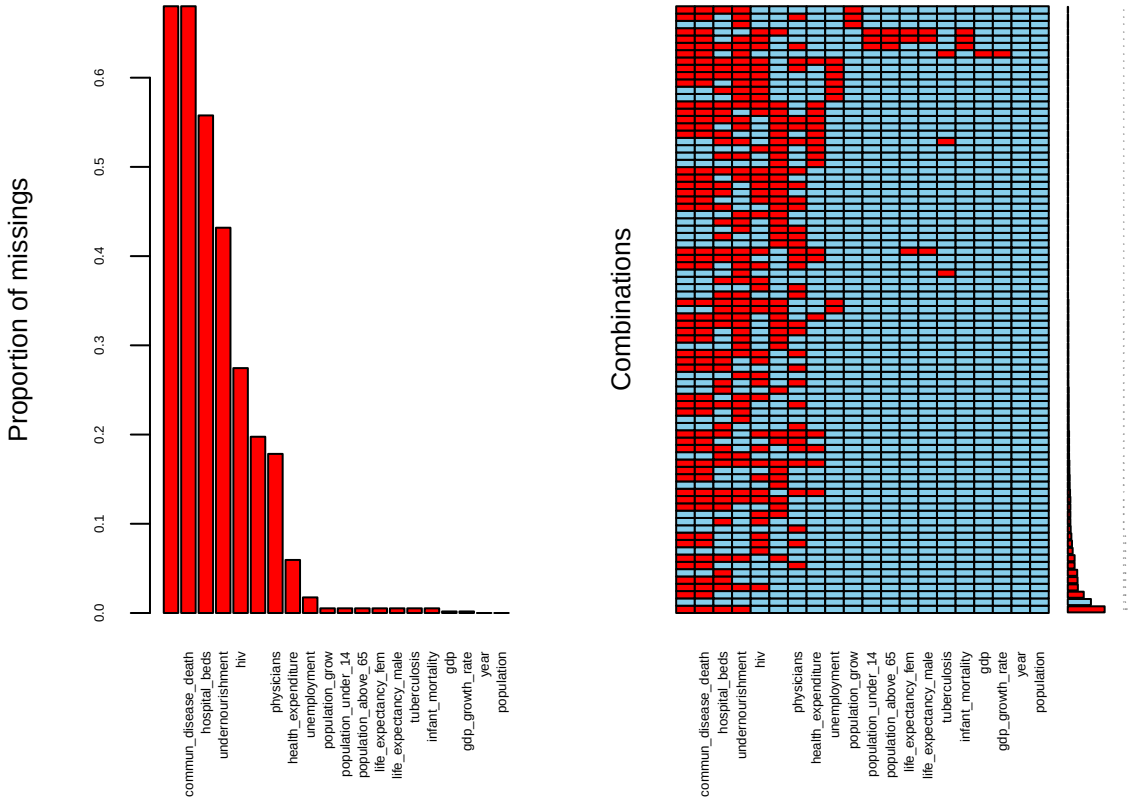
Sobre un total de 572 files, els casos amb valors buits (NA) són:

```
# Estadístiques de valors buits
vb <- colSums(is.na(health_data))
kable(vb[vb>0],col.names="Valors buits")
```

	Valors buits
population_grow	3
population_under_14	3
population_above_65	3
gdp	1
gdp_growth_rate	1
unemployment	10
education_gov_expenditure	113
health_expenditure	34
life_expectancy_fem	3
life_expectancy_male	3
non_commun_disease_death	389
commun_disease_death	389
tuberculosis	3
hiv	157
infant_mortality	3
undernourishment	247
hospital_beds	319
physicians	102

Mostrem la proporció de valors buits de les diferents variables i les seves possibles combinacions pel dataset complert (anys 2005, 2010 i 2019).

```
health_miss <- health_data[, !(colnames(health_data) %in% c("country_code", "country"))]

aggr(health_miss, numbers=T, sortVars=T, labels=names(health_miss), cex.axis=0.4,
     cex.lab=0.8, cex.numbers=0.1, prop=c(T,F))
```

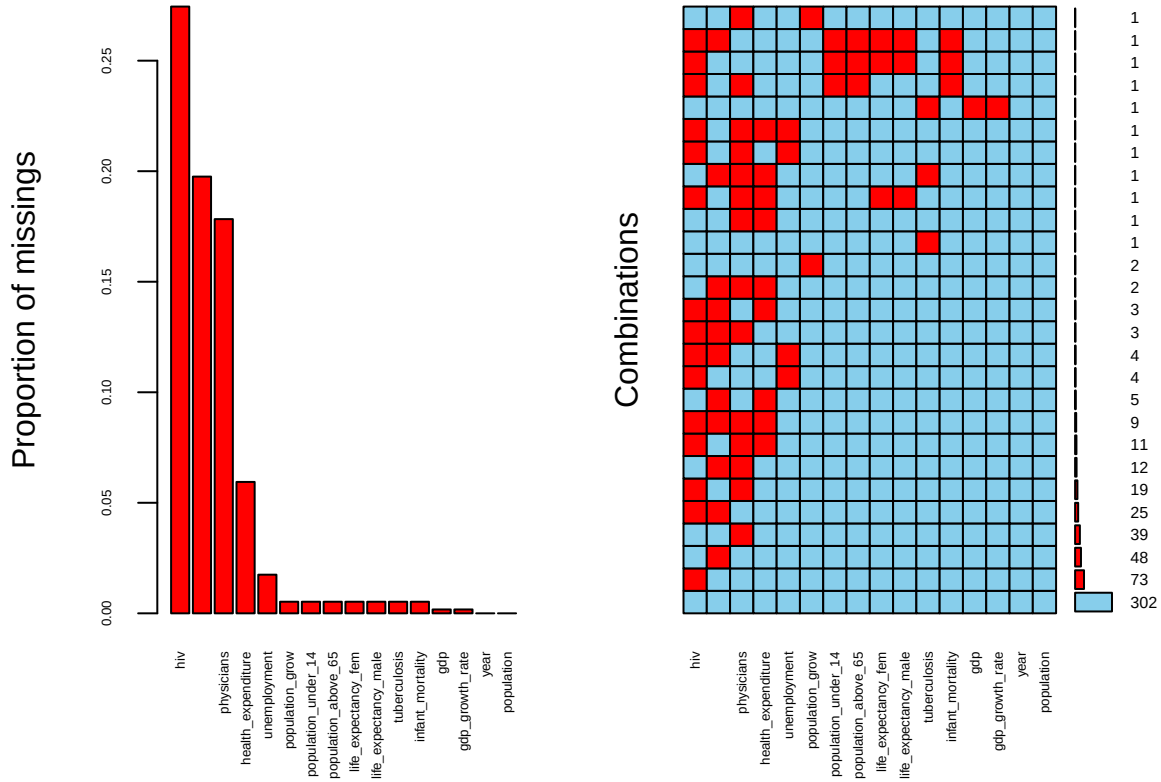



```
##
##  Variables sorted by number of missings:
##                    Variable       Count
##   non_commun_disease_death 0.680069930
##       commun_disease_death 0.680069930
##              hospital_beds 0.557692308
##          undernourishment 0.431818182
##                       hiv 0.274475524
##  education_gov_expenditure 0.197552448
##                physicians 0.178321678
##        health_expenditure 0.059440559
##              unemployment 0.017482517
##           population_grow 0.005244755
##        population_under_14 0.005244755
##        population_above_65 0.005244755
##        life_expectancy_fem 0.005244755
##       life_expectancy_male 0.005244755
##               tuberculosis 0.005244755
##           infant_mortality 0.005244755
##                       gdp 0.001748252
```

```
##            gdp_growth_rate 0.001748252
##                       year 0.000000000
##                 population 0.000000000
```

En vista de la important proporció de buits existent a les columnes non_commun_disease_death, commun_disease_death, hospital_beds i undernourishment, que oscil · la entre el 43 i el 68%, decidim realitzar una selecció de les dades, eliminant aquestes columnes de l'estudi, de manera que els tests i models que realitzem es sustentin en dades de major qualitat.

```
health_miss <- health_data[, !(colnames(health_data) %in% c("country_code", "country",
  "non_commun_disease_death", "commun_disease_death", "hospital_beds",
  "undernourishment"))]

aggr(health_miss, numbers=T, sortVars=T, labels=names(health_miss), cex.axis=0.4,
     cex.lab=1, cex.numbers=0.5, prop=c(T,F))
```




```
##
##  Variables sorted by number of missings:
##                   Variable       Count
##                        hiv 0.274475524
##  education_gov_expenditure 0.197552448
##                physicians 0.178321678
##        health_expenditure 0.059440559
##              unemployment 0.017482517
##           population_grow 0.005244755
##       population_under_14 0.005244755
##       population_above_65 0.005244755
##       life_expectancy_fem 0.005244755
```

```
##       life_expectancy_male 0.005244755
##               tuberculosis 0.005244755
##           infant_mortality 0.005244755
##                        gdp 0.001748252
##            gdp_growth_rate 0.001748252
##                       year 0.000000000
##                 population 0.000000000
```

Un cop fet això, implementarem el mètode *missForest* per imputar valors buits a la resta de columnes del dataset.

```
# missForest
missForest.imp <- missForest(health_miss, variablewise=T)
```

```
# Conservem el dataframe original canviant-li el nom a health_data_na.
health_data_na <- health_data

# Construim ara el dataframe de treball amb les dades imputades gràcies a aplicar el
# mètode MissForest, utilitzant per a aquest el nom de health_data:

health_data <- data.frame(health_data[, c("country_code", "country")], missForest.imp$ximp)

# Verifiquem que efectivament el mètode missForest ha omplert els valors buits.
colSums(is.na(health_data))
```

```
##              country_code                   country                      year
##                         0                         0                         0
##                population           population_grow       population_under_14
##                         0                         0                         0
##       population_above_65                       gdp           gdp_growth_rate
##                         0                         0                         0
##              unemployment education_gov_expenditure        health_expenditure
##                         0                         0                         0
##       life_expectancy_fem      life_expectancy_male              tuberculosis
##                         0                         0                         0
##                       hiv          infant_mortality                physicians
##                         0                         0                         0
```

```
# Anàlisi estadístic descriptiu de health_data després d'aplicar MissForest
summary(health_data)
```

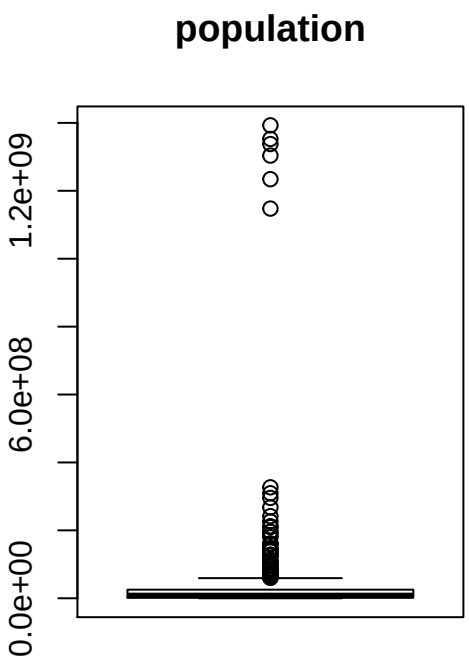
```
##  country_code         country                year        population
##  Length:572         Length:572         Min.   :2005   Min.   :1.978e+04
##  Class :character   Class :character   1st Qu.:2005   1st Qu.:2.073e+06
##  Mode  :character   Mode  :character   Median :2010   Median :7.655e+06
##                                        Mean   :2011   Mean   :3.660e+07
##                                        3rd Qu.:2019   3rd Qu.:2.530e+07
##                                        Max.   :2019   Max.   :1.393e+09
##  population_grow  population_under_14 population_above_65      gdp
##  Min.   :-3.980   Min.   :11.90       Min.   : 0.690      Min.   :     112
##  1st Qu.: 0.480   1st Qu.:18.99       1st Qu.: 3.317      1st Qu.:    6298
##  Median : 1.340   Median :28.14       Median : 5.745      Median :   23372
```

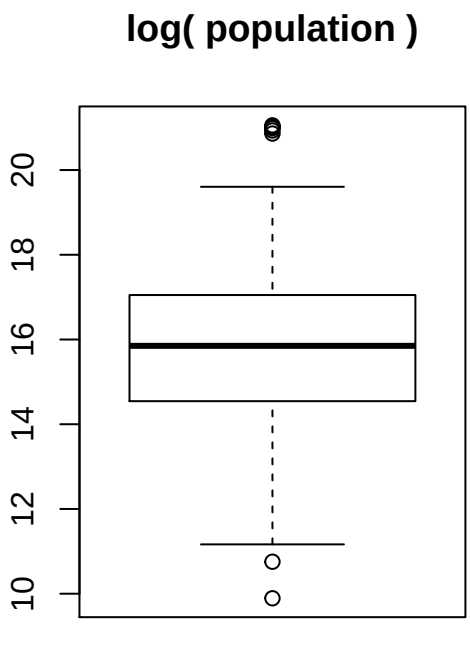
```
##  Mean   : 1.455   Mean   :29.08       Mean   : 7.858      Mean   :  336864
##  3rd Qu.: 2.362   3rd Qu.:38.85       3rd Qu.:11.920      3rd Qu.:  147658
##  Max.   :13.870   Max.   :50.04       Max.   :27.580      Max.   :19485394
##  gdp_growth_rate    unemployment    education_gov_expenditure
##  Min.   :-14.000   Min.   : 0.100   Min.   : 1.000
##  1st Qu.:  2.100   1st Qu.: 4.200   1st Qu.: 3.500
##  Median :  3.900   Median : 6.800   Median : 4.465
##  Mean   :  4.376   Mean   : 8.289   Mean   : 4.596
##  3rd Qu.:  6.500   3rd Qu.:10.862   3rd Qu.: 5.400
##  Max.   : 64.000   Max.   :37.200   Max.   :12.800
##  health_expenditure life_expectancy_fem life_expectancy_male  tuberculosis
##  Min.   : 1.300     Min.   :44.60       Min.   :40.57        Min.   :   0.0
##  1st Qu.: 4.488     1st Qu.:66.90       1st Qu.:62.81        1st Qu.:  15.0
##  Median : 6.085     Median :75.47       Median :69.39        Median :  51.5
##  Mean   : 6.288     Mean   :72.70       Mean   :67.84        Mean   : 131.6
##  3rd Qu.: 8.000     3rd Qu.:79.40       3rd Qu.:74.32        3rd Qu.: 181.0
##  Max.   :17.100     Max.   :87.70       Max.   :82.30        Max.   :1280.0
##       hiv          infant_mortality   physicians
##  Min.   : 0.100   Min.   :  1.60   Min.   :0.000
##  1st Qu.: 0.200   1st Qu.:  8.30   1st Qu.:0.300
##  Median : 0.500   Median : 19.20   Median :1.200
##  Mean   : 1.761   Mean   : 30.53   Mean   :1.546
##  3rd Qu.: 1.500   3rd Qu.: 51.10   3rd Qu.:2.580
##  Max.   :27.400   Max.   :141.90   Max.   :8.200
```

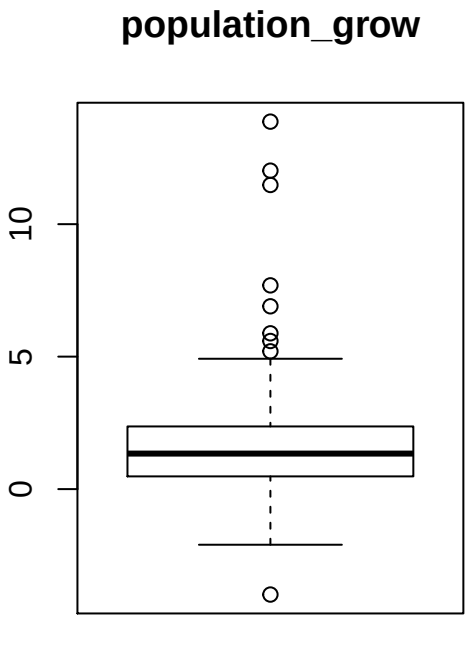
3.2. Identificació i tractament de valors extrems.

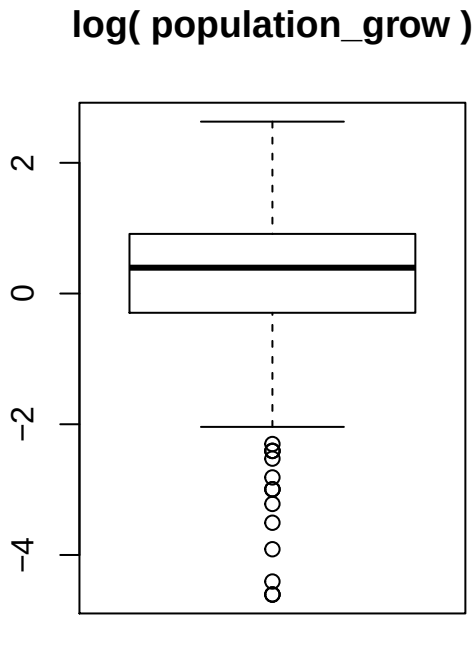
Mostrem dos gràfics de boxplot per cada variable, un amb la variable no transformada i l'altre gràfic aplicant una transformació de la variable amb el logaritme natural, per tal d'identificar els valors extrems. Analitzarem posteriorment els valors extrems tenint en compte el logaritme natural.

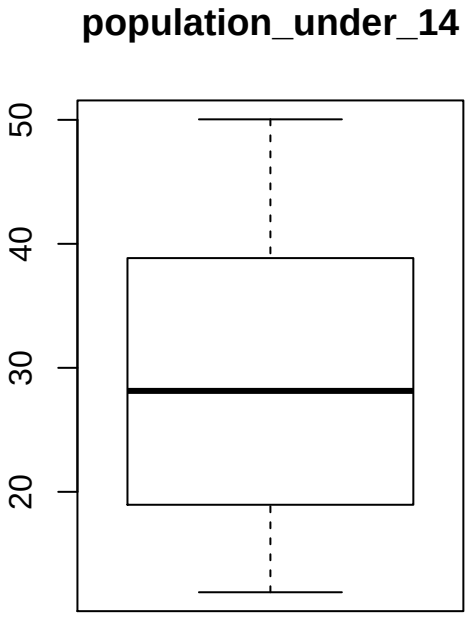
```r
# Boxplot de les variables per l'anàlisi de valors extrems

par(mfrow=c(1, 2))
for (i in 4:ncol(health_data)) {
  if (is.numeric(health_data[, i])){
  boxplot(health_data[, i], main = colnames(health_data)[i], width=100)
  boxplot(log(health_data[, i]), main =paste("log(", colnames(health_data)[i],")"),
          width=100)
  cat("  \n  \n")
  }
}
```

