1. Introduction

This file s a sample of statistical analyzis of data in R. It is a part of DAP project - input is a csv project with dataframe crated after retrievial of from MongoDB and before submission to PosgresSQL DB.

In [32]:

```
#Libraries
library(car)
```

2. Getting the Data

In [4]:

```
data <- as.data.frame <- (read.table("df_for_R.csv", sep=",", header=TRUE))
data
data.head
```

A data.frame: 39 × 6

X	Country	Population	Obesity_percentage	covid_deaths	covid_death_percentage
<int>	<chr>	<int>	<dbl>	<int>	<dbl>
0	Austria	9006398	21.9	10152	1.127199e-03
1	Australia	25499884	30.4	910	3.568644e-05
2	Belgium	11589623	24.5	24140	2.082898e-03
3	Canada	37742154	31.3	24110	6.388083e-04
4	Chile	19116201	28.8	26073	1.363922e-03
5	Colombia	50882891	22.1	72725	1.429262e-03
6	Czech Republic	10708981	28.5	29141	2.721174e-03
7	Denmark	5792202	21.3	2482	4.285072e-04
8	Estonia	1326535	23.8	1148	8.654125e-04
9	Finland	5540720	24.9	911	1.644191e-04
10	France	65273511	23.2	104077	1.594475e-03
11	Germany	83783942	25.7	82588	9.857259e-04
12	Greece	10423054	27.4	10242	9.826295e-04
13	Hungary	9660351	28.6	27172	2.812734e-03
14	Iceland	341243	23.1	29	8.498343e-05
15	Ireland	4937786	26.9	4896	9.915375e-04
16	Israel	8655535	26.7	6361	7.349055e-04
17	Italy	60461826	22.9	120256	1.988957e-03
18	Japan	126476461	4.4	10052	7.947724e-05
19	Korea	51269185	4.9	1825	3.559643e-05
20	Latvia	1886198	25.7	2118	1.122894e-03
21	Lithuania	2722289	28.4	3900	1.432618e-03
22	Luxembourg	625978	24.2	792	1.265220e-03
23	Mexico	128932753	28.4	215918	1.674656e-03
24	Netherlands	17134872	23.1	17339	1.011913e-03
25	New Zealand	4822233	32.0	26	5.391693e-06
26	Norway	5421241	25.0	753	1.388981e-04
27	Poland	37846611	25.6	66533	1.757965e-03
28	Portugal	10196709	23.2	16973	1.664557e-03
29	Slovakia	5459642	22.4	11611	2.126696e-03
30	Slovenia	2078938	22.5	4236	2.037579e-03
31	Spain	46754778	27.1	77943	1.667060e-03
32	Sweden	10099265	22.1	14000	1.386239e-03
33	Switzerland	8654622	21.2	10617	1.226743e-03
34	Turkey	84339067	32.2	39398	4.671382e-04
35	United Kingdom	67886011	29.5	127734	1.881595e-03
36	United States	331002651	37.3	574340	1.735152e-03
37	China	1439323776	6.6	4845	3.366164e-06
38	India	1380004385	3.8	204832	1.484285e-04

```
Error in eval(expr, envir, enclos): object 'data.head' not found
Traceback:
```

2.1. Removiong Observation - with posible data-entry errors

Threre is no entries with epossible data entry errors:

3. Descriptive statistics

In [5]:

```
summary(data)
```

```
       X           Country            Population        Obesity_percentage
 Min.   : 0.0   Length:39          Min.   :3.412e+05   Min.   : 3.80     
 1st Qu.: 9.5   Class :character   1st Qu.:5.500e+06   1st Qu.:22.45     
 Median :19.0   Mode  :character   Median :1.071e+07   Median :24.90     
 Mean   :19.0                      Mean   :1.073e+08   Mean   :23.89     
 3rd Qu.:28.5                      3rd Qu.:5.587e+07   3rd Qu.:28.40     
 Max.   :38.0                      Max.   :1.439e+09   Max.   :37.30     
  covid_deaths     covid_death_percentage
 Min.   :    26   Min.   :3.366e-06     
 1st Qu.:  3191   1st Qu.:4.478e-04     
 Median : 11611   Median :1.127e-03     
 Mean   : 50082   Mean   :1.126e-03     
 3rd Qu.: 52966   3rd Qu.:1.671e-03     
 Max.   :574340   Max.   :2.813e-03     
```

