6315305_assignment_eda

Andreas Sakapetis

November 30, 2018

```
#loading the library "readxl" to be able to use the function read_xlsx().
library(readxl)
#loading the library "formattable" to introduce a table with all the variables.
library(formattable)
#loading the library "ggplot2" to create usefull plots.
library(ggplot2)
#loading the library "GGally" to create usefull plots.
library(GGally)
#loading the library "tidyverse" to perform varius manipulations.
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------- tidyverse
```

```
## v tibble  1.4.2     v purrr   0.2.5
## v tidyr   0.8.2     v dplyr   0.7.8
## v readr   1.1.1     v stringr 1.3.1
## v tibble  1.4.2     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------------------------------------- tidyverse_conflicts()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#importing data from local folder to Rstudio in ob.data variable.
ob.data <- read_xlsx("C:\\Users\\Andreas\\Desktop\\Assignment A\\PROBESITY dataset.xlsx", sheet = 1)
#import the first sheet with the variable abbreviations.
var_explain = read_xlsx("C:\\Users\\Andreas\\Desktop\\Assignment A\\PROBESITY dataset.xlsx", sheet = 2)
```

```
#view the names of the columns in the ob.data dataset.
names(ob.data)
```

```
##  [1] "country"         "year"            "ctry_dum"
##  [4] "code"            "euro45"          "eu"
##  [7] "rgdp"            "pop"             "hc"
## [10] "urban"           "oga"             "demeaned_lhc"
## [13] "demeaned_lurban" "lrgdppop_hp"     "demeaned_loga"
## [16] "PFOV20P"         "PFOB20P"         "PMOV20P"
## [19] "PMOB20P"
```

```
#Use the function head() to examine at the first few rows of the ob.data dataset.
head(ob.data)
```

```
## # A tibble: 6 x 19
##   country  year ctry_dum code  euro45    eu   rgdp   pop    hc urban   oga
##   <chr>   <dbl>    <dbl> <chr>  <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 Albania  1975        1 ALB        1     0  8762.  2.41  1.70  32.7  28.9
## 2 Albania  1976        1 ALB        1     0  9096.  2.46  1.74  32.9  31.7
## 3 Albania  1977        1 ALB        1     0  9488.  2.52  1.77  33.1  29.9
## 4 Albania  1978        1 ALB        1     0  9868.  2.57  1.81  33.3  25.3
## 5 Albania  1979        1 ALB        1     0 10175.  2.63  1.84  33.5  31.5
## 6 Albania  1980        1 ALB        1     0 10771.  2.68  1.88  33.8  32.0
```

```
## # ... with 8 more variables: demeaned_lhc <dbl>, demeaned_lurban <dbl>,
## #   lrgdppop_hp <dbl>, demeaned_loga <dbl>, PFOV20P <dbl>, PFOB20P <dbl>,
## #   PMOV20P <dbl>, PMOB20P <dbl>
```

```
#Use the function tail() to examine the last few rows of the ob.data dataset.
tail(ob.data)
```

```
## # A tibble: 6 x 19
##   country  year ctry_dum code  euro45    eu    rgdp   pop    hc urban   oga
##   <chr>   <dbl>    <dbl> <chr>  <dbl> <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 United~  2011      164 GBR        1     1  2.30e6  63.2  3.71  81.6  88.7
## 2 United~  2012      164 GBR        1     1  2.39e6  63.6  3.72  81.8  88.1
## 3 United~  2013      164 GBR        1     1  2.50e6  64.0  3.73  82.1  87.2
## 4 United~  2014      164 GBR        1     1  2.59e6  64.3  3.73  82.3  87.3
## 5 United~  2015      164 GBR        1     1 NA       NA   NA     NA    NA
## 6 United~  2016      164 GBR        1     1 NA       NA   NA     NA    NA
## # ... with 8 more variables: demeaned_lhc <dbl>, demeaned_lurban <dbl>,
## #   lrgdppop_hp <dbl>, demeaned_loga <dbl>, PFOV20P <dbl>, PFOB20P <dbl>,
## #   PMOV20P <dbl>, PMOB20P <dbl>
```

The dataset consists of 19 variables, with a total of 1890 observations.The data includes information about prevalence rates for female and male adults (aged 20 and over) from 45 countries in the 1975-2016 period. There are 3054 missing values, spread across different variables. Within the datase exist variables that the type is either numeric or character. However, the variable euro45 and eu are categorical variables. More specific for the variable euro45 if the value is 1 this means that the country is a member of the european union, else if the value is 0 then the country is not. As for the eu variable if the the value is 1, this is translated that the country is a region of the European continent, else if the value is 0 the country is outised the European continent. Moreover, within the dataset are variables that are calculated in order to assist the researchers of the study to make conlusions such as urban, demeaned_lhc, demeaned_urban, irgdppop_lhp and loga.

```
#This function counts missing values
sapply(ob.data, function(x) sum(length(which(is.na(x)))))
```

```
##         country            year        ctry_dum            code 
##               0               0               0               0 
##          euro45              eu            rgdp             pop 
##               0               0             375             375 
##              hc           urban             oga    demeaned_lhc 
##             525             105             332             525 
## demeaned_lurban     lrgdppop_hp   demeaned_loga         PFOV20P 
##             105             375             337               0 
##         PFOB20P         PMOV20P         PMOB20P 
##               0               0               0
```

```
#create a table with all the variables
widget.formattable = formattable(var_explain)
#display the table
widget.formattable
```

Variable

Description

euro45

1 if the country belongs to Europe, 0 otherwise

eu

2 if the country belongs to the EU, 0 otherwise

rgdp

Expenditure-side real GDP at chained PPPs (in mil. 2011US$)

pop

Population (in millions)

hc

Human capital index, based on years of schooling and returns to education; see Human capital in PWT9.

urban

Urbanization rate= (Urban Pop/Total Pop)x100%

oga

overall globalization index

demeaned_lhc

Demeaned log of hc

demeaned_lurban

Demeaned log of urban

lrgdppop_hp

Hodrick-Prescott filtered log of rgdp/pop

loga

Log of OGA

PFOV20P

prevalence females_overweight_20+

PFOB20P

prevalence females_obesity_20+

PMOV20P

prevalence males_overweight_20+

PMOB20P

prevalence males_obesity_20+

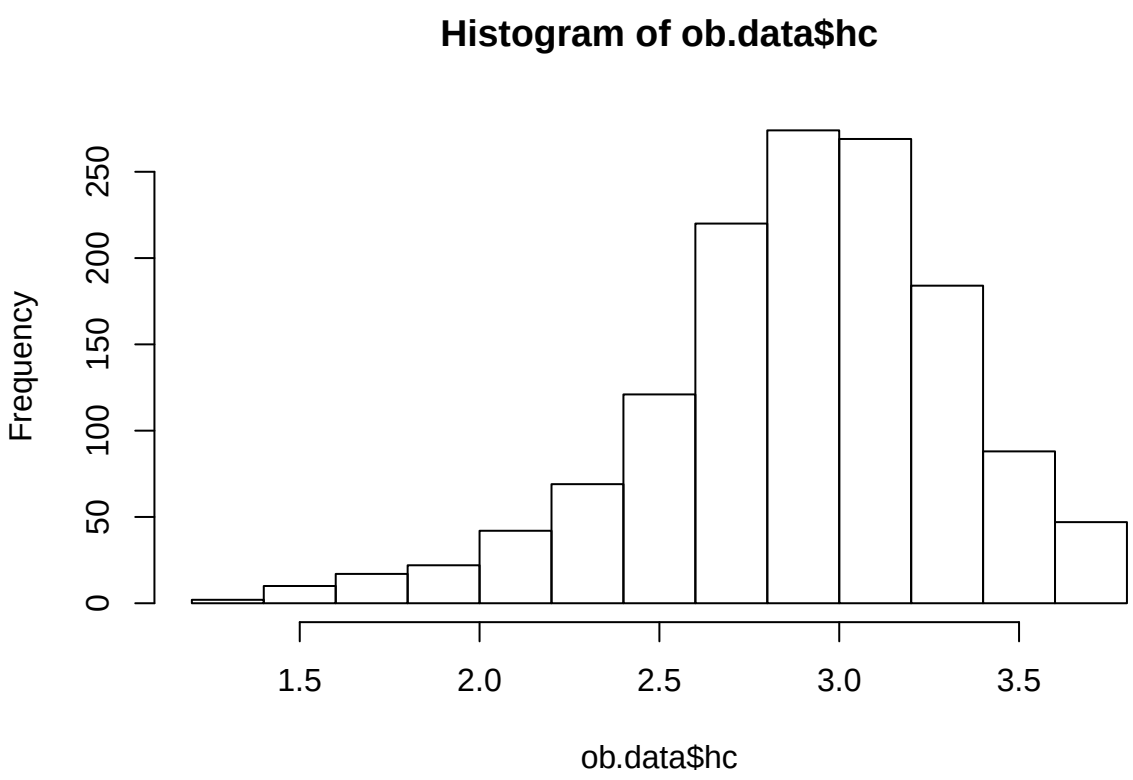
```
#With sapply function it is possible to examine the type of the variables.
sapply(ob.data,class)
```

```
##         country            year        ctry_dum            code
##     "character"       "numeric"       "numeric"     "character"
##          euro45              eu            rgdp             pop
##       "numeric"       "numeric"       "numeric"       "numeric"
##              hc           urban             oga    demeaned_lhc
##       "numeric"       "numeric"       "numeric"       "numeric"
## demeaned_lurban     lrgdppop_hp   demeaned_loga         PFOV20P
##       "numeric"       "numeric"       "numeric"       "numeric"
##         PFOB20P         PMOV20P         PMOB20P
##       "numeric"       "numeric"       "numeric"
```

```r
#acquire summary statistics for the ob.data dataset.
summary(ob.data)
```

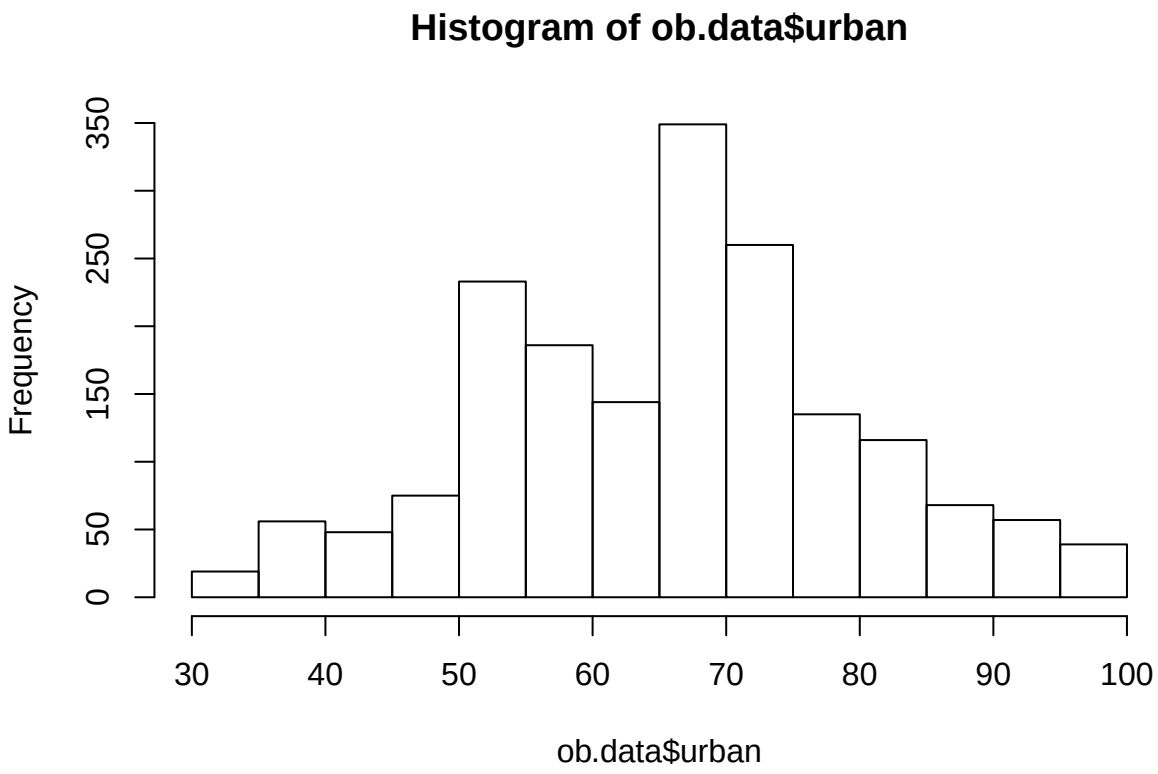
```
##    country                year         ctry_dum          code
##  Length:1890        Min.   :1975   Min.   :  1.00   Length:1890
##  Class :character   1st Qu.:1985   1st Qu.: 46.00   Class :character
##  Mode  :character   Median :1996   Median : 83.00   Mode  :character
##                     Mean   :1996   Mean   : 83.62
##                     3rd Qu.:2006   3rd Qu.:125.00
##                     Max.   :2016   Max.   :164.00
##
##      euro45         eu              rgdp              pop
##  Min.   :1   Min.   :0.0000   Min.   :   1990   Min.   :  0.218
##  1st Qu.:1   1st Qu.:0.0000   1st Qu.:  32217   1st Qu.:  3.562
##  Median :1   Median :1.0000   Median : 139239   Median :  7.869
##  Mean   :1   Mean   :0.6222   Mean   : 393465   Mean   : 18.565
##  3rd Qu.:1   3rd Qu.:1.0000   3rd Qu.: 339101   3rd Qu.: 16.565
##  Max.   :1   Max.   :1.0000   Max.   :3706587   Max.   :148.436
##                               NA's   :375       NA's   :375
##        hc            urban            oga          demeaned_lhc
##  Min.   :1.359   Min.   :31.29   Min.   :23.92   Min.   :-0.3662
##  1st Qu.:2.658   1st Qu.:55.26   1st Qu.:51.28   1st Qu.:-0.0568
##  Median :2.934   Median :66.84   Median :68.15   Median : 0.0074
##  Mean   :2.892   Mean   :65.97   Mean   :65.05   Mean   : 0.0000
##  3rd Qu.:3.181   3rd Qu.:74.06   3rd Qu.:80.83   3rd Qu.: 0.0604
##  Max.   :3.734   Max.   :97.82   Max.   :92.84   Max.   : 0.2410
##  NA's   :525     NA's   :105     NA's   :332     NA's   :525
##  demeaned_lurban     lrgdppop_hp       demeaned_loga        PFOV20P
##  Min.   :-0.45791   Min.   :-0.5599   Min.   :-0.6463   Min.   :0.3202
##  1st Qu.:-0.01706   1st Qu.:-0.0452   1st Qu.:-0.1549   1st Qu.:0.4434
##  Median : 0.00358   Median : 0.0007   Median : 0.0338   Median :0.4805
##  Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000   Mean   :0.4776
##  3rd Qu.: 0.02700   3rd Qu.: 0.0493   3rd Qu.: 0.1437   3rd Qu.:0.5189
##  Max.   : 0.33594   Max.   : 1.0369   Max.   : 0.5158   Max.   :0.7102
##  NA's   :105        NA's   :375       NA's   :337
##     PFOB20P           PMOV20P          PMOB20P
##  Min.   :0.06894   Min.   :0.3082   Min.   :0.03642
##  1st Qu.:0.15248   1st Qu.:0.4894   1st Qu.:0.10283
##  Median :0.18503   Median :0.5478   Median :0.13820
##  Mean   :0.18318   Mean   :0.5468   Mean   :0.14422
##  3rd Qu.:0.21531   3rd Qu.:0.6115   3rd Qu.:0.18237
##  Max.   :0.40689   Max.   :0.7170   Max.   :0.28287
##
```

```r
hist(ob.data$hc)
```



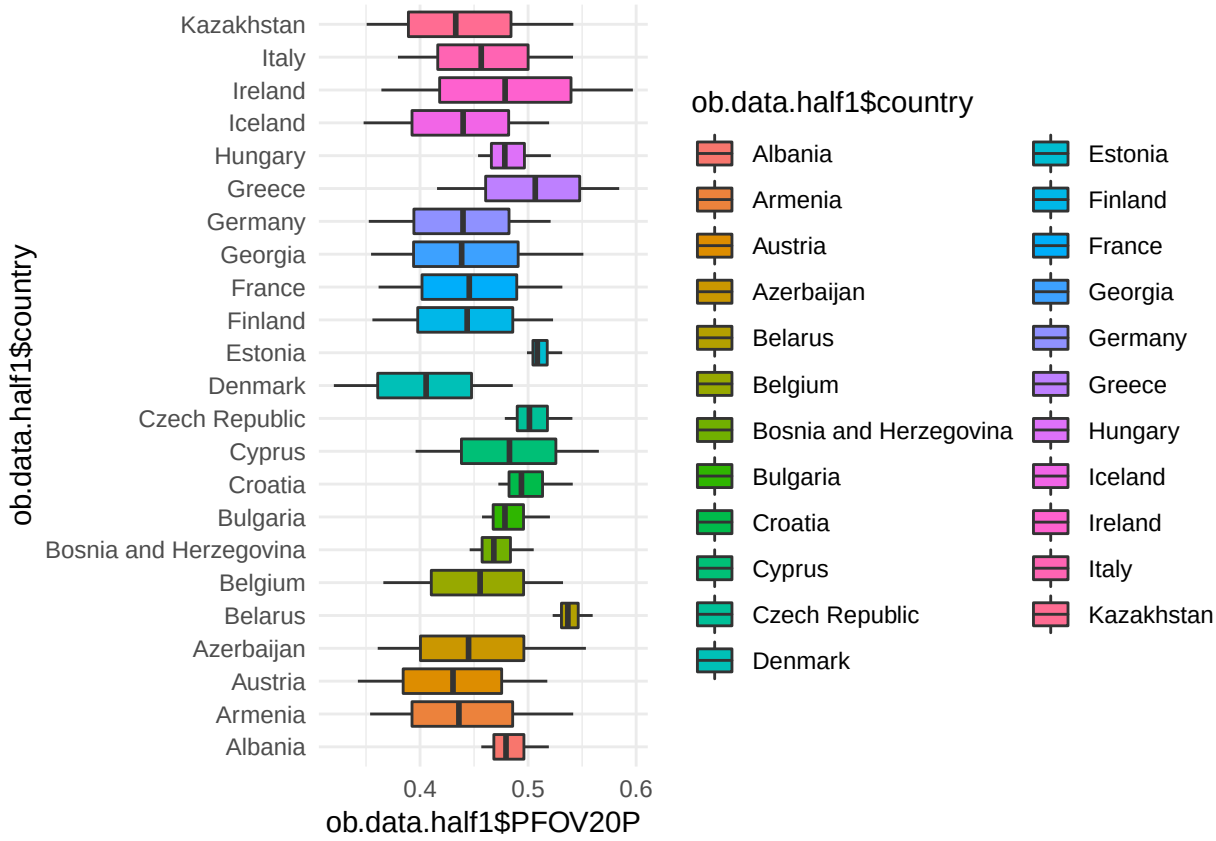
The histogram for Human capital index, based on years of schooling and returns to education reveals that there is a negative skewness.

```
hist(ob.data$urban)
```

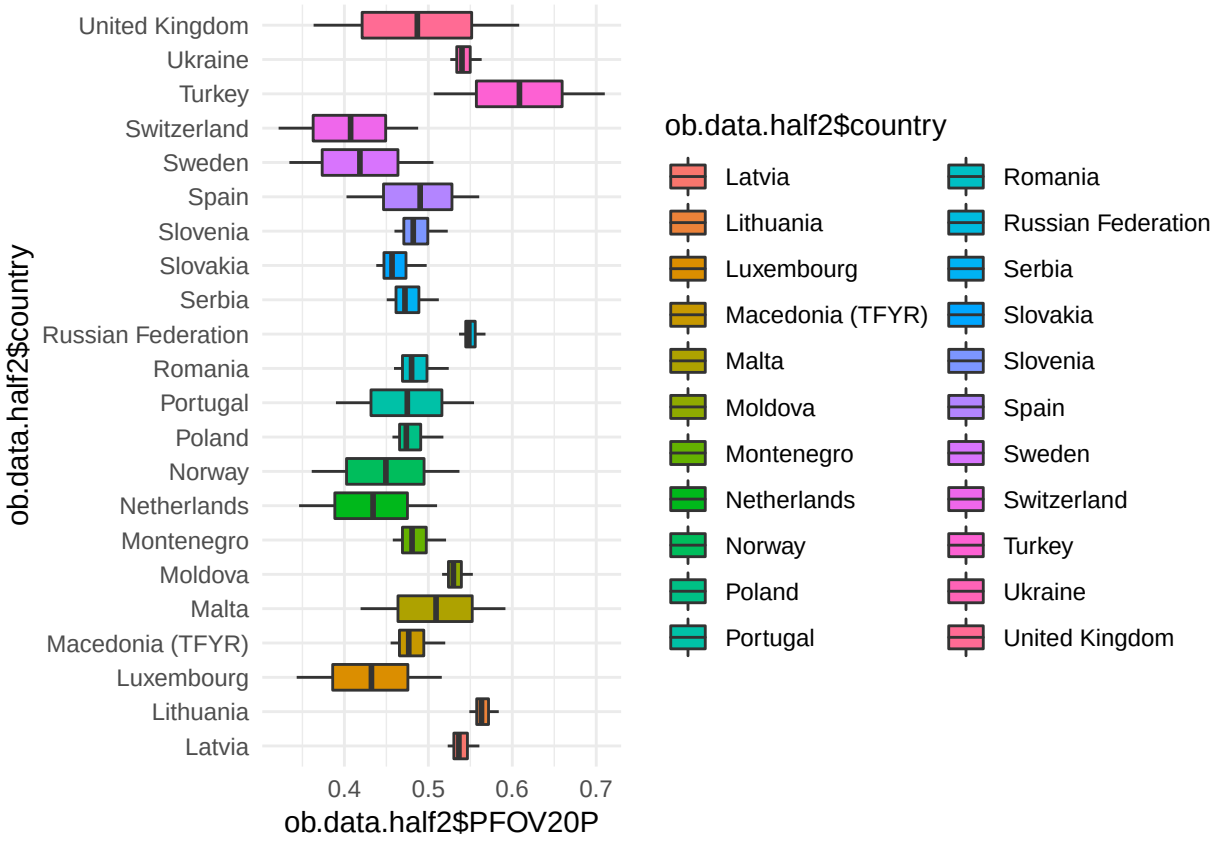



```
#With this lines the dataset is splitted in two for better visualization
#purposes. First, making sure that the dataset is sorted alphabetically by the
#country variable. Then the dataset is splitted in two halves. Manual inspection
#assisted the effort to distinguish the precise row that the dataset needs to be
#splitted.

#Alphabetically sorting the dataset based on the country variable.
ob.data <- ob.data %>%
  arrange(ob.data$country)

#This is the first half.
ob.data.half1 <-ob.data[1:966,]

#This is the second half.
ob.data.half2 <-ob.data[967:1890,]
```

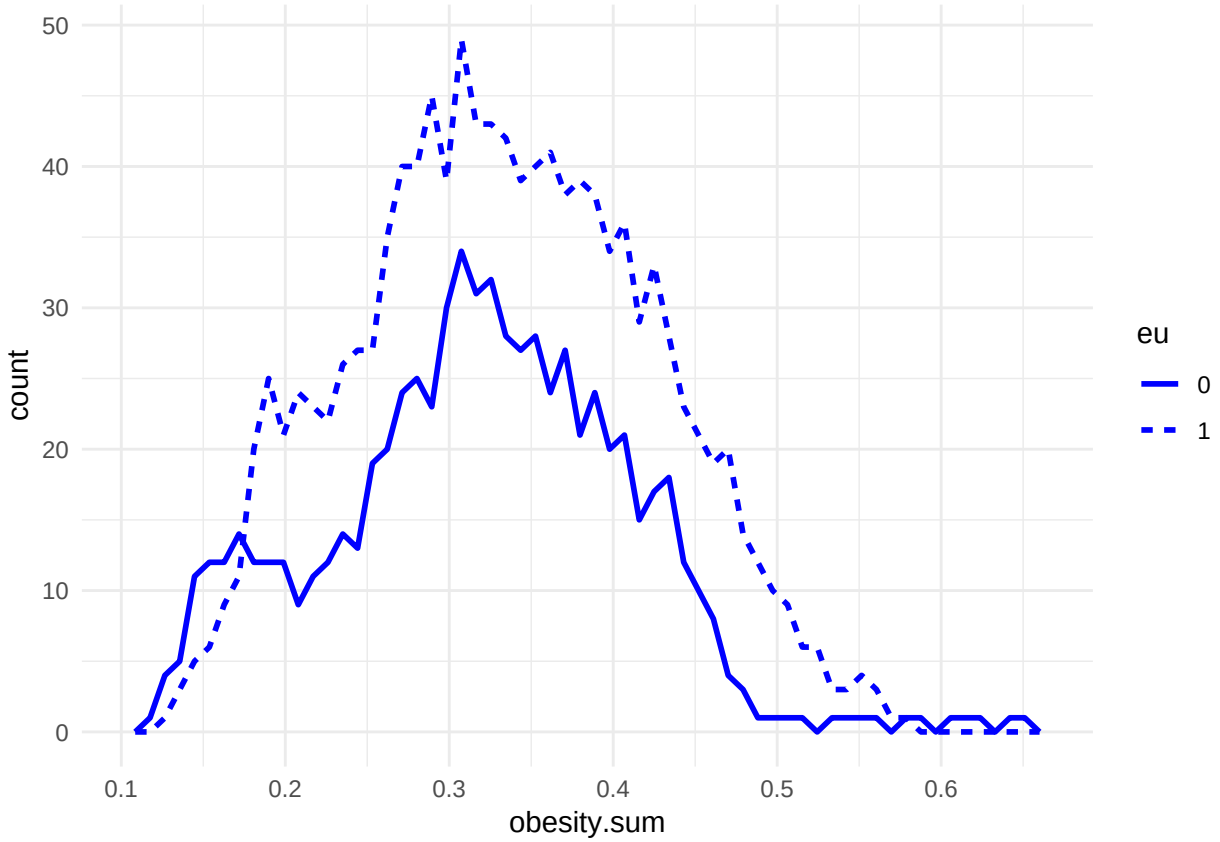
```
ob.data.half1 %>%
  ggplot(aes(x = ob.data.half1$country, y = ob.data.half1$PFOV20P, fill = ob.data.half1$country)) +
  geom_boxplot() +
  coord_flip() +
  theme_minimal()
```




```
ob.data.half2 %>%
  ggplot(aes(x = ob.data.half2$country, y = ob.data.half2$PFOV20P, fill = ob.data.half2$country)) +
  geom_boxplot() +
  coord_flip() +
  theme_minimal()
```

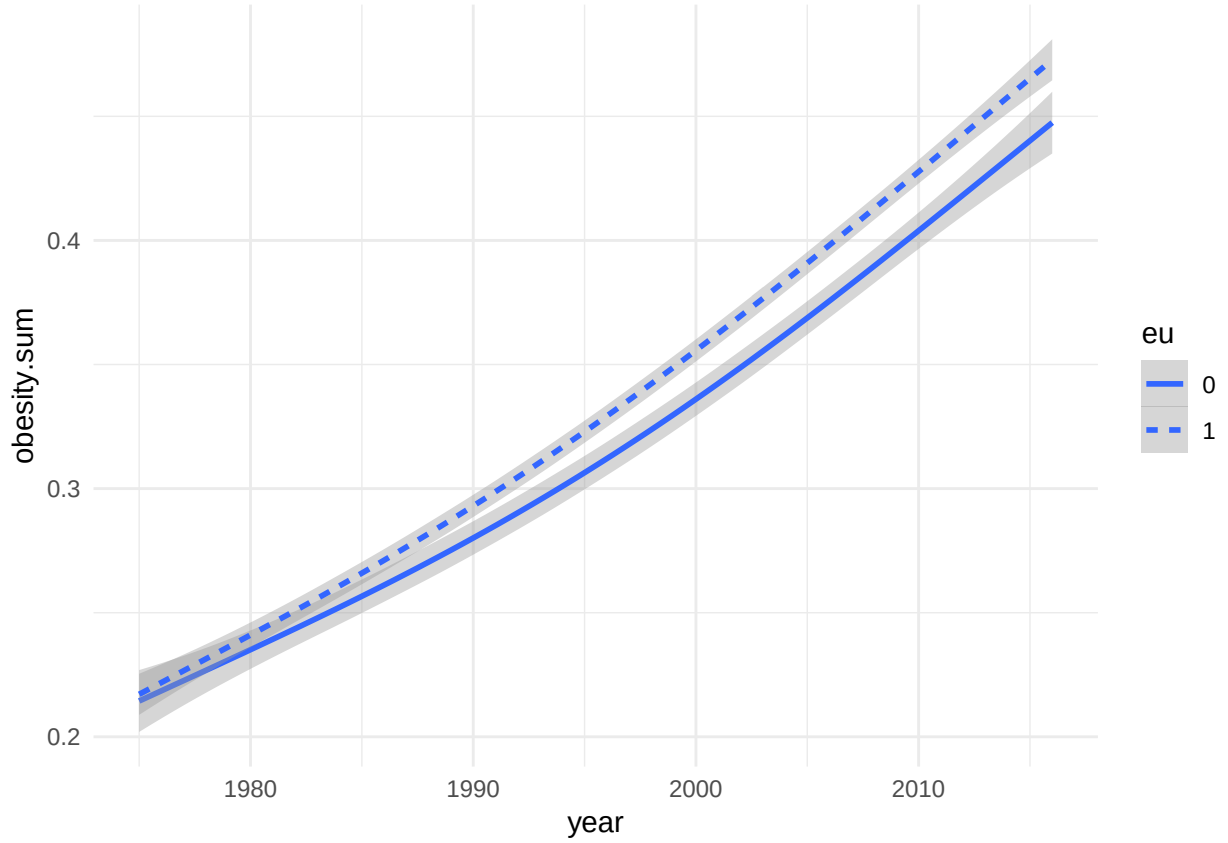



```
#creating a more refined dataframe with the variables that seem to be more
#interesting.
euro.obese <-
  ob.data %>%
  select(year, code, eu, PFOV20P, PFOB20P, PMOV20P, PMOB20P) %>%
  mutate(obesity.sum = PFOB20P + PMOB20P)

#Transforming the eu variable to factor.
euro.obese$eu <- as.factor(euro.obese$eu)

#Creating a plot to examine the frequency of obesity in the Euopean Continent and
#outside the European Continent.
frequency.obese <-
  euro.obese %>%
  ggplot(aes(x = obesity.sum, linetype  = eu))+
  geom_freqpoly(size = 1, bins = 60 , color = "blue")+
  theme_minimal()

frequency.obese
```



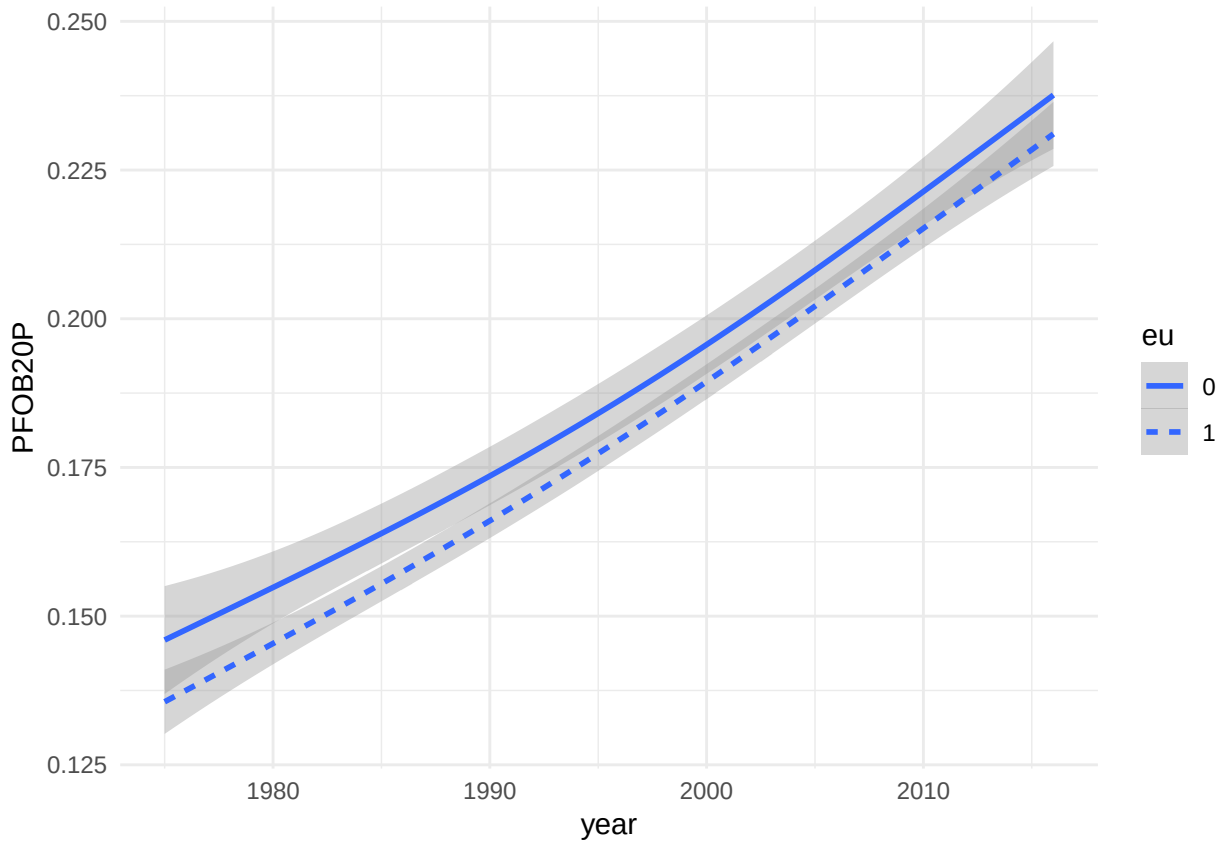

```
#Creating a plot to examine how obesity transforms into the given timeline
#separatedi n the Euopean Continent and outside the European Continent.
total.obesity <-
  euro.obese  %>%
  ggplot(mapping = aes(x = year, y = obesity.sum, linetype = eu))+
  geom_smooth()+
  theme_minimal()

total.obesity
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



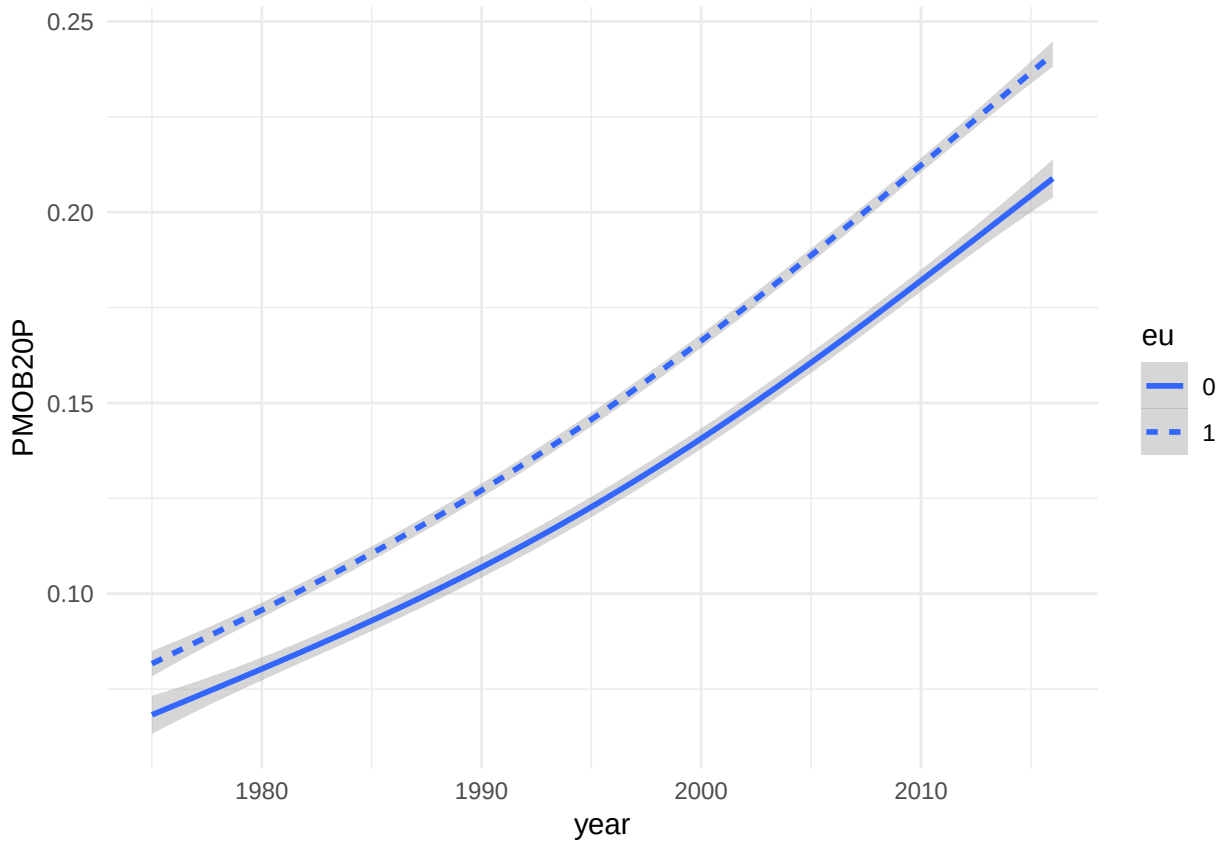

```
#Creating a plot to examine how female obesity transforms into the given timeline
#separatedi n the Euopean Continent and outside the European Continent.
female.obesity <-
  euro.obese %>%
  ggplot(mapping = aes(x = year, y = PFOB20P, linetype = eu))+
  geom_smooth()+
  theme_minimal()

female.obesity
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

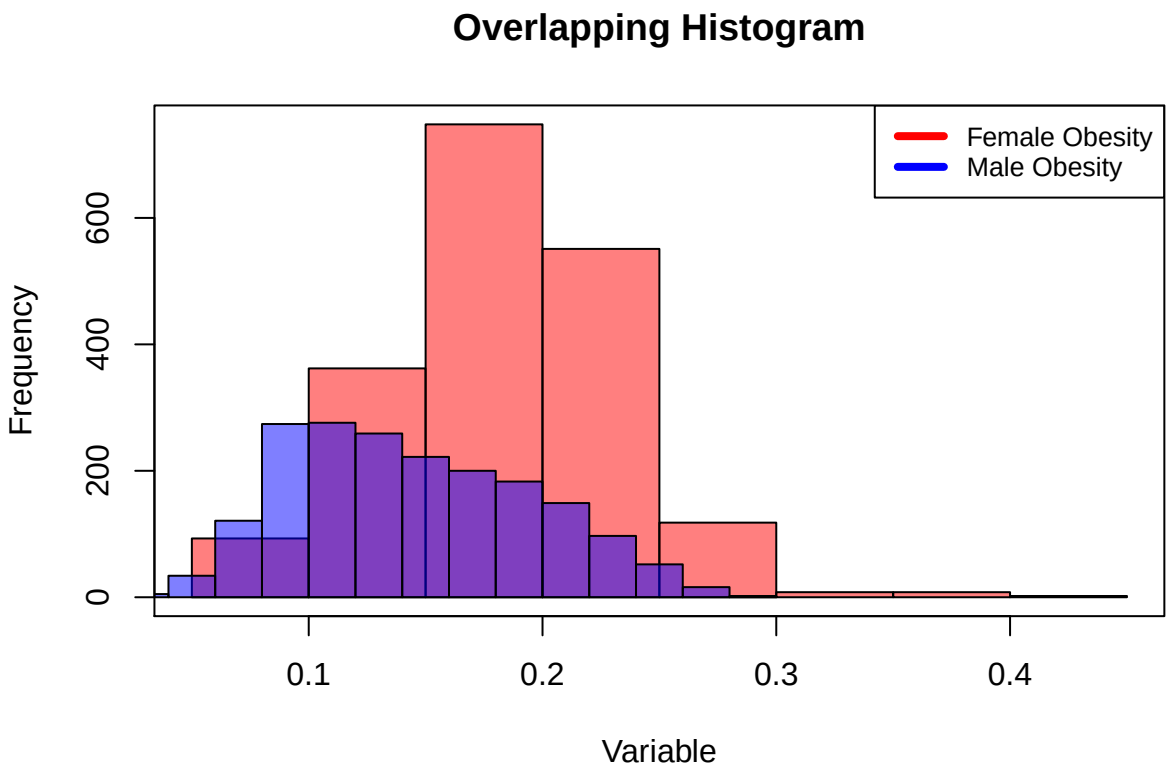



```
#Creating a plot to examine how male obesity transforms into the given timeline
#separatedi n the Euopean Continent and outside the European Continent.
male.obesity <-
  euro.obese  %>%
  ggplot(mapping = aes(x = year, y = PMOB20P, linetype = eu))+
  geom_smooth()+
  theme_minimal()

male.obesity
```

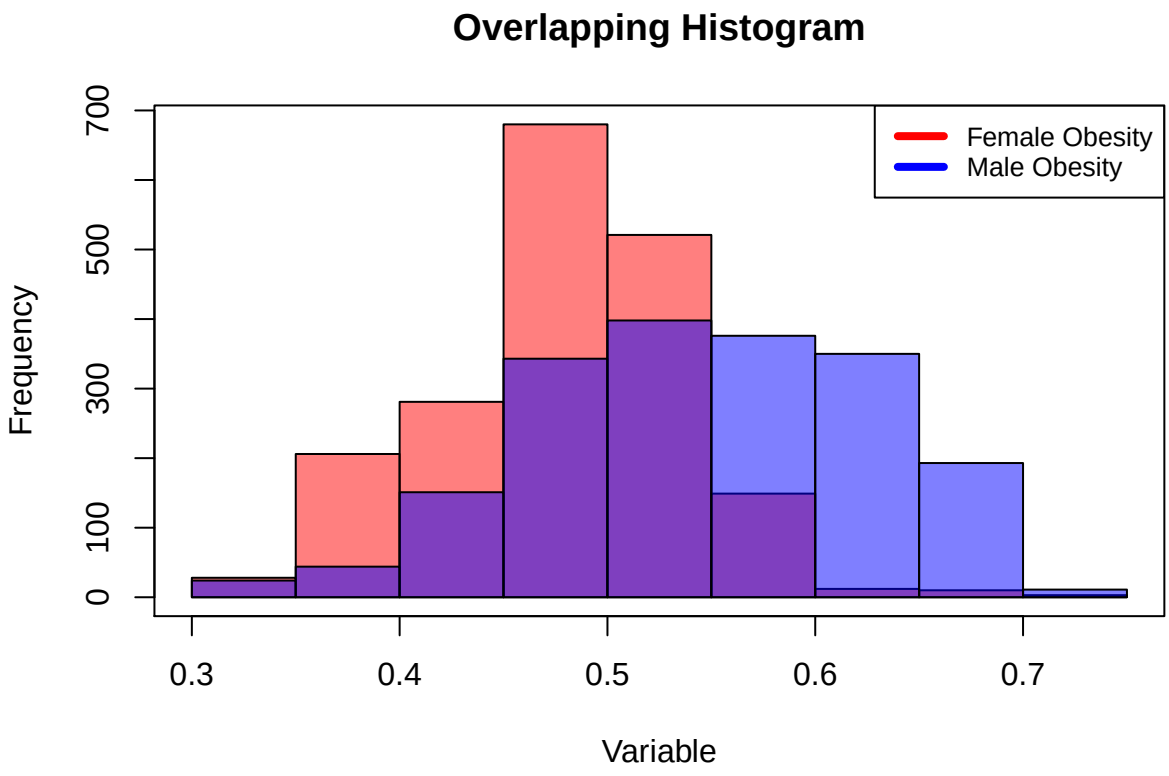
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
#Creating a plot to compare Female and Male obesity.
hist(euro.obese$PFOB20P, col=rgb(1,0,0,0.5), main= "Overlapping Histogram", xlab="Variable")
hist(euro.obese$PMOB20P, col=rgb(0,0,1,0.5), add=T)
box()
legend("topright", legend=c("Female Obesity", "Male Obesity"),
       col=c("red", "blue"), lwd=4, cex=0.8)
```

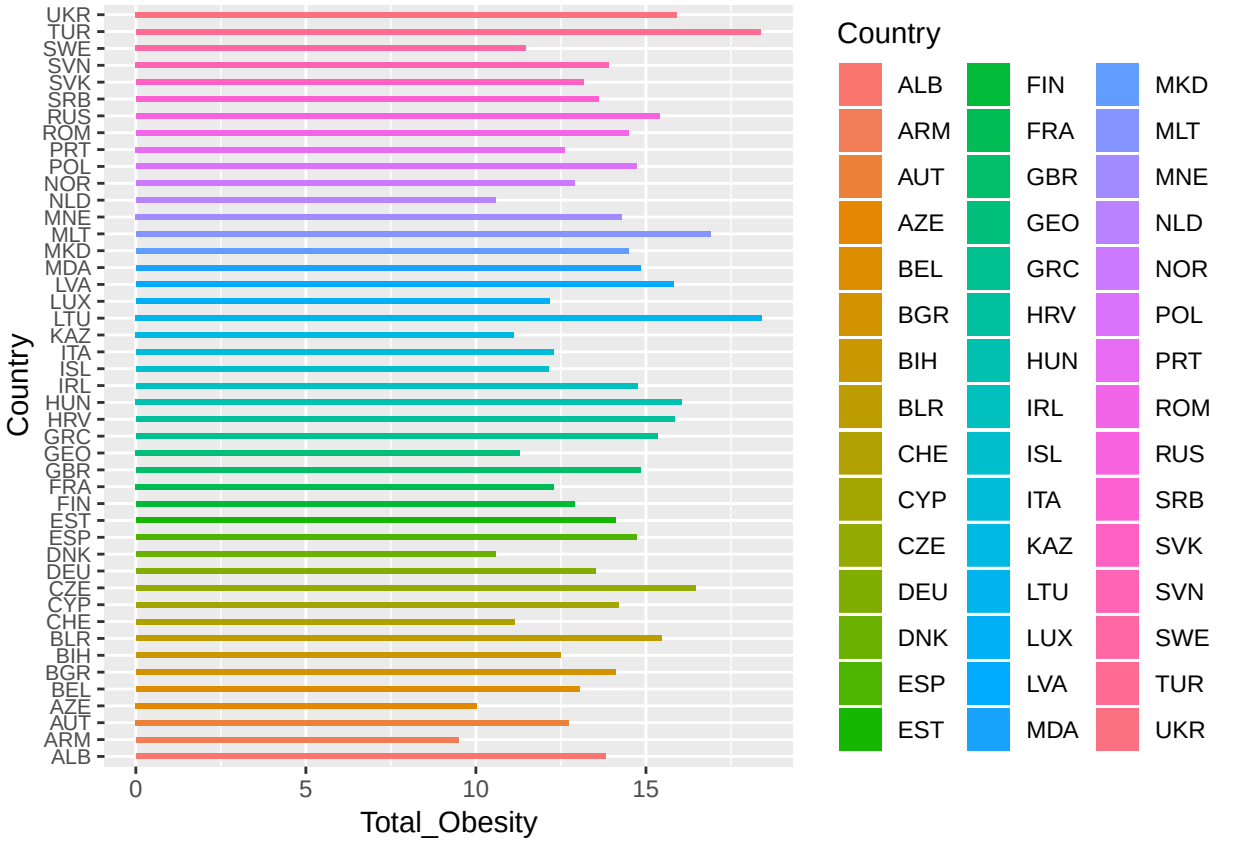



```
#Creating a plot to compare Females and Males that are overweight.
hist(euro.obese$PFOV20P, col=rgb(1,0,0,0.5), main= "Overlapping Histogram", xlab="Variable")
hist(euro.obese$PMOV20P, col=rgb(0,0,1,0.5), add=T)
box()
legend("topright", legend=c("Female Obesity", "Male Obesity"),
       col=c("red", "blue"), lwd=4, cex=0.8)
```



From the careful obseravation of varius graphics it is possible to conclude that the total population that is obese increases throughout the years. Taking into consideration the given data, the population in countries of the European Continent tends to be more obese than the population in the countries outside the European Continent. Another important fact is that obesity is more prone to females than males.

```
#Summarizing the countries based on obesity.sum.
obese.country <-
  euro.obese %>%
  group_by(code) %>%
  summarize(
    sum(obesity.sum)
  )

colnames(obese.country) <- c("Country", "Total_Obesity")

#Create a plot to examine which are the most obese countries.
obesity.plot <-
  obese.country %>%
  ggplot(mapping = aes(x = Country, y = Total_Obesity, fill= Country))+
  geom_bar(stat = "identity", width = 0.3) +
  theme(axis.text.y = element_text(size = rel(0.9), angle = 0))+
  coord_flip()

obesity.plot
```



As it possible to Observe from the graphic there are countries of the European continent that surpass in obesity countries from outside the European continent. This fact comes to a total agreement with the conclusions made so far from the previous graphics.

Firstly, the analyst loaded the necessary libraries readxl, formattable, ggplot2, ggally, tidyverse each designated for a different purpose. Then the analyst proceeds with loading the data from a local file location. In order to get a firm, grasp of the data the analyst used a series of functions such as: names, head and tail.

After the analyst used an sapply, a sum, a length, a which and an is.na function in conjunction to find the missing values of the dataset. The function formattable was utilized to load the second sheet of excel file which contained information about the variables of the dataset. Then With sapply function it is possible to examine the type of the variables.

Summary statistics were calculated by summary function. Moreover, various histograms were utilized to see the tendencies of the variables and their distribution. For better visualization purposes the analyst divided the dataset in two halves based on alphabetic order.

Thus, the boxplots were created to be able to examine the countries based on obese population, no outliers were spotted. Furthermore, a frequency diagram was created to analyze the frequency of obese people in and outside the European Continent.

Apart from that. The created line diagrams depict how obesity is related to the timeline provided by the dataset. The overlapping histograms were created to show the difference between men and women independent of the countries. Finally the bar plot depicts the total obesity for the countriescalculated as the total sum of women and men.