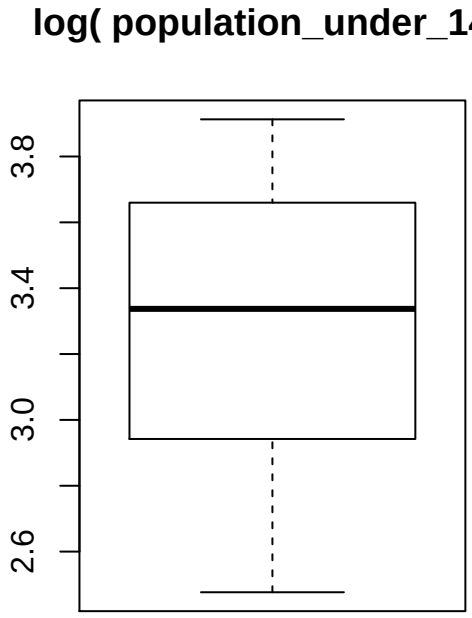

population
0.0e+00
6.0e+08
1.2e+09

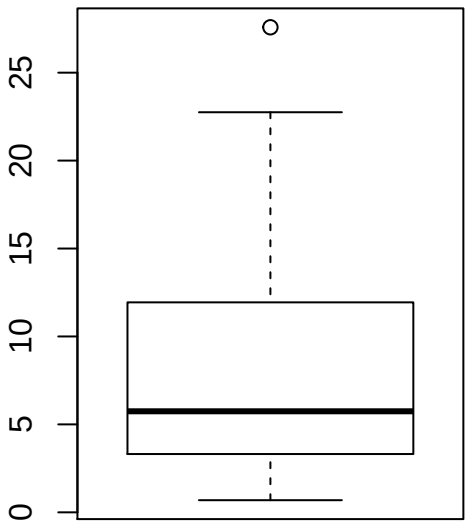

log(population)
10
12
14
16
18
20

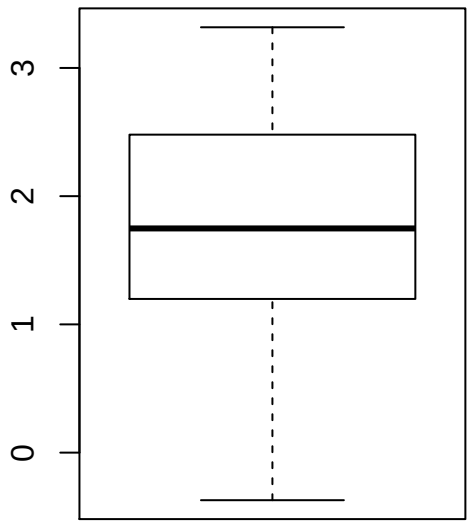

population_grow
0
5
10

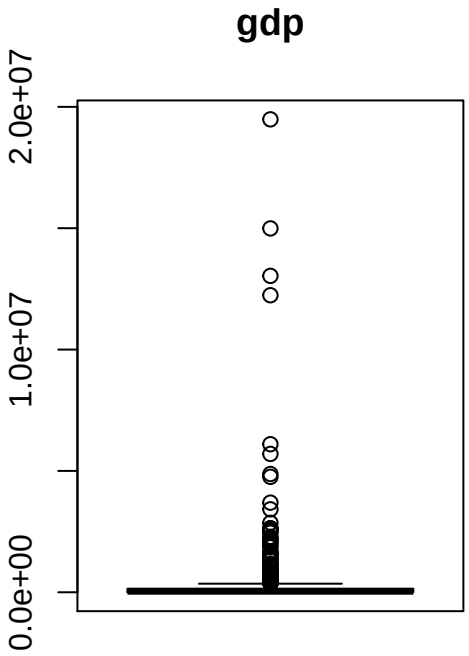

log(population_grow)
−4
−2
0
2

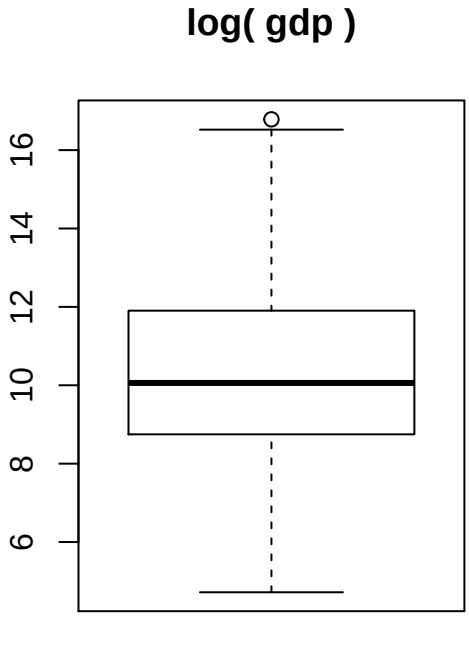

population_under_14
50
40
30
20

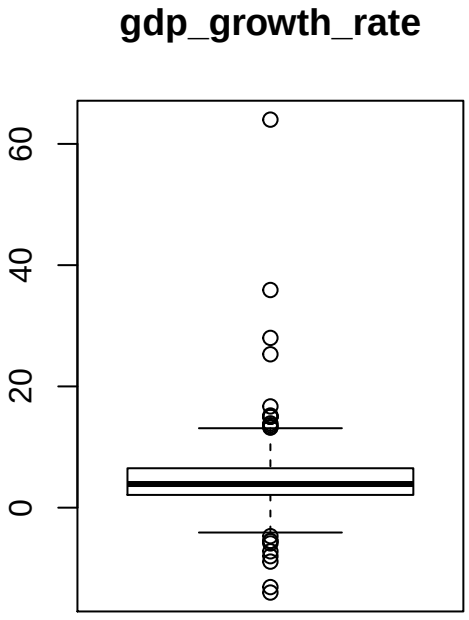

log(population_under_14)
3.8
3.4
3.0
2.6

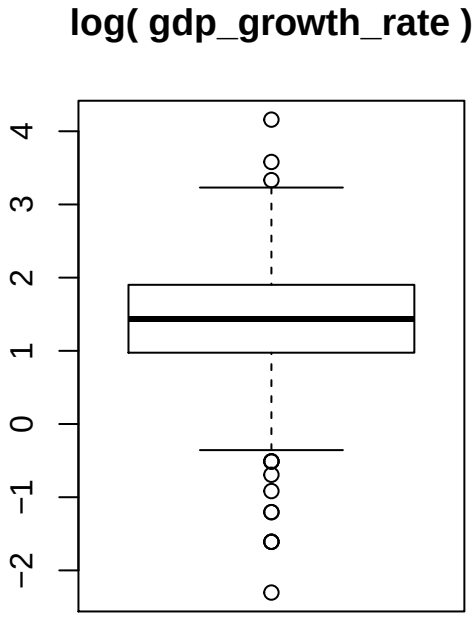

population_above_65
25
20
15
10
5
0

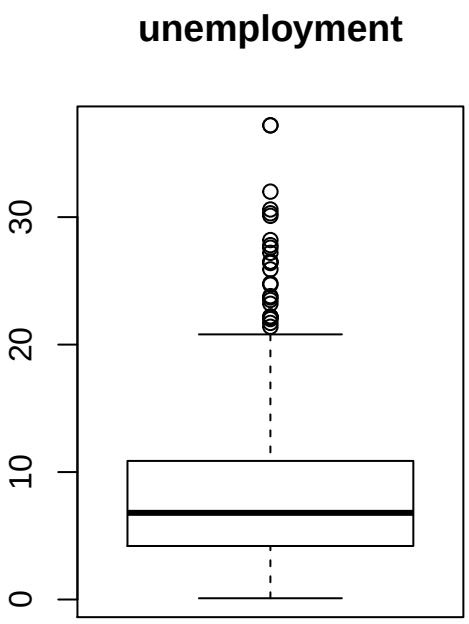

log(population_above_65)
3
2
1
0

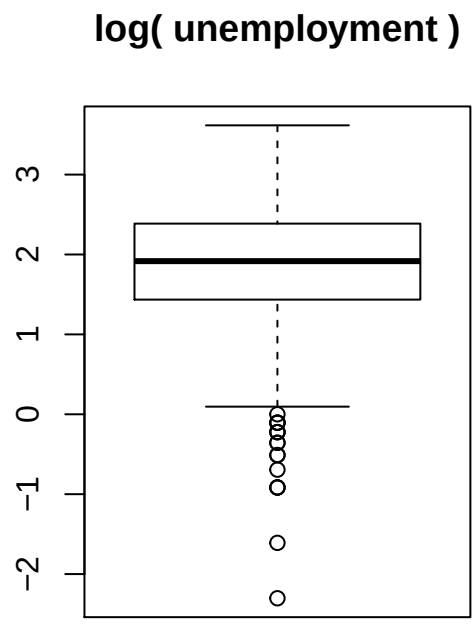

gdp
0.0e+00
1.0e+07
2.0e+07

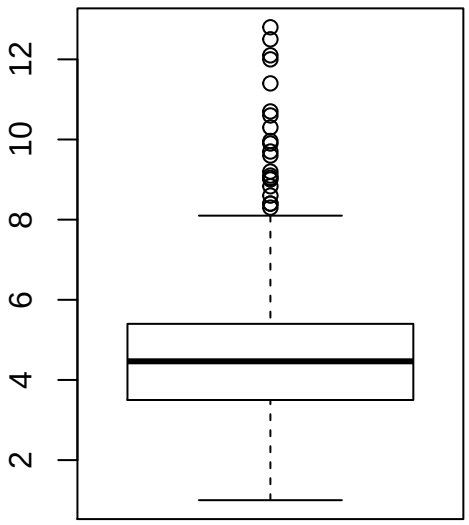

log(gdp)
6
8
10
12
14
16

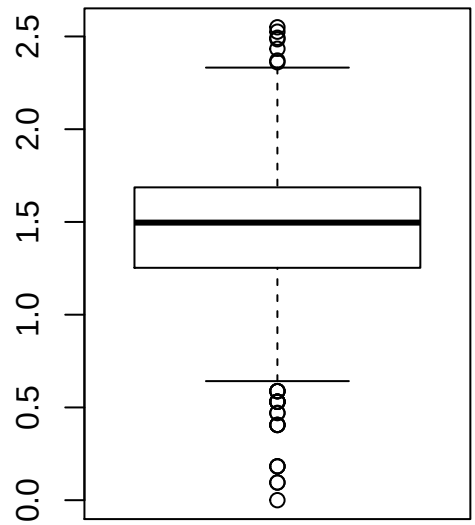

gdp_growth_rate
0
20
40
60

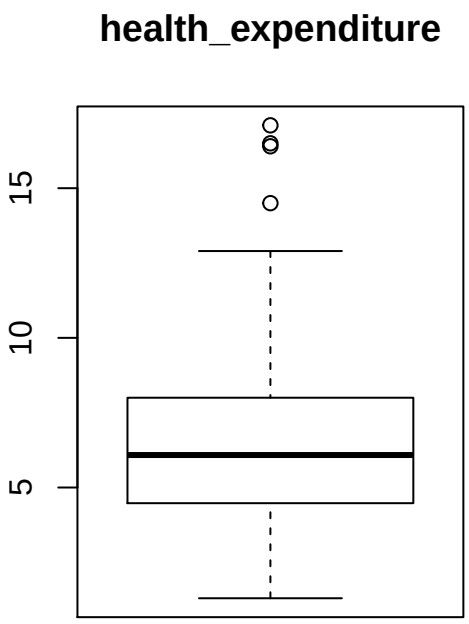

log(gdp_growth_rate)
−2
−1
0
1
2
3
4

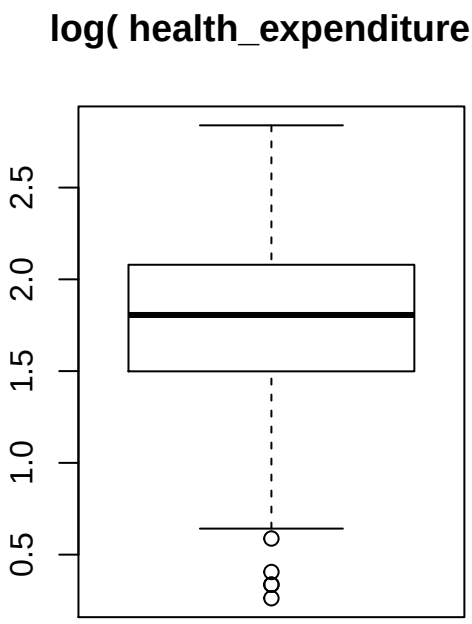

unemployment
0
10
20
30

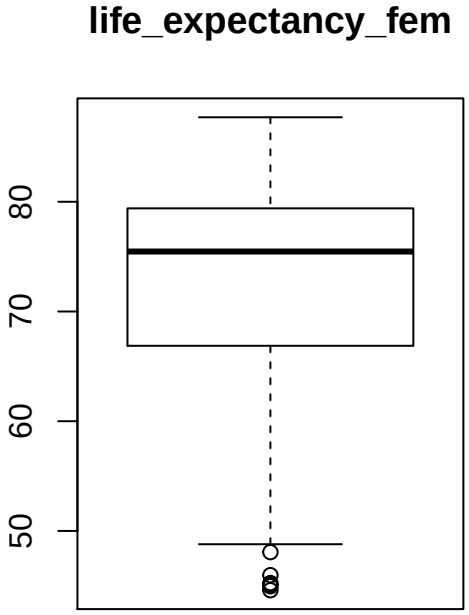

log(unemployment)
−2
−1
0
1
2
3

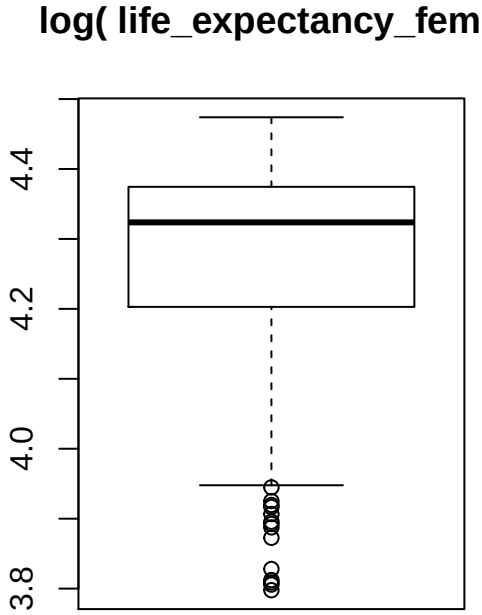

education_gov_expenditure
2
4
6
8
10
12

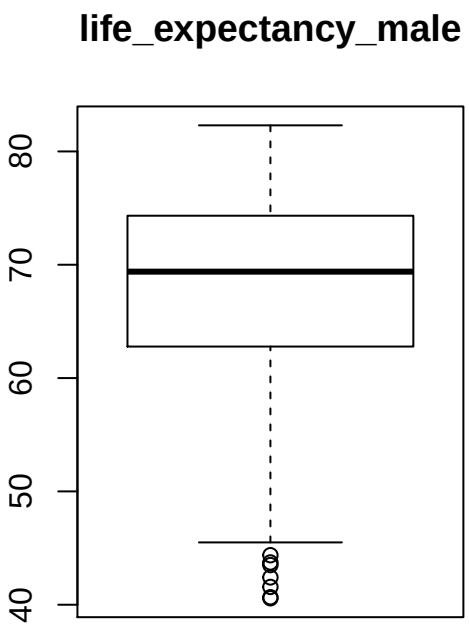

log(education_gov_expenditure
0.0
0.5
1.0
1.5
2.0
2.5

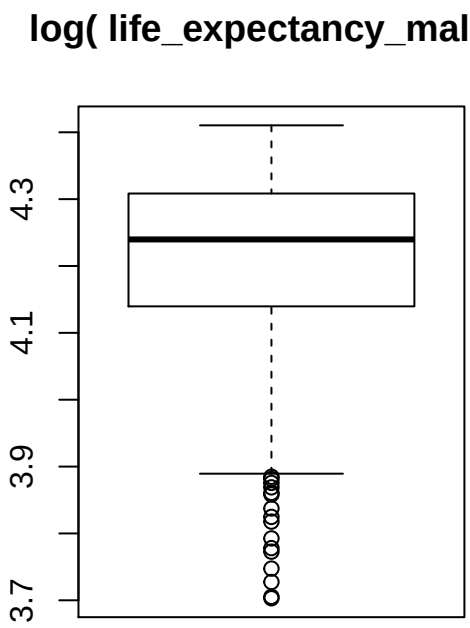

health_expenditure
5
10
15

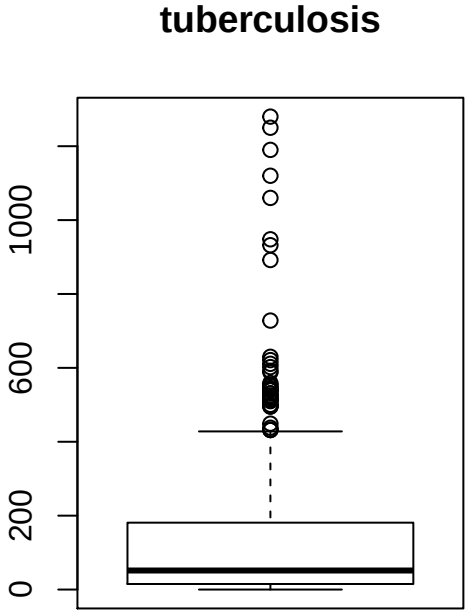

log(health_expenditure)
0.5
1.0
1.5
2.0
2.5

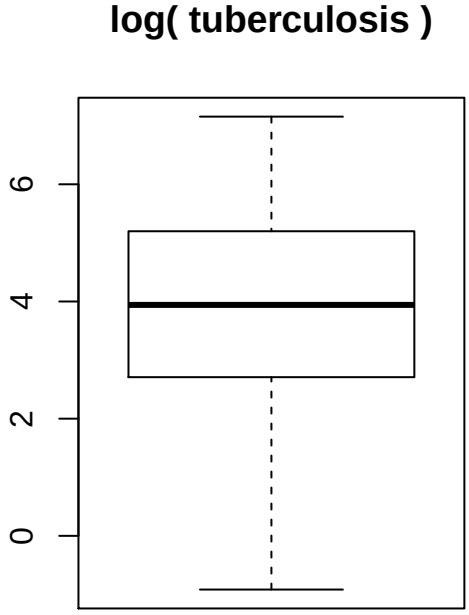

life_expectancy_fem
50
60
70
80

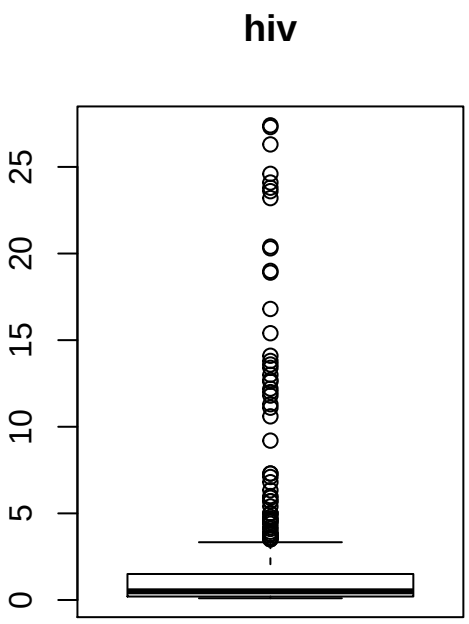

log(life_expectancy_fem)
3.8
4.0
4.2
4.4

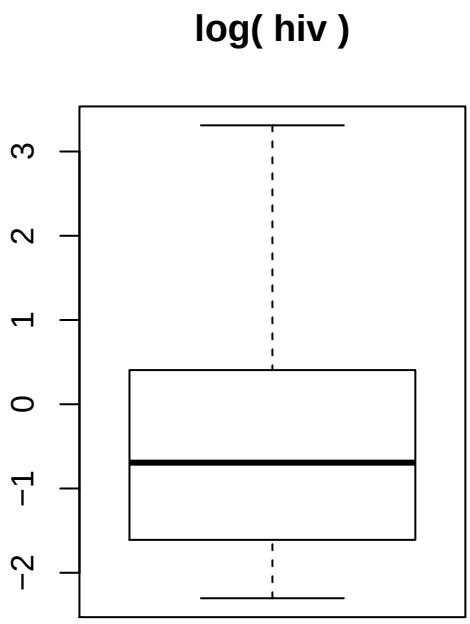

life_expectancy_male
80
70
60
50
40

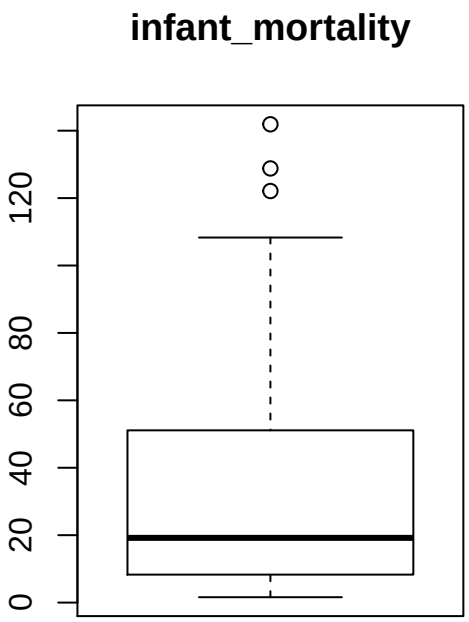

log(life_expectancy_male)
4.3
4.1
3.9
3.7

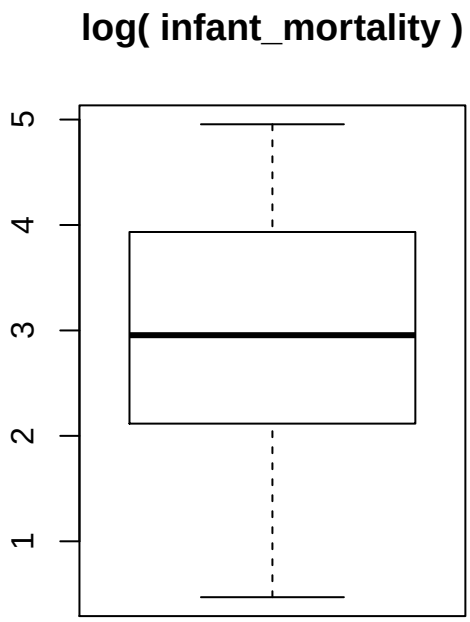

tuberculosis
1000
600
200
0

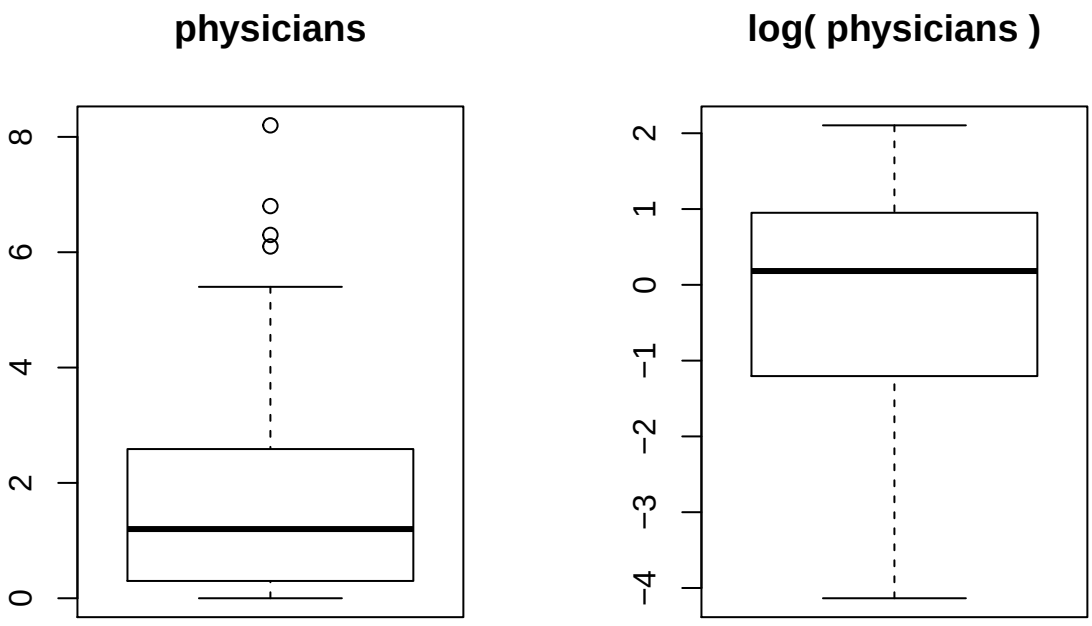

log(tuberculosis)
6
4
2
0


hiv
0
5
10
15
20
25


log(hiv)
−2
−1
0
1
2
3


infant_mortality
0
20
40
60
80
120


log(infant_mortality)
1
2
3
4
5



Llistem els outliers que s'obtenen a partir de la transformació logarítmica. Amb aquesta transformació aconseguim reduir el nombre de dades que es consideren outliers, per tal d'analitzar-ne la seva validesa.

```
# Mostrem outliers considerant la transformació logarítmica.

log_outliers <- function(x) {
  var_b <- boxplot.stats(log(x))
  log(x) %in% var_b$out
}

out_list <- lapply(4:ncol(health_data), function (i, df){
  df[log_outliers(df[,i])==T, c(1,2,3,i)]
}, df=health_data)
```