In [6]:

```
str(data)
```

```
'data.frame':	39 obs. of  6 variables:
 $ X                      : int  0 1 2 3 4 5 6 7 8 9 ...
 $ Country                : chr  "Austria" "Australia" "Belgium" "Canada" ...
 $ Population             : int  9006398 25499884 11589623 37742154 19116201 50882891 10708981 5792202 1326535 5540720 ...
 $ Obesity_percentage     : num  21.9 30.4 24.5 31.3 28.8 22.1 28.5 21.3 23.8 24.9 ...
 $ covid_deaths           : int  10152 910 24140 24110 26073 72725 29141 2482 1148 911 ...
 $ covid_death_percentage: num  1.13e-03 3.57e-05 2.08e-03 6.39e-04 1.36e-03 ...
```

4. Visualisation - Numerical Variables

Options to fit figures in paper

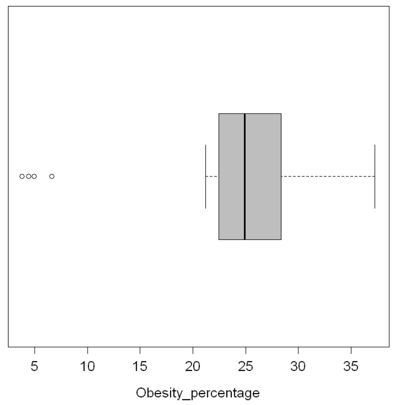
In [8]:

```
#options(scipen=5)
attach(data)
#options(repr.plot.width=6, repr.plot.height=3)
```

4.1. Boxplot - numerical variables

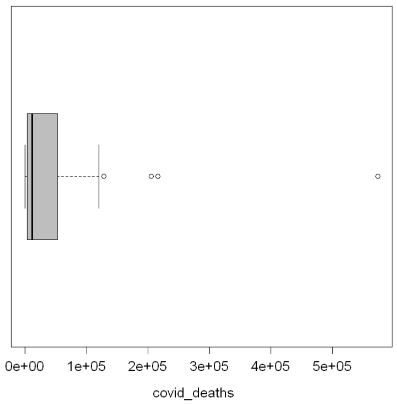
In [12]:

```
boxplot (Obesity_percentage, breaks=40, ylab='',xlab='Obesity_percentage', col='grey', cex.lab=1.25, cex.axis=1.25,
         horizontal=TRUE)
```



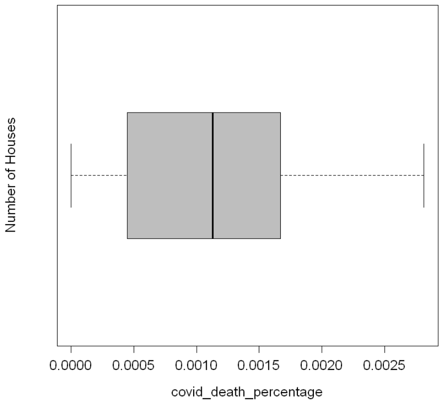
In [13]:

```
boxplot (covid_deaths    , breaks=45, ylab=' ',xlab='covid_deaths    ', col='grey', cex.lab=1.25, cex.axis=1.25
      ,horizontal=TRUE)
```



In [14]:

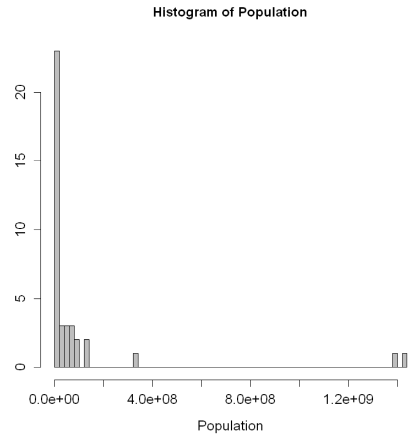
```
boxplot (covid_death_percentage,breaks=30, ylab='Number of Houses',xlab='covid_death_percentage', col='grey', cex.lab=1.25, cex.axis=1.25,horizontal=TRUE)
```



4.2. Histograms - numerical variables

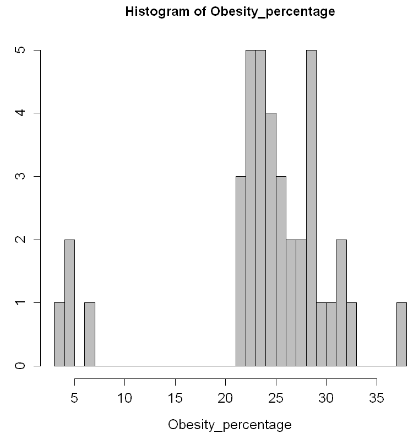
In [17]:

```
hist (Population, breaks=100, ylab='',xlab='Population', col='grey', cex.lab=1.25, cex.axis=1.25)
```