valors extrems

	country_code	country	year	population
109	CHN	China	2005	1303720000
110	CHN	China	2010	1337705000
111	CHN	China	2019	1392730000
236	IND	India	2005	1147609927
237	IND	India	2010	1234281170
238	IND	India	2019	1352617328
275	KNA	Saint Kitts and Nevis	2005	46857
429	PLW	Palau	2005	19781

valors extrems

	country_code	country	year	population_grow
34	BIH	Bosnia and Herzegovina	2005	0.0300
118	CUB	Cuba	2005	0.1000
209	GUY	Guyana	2005	0.0600
257	JPN	Japan	2005	0.0100
258	JPN	Japan	2010	0.0200
322	MDA	Moldova (the Republic of)	2010	0.0122
331	MKD	Republic of North Macedonia	2010	0.0800
332	MKD	Republic of North Macedonia	2019	0.0500
353	MUS	Mauritius	2019	0.0500
419	POL	Poland	2019	0.0000
427	PRT	Portugal	2010	0.0500
443	RUS	Russian Federation (the)	2010	0.0400
469	SVK	Slovakia	2005	0.0100
470	SVK	Slovakia	2010	0.0900
520	TON	Tonga	2010	0.0900
540	URY	Uruguay	2005	0.0100
546	VCT	Saint Vincent and the Grenadines	2005	0.0900

valors extrems

	country_code	country	year	gdp
539	USA	United States of America (the)	2019	19485394

valors extrems

	country_code	country	year	gdp_growth_rate
31	AZE	Azerbaijan	2005	28.0
33	AZE	Azerbaijan	2019	0.1
38	BRB	Barbados	2010	0.3
61	BRN	Brunei Darussalam	2005	0.4
84	BLZ	Belize	2019	0.5
159	ESP	Spain	2010	0.0
200	GRC	Greece	2005	0.6
293	LBN	Lebanon	2019	0.6
295	LCA	Saint Lucia	2010	0.2
317	LBY	Libya	2019	64.0
332	MKD	Republic of North Macedonia	2019	0.2
429	PLW	Palau	2005	0.2
440	SRB	Serbia	2010	0.6
510	TLS	Timor-Leste	2005	35.9
532	UKR	Ukraine	2010	0.3

valors extrems

	country_code	country	year	unemployment
58	BEN	Benin	2005	0.9
59	BEN	Benin	2010	1.0
79	BLR	Belarus	2005	0.9
81	BLR	Belarus	2019	0.5
267	KHM	Cambodia	2010	0.4
268	KHM	Cambodia	2019	0.2
289	LAO	Lao People's Democratic Republic (the)	2010	0.7
290	LAO	Lao People's Democratic Republic (the)	2019	0.7
336	MMR	Myanmar	2005	0.8
337	MMR	Myanmar	2010	0.8
338	MMR	Myanmar	2019	0.8
376	NER	Niger (the)	2010	0.9
377	NER	Niger (the)	2019	0.4
415	PAK	Pakistan	2010	0.6
434	QAT	Qatar	2010	0.4
435	QAT	Qatar	2019	0.1
505	THA	Thailand	2010	0.6

valors extrems

	country_code	country	year	education_gov_expenditure
42	BGD	Bangladesh	2019	1.5
76	BWA	Botswana	2005	10.7
89	COD	Congo (the Democratic Republic of the)	2010	1.5
90	COD	Congo (the Democratic Republic of the)	2019	1.5
91	CAF	Central African Republic (the)	2005	1.7
92	CAF	Central African Republic (the)	2010	1.2
93	CAF	Central African Republic (the)	2019	1.2
94	COG	Congo (the)	2005	1.8
118	CUB	Cuba	2005	10.6
119	CUB	Cuba	2010	12.8
141	DOM	Dominican Republic (the)	2005	1.8
147	ECU	Ecuador	2005	1.2
172	FSM	Micronesia (Federated States of)	2019	12.5
191	GMB	Gambia (the)	2005	1.1
194	GIN	Guinea	2005	1.8
266	KHM	Cambodia	2005	1.7
267	KHM	Cambodia	2010	1.5
269	KIR	Kiribati	2005	12.0
289	LAO	Lao People's Democratic Republic (the)	2010	1.7
292	LBN	Lebanon	2010	1.6
298	LKA	Sri Lanka	2010	1.7
303	LSO	Lesotho	2005	12.1
304	LSO	Lesotho	2010	11.4
457	SDN	Sudan (the)	2005	1.6
485	SSD	South Sudan	2019	1.0
498	TCD	Chad	2005	1.7
567	ZMB	Zambia	2005	1.7
568	ZMB	Zambia	2010	1.1
571	ZWE	Zimbabwe	2010	1.8

valors extrems

	country_code	country	year	health_expenditure
197	GNQ	Equatorial Guinea	2005	1.4
198	GNQ	Equatorial Guinea	2010	1.5
434	QAT	Qatar	2010	1.8
510	TLS	Timor-Leste	2005	1.3
511	TLS	Timor-Leste	2010	1.4

valors extrems

	country_code	country	year	life_expectancy_fem
91	CAF	Central African Republic (the)	2005	45.97
92	CAF	Central African Republic (the)	2010	48.79
303	LSO	Lesotho	2005	45.10
304	LSO	Lesotho	2010	48.07
357	MWI	Malawi	2005	50.38
378	NGA	Nigeria	2005	49.03
379	NGA	Nigeria	2010	51.66
472	SLE	Sierra Leone	2005	45.25
473	SLE	Sierra Leone	2010	50.23
495	SWZ	Eswatini	2005	44.60
496	SWZ	Eswatini	2010	49.19
498	TCD	Chad	2005	49.73
567	ZMB	Zambia	2005	50.66
570	ZWE	Zimbabwe	2005	44.96

valors extrems

	country_code	country	year	life_expectancy_male
16	AGO	Angola	2005	47.89
91	CAF	Central African Republic (the)	2005	43.50
92	CAF	Central African Republic (the)	2010	45.82
100	CIV	Côte d'Ivoire	2005	48.66
303	LSO	Lesotho	2005	40.57
304	LSO	Lesotho	2010	42.41
357	MWI	Malawi	2005	45.50
366	MOZ	Mozambique	2005	48.21
378	NGA	Nigeria	2005	47.50
472	SLE	Sierra Leone	2005	43.73
473	SLE	Sierra Leone	2010	48.50
495	SWZ	Eswatini	2005	40.66
496	SWZ	Eswatini	2010	44.37
498	TCD	Chad	2005	47.38
567	ZMB	Zambia	2005	46.40
570	ZWE	Zimbabwe	2005	41.57

valors extrems

	country_code	country	year	tuberculosis
275	KNA	Saint Kitts and Nevis	2005	0

valors extrems

	country_code	country	year	physicians
47	BFA	Burkina Faso	2010	0
55	BDI	Burundi	2005	0
56	BDI	Burundi	2010	0
57	BDI	Burundi	2019	0
58	BEN	Benin	2005	0
92	CAF	Central African Republic (the)	2010	0
161	ETH	Ethiopia	2005	0
162	ETH	Ethiopia	2010	0
300	LBR	Liberia	2005	0
301	LBR	Liberia	2010	0
302	LBR	Liberia	2019	0
303	LSO	Lesotho	2005	0
357	MWI	Malawi	2005	0
358	MWI	Malawi	2010	0
359	MWI	Malawi	2019	0
366	MOZ	Mozambique	2005	0
367	MOZ	Mozambique	2010	0
375	NER	Niger (the)	2005	0
376	NER	Niger (the)	2010	0
377	NER	Niger (the)	2019	0
408	PNG	Papua New Guinea	2005	0
435	QAT	Qatar	2019	0
472	SLE	Sierra Leone	2005	0
473	SLE	Sierra Leone	2010	0
474	SLE	Sierra Leone	2019	0
479	SOM	Somalia	2010	0
480	SOM	Somalia	2019	0
498	TCD	Chad	2005	0
499	TCD	Chad	2010	0
500	TCD	Chad	2019	0
501	TGO	Togo	2005	0
503	TGO	Togo	2019	0
528	TZA	Tanzania, United Republic of	2005	0
530	TZA	Tanzania, United Republic of	2019	0

Analitzant els valors dels outliers, veiem que la majoria són valors legítims, excepte un valor que es mostra a continuació, que és un error provinent de la font de dades des de la que es va fer web scraping. El que farem és substituir aquest outlier no legítim cercant una font fiable (https://tradingeconomics.com/libya/gdp-growth-annual) i introduint manualment la dada correcta.

```
health_data$gdp_growth_rate[health_data$country == "Libya" & health_data$year == "2019"] <- 4.0
```

4. Anàlisi de les dades.

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

D'una banda, els grups de dades que interessa comparar són directament les diferents columnes del dataset entre elles, ja que l'objectiu d'aquest estudi és trobar correlacions entre els indicadors socioeconòmics dels països i els indicadors de salut pública que afecten la població, per tal de construir un model que pugui predir l'impacte en l'esperança de vida de la població segons diferents variables.

D'altra banda, també compararem les dades del dataframe agrupades en 5 grups diferents segons el nivell de despesa sanitària pública per càpita.

```
# Creem quatre noves variables calculades, que mostrin les variables gdp, gdp_growth_rate,
# education_gov_expenditure, i health_expenditure per càpita:

health_data$gdp_capita <- (health_data$gdp / health_data$population) * 1e6

health_data$gdp_growth_capita <-
  ((health_data$gdp_growth_rate / 100) * health_data$gdp * 1e6) /
  health_data$population

health_data$educ_exp_capita <-
  ((health_data$education_gov_expenditure/100) * health_data$gdp * 1e6) /
  health_data$population

health_data$health_exp_capita <-
  ((health_data$health_expenditure/100) * health_data$gdp * 1e6) / health_data$population

summary(health_data)
```

```
##  country_code         country               year        population
##  Length:572         Length:572         Min.   :2005   Min.   :1.978e+04
##  Class :character   Class :character   1st Qu.:2005   1st Qu.:2.073e+06
##  Mode  :character   Mode  :character   Median :2010   Median :7.655e+06
##                                        Mean   :2011   Mean   :3.660e+07
##                                        3rd Qu.:2019   3rd Qu.:2.530e+07
##                                        Max.   :2019   Max.   :1.393e+09
##  population_grow  population_under_14 population_above_65      gdp
##  Min.   :-3.980   Min.   :11.90       Min.   : 0.690      Min.   :     112
##  1st Qu.: 0.480   1st Qu.:18.99       1st Qu.: 3.317      1st Qu.:    6298
##  Median : 1.340   Median :28.14       Median : 5.745      Median :   23372
##  Mean   : 1.455   Mean   :29.08       Mean   : 7.858      Mean   :  336864
##  3rd Qu.: 2.362   3rd Qu.:38.85       3rd Qu.:11.920      3rd Qu.:  147658
##  Max.   :13.870   Max.   :50.04       Max.   :27.580      Max.   :19485394
##  gdp_growth_rate    unemployment    education_gov_expenditure
##  Min.   :-14.000   Min.   : 0.100   Min.   : 1.000
##  1st Qu.:  2.100   1st Qu.: 4.200   1st Qu.: 3.500
##  Median :  3.900   Median : 6.800   Median : 4.465
##  Mean   :  4.271   Mean   : 8.289   Mean   : 4.596
##  3rd Qu.:  6.425   3rd Qu.:10.862   3rd Qu.: 5.400
##  Max.   : 35.900   Max.   :37.200   Max.   :12.800
##  health_expenditure life_expectancy_fem life_expectancy_male  tuberculosis
##  Min.   : 1.300     Min.   :44.60       Min.   :40.57        Min.   :   0.0
```

```
##  1st Qu.: 4.488     1st Qu.:66.90       1st Qu.:62.81        1st Qu.:  15.0
##  Median : 6.085     Median :75.47       Median :69.39        Median :  51.5
##  Mean   : 6.288     Mean   :72.70       Mean   :67.84        Mean   : 131.6
##  3rd Qu.: 8.000     3rd Qu.:79.40       3rd Qu.:74.32        3rd Qu.: 181.0
##  Max.   :17.100     Max.   :87.70       Max.   :82.30        Max.   :1280.0
##       hiv          infant_mortality   physicians      gdp_capita
##  Min.   : 0.100   Min.   :  1.60   Min.   :0.000   Min.   :    90.75
##  1st Qu.: 0.200   1st Qu.:  8.30   1st Qu.:0.300   1st Qu.:  1318.87
##  Median : 0.500   Median : 19.20   Median :1.200   Median :  4200.13
##  Mean   : 1.761   Mean   : 30.53   Mean   :1.546   Mean   : 12107.31
##  3rd Qu.: 1.500   3rd Qu.: 51.10   3rd Qu.:2.580   3rd Qu.: 14796.79
##  Max.   :27.400   Max.   :141.90   Max.   :8.200   Max.   :104964.37
##  gdp_growth_capita  educ_exp_capita    health_exp_capita
##  Min.   :-1537.25   Min.   :   2.924   Min.   :    4.875
##  1st Qu.:   43.64   1st Qu.:  55.862   1st Qu.:   64.483
##  Median :  137.60   Median : 191.675   Median :  249.534
##  Mean   :  423.12   Mean   : 610.703   Mean   :  902.114
##  3rd Qu.:  498.48   3rd Qu.: 713.700   3rd Qu.:  903.858
##  Max.   :13220.22   Max.   :5902.146   Max.   :10199.357
```

Establirem diferents categories de dades segons la despesa sanitària pública. Per això es crea una nova variable que representa el tram de despesa sanitària pública. Es calcula prenent igual freqüència d'ocurrència en cada tram per crear un total de 5 intervals.

```
health_data$health_expend_category <- cut2(health_data$health_exp_capita, g=5)

health_data[1:20,c("country_code", "country", "health_expend_category")]
```

country_code	country	health_expend_category
ARE	United Arab Emirates (the)	[388.99, 1246.4)
ARE	United Arab Emirates (the)	[1246.40,10199.4]
ARE	United Arab Emirates (the)	[1246.40,10199.4]
AFG	Afghanistan	[4.87, 50.5)
AFG	Afghanistan	[4.87, 50.5)
AFG	Afghanistan	[50.50, 159.1)
ATG	Antigua and Barbuda	[388.99, 1246.4)
ATG	Antigua and Barbuda	[388.99, 1246.4)
ATG	Antigua and Barbuda	[388.99, 1246.4)
ALB	Albania	[159.14, 389.0)
ALB	Albania	[159.14, 389.0)
ALB	Albania	[159.14, 389.0)
ARM	Armenia	[50.50, 159.1)
ARM	Armenia	[159.14, 389.0)
ARM	Armenia	[159.14, 389.0)
AGO	Angola	[50.50, 159.1)
AGO	Angola	[50.50, 159.1)
AGO	Angola	[50.50, 159.1)
ARG	Argentina	[388.99, 1246.4)
ARG	Argentina	[388.99, 1246.4)

```
# Convertim aquests intervals en nivells categòrics
levels(health_data$health_expend_category) <- c("very low", "low", "medium", "high", "very high")
```

```
# Creem subgrups
cat_very_low_fem <-
  health_data$life_expectancy_fem[health_data$health_expend_category=="very low"]
cat_low_fem <-
  health_data$life_expectancy_fem[health_data$health_expend_category=="low"]
cat_medium_fem <-
  health_data$life_expectancy_fem[health_data$health_expend_category=="medium"]
cat_high_fem <-
  health_data$life_expectancy_fem[health_data$health_expend_category=="high"]
cat_very_high_fem <-
  health_data$life_expectancy_fem[health_data$health_expend_category=="very high"]

cat_very_low_male <-
  health_data$life_expectancy_male[health_data$health_expend_category=="very low"]
cat_low_male <-
  health_data$life_expectancy_male[health_data$health_expend_category=="low"]
cat_medium_male <-
  health_data$life_expectancy_male[health_data$health_expend_category=="medium"]
cat_high_male <-
  health_data$life_expectancy_male[health_data$health_expend_category=="high"]
cat_very_high_male <-
  health_data$life_expectancy_male[health_data$health_expend_category=="very high"]
```

4.2. Comprovació de la normalitat i homogeneïtat de la variància.

Comprovació de la normalitat de les variables del dataset

Comprovem amb proves estadístiques la no normalitat de cadascuna de les variables de health_data. Utilitzarem la prova de Shapiro-Wilk, mitjançant la funció shapiro.test().

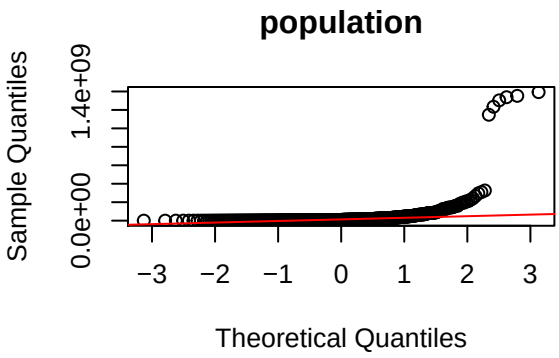
```
lshap <- sapply(health_data[, !(colnames(health_data) %in% c("country_code", "country", "year",
  "health_expend_category"))], shapiro.test)
```

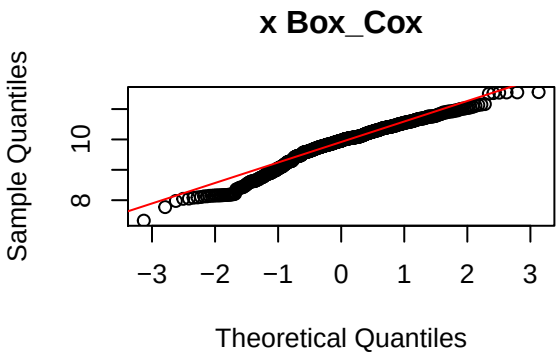
	statistic	p.value
population	c(W = 0.224805695864548)	2.61199611414149e-43
population_grow	c(W = 0.856648816515033)	2.09227247124823e-22
population_under_14	c(W = 0.940801412476724)	2.58122508887395e-14
population_above_65	c(W = 0.865321511204978)	9.03529227264668e-22
gdp	c(W = 0.223298491255859)	2.44779060132266e-43
gdp_growth_rate	c(W = 0.904386002338808)	1.85848345262076e-18
unemployment	c(W = 0.870134905848752)	2.0985898367291e-21
education_gov_expenditure	c(W = 0.93783097725515)	1.03549136314357e-14
health_expenditure	c(W = 0.97032983679328)	2.35252177009053e-09
life_expectancy_fem	c(W = 0.926139507634802)	3.72622713328277e-16
life_expectancy_male	c(W = 0.954150796228931)	2.427215142594e-12
tuberculosis	c(W = 0.683257747888483)	2.491997322588e-31
hiv	c(W = 0.431511393877858)	6.26568542903413e-39
infant_mortality	c(W = 0.862335257593656)	5.41743670223244e-22
physicians	c(W = 0.904151214239933)	1.76394639791389e-18
gdp_capita	c(W = 0.697428853997288)	9.01265856754629e-31
gdp_growth_capita	c(W = 0.46417985056566)	4.02904770137275e-38
educ_exp_capita	c(W = 0.651030488572157)	1.57478069623714e-32
health_exp_capita	c(W = 0.60351163461669)	3.81006383467123e-34

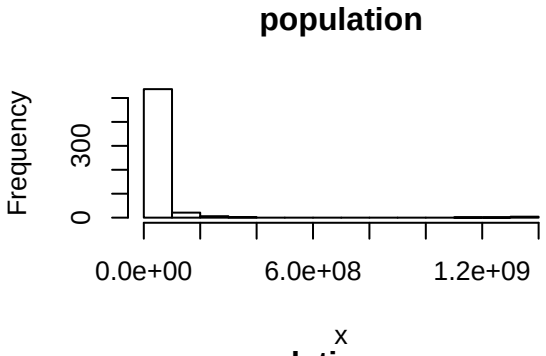
La prova de Shapiro-Wilk dona com a resultat un p-valor menor al nivell de significació alpha = 0.05 per a cadascuna de les variables numèriques del dataset, per tant, es rebutja la hipòtesi nul · la i es conclou que cap de les variables segueix una distribució normal de manera significativa.

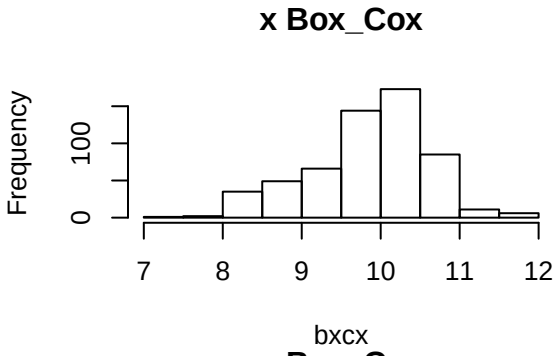
Intentem ara aplicar alguna transformació a les dades per veure si podem així obtenir una normalització, amb la intenció d'aplicar testos paramètrics posteriorment a les variables transformades:

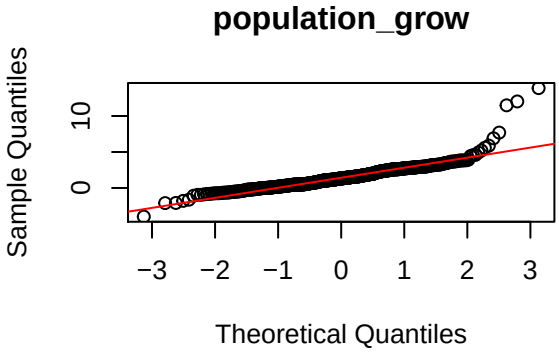
```
# Transformació de Box-Cox

for (i in 4:ncol(health_data)) {
  if (is.numeric(health_data[, i])){
  x <- health_data[, i]
  bxcx <- BoxCox(x, lambda=BoxCoxLambda(x))

  shapiro.test(bxcx)

  par(mfrow=c(2,2))

  # Gràfica Q-Q per Fare sense transformat
  qqnorm(x, main=colnames(health_data)[i])
  qqline(x, col=2)

  # Gràfica Q-Q per Fare transformat
  qqnorm(bxcx, main="x Box_Cox")
  qqline(bxcx, col=2)

  hist(x, main=colnames(health_data)[i])
  hist(bxcx, main="x Box_Cox")
  par(mfrow=c(1,1))

  }
}
```

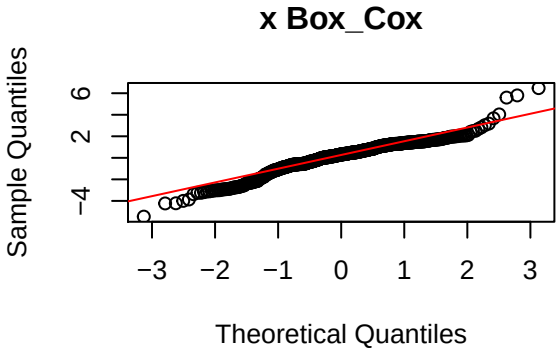

population
Sample Quantiles
0.0e+00
1.4e+09
−3
−2
−1
0
1
2
3
Theoretical Quantiles

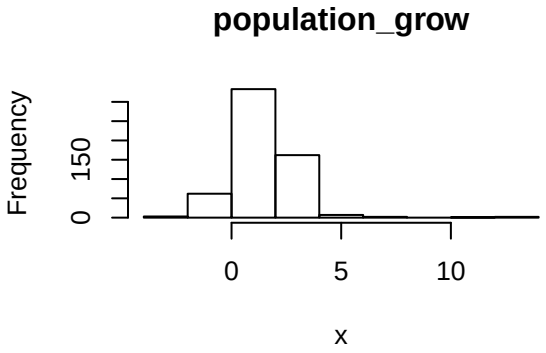

x Box_Cox
Sample Quantiles
8
10
−3
−2
−1
0
1
2
3
Theoretical Quantiles

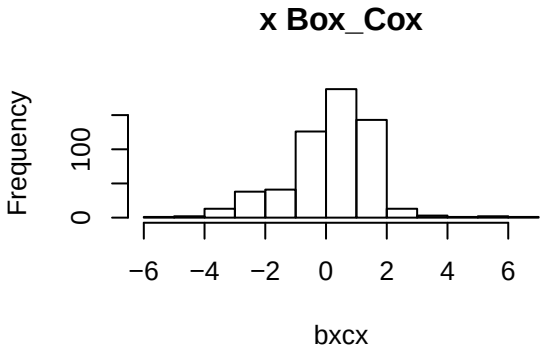

population
Frequency
0
300
0.0e+00
6.0e+08
1.2e+09
x

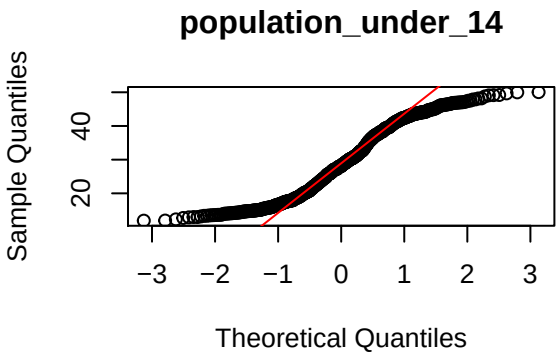

x Box_Cox
Frequency
0
100
7
8
9
10
11
12
bxcx

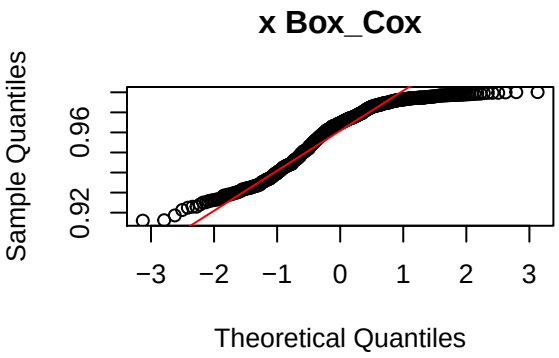

population_grow
Sample Quantiles
0
10
−3
−2
−1
0
1
2
3
Theoretical Quantiles

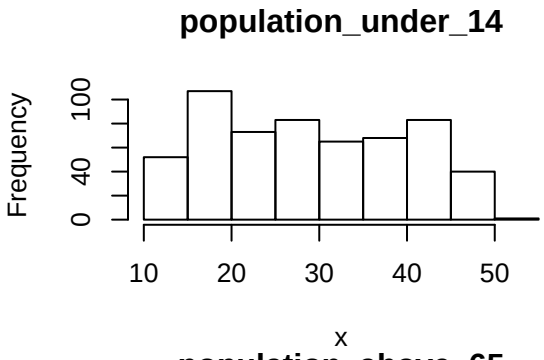

x Box_Cox
Sample Quantiles
−4
2
6
−3
−2
−1
0
1
2
3
Theoretical Quantiles

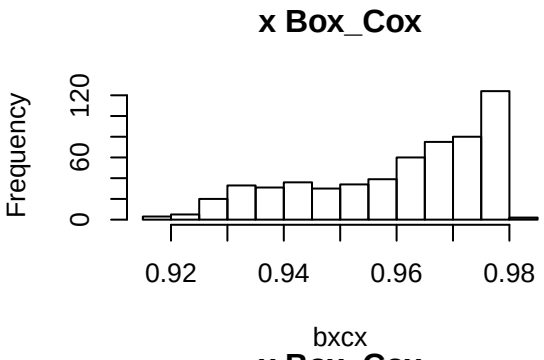

population_grow
Frequency
0
150
0
5
10
x

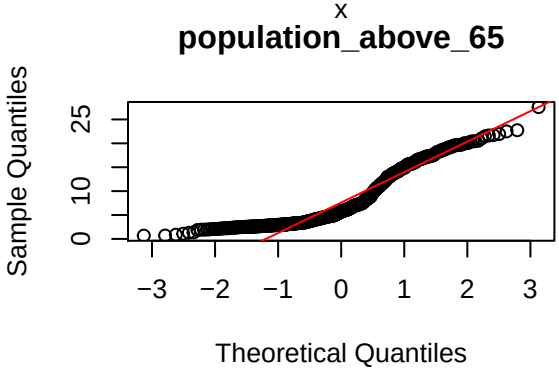

x Box_Cox
Frequency
0
100
−6
−4
−2
0
2
4
6
bxcx

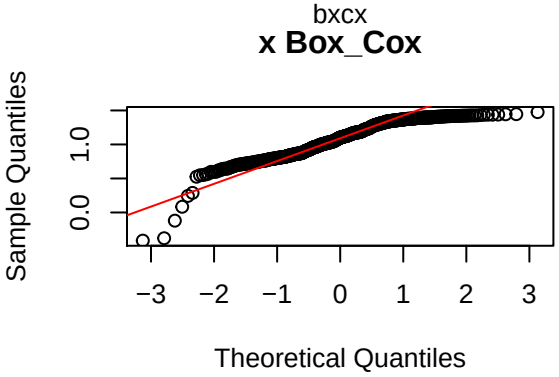

population_under_14
Sample Quantiles
20
40
−3
−2
−1
0
1
2
3
Theoretical Quantiles

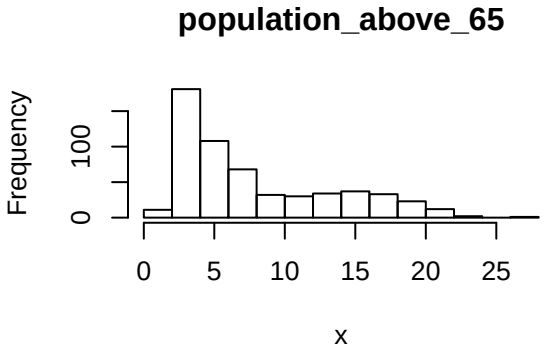

x Box_Cox
Sample Quantiles
0.92
0.96
−3
−2
−1
0
1
2
3
Theoretical Quantiles

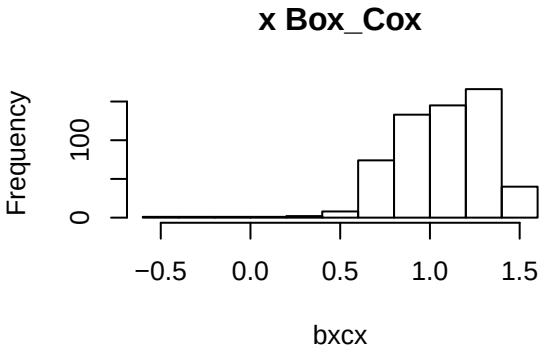

population_under_14
Frequency
0
40
100
10
20
30
40
50
x

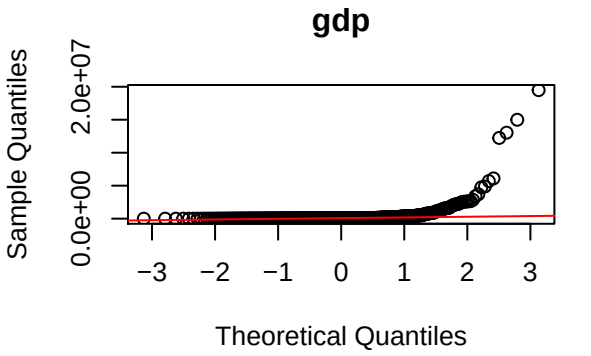

x Box_Cox
Frequency
0
60
120
0.92
0.94
0.96
0.98
bxcx

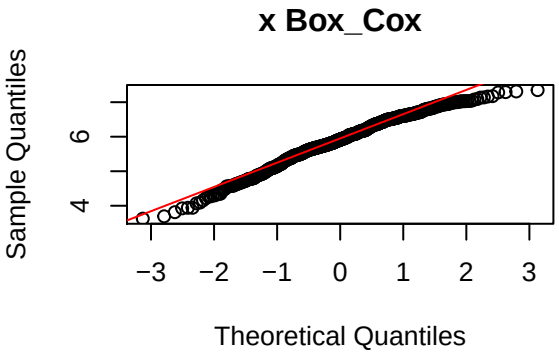

population_above_65
Sample Quantiles
0
10
25
−3
−2
−1
0
1
2
3
Theoretical Quantiles

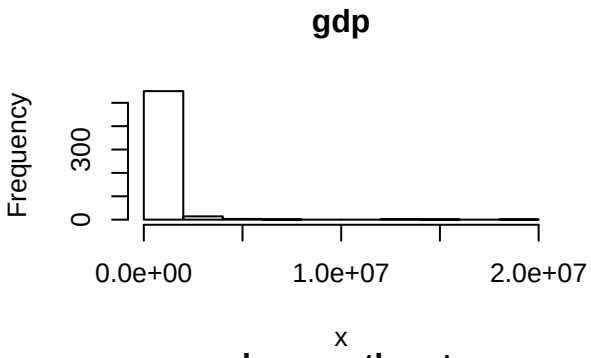

x Box_Cox
Sample Quantiles
0.0
1.0
−3
−2
−1
0
1
2
3
Theoretical Quantiles

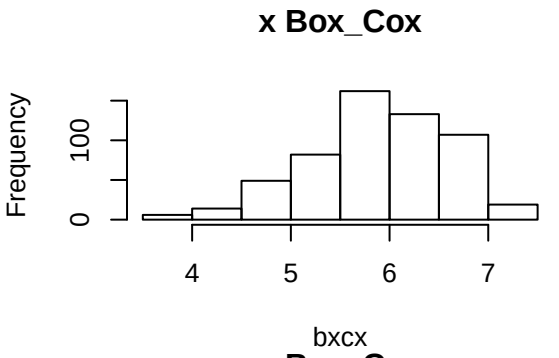

population_above_65
Frequency
0
100
0
5
10
15
20
25
x

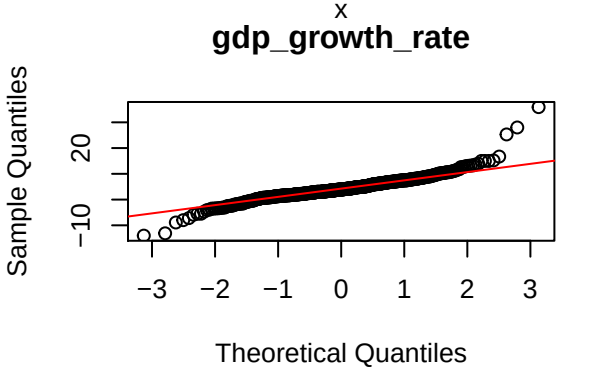

x Box_Cox
Frequency
0
100
−0.5
0.0
0.5
1.0
1.5
bxcx

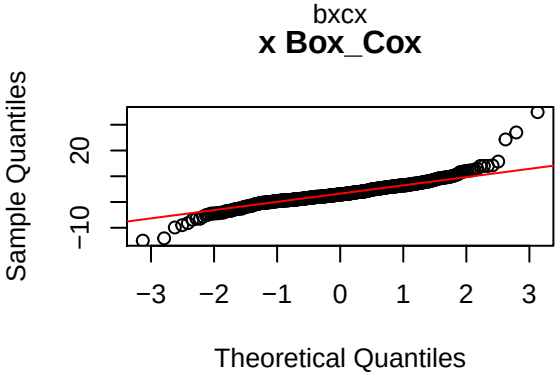

gdp
Sample Quantiles
0.0e+00
2.0e+07
−3
−2
−1
0
1
2
3
Theoretical Quantiles

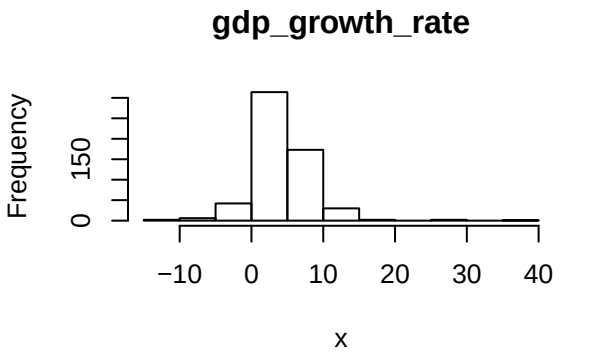

x Box_Cox
Sample Quantiles
4
6
−3
−2
−1
0
1
2
3
Theoretical Quantiles

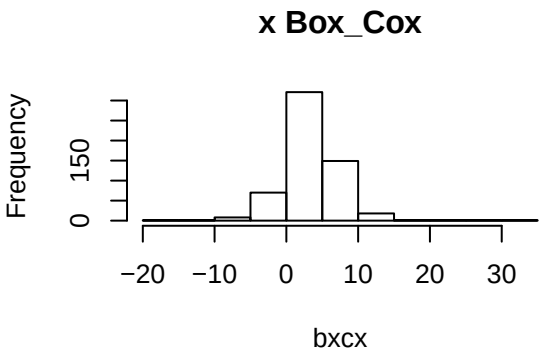

gdp
Frequency
0
300
0.0e+00
1.0e+07
2.0e+07
x

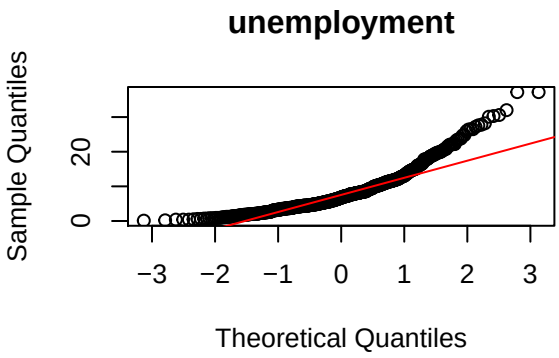

x Box_Cox
Frequency
0
100
4
5
6
7
bxcx

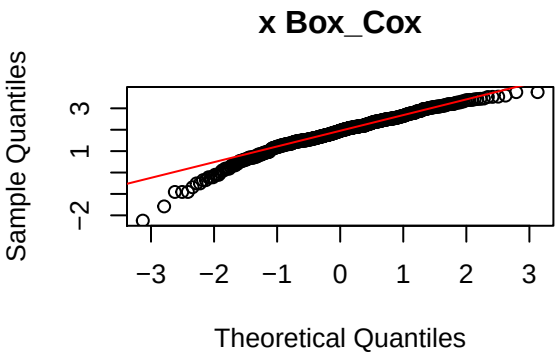

gdp_growth_rate
Sample Quantiles
−10
20
−3
−2
−1
0
1
2
3
Theoretical Quantiles

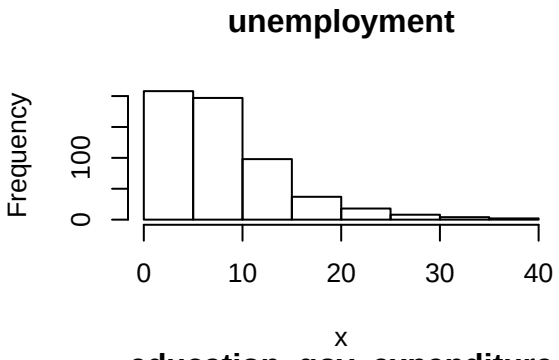

x Box_Cox
Sample Quantiles
−10
20
−3
−2
−1
0
1
2
3
Theoretical Quantiles

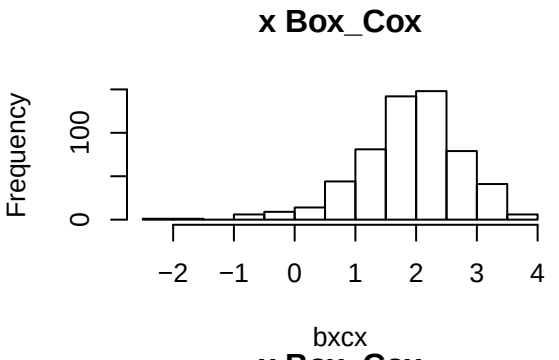

gdp_growth_rate
Frequency
0
150
−10
0
10
20
30
40
x

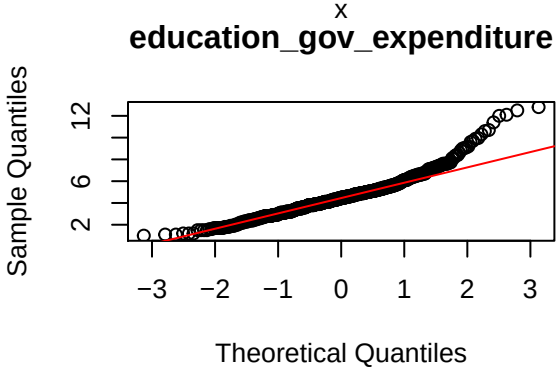

x Box_Cox
Frequency
0
150
−20
−10
0
10
20
30
bxcx

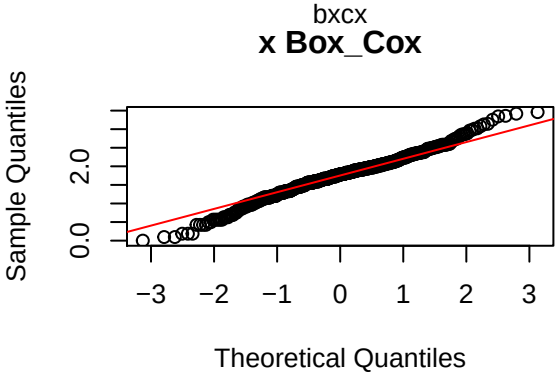

unemployment
Sample Quantiles
Theoretical Quantiles

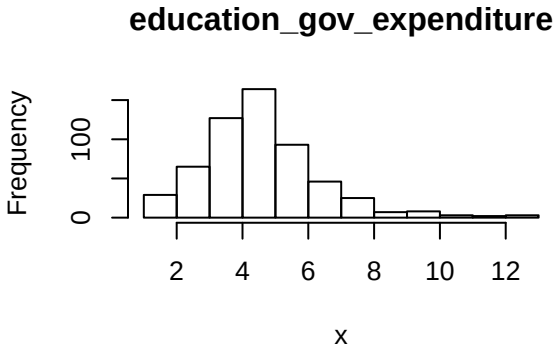

x Box_Cox
Sample Quantiles
Theoretical Quantiles

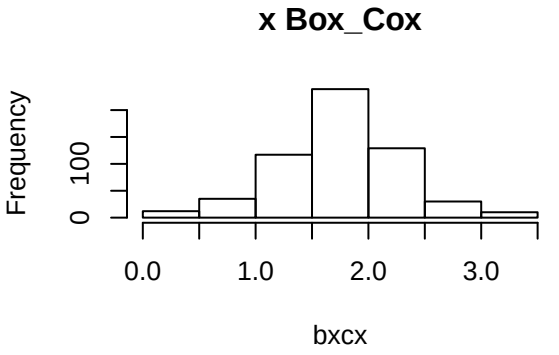

unemployment
Frequency
x

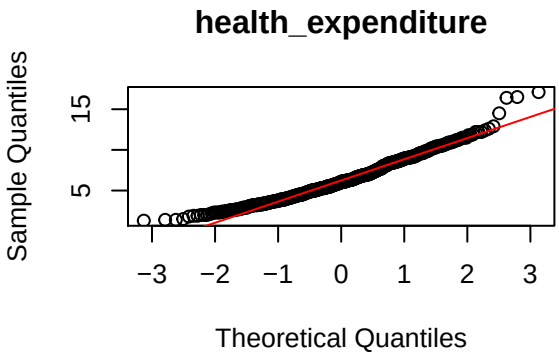

x Box_Cox
Frequency
bxcx

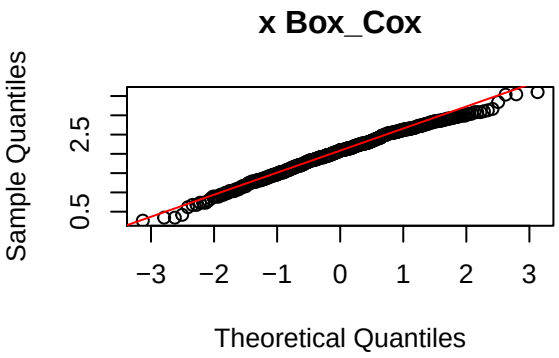

education_gov_expenditure
Sample Quantiles
Theoretical Quantiles


x Box_Cox
Sample Quantiles
Theoretical Quantiles

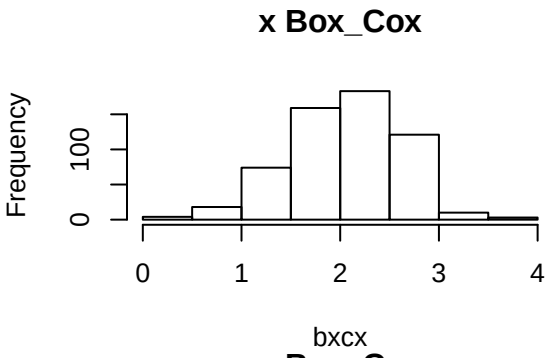

education_gov_expenditure
Frequency
x

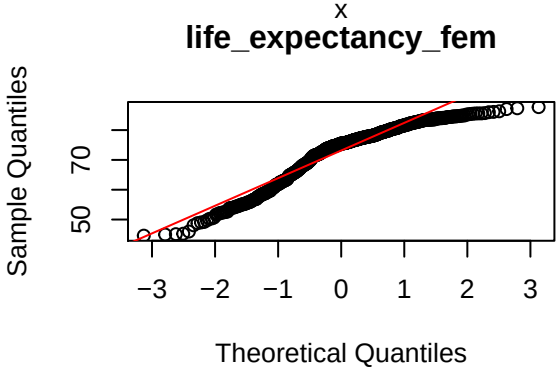

x Box_Cox
Frequency
bxcx

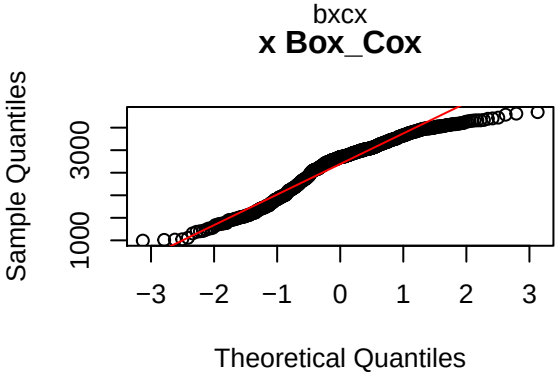

health_expenditure
Sample Quantiles
Theoretical Quantiles

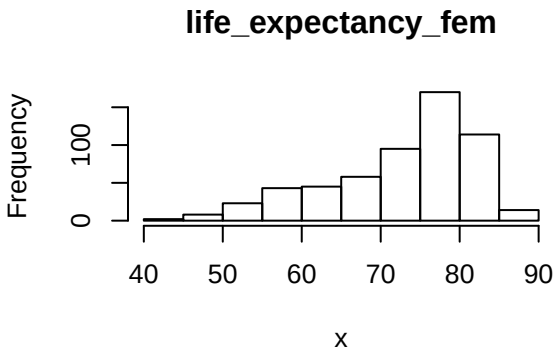

x Box_Cox
Sample Quantiles
Theoretical Quantiles

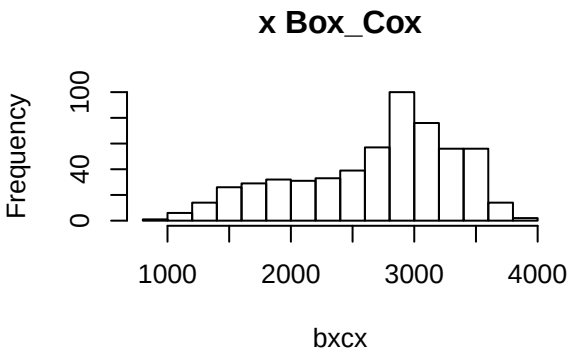

health_expenditure
Frequency
x

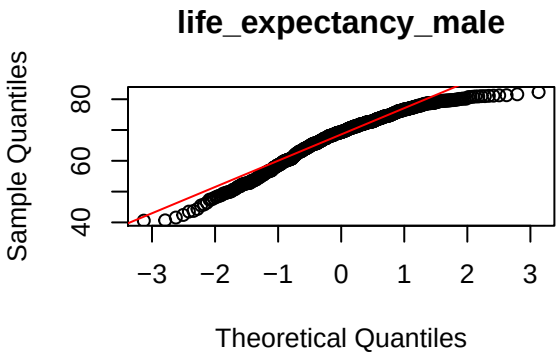

x Box_Cox
Frequency
bxcx

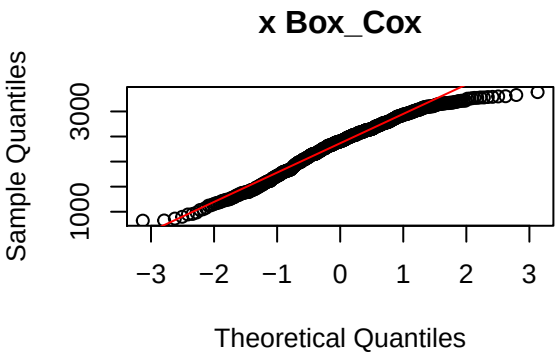

life_expectancy_fem
Sample Quantiles
Theoretical Quantiles

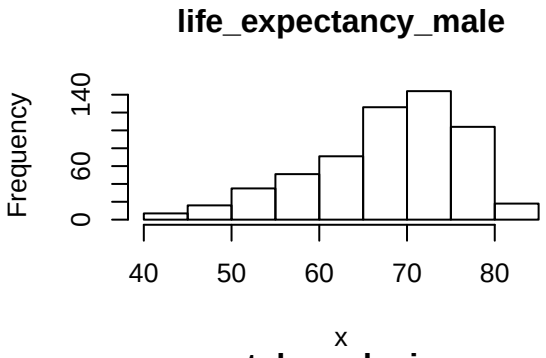

x Box_Cox
Sample Quantiles
Theoretical Quantiles

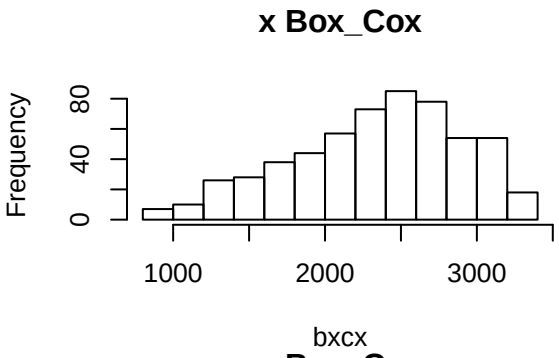

life_expectancy_fem
Frequency
x

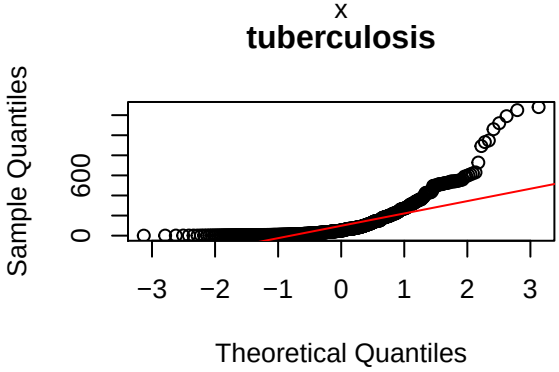

x Box_Cox
Frequency
bxcx

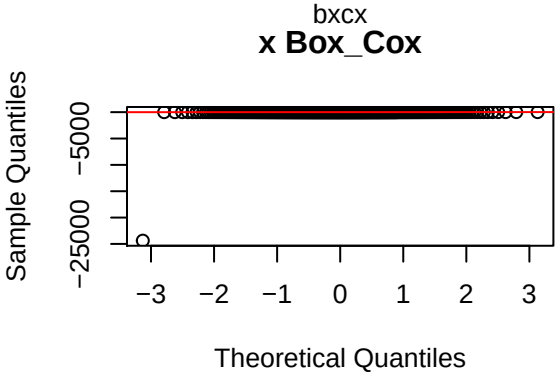

life_expectancy_male
Sample Quantiles
Theoretical Quantiles
40
60
80
−3
−2
−1
0
1
2
3

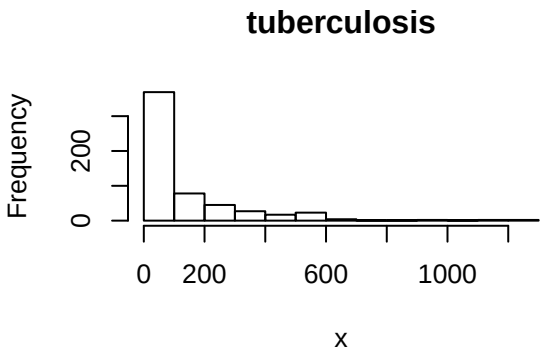

x Box_Cox
Sample Quantiles
Theoretical Quantiles
1000
3000
−3
−2
−1
0
1
2
3

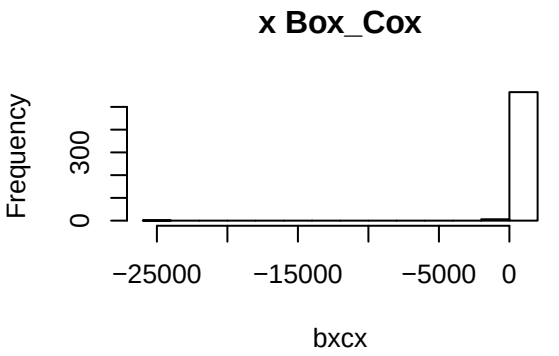

life_expectancy_male
Frequency
x
0
60
140
40
50
60
70
80

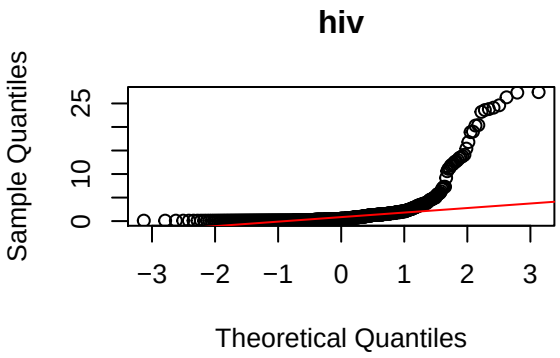

x Box_Cox
Frequency
bxcx
0
40
80
1000
2000
3000

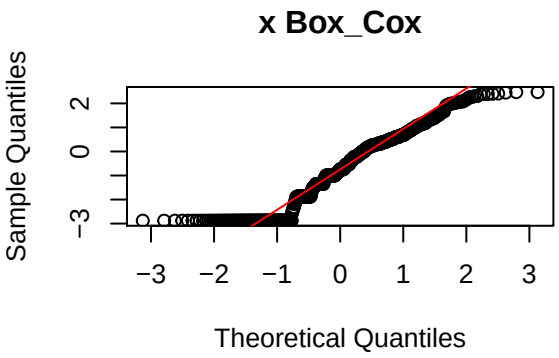

tuberculosis
Sample Quantiles
Theoretical Quantiles
0
600
−3
−2
−1
0
1
2
3

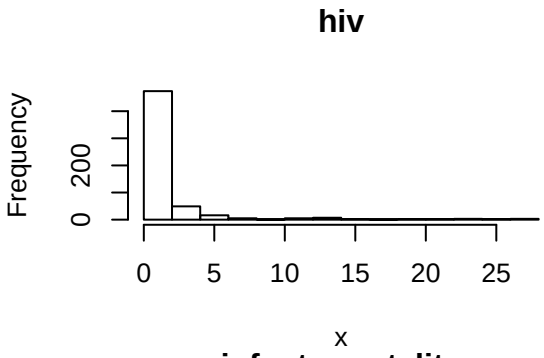

x Box_Cox
Sample Quantiles
Theoretical Quantiles
−25000
−5000
−3
−2
−1
0
1
2
3

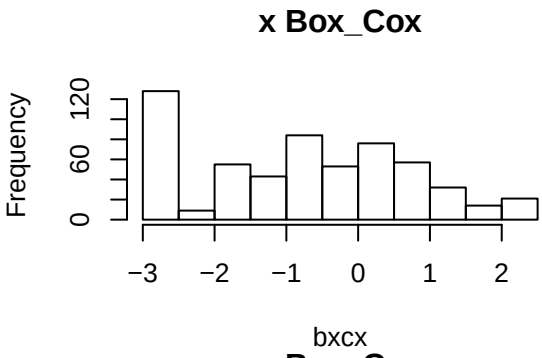

tuberculosis
Frequency
x
0
200
0
200
600
1000

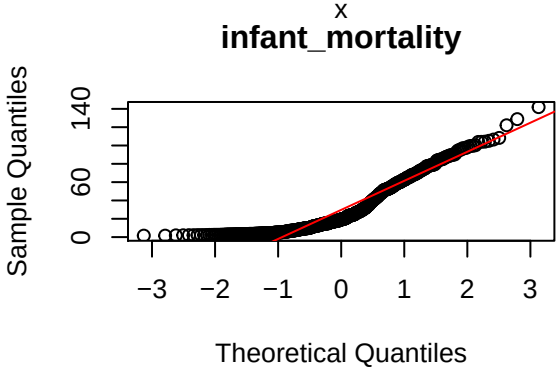

x Box_Cox
Frequency
bxcx
0
300
−25000
−15000
−5000
0

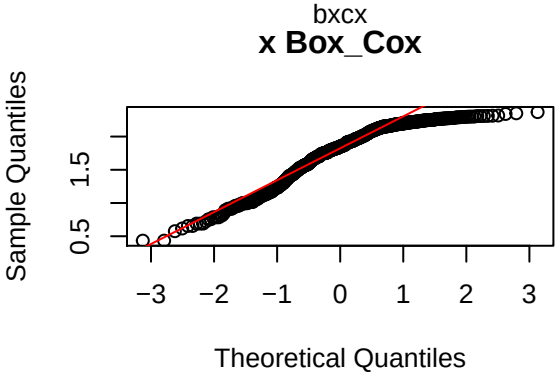

hiv
Sample Quantiles
Theoretical Quantiles

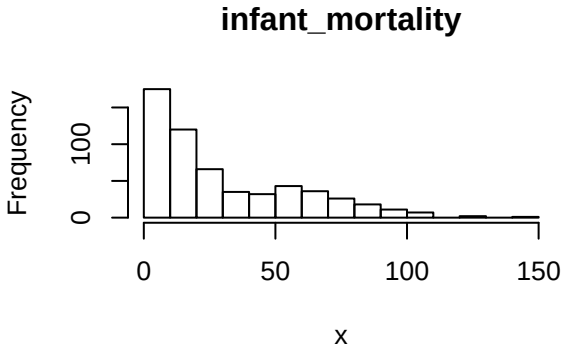

x Box_Cox
Sample Quantiles
Theoretical Quantiles

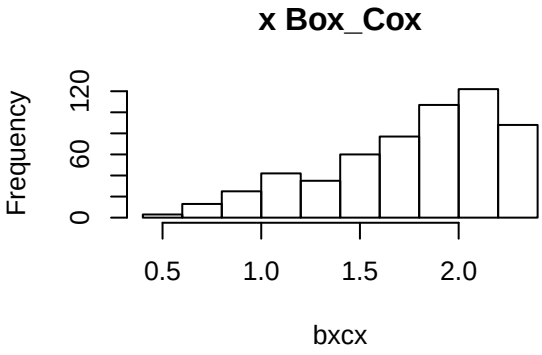

hiv
Frequency
x

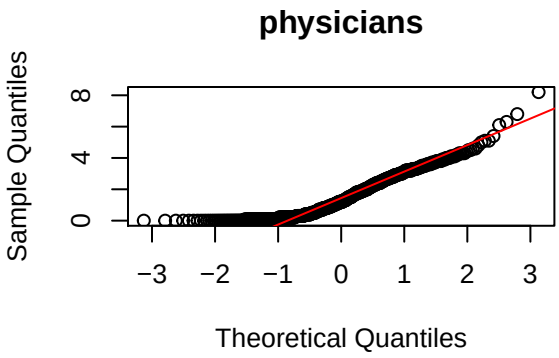

x Box_Cox
Frequency
bxcx

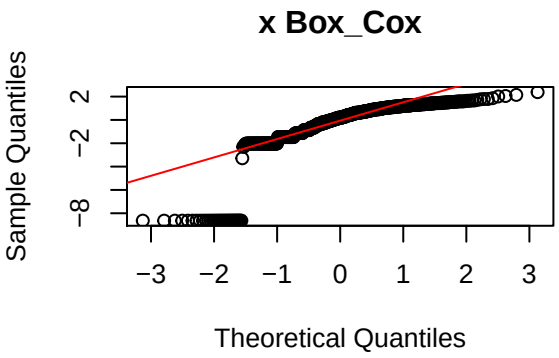

infant_mortality
Sample Quantiles
Theoretical Quantiles

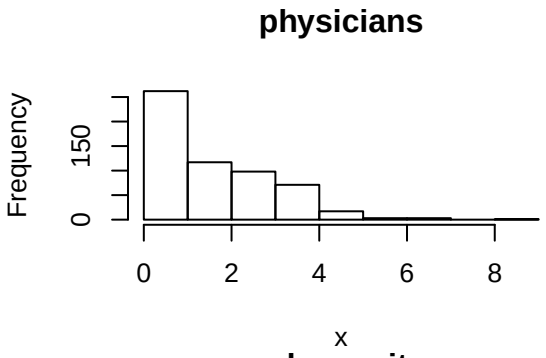

x Box_Cox
Sample Quantiles
Theoretical Quantiles

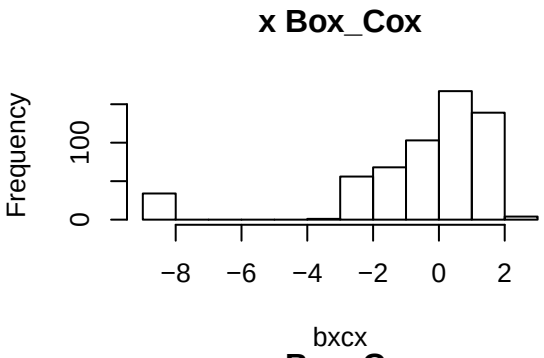

infant_mortality
Frequency
x

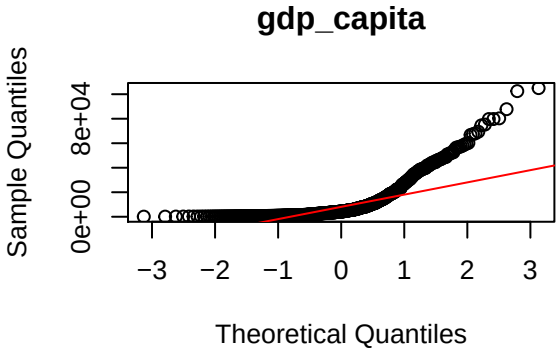

x Box_Cox
Frequency
bxcx

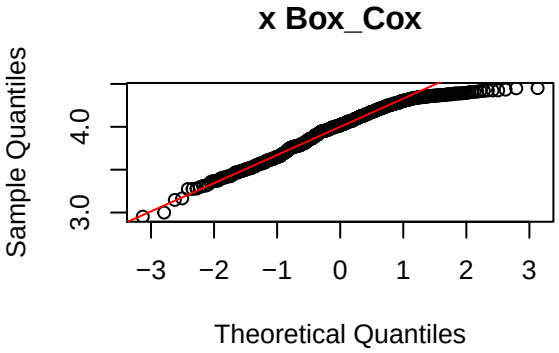

physicians
Sample Quantiles
Theoretical Quantiles
0
4
8
−3
−2
−1
0
1
2
3

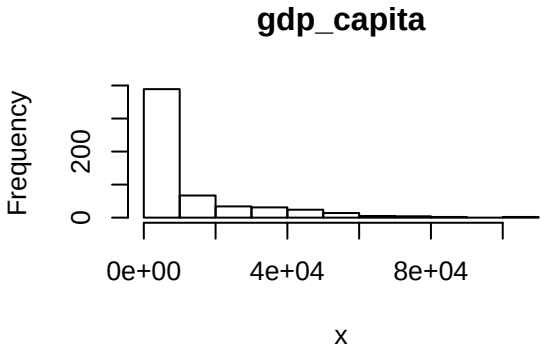

x Box_Cox
Sample Quantiles
Theoretical Quantiles
−8
−2
2
−3
−2
−1
0
1
2
3

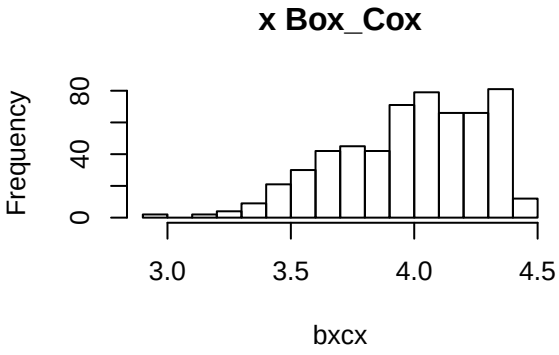

physicians
Frequency
x
0
150
0
2
4
6
8

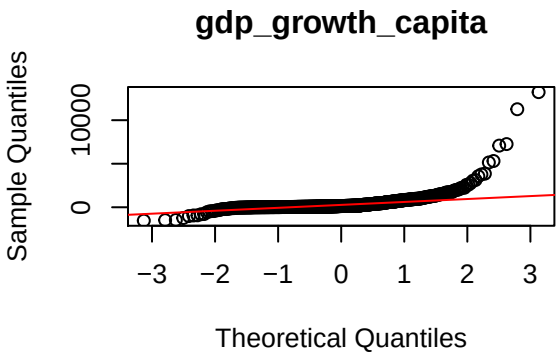

x Box_Cox
Frequency
bxcx
0
100
−8
−6
−4
−2
0
2

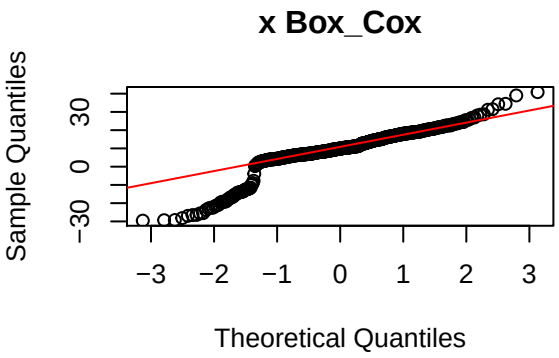

gdp_capita
Sample Quantiles
Theoretical Quantiles
0e+00
8e+04
−3
−2
−1
0
1
2
3

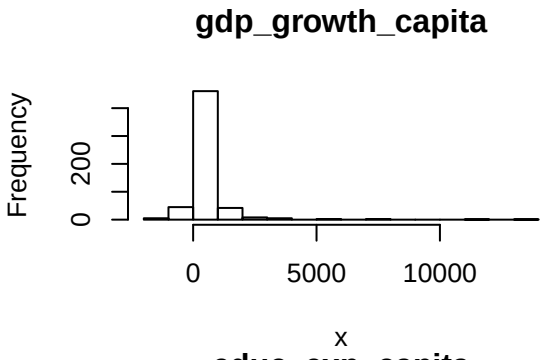

x Box_Cox
Sample Quantiles
Theoretical Quantiles
3.0
4.0
−3
−2
−1
0
1
2
3

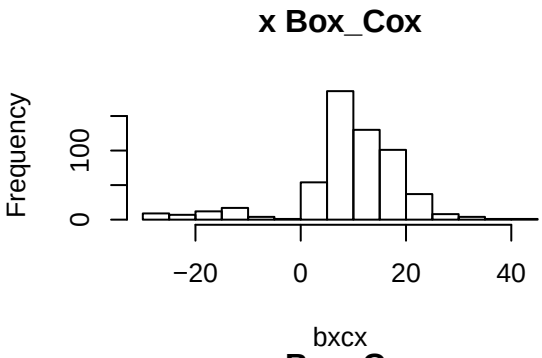

gdp_capita
Frequency
x
0
200
0e+00
4e+04
8e+04

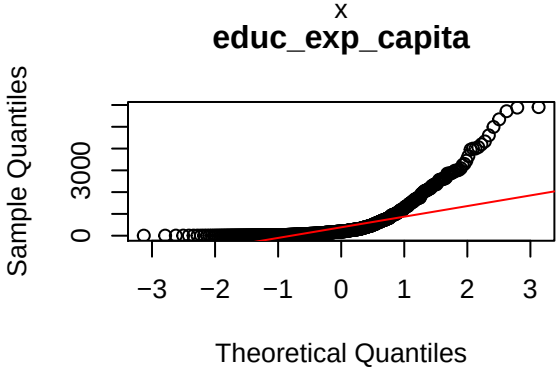

x Box_Cox
Frequency
bxcx
0
40
80
3.0
3.5
4.0
4.5

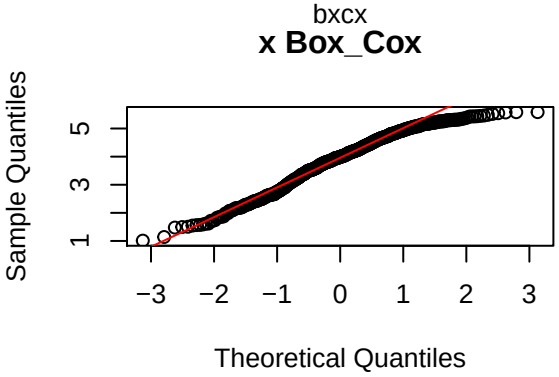

gdp_growth_capita
Sample Quantiles
Theoretical Quantiles

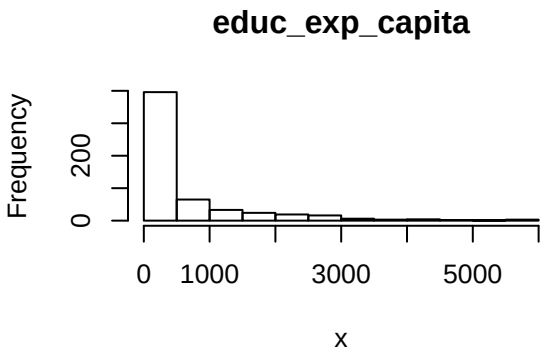

x Box_Cox
Sample Quantiles
Theoretical Quantiles

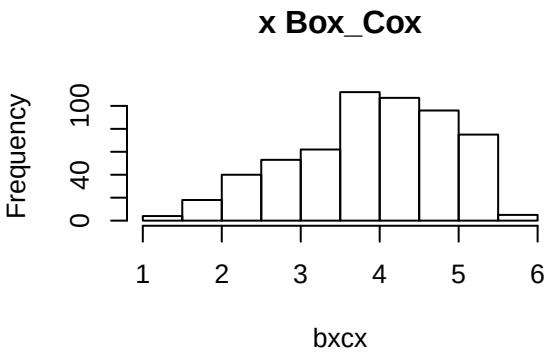

gdp_growth_capita
Frequency
x

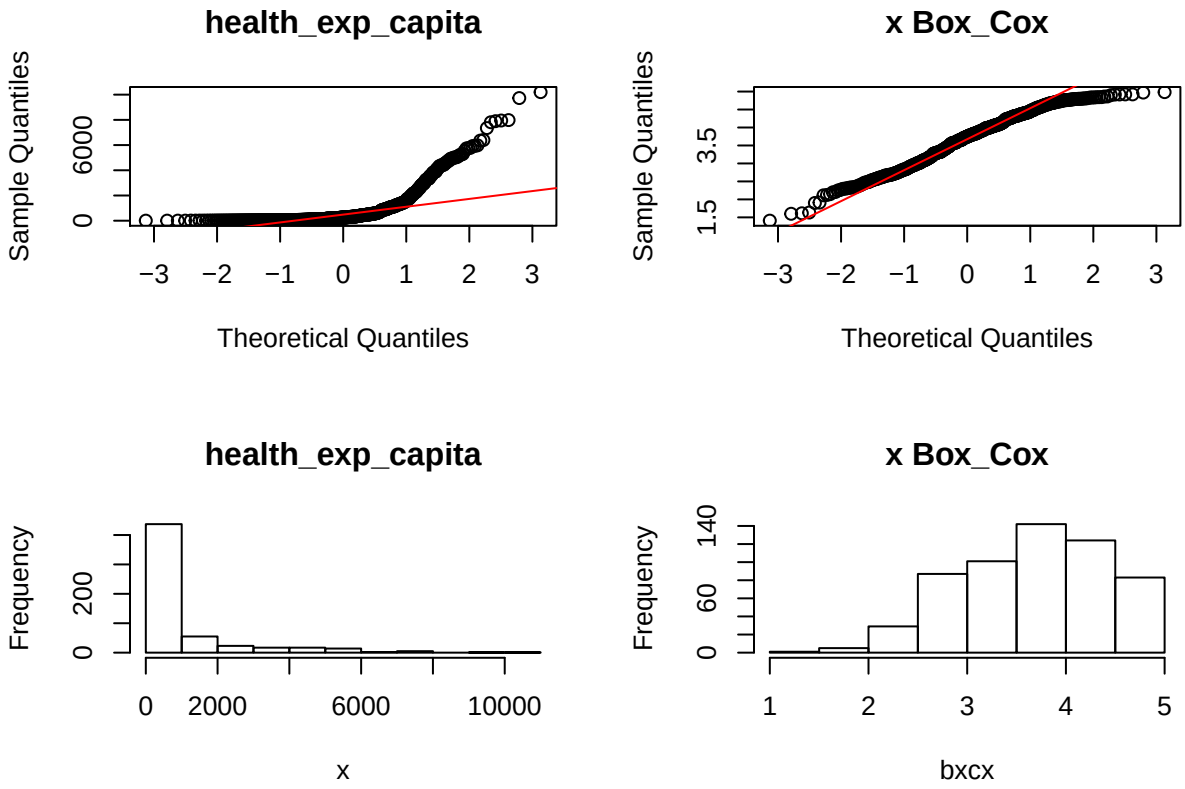

x Box_Cox
Frequency
bxcx


educ_exp_capita
Sample Quantiles
Theoretical Quantiles


x Box_Cox
Sample Quantiles
Theoretical Quantiles


educ_exp_capita
Frequency
x


x Box_Cox
Frequency
bxcx



Observant els gràfics resultats, s'intueix que les variables no semblen seguir una distribució normal, ja que els quantils de la mostra s'allunyen dels quantils teòrics de la distribució normal i els histogrames no semblen adaptar-se a una distribució normal en la majoria de casos.

Apliquem el test de Shapiro a totes les variables transformades per veure si alguna segueix una distribució normal un cop aplicada la transformació Box-Cox:

```
test_box_cox <- function(x){
  tryCatch({
    y <- (BoxCox(x,lambda=BoxCoxLambda(x)))
    shapiro.test(y)
  }, error=function(cond) {
      return(NA)})
}

lshap <- sapply(health_data[,4:ncol(health_data)], test_box_cox)

lshap.df <-  as.data.frame(sapply(lshap[!is.na(lshap)], function(x) {as.character(x$p.value)}))
colnames(lshap.df) <- c("p-value")

lshap.df
```

	p-value
population	5.33231572943436e-10
population_grow	1.66986864121624e-12
population_under_14	3.30511182204066e-18
population_above_65	7.29195365020304e-17
gdp	1.18414632276054e-06
gdp_growth_rate	1.85846405934116e-18
unemployment	1.14074265042466e-09
education_gov_expenditure	5.37877587612958e-05
health_expenditure	0.00774826569829806
life_expectancy_fem	4.84845748496625e-13
life_expectancy_male	1.35325818982158e-08
tuberculosis	8.59559045008251e-47
hiv	8.72364995745125e-15
infant_mortality	1.42666788278048e-16
physicians	6.4729166592922e-31
gdp_capita	2.09368428262233e-11
gdp_growth_capita	2.48212564749547e-21
educ_exp_capita	1.68697964091817e-09
health_exp_capita	5.95948979095356e-08

La prova de Shapiro-Wilk dona com a resultat un p-valor menor al nivell de significació alpha = 0.05 i es rebutja la hipòtesi nul · la. Podem concloure que la transformació de Box-Cox no ens permet obtenir distribucions normals de les variables transformades.

Provem d'aplicar altres transformacions, com ara una transformació amb logaritme natural ln(x+1) a aquelles variables que no prenen valors negatius.

```
# Transformació amb logaritme natural ln(x+1)
# Apliquem de nou la prova de Shapiro-Wilk

alpha <- 0.05
col.names = colnames(health_data)

df1 <- data.frame()

for (i in 1:ncol(health_data)) {
  if (is.numeric(health_data[,i])) {

    p_val <- shapiro.test(log(health_data[,i] + 1))$p.value

    df1 <- rbind(df1,data.frame(variable = col.names[i],p.value = p_val))

  }
}
```

transformació logaritme natural ln(x+1)

variable	p.value
year	1.93390306054842e-27
population	1.44082311399075e-05
population_grow	3.70247085819278e-18
population_under_14	4.13120209437311e-14
population_above_65	1.21859150365859e-13
gdp	0.022874218863429
gdp_growth_rate	1.61969476980565e-18
unemployment	0.0667112002176143
education_gov_expenditure	0.000123839689027428
health_expenditure	0.00733962110012018
life_expectancy_fem	2.08590783070286e-19
life_expectancy_male	1.49295649564815e-16
tuberculosis	5.7748123202955e-07
hiv	4.02228403468883e-27
infant_mortality	6.67178147791763e-11
physicians	1.31487675046107e-14
gdp_capita	5.39532011884959e-07
gdp_growth_capita	4.42221519831865e-05
educ_exp_capita	2.98897434524832e-06
health_exp_capita	2.76810192563359e-07

Veiem que la transformació amb ln(x+1) només permet obtenir una distribució normal per **unemployment**, però aquesta és una variable que es relaciona poc amb les variables de salut d'interès, com veurem mitjançant els testos de correlació.

Per últim, provem les transformacions x^2 $x^{0.5}$ $\frac{1}{x}$

```
# Transformacions amb x^2, x^0.5 i 1/x
# Apliquem de nou la prova de Shapiro-Wilk

alpha <- 0.05
col.names = colnames(health_data)

df1 <- data.frame()

for (i in 4:ncol(health_data)) {
  if (is.numeric(health_data[,i])) {

    p_val_2 = shapiro.test((health_data[,i])^2)$p.value
    p_val_05 = shapiro.test((health_data[,i])^0.5)$p.value
    p_val_frac = shapiro.test(1/(health_data[,i]))$p.value

    df1 <- rbind(df1,data.frame(variable = col.names[i],p.valor2 = p_val_2,
                                p.valor05 = p_val_05,p.valorfrac = p_val_frac))


  }
}
```

transformacions x^2, $x^{0.5}$ i $\frac{1}{x}$ (p-valor)

	x^2	$x^{0.5}$	$\frac{1}{x}$
population	1.11174052061957e-45	4.45250112196058e-33	4.95062120162108e-41
population_grow	1.19273119673814e-42	2.48416049313479e-09	NaN
population_under_14	6.39247783130799e-18	1.30663283531676e-13	3.30219479433097e-18
population_above_65	8.36308383367867e-29	9.94559098418343e-17	3.67221730448156e-26
gdp	6.28832556782952e-46	1.19071249162523e-34	9.2626891752471e-41
gdp_growth_rate	7.74887206327507e-41	3.55005875741414e-07	NaN
unemployment	1.84226857300485e-34	3.40049125297959e-07	1.76142712700465e-42
education_gov_expenditure	6.16431515033913e-29	7.20890762866251e-06	4.04125739743022e-26
health_expenditure	5.65863434182227e-24	0.229011008725573	2.48790577424472e-24
life_expectancy_fem	4.85093340643569e-13	8.50879099847691e-18	9.5198069936323e-23
life_expectancy_male	1.35401317250285e-08	1.86569784343562e-14	6.44510686223952e-21
tuberculosis	5.61853111069281e-41	8.24532828415412e-20	NaN
hiv	2.21631192918501e-43	4.58271028174038e-30	3.96570678656274e-27
infant_mortality	1.29027055053794e-31	4.1518111359678e-14	3.4421312673684e-28
physicians	8.22647624544166e-32	3.51079225985604e-10	NaN
gdp_capita	6.03975566470943e-39	3.79911193330601e-21	6.33206999224329e-35
gdp_growth_capita	1.55128829953758e-45	1.21500239336428e-24	NaN
educ_exp_capita	1.81288907367127e-40	5.97249558166292e-22	3.8347813239709e-36
health_exp_capita	1.09895840140002e-40	6.6724267317477e-25	6.56938396168317e-35

Les transformacions x^2 i $\frac{1}{x}$ tampoc no ens permeten obtenir distribucions normals de les variables transformades, i amb $x^{0.5}$ només es podria normalitzar **health_expenditure**.

Per tant, haurem de considerar que les variables no segueixen una distribució normal, ni tan sols transformades, de manera que haurem d'aplicar tests no paramètrics.

Comprovació de la normalitat de les variables life_expectancy_fem/male segons les categories de despesa sanitària.

Comprovem amb la prova de Shapiro-Wilk la normalitat de la variable life_expectancy per cada grup.

```
shapiro.test(cat_very_low_fem)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cat_very_low_fem
## W = 0.99364, p-value = 0.8805
```

```
shapiro.test(cat_low_fem)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cat_low_fem
## W = 0.92715, p-value = 1.049e-05
```

```r
shapiro.test(cat_medium_fem)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cat_medium_fem
## W = 0.70846, p-value = 8.376e-14
```

```r
shapiro.test(cat_high_fem)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cat_high_fem
## W = 0.7209, p-value = 2.007e-13
```

```r
shapiro.test(cat_very_high_fem)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cat_very_high_fem
## W = 0.94056, p-value = 7.242e-05
```

```r
shapiro.test(cat_very_low_male)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cat_very_low_male
## W = 0.98253, p-value = 0.1399
```

```r
shapiro.test(cat_low_male)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cat_low_male
## W = 0.91903, p-value = 3.558e-06
```

```r
shapiro.test(cat_medium_male)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cat_medium_male
## W = 0.81849, p-value = 1.366e-10
```

```r
shapiro.test(cat_high_male)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cat_high_male
## W = 0.85014, p-value = 2.236e-09
```

```r
shapiro.test(cat_very_high_male)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cat_very_high_male
## W = 0.94746, p-value = 0.0002126
```

La prova de Shapiro-Wilk dona com a resultat un p-valor major al nivell de significació alpha = 0.05 pels subgrups cat_very_low_fem i cat_very_low_male, de manera que es pot considerar que segueixen una distribució normal. Però la resta de subgrups no segueixen una distribució normal. Per tant, si fem comparacions entre la totalitat dels grups haurem de considerar proves no paramètriques.

Comprovació de la homoscedasticitat entre les variables life_expectancy_fem/male segons les categories de despesa sanitària.

Per comprovar la homoscedasticitat entre diversos grups de dades, atès que les dades no compleixen la condició de normalitat, apliquem proves no paramètriques:

```r
# Comprovació de homoscedasticitat entre les variables life_expectancy_fem i
# life_expectancy_male segons els diferents grups de despesa sanitària pública.

fligner.test(life_expectancy_fem ~ health_expend_category, data=health_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  life_expectancy_fem by health_expend_category
## Fligner-Killeen:med chi-squared = 93.664, df = 4, p-value < 2.2e-16
```

```r
fligner.test(life_expectancy_male ~ health_expend_category, data=health_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  life_expectancy_male by health_expend_category
## Fligner-Killeen:med chi-squared = 59.72, df = 4, p-value = 3.322e-12
```

Del test de Fligner-Killeen obtenim un p_valor < 0.05, de manera que es conclou que les variables life_expectancy_fem i life_expectancy_male presenten variànces estadísticament diferents segons els diferents grups de despesa sanitària pública.

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.

Comparació entre distribucions de les variables life_expectancy_fem/male entre els diferents grups segons despesa pública sanitària per càpita.

Un cop comprovat que les variables no cumpleixen les suposicions per l'aplicació de tests paramètrics a la secció 4.2, s'implementa el test de Wilcoxon per comprovar la similitud de distribucions entre alguns dels grups, dos a dos.

```
wilcox.test(life_expectancy_fem ~ health_expend_category, data=health_data,
            subset=health_expend_category %in% c("very low", "low"))
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  life_expectancy_fem by health_expend_category
## W = 2777, p-value = 4.852e-14
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(life_expectancy_male ~ health_expend_category, data=health_data,
            subset=health_expend_category %in% c("very low", "low"))
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  life_expectancy_male by health_expend_category
## W = 2868, p-value = 1.919e-13
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(life_expectancy_fem ~ health_expend_category, data=health_data,
            subset=health_expend_category %in% c("high", "very high"))
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  life_expectancy_fem by health_expend_category
## W = 1120.5, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(life_expectancy_male ~ health_expend_category, data=health_data,
            subset=health_expend_category %in% c("high", "very high"))
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  life_expectancy_male by health_expend_category
## W = 1304.5, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

Amb el test de Krustal_Wallis comprovarem la similitud de distribucions entre tots els grups.

```r
kruskal.test(life_expectancy_fem ~ health_expend_category, data=health_data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  life_expectancy_fem by health_expend_category
## Kruskal-Wallis chi-squared = 418.07, df = 4, p-value < 2.2e-16
```

```r
kruskal.test(life_expectancy_male ~ health_expend_category, data=health_data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  life_expectancy_male by health_expend_category
## Kruskal-Wallis chi-squared = 398.48, df = 4, p-value < 2.2e-16
```

S'observa doncs una diferència estadísticament significativa entre les distribucions d'esperança de vida per diferents rangs de despesa sanitària pública. La mitja de l'esperança de vida és notablement major quan major és la despesa sanitària pública, i la variància es redueix, fent disminuir la dispersió de l'esperança de vida entre individus.

Observem que es podria fer una anàlisi en més profunditat si a part dels 5 grups creats segons el rang de despesa sanitària per cada sexe, també apliquéssim alhora una divisió segons els 3 anys dels quals tenim dades (2005, 2010 i 2019). Amb això es podria comprovar si cadascun dels grups segons despesa i sexe ha patit algun tipus d'evolució durant els 3 anys, o bé si l'esperança de vida ha quedat estancada.

Comprovació de la correlació entre variables del dataset - Test de correlació

A la secció 4.2 hem comprovat que cap de les variables del dataset segueix una distribució normal, de manera que basarem l'anàlisi de correlacions en el mètode de Spearman, que és un mètode no paramètric:

```r
corr_matrix <- rcorr(as.matrix(health_data[, !(colnames(health_data) %in% c("country_code",
      "country", "year", "health_expend_category"))]), type="spearman")

# En millorem la seva visualització
flat_cor_mat <- function(cor_r, cor_p){
  cor_r <- rownames_to_column(as.data.frame(cor_r), var = "row")
  cor_r <- gather(cor_r, column, cor, -1)
  cor_p <- rownames_to_column(as.data.frame(cor_p), var = "row")
  cor_p <- gather(cor_p, column, p, -1)
  cor_p_matrix <- left_join(cor_r, cor_p, by = c("row", "column"))
  cor_p_matrix
}

corr_matrix_flat <- flat_cor_mat(corr_matrix$r, corr_matrix$P)
```

Per aquelles relacions entre parells de variables on el p-valor és menor que el nivell de significació alpha = 0.05, l'hipòtesi nul · la és rebutjada i per tant es pot afirmar que el coeficient de correlació és significativament diferent de 0, i per tant, existeix una certa correlació entre els parells de variables, i aquest nivell de correlació es descriu mitjançant el coeficient de correlació obtingut.

Filtrem doncs la taula de correlacions en funció de si el p-valor és inferior a 0.05 i a la vegada si el coeficient de correlació és major a 0.7 o menor a -0.7, per tal d'obtenir un cert nivell mínim de correlació entre parells de variables:

```r
# Parells de variables on el p-valor és inferior a 0.05 i el coefficient de correlació
# és major a 0.7 o menor a -0.7.

corr_matrix_filtered <- corr_matrix_flat[which((corr_matrix_flat$p < 0.05) &
                    ((corr_matrix_flat$cor >= 0.7) | (corr_matrix_flat$cor <= -0.7))),]
```

	row	column	cor	p
5	gdp	population	0.7018366	0
22	population_under_14	population_grow	0.7214206	0
23	population_above_65	population_grow	-0.8177937	0
40	population_grow	population_under_14	0.7214206	0
42	population_above_65	population_under_14	-0.8660948	0
48	life_expectancy_fem	population_under_14	-0.8670608	0
49	life_expectancy_male	population_under_14	-0.7981425	0
52	infant_mortality	population_under_14	0.8911321	0
53	physicians	population_under_14	-0.8513685	0
54	gdp_capita	population_under_14	-0.8003325	0
56	educ_exp_capita	population_under_14	-0.7888940	0
57	health_exp_capita	population_under_14	-0.8177001	0
59	population_grow	population_above_65	-0.8177937	0
60	population_under_14	population_above_65	-0.8660948	0
67	life_expectancy_fem	population_above_65	0.8079803	0
68	life_expectancy_male	population_above_65	0.7069169	0
71	infant_mortality	population_above_65	-0.8002756	0
72	physicians	population_above_65	0.8103749	0
76	health_exp_capita	population_above_65	0.7256508	0
77	population	gdp	0.7018366	0
174	population_under_14	life_expectancy_fem	-0.8670608	0
175	population_above_65	life_expectancy_fem	0.8079803	0
182	life_expectancy_male	life_expectancy_fem	0.9574130	0
183	tuberculosis	life_expectancy_fem	-0.7618016	0
185	infant_mortality	life_expectancy_fem	-0.9506028	0
186	physicians	life_expectancy_fem	0.8228037	0
187	gdp_capita	life_expectancy_fem	0.8437640	0
189	educ_exp_capita	life_expectancy_fem	0.8263818	0
190	health_exp_capita	life_expectancy_fem	0.8648388	0
193	population_under_14	life_expectancy_male	-0.7981425	0
194	population_above_65	life_expectancy_male	0.7069169	0
200	life_expectancy_fem	life_expectancy_male	0.9574130	0
202	tuberculosis	life_expectancy_male	-0.7980531	0
204	infant_mortality	life_expectancy_male	-0.9138937	0
205	physicians	life_expectancy_male	0.7438026	0
206	gdp_capita	life_expectancy_male	0.8360465	0
208	educ_exp_capita	life_expectancy_male	0.8182075	0
209	health_exp_capita	life_expectancy_male	0.8441780	0
219	life_expectancy_fem	tuberculosis	-0.7618016	0
220	life_expectancy_male	tuberculosis	-0.7980531	0
223	infant_mortality	tuberculosis	0.7708805	0
225	gdp_capita	tuberculosis	-0.7299326	0

(continued)

	row	column	cor	p
227	educ_exp_capita	tuberculosis	-0.7356304	0
228	health_exp_capita	tuberculosis	-0.7506836	0
250	population_under_14	infant_mortality	0.8911321	0
251	population_above_65	infant_mortality	-0.8002756	0
257	life_expectancy_fem	infant_mortality	-0.9506028	0
258	life_expectancy_male	infant_mortality	-0.9138937	0
259	tuberculosis	infant_mortality	0.7708805	0
262	physicians	infant_mortality	-0.8271644	0
263	gdp_capita	infant_mortality	-0.8684935	0
265	educ_exp_capita	infant_mortality	-0.8639795	0
266	health_exp_capita	infant_mortality	-0.8826174	0
269	population_under_14	physicians	-0.8513685	0
270	population_above_65	physicians	0.8103749	0
276	life_expectancy_fem	physicians	0.8228037	0
277	life_expectancy_male	physicians	0.7438026	0
280	infant_mortality	physicians	-0.8271644	0
282	gdp_capita	physicians	0.7424032	0
284	educ_exp_capita	physicians	0.7305911	0
285	health_exp_capita	physicians	0.7767078	0
288	population_under_14	gdp_capita	-0.8003325	0
295	life_expectancy_fem	gdp_capita	0.8437640	0
296	life_expectancy_male	gdp_capita	0.8360465	0
297	tuberculosis	gdp_capita	-0.7299326	0
299	infant_mortality	gdp_capita	-0.8684935	0
300	physicians	gdp_capita	0.7424032	0
302	gdp_growth_capita	gdp_capita	0.7177587	0
303	educ_exp_capita	gdp_capita	0.9771321	0
304	health_exp_capita	gdp_capita	0.9711640	0
320	gdp_capita	gdp_growth_capita	0.7177587	0
326	population_under_14	educ_exp_capita	-0.7888940	0
333	life_expectancy_fem	educ_exp_capita	0.8263818	0
334	life_expectancy_male	educ_exp_capita	0.8182075	0
335	tuberculosis	educ_exp_capita	-0.7356304	0
337	infant_mortality	educ_exp_capita	-0.8639795	0
338	physicians	educ_exp_capita	0.7305911	0
339	gdp_capita	educ_exp_capita	0.9771321	0
342	health_exp_capita	educ_exp_capita	0.9665048	0
345	population_under_14	health_exp_capita	-0.8177001	0
346	population_above_65	health_exp_capita	0.7256508	0
352	life_expectancy_fem	health_exp_capita	0.8648388	0
353	life_expectancy_male	health_exp_capita	0.8441780	0
354	tuberculosis	health_exp_capita	-0.7506836	0
356	infant_mortality	health_exp_capita	-0.8826174	0
357	physicians	health_exp_capita	0.7767078	0
358	gdp_capita	health_exp_capita	0.9711640	0
360	educ_exp_capita	health_exp_capita	0.9665048	0

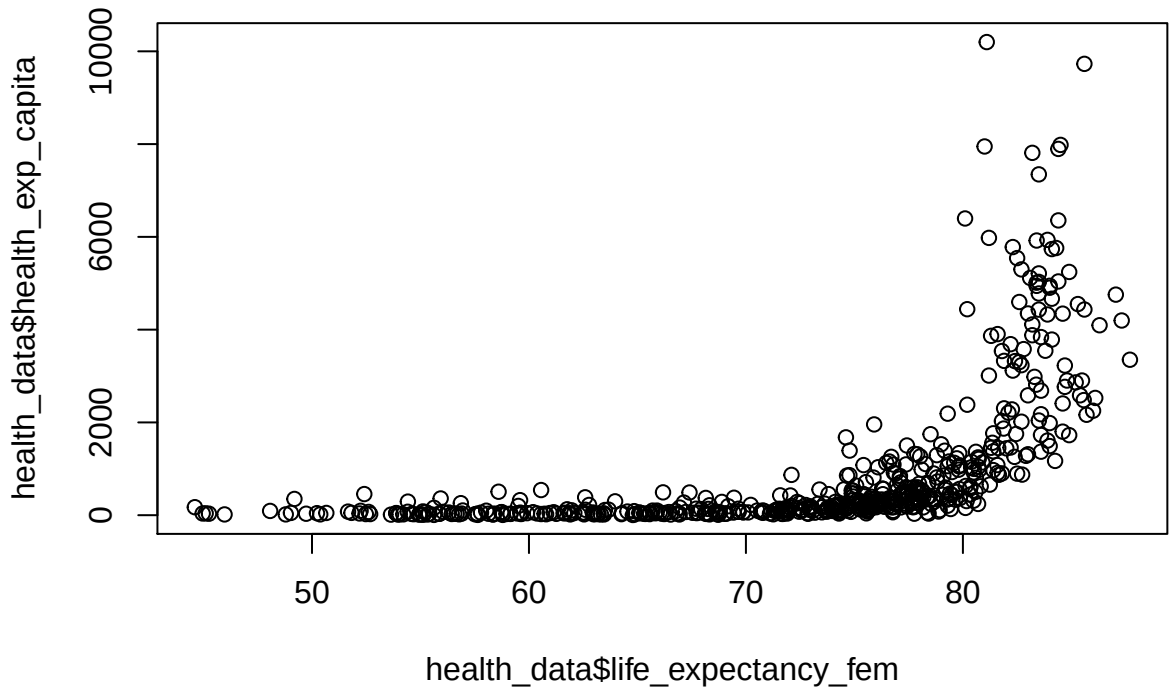
D'aquesta anàlisi obtenim diferents correlacions d'interès, que queden descrites a la secció 6.

Entre elles, veiem que existeix una relació significativa entre major health_exp_capita, educ_exp_capita i physicians, i una major life_expectancy_fem/male. I a major proporció de population_under_14 i infant_mortality, menor life_expectancy_fem.

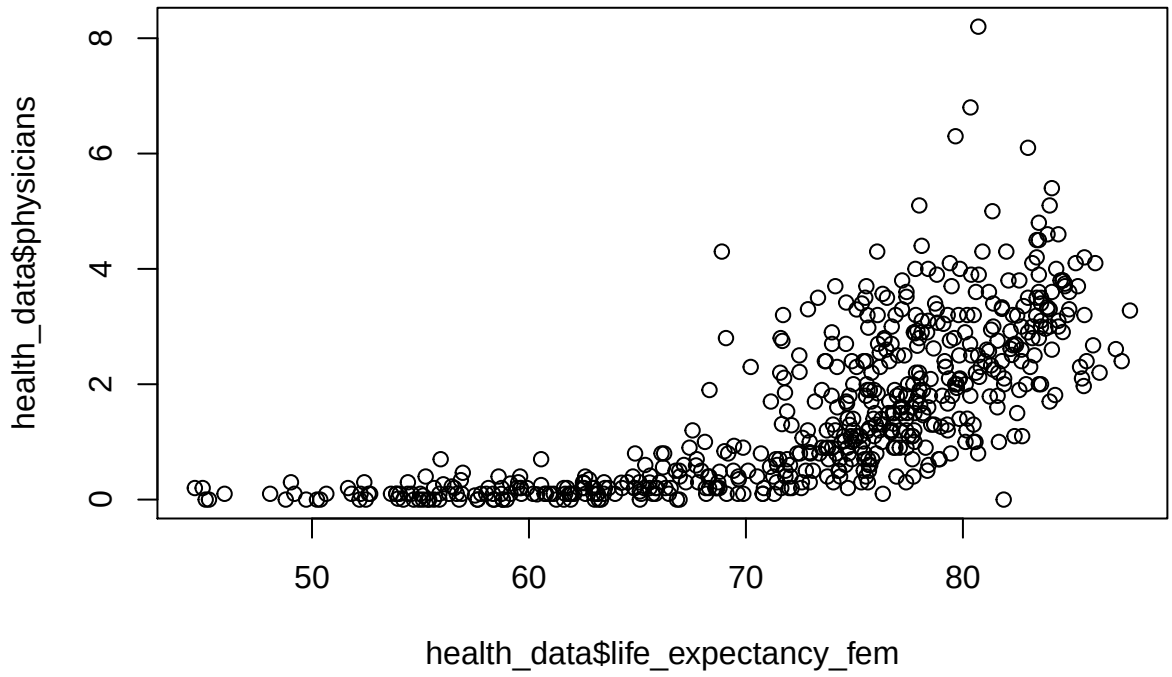
Model de Regressió

Analitzar aquestes correlacions ens permet implementar un model de regressió, que permeti estimar el life_expectancy_fem en funció de nous valors per les variables "physicians", "health_exp_capita" i "infant_mortality" (que són variables sobre les que els governs poden actuar directament).

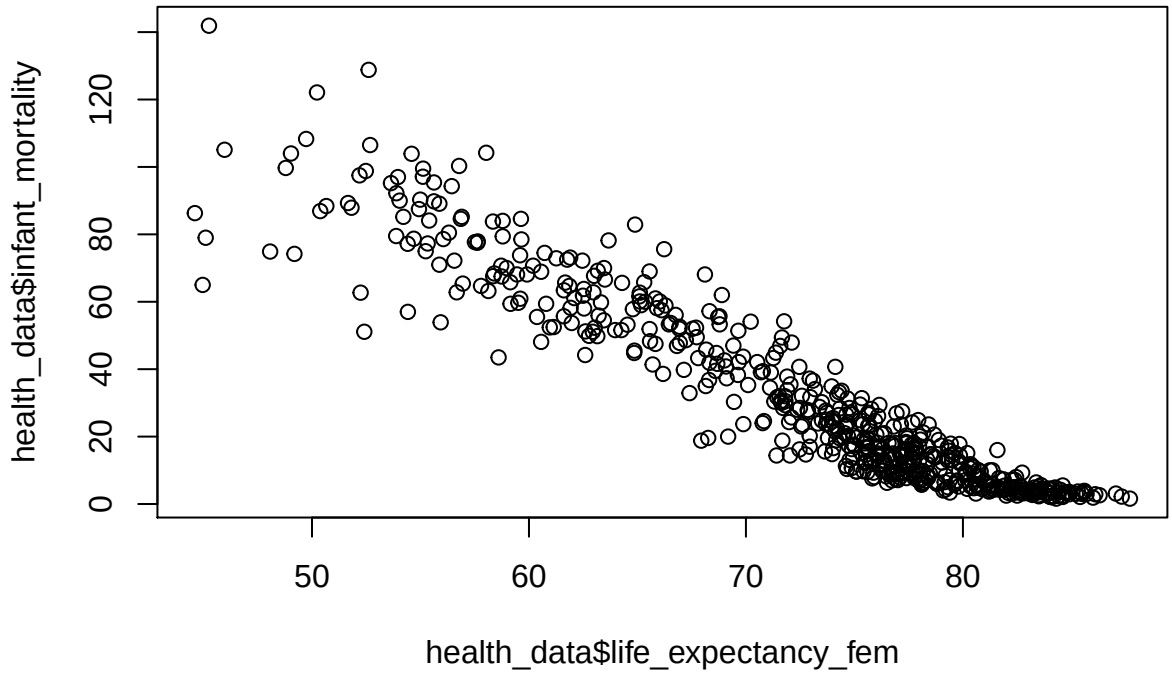
```
plot(health_data$life_expectancy_fem, health_data$health_exp_capita)
```




```
plot(health_data$life_expectancy_fem, health_data$physicians)
```




```r
plot(health_data$life_expectancy_fem, health_data$infant_mortality)
```




```r
health_exp_cap = health_data$health_exp_capita
physn = health_data$physicians
inf_mort = health_data$infant_mortality

# Variable a predir:
life_exp_fem = health_data$life_expectancy_fem

ml <- lm(life_exp_fem ~
```

```
            health_exp_cap
          + physn
          + I(physn^2)
          + inf_mort
        , data=health_data)
summary(ml)
```

```
##
## Call:
## lm(formula = life_exp_fem ~ health_exp_cap + physn + I(physn^2) +
##     inf_mort, data = health_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.6456  -1.4927   0.2127   1.8213   8.2361
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     7.864e+01  4.546e-01 172.997  < 2e-16 ***
## health_exp_cap  5.346e-04  9.154e-05   5.840 8.79e-09 ***
## physn           1.775e+00  3.022e-01   5.873 7.30e-09 ***
## I(physn^2)     -2.316e-01  5.282e-02  -4.384 1.39e-05 ***
## inf_mort       -2.677e-01  6.622e-03 -40.423  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.792 on 567 degrees of freedom
## Multiple R-squared:  0.9109, Adjusted R-squared:  0.9102
## F-statistic:  1448 on 4 and 567 DF,  p-value: < 2.2e-16
```

S'obté un model amb un coeficient de determinació R^2 de 0.9109, cosa que podem considerar acceptable.

Aquest model permet predir l'esperança de vida femenina establint nous valors per les variables "health_exp_capita", "physicians" i "infant_mortality".

```
# Indicadors de partida
newdata <- data.frame(
  health_exp_cap = 250,
  physn = 1.2,
  inf_mort = 19
  )

# Predicció de life_expectancy_fem
predict(ml, newdata)
```

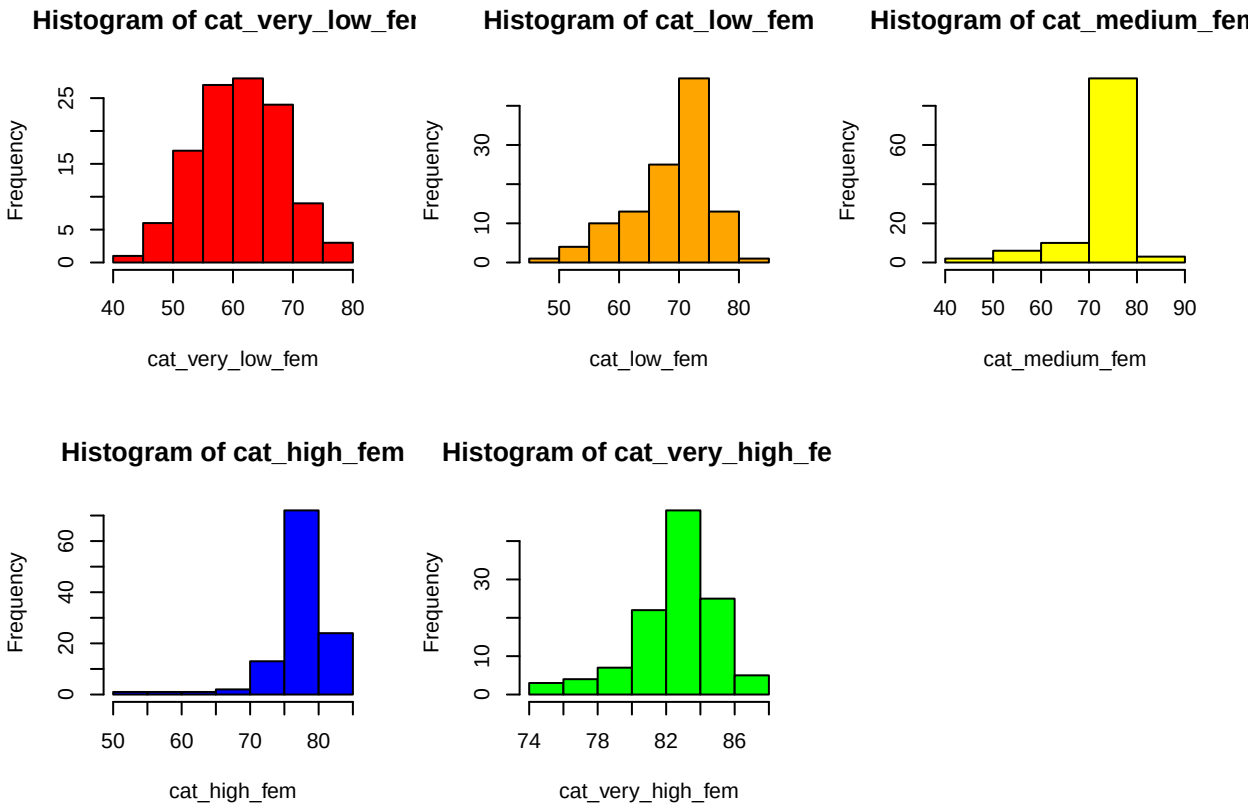
```
##        1
## 75.48384
```

Amb aquest model podem observar com amb lleus millores dels 3 paràmetres "health_exp_capita", "physicians" i "infant_mortality", s'obtenen ràpidament millores notables en l'esperança de vida femenina. Per exemple, en la predicció feta anteriorment observem que fixant valors d'aquests paràmetres propers a les medianes de cada variable en qüestió, obtenim una esperança de vida superior a la mediana (75.48734 years > median(life_expectancy_fem) = 72.70 years). Degut a que l'esperança de vida masculina té una elevada correlació amb la femenina, es comprovaria el mateix efecte si analitzessim un model per life_expectancy_male.

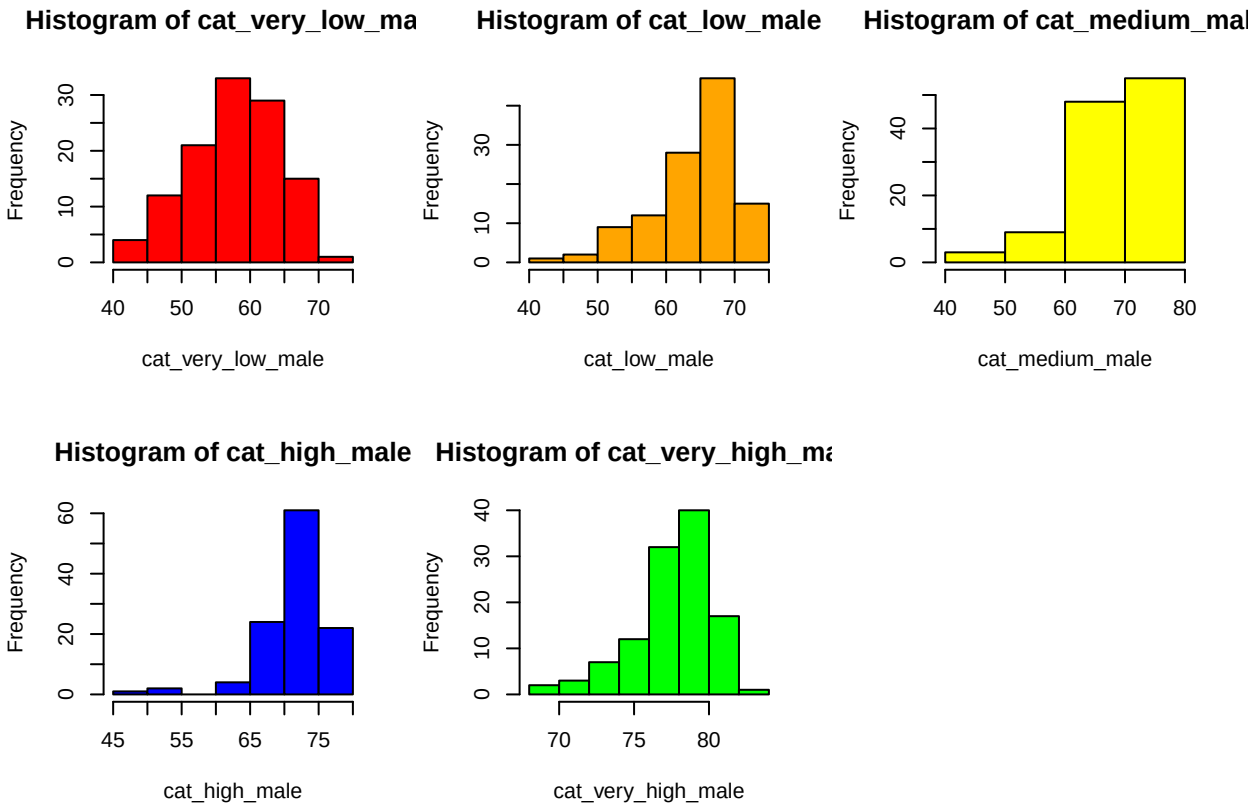
5 Representació dels resultats a partir de taules i gràfiques.

Histogrames de Life expectancy en funció del rang de despesa sanitària pública

```
par(mfrow=c(2,3))
hist(cat_very_low_fem, breaks=5, col="red")
hist(cat_low_fem, breaks=5, col="orange")
hist(cat_medium_fem, breaks=5, col="yellow")
hist(cat_high_fem, breaks=5, col="blue")
hist(cat_very_high_fem, breaks=5, col="green")
par(mfrow=c(2,3))
```



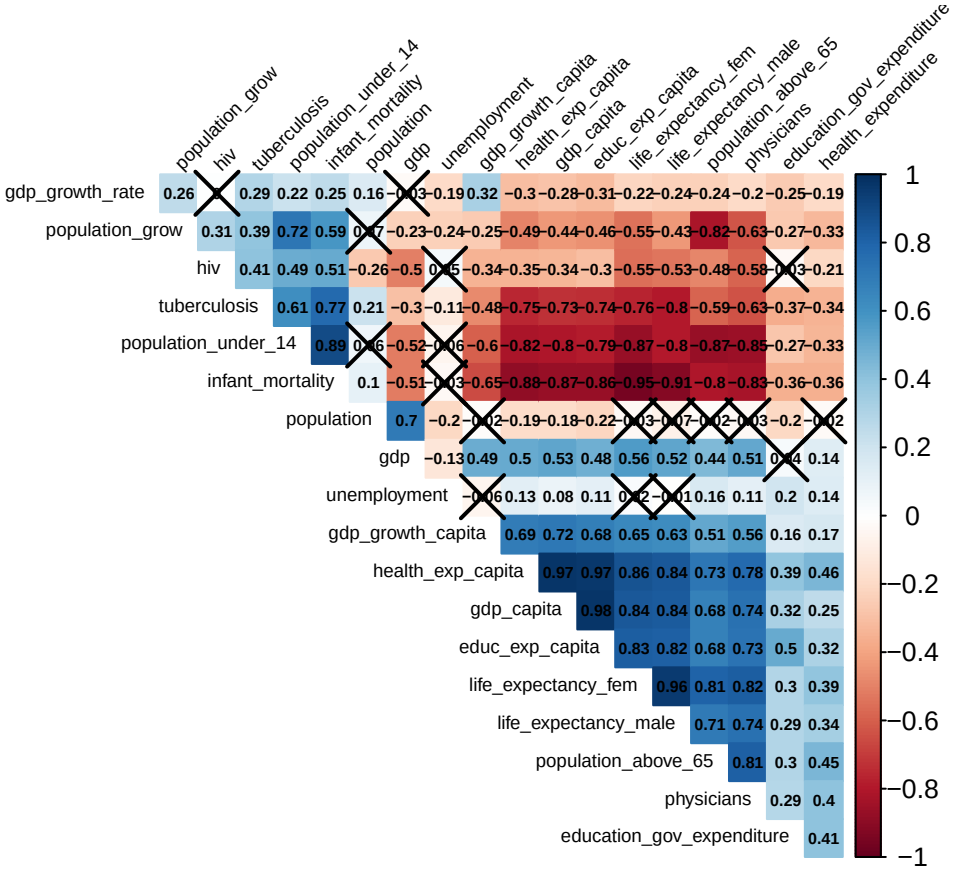

```
hist(cat_very_low_male, breaks=5, col="red")
hist(cat_low_male, breaks=5, col="orange")
hist(cat_medium_male, breaks=5, col="yellow")
hist(cat_high_male, breaks=5, col="blue")
hist(cat_very_high_male, breaks=5, col="green")
par(mfrow=c(1,1))
```



Gràcies als testos estadístics s'ha observat una diferència estadísticament significativa entre les distribucions d'esperança de vida per diferents rangs de despesa sanitària pública. La mitja de l'esperança de vida és notablement major quan major és la despesa sanitària pública, i la variància es redueix, fent disminuir la dispersió de l'esperança de vida entre individus.

Visualització dels coeficients de correlació entre les variables del dataset (no agrupades)

```
# Visualitzem els coeficients de correlació entre variables del dataset, marcant els
# resultats del p-valor:

corrplot(corr_matrix$r, method = "color",
         type = "upper",order = "hclust",
         addCoef.col = "black",
         tl.col = "black", tl.srt = 45,
         p.mat = corr_matrix$P, sig.level=0.05,
         diag = FALSE, tl.cex=0.55, number.cex=0.5
         )
```

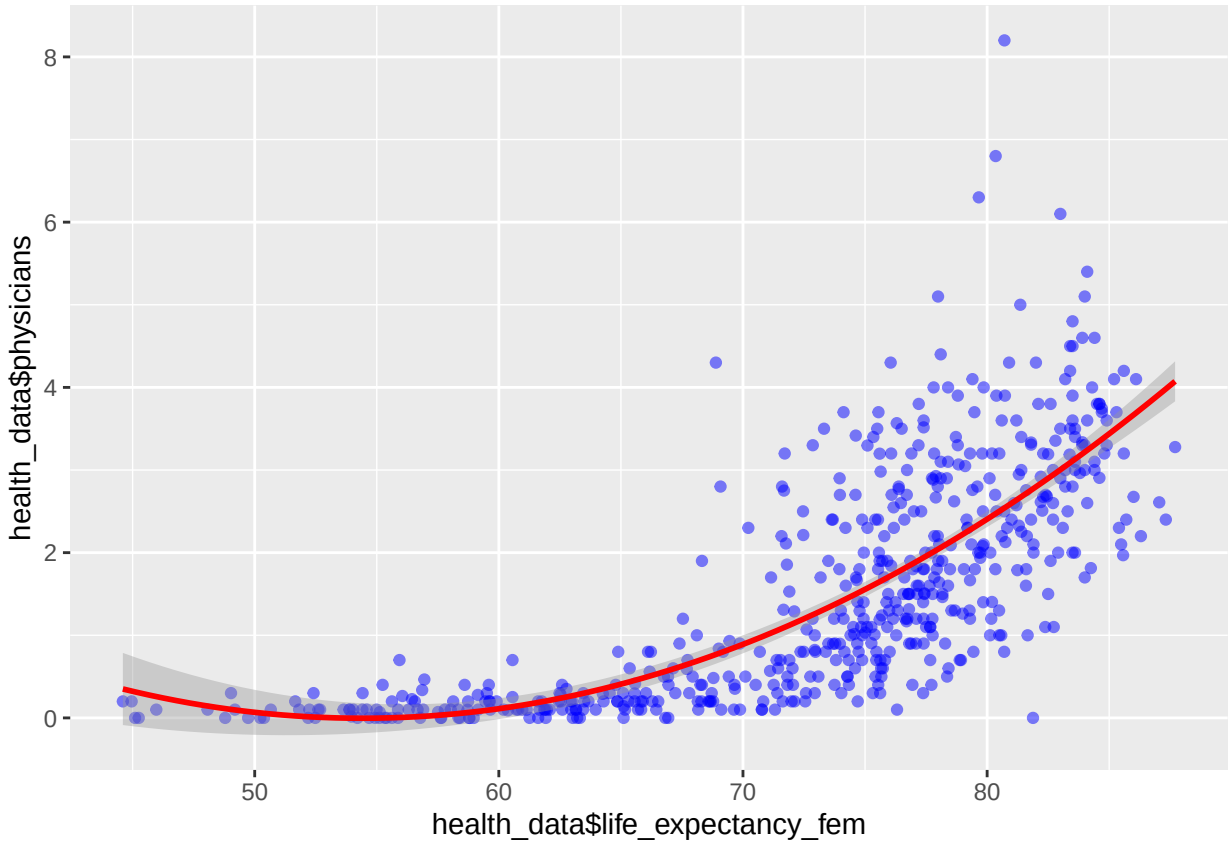


Les correlacions observades es descriuen a la secció 6.

Model de Regressió - Estimació de Life expectancy

Mostrem la gràfica del model de regressió tenint en compte únicament el terme amb major grau del polinomi obtingut, en aquest cas, health_data$physicians:

```
ggplot(data=health_data, aes(x=health_data$life_expectancy_fem, y=health_data$physicians,
       color=health_data$infant_mortality)) + geom_point(color="blue", alpha=0.5) +
       stat_smooth(method='lm', formula = y~poly(x,2), colour="red")
```



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

Gràcies a l'anàlisi de les dades d'esperança de vida segons el rang de despesa sanitària pública, hem pogut comprovar que existeix una diferència estadísticament significativa entre les distribucions segons aquests rangs de despesa. No només la mitja de l'esperança de vida és notablement major quan major és la despesa sanitària pública, sinó que a mesura que la despesa augmenta, la variància de l'esperança de vida es redueix. A efectes pràctics, això significa una menor dispersió del valor d'esperança de vida entre la població de països amb major despesa sanitària, i per tant una menor desigualtat en anys viscuts pels individus, cosa que podria traduir-se en una qualitat de vida més homogènia i igualitària entre la població.

De l'anàlisi mitjançant el test de correlació hem identificat diferents correlacions d'interès entre variables:

- La població total i el gdp total no semblen relacionar-se significativament amb la resta de variables analitzades. És a dir, a priori viure en un país més o menys poblat o amb més o menys gdp total no és un indicador de millors o pitjors condicions de salut.
- Les dades globals de gdp, gdp_growth_rate, education_gov_expenditure i health_expenditure tenen coeficients de correlació baixos en relació amb la resta de variables del dataset, però en canvi existeixen coeficients més elevats en el cas d'aquestes variables considerades per càpita.
- Existeix una relació significativa entre major health_exp_capita, educ_exp_capita i physicians, i una major life_expectancy_fem/male. I a major proporció de population_under_14 i infant_mortality, menor life_expectancy_fem.
- Existeix una relació significativa entre major gdp_capita, health_exp_capita, educ_exp_capita i physicians, i menor infant_mortality. Tanmateix, a major prevalença de tuberculosi, major infant_mortality.
- La prevalença de tuberculosi es relaciona més fortament amb la reducció de l'esperança de vida que la prevalença de VIH. Això es pot deure al fet que per la tuberculosi existeix un tractament efectiu, però els països amb menor esperança de vida (i alhora amb menys recursos econòmics) tenen més dificultats d'accés a aquest tractament comparat amb els països amb major esperança de vida (desenvolupats), i alhora, al fet que ni els països desenvolupats ni els no desenvolupats encara no disposen d'una cura per l'VIH.

Aquesta anàlisi de correlacions ens permet establir models de regressió. En aquest cas hem optat per implementar un model de regressió que permeti estimar l'esperança de vida femenina segons 3 paràmetres d'entrada: "health_exp_capita", "physicians" i "infant_mortality", que són variables sobre les que cada govern pot realitzar actuacions de millora. Analitzant el model podem concloure que lleus increments en despesa sanitària pública, nombre de metges i en l'aplicació de mesures per la reducció de la mortalitat infantil produirien notables increments en l'esperança de vida de la població, cosa que reduiria la desigualtat en qualitat de vida entre territoris.

```r
# Generació del fitxer CSV final

  write.csv(health_data,"world_health_indicators_final.csv",
            row.names = FALSE, quote=FALSE,na="")
```

Contribucions

Contribucions	Firma
Investigació prèvia	ADS,XVD
Redacció de les respostes	ADS,XVD
Desenvolupament codi	ADS,XVD