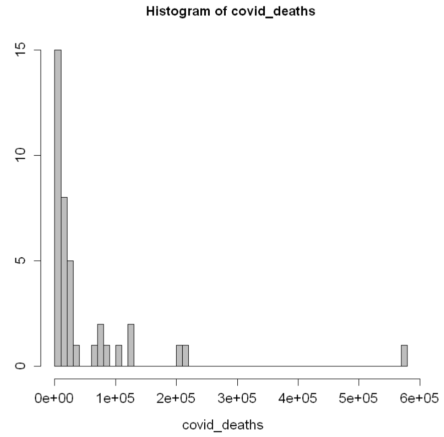
In [18]:

```
hist (Obesity_percentage, breaks=45, ylab='',xlab='Obesity_percentage', col='grey', cex.lab=1.25, cex.axis=1.25)
```



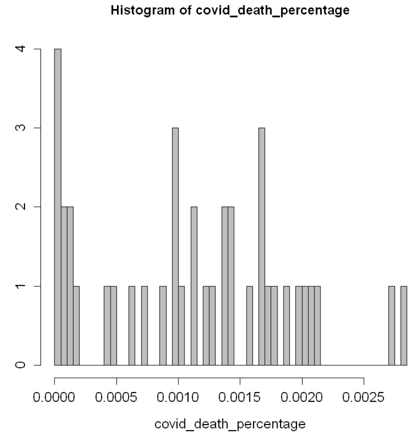
In [20]:

```
hist (covid_deaths    ,breaks=45, ylab='  ',xlab='covid_deaths    ', col='grey', cex.lab=1.25, cex.axis=1.25)
```



In [21]:

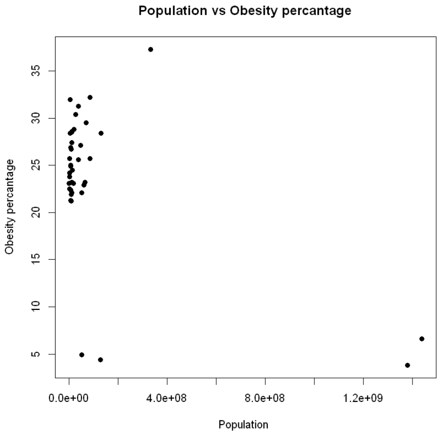
```
hist (covid_death_percentage,breaks=50, ylab=' ',xlab='covid_death_percentage', col='grey', cex.lab=1.25, cex.axis=1.25)
```



4.3. Visualisation -Scatterplots (dependency between columns)

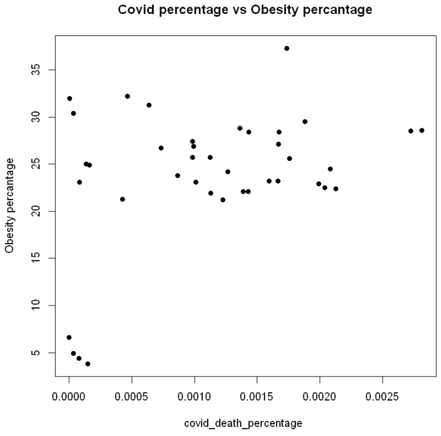
In [22]:

```
plot(Population, Obesity_percentage, main="Population vs Obesity percantage",
   xlab="Population ", ylab="Obesity percantage", pch=19)
```



In [23]:

```
plot(covid_death_percentage, Obesity_percentage, main="Covid percentage vs Obesity percantage",
   xlab="covid_death_percentage ", ylab="Obesity percantage", pch=19)
```



5. Investigation of correlation between values column

In [33]:

```
data.cor <- cor(data[,c(3:6)])
```

In [45]:

```
install.packages("corrplot")
```

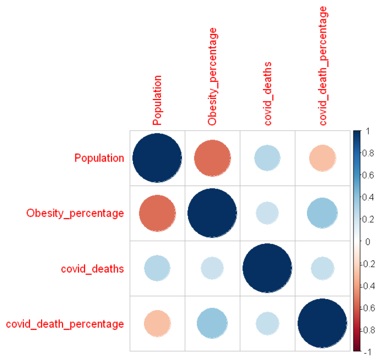
```
Warning message:
"package 'corrplot' is in use and will not be installed"
```

In [41]:

```
library(corrplot)
```

In [42]:

```
options(repr.plot.width=7, repr.plot.height=7)
corrplot(data.cor)
```



In [ ]: