Data Mining Project (MaBAn 2020)

Predicting obesity levels according to daily habits

by : Ángel Tomás-Ripoll & Laurence Tréteault-Falsafi

Contents

Introduction **1**

Data Pre-Processing **4**

Exploratory Data Analysis **6**

Introduction

For this project, our objective is to predict the expected weight level (in Kg) for a given person depending on certain daily habits (eating and physical activity) and on the person's age, gender and height.

To do this, we found a quite interesting dataset (click here : http://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+) containing 2111 observations and 17 variables (mainly categorical).

Please, find here a manually created metadata table :

```r
# To adjust the page margins when knitting to PDF :

library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=45),tidy=TRUE)
```

```r
# Used packages :
library(pander)
library(dplyr)
library(gt)
library(car)
library(ggplot2)
library(gridExtra)

# Working Directory :
setwd("~/GitHub/CVTDM_Project_MaBAn_2020")
```

```
# Reading the data :
obesity <- read.csv("Obesity.csv", header = T,
    sep = ",")
attach(obesity)



# Small metadata table :

tibble_table <- tibble(`Variable Name` = c(colnames(obesity)[1:14],
    "", colnames(obesity)[15:17]), Description = c("Gender",
    "Age", "Height", "Weight", "Has a family member suffered or suffers from overweight?",
    "Do you eat high caloric food frequently?",
    "Do you usually eat vegetables in your meals?",
    "How many main meals do you have daily?",
    "Do you eat any food between meals?", "Do you smoke?",
    "How much water do you drink daily?", "Do you monitor the calories you eat daily?",
    "How often do you have physical activity?",
    "How much time do you use technological devices such as",
    "cell phone videogames, television, computer and others?",
    "How often do you drink alcohol?", "Which transportation do you usually use?",
    "Obesity level based on calculation of Mass Body Index"))

metadata <- gt(data = tibble_table)

metadata %>% tab_header(title = md("**Metadata**"),
    subtitle = "from the dataset we are using") %>%

tab_source_note(source_note = "Based on information in :

  https://www.sciencedirect.com/science/article/pii/S2352340919306985")
```

Metadata

from the dataset we are using

Variable Name	Description
Gender	Gender
Age	Age
Height	Height
Weight	Weight
family_history_with_overweight	Has a family member suffered or suffers from overweight?
FAVC	Do you eat high caloric food frequently?
FCVC	Do you usually eat vegetables in your meals?
NCP	How many main meals do you have daily?
CAEC	Do you eat any food between meals?
SMOKE	Do you smoke?
CH2O	How much water do you drink daily?
SCC	Do you monitor the calories you eat daily?
FAF	How often do you have physical activity?
TUE	How much time do you use technological devices such as
	cell phone videogames, television, computer and others?
CALC	How often do you drink alcohol?
MTRANS	Which transportation do you usually use?
NObeyesdad	Obesity level based on calculation of Mass Body Index

Based on information in :
https://www.sciencedirect.com/science/article/pii/S2352340919306985

Here is a small overview of the first observations :

```r
pander(head(obesity))
```

Table continues below

Gender	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC
Female	21	1.62	64	yes	no	2
Female	21	1.52	56	yes	no	3
Male	23	1.8	77	yes	no	2
Male	27	1.8	87	no	no	3
Male	22	1.78	89.8	no	no	2
Male	29	1.62	53	no	yes	2

Table continues below

NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC
3	Sometimes	no	2	no	0	1	no
3	Sometimes	yes	3	yes	3	0	Sometimes
3	Sometimes	no	2	no	2	1	Frequently
3	Sometimes	no	2	no	2	0	Frequently
1	Sometimes	no	2	no	0	0	Sometimes
3	Sometimes	no	2	no	0	0	Sometimes

MTRANS	NObeyesdad
Public_Transportation	Normal_Weight
Public_Transportation	Normal_Weight
Public_Transportation	Normal_Weight
Walking	Overweight_Level_I
Public_Transportation	Overweight_Level_II
Automobile	Normal_Weight

The variable of interest is the fourth one, the “Weight”, so it will be our dependent variable.

We were “lucky” on the fact that this dataset has a quite high level of quality, because it has no missing observations, and our subsequent exploratory analysis will tell us if there are outliers to be handled with.

Once we are done with a Data Exploratory Analysis and with a proper Data Pre-Processing, we will develop several models in order to accurately predict the level of weight of each individual.

The models will be :

1. **Multiple Linear Regression** (not ANOVA since "Age" and "Height" are numerical)
2. **Classification tree** (complemented with a random forest / boosted trees / bagged trees)
3. **k-Nearest Neighbors**
4. **Ensemble Method**

We will deploy the best model based on error metrics and prediction performance.

At the very end, we will make a Shiny App available, in which any user can fill-in a questionnaire concerning daily habits, age and height. Then, the App will tell the user what is the expected weight according to those characteristics, and will present the result in two forms :

- The expected weight in Kg.

- The expected obesity level based on the Body Mass Index, following the classification comming from the World Health Organisation.

The user will also be able to select the type of model that will predict the results. That way, it will be interesting to see with just a few clicks how each model will yield different results.

Data Pre-Processing

The first thing to do is to change the column names so that they are more visually meaningful!

```r
# Changing column names:

names(obesity)[5] = "family_history"
names(obesity)[6] = "eat_caloric"
names(obesity)[7] = "vegetables"
names(obesity)[8] = "main_meals"
names(obesity)[9] = "food_inbetween"
names(obesity)[12] = "monitor_cal"
names(obesity)[13] = "physical_act"
names(obesity)[14] = "tech_devices"
names(obesity)[15] = "alcohol"
```

```r
# Checking the dataset structure :

str(obesity)
```

```
## 'data.frame':    2111 obs. of  17 variables:
##  $ Gender        : chr  "Female" "Female" "Male" "Male" ...
##  $ Age           : num  21 21 23 27 22 29 23 22 24 22 ...
##  $ Height        : num  1.62 1.52 1.8 1.8 1.78 1.62 1.5 1.64 1.78 1.72 ...
##  $ Weight        : num  64 56 77 87 89.8 53 55 53 64 68 ...
##  $ family_history: chr  "yes" "yes" "yes" "no" ...
##  $ eat_caloric   : chr  "no" "no" "no" "no" ...
##  $ vegetables    : num  2 3 2 3 2 2 3 2 3 2 ...
##  $ main_meals    : num  3 3 3 3 1 3 3 3 3 3 ...
##  $ food_inbetween: chr  "Sometimes" "Sometimes" "Sometimes" "Sometimes" ...
##  $ SMOKE         : chr  "no" "yes" "no" "no" ...
##  $ CH2O          : num  2 3 2 2 2 2 2 2 2 2 ...
##  $ monitor_cal   : chr  "no" "yes" "no" "no" ...
##  $ physical_act  : num  0 3 2 2 0 0 1 3 1 1 ...
##  $ tech_devices  : num  1 0 1 0 0 0 0 0 1 1 ...
##  $ alcohol       : chr  "no" "Sometimes" "Frequently" "Frequently" ...
##  $ MTRANS        : chr  "Public_Transportation" "Public_Transportation" "Public_Transportation" "Walk
##  $ NObeyesdad    : chr  "Normal_Weight" "Normal_Weight" "Normal_Weight" "Overweight_Level_I" ...
```

As we see here, all the categorical variables are treated as `character`.

Therefore, we will first of all convert all the categorical variables to `factor` type.

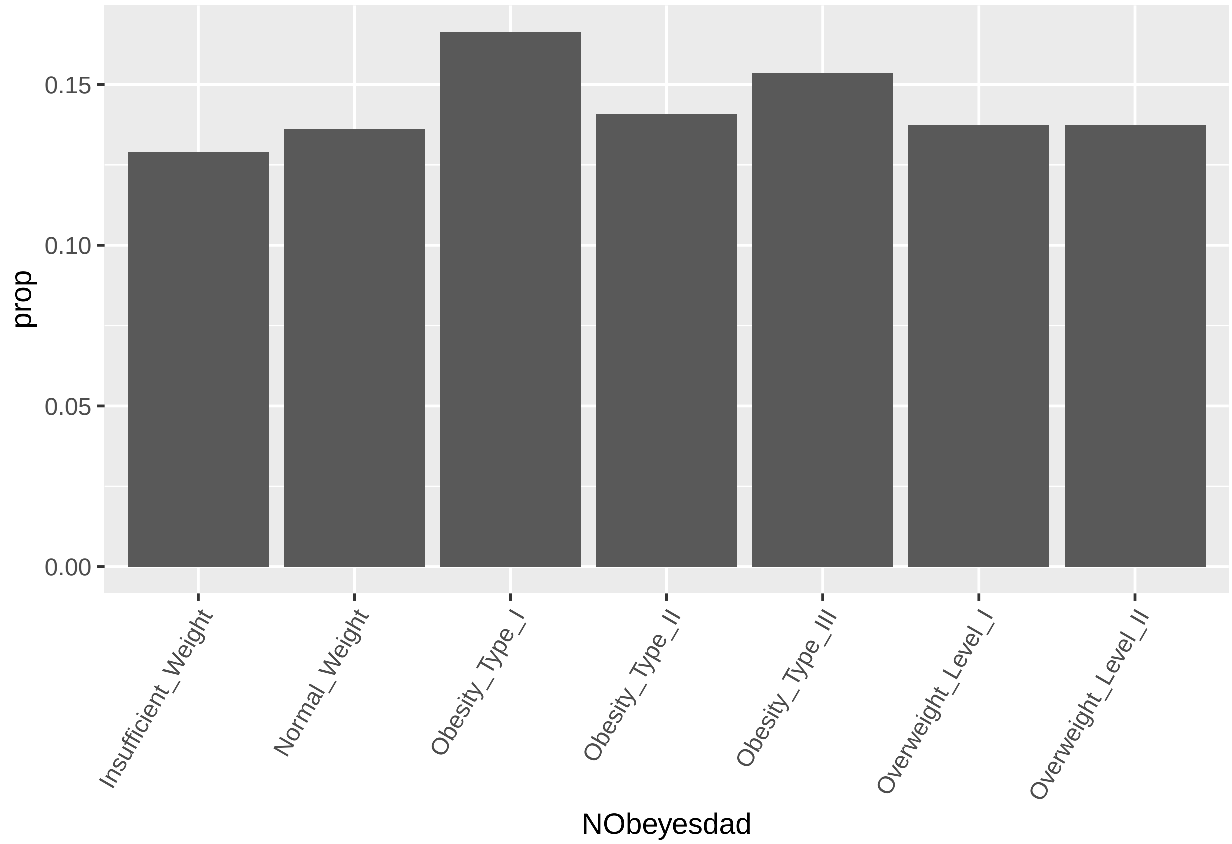
```r
# Converting character variables to factor :

obesity$Gender = as.factor(obesity$Gender)
obesity$family_history = as.factor(obesity$family_history)
obesity$eat_caloric = as.factor(obesity$eat_caloric)
obesity$food_inbetween = as.factor(obesity$food_inbetween)
obesity$SMOKE = as.factor(obesity$SMOKE)
obesity$monitor_cal = as.factor(obesity$monitor_cal)
obesity$alcohol = as.factor(obesity$alcohol)
obesity$MTRANS = as.factor(obesity$MTRANS)
obesity$NObeyesdad = as.factor(obesity$NObeyesdad)
```

Now, since many variables are in fact numerical and continuous between a range (for example `vegetables`, inside the range 1 to 3), we will transform them into categorical. This is, somehow, BINNING. For this, we will follow the names given in the information file refered to earlier (https://www.sciencedirect.com/science/article/pii/S2352340919306985).

```
# Binning some numerical variables :

obesity$vegetables[obesity$vegetables <= 1] <- "Never"

obesity$vegetables[obesity$vegetables > 1 & obesity$vegetables <=
    2] <- "Sometimes"

obesity$vegetables[obesity$vegetables > 2 & obesity$vegetables <=
    3] <- "Always"



obesity$main_meals[obesity$main_meals >= 1 & obesity$main_meals <
    3] <- "Btw_1_&_2"

obesity$main_meals[obesity$main_meals == 3] <- "Three"

obesity$main_meals[obesity$main_meals > 3 & obesity$main_meals <=
    4] <- "More_than_3"
```

Exploratory Data Analysis

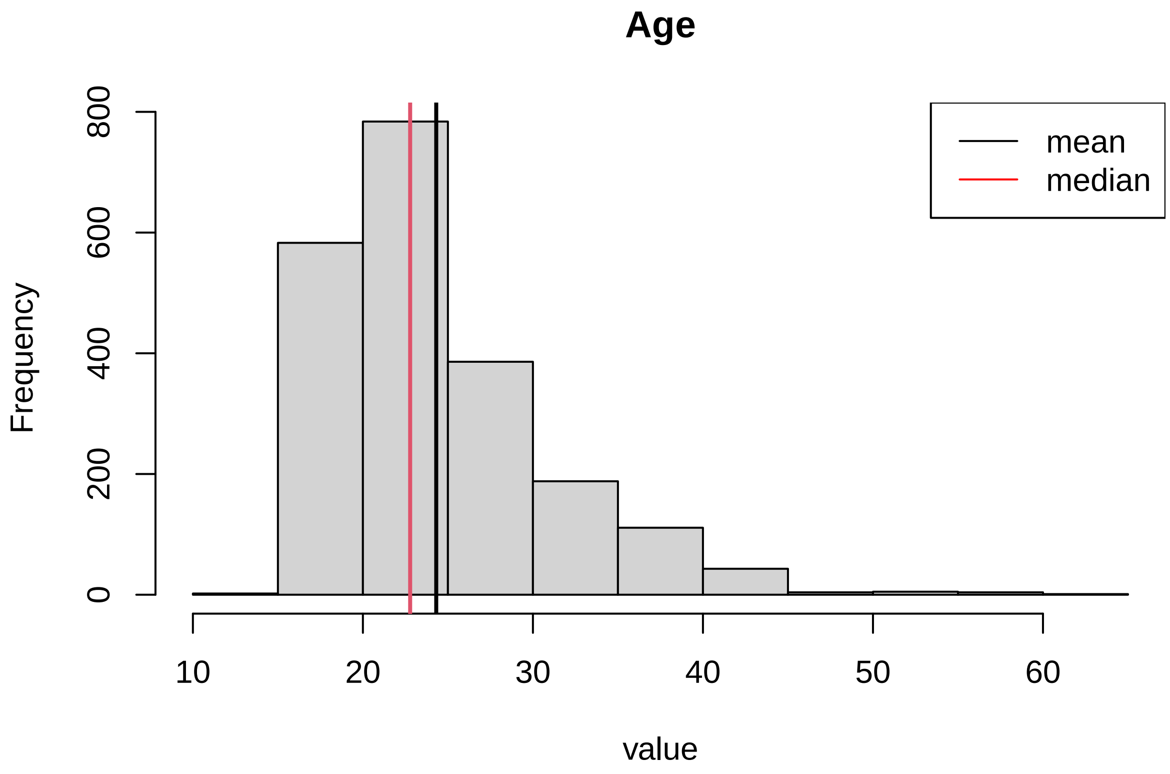
```
ggplot(data = obesity, aes(x = NObeyesdad)) +
    geom_bar(aes(y = ..prop.., group = 1)) + theme(axis.text.x = element_text(angle = 60,
    hjust = 1))
```

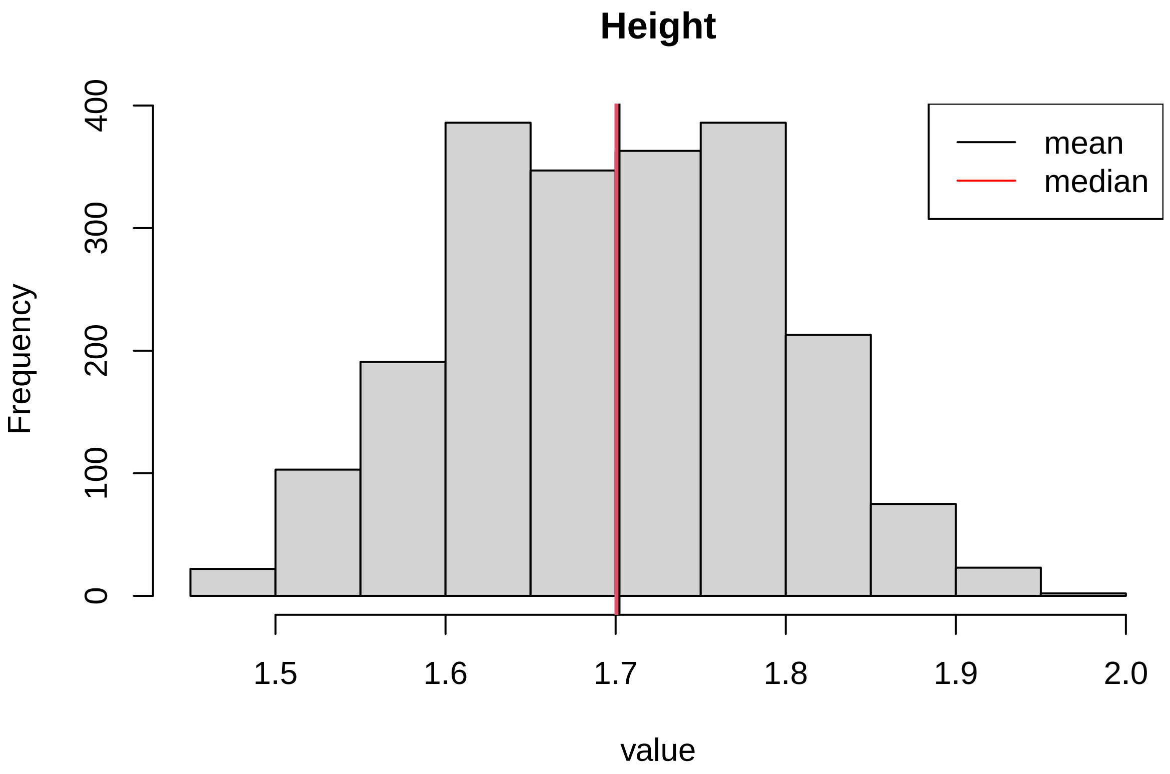


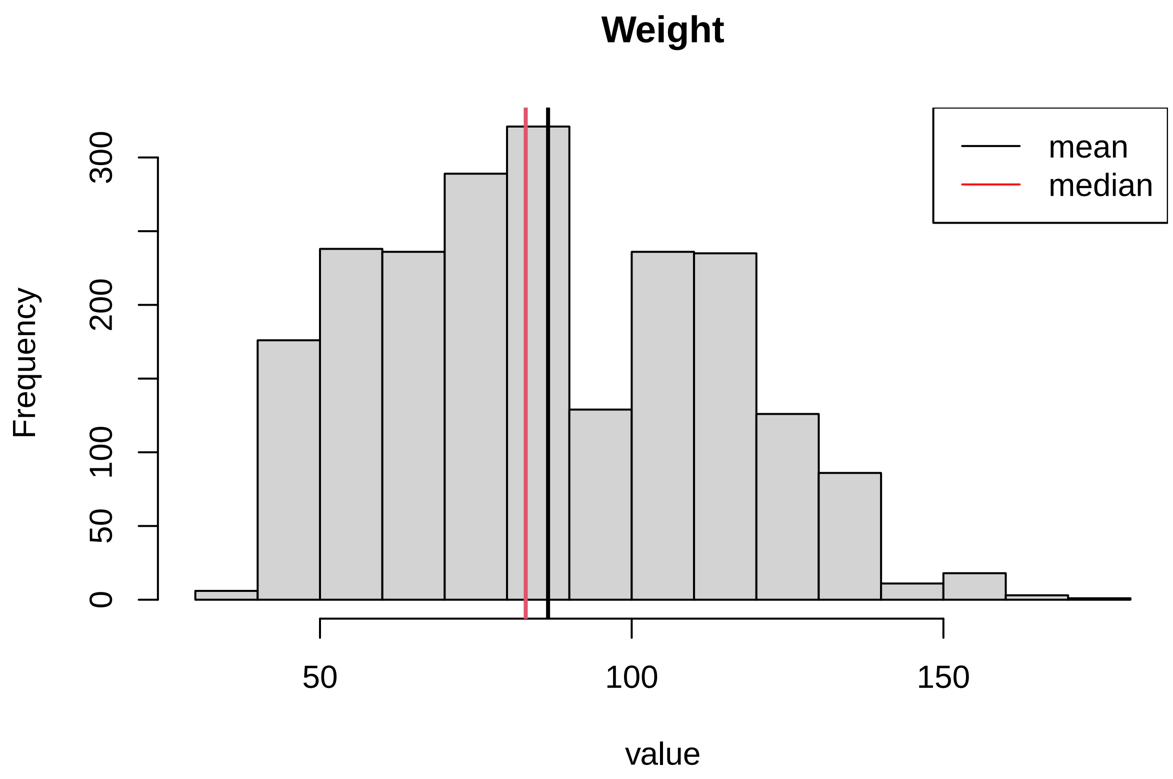
We see that the distribution of observations across the different weights is quite uniform, meaning that we do not have an unbalanced data set with respect to our variable of interest (the weight).

Let's now look at some histograms for all the continuous variables in our dataset.

```r
# Creating histograms :

for (i in 2:4) {
    hist(obesity[, i], breaks = 10, main = names(obesity[i]),
        xlab = "value", freq = T)
    abline(v = mean(obesity[, i]), col = 1, lwd = 2)
    abline(v = median(obesity[, i]), col = 2,
        lwd = 2)
    legend("topright", legend = c("mean", "median"),
        col = c("black", "red"), lty = 1)
}
```


Age
mean
median
Frequency
0
200
400
600
800
10
20
30
40
50
60
value

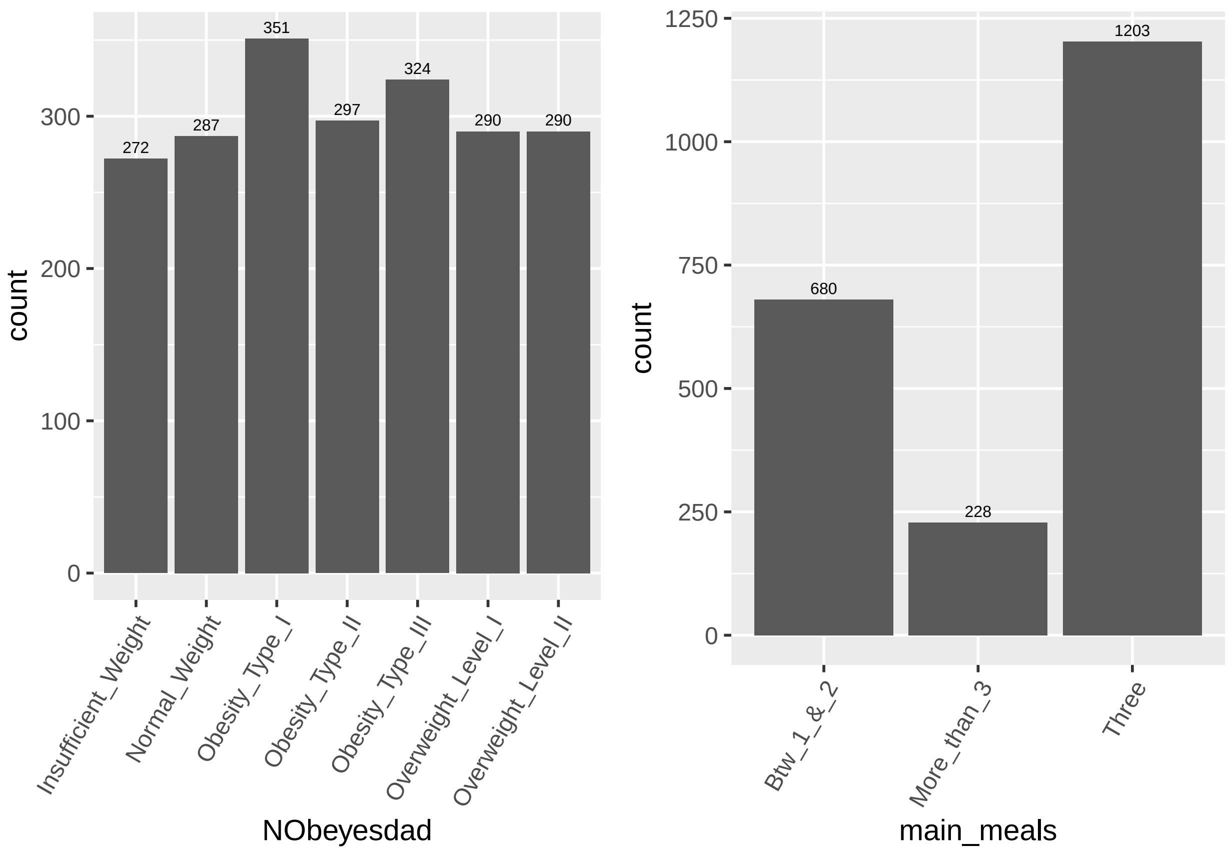

Height
mean
median
Frequency
0
100
200
300
400
1.5
1.6
1.7
1.8
1.9
2.0
value



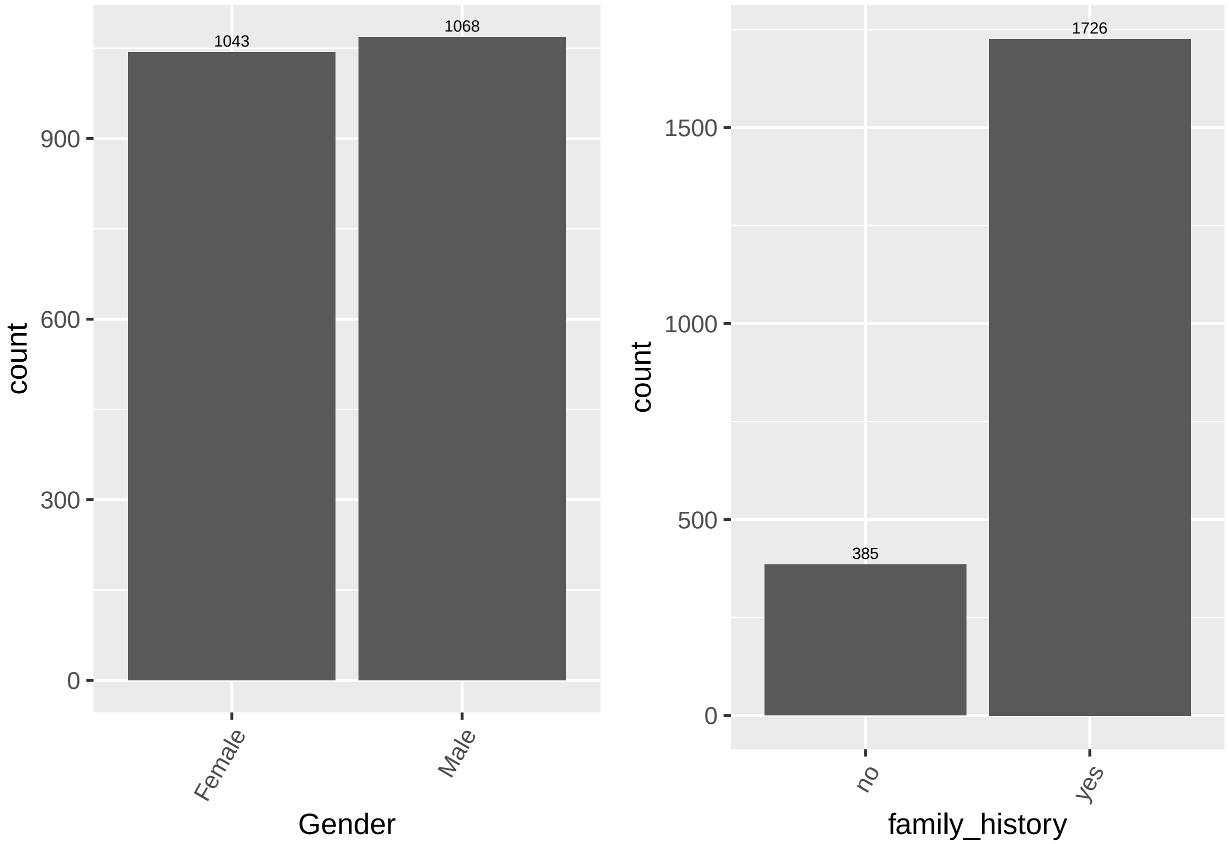
Now, let's do some barplots in order to get an idea of the distribution of each of the categorical variables.

```
# Barplots :

plot_1 = ggplot(data = obesity, aes(x = NObeyesdad)) +
    geom_bar(aes(y = ..count.., group = 1)) +
    theme(axis.text.x = element_text(angle = 60,
        hjust = 1)) + geom_text(stat = "count",
    aes(label = ..count..), vjust = -0.5, size = 2.2)

plot_2 = ggplot(data = obesity, aes(x = main_meals)) +
    geom_bar(aes(y = ..count.., group = 1)) +
    theme(axis.text.x = element_text(angle = 60,
        hjust = 1)) + geom_text(stat = "count",
    aes(label = ..count..), vjust = -0.5, size = 2.2)

plot_3 = ggplot(data = obesity, aes(x = Gender)) +
    geom_bar(aes(y = ..count.., group = 1)) +
    theme(axis.text.x = element_text(angle = 60,
        hjust = 1)) + geom_text(stat = "count",
    aes(label = ..count..), vjust = -0.5, size = 2.2)
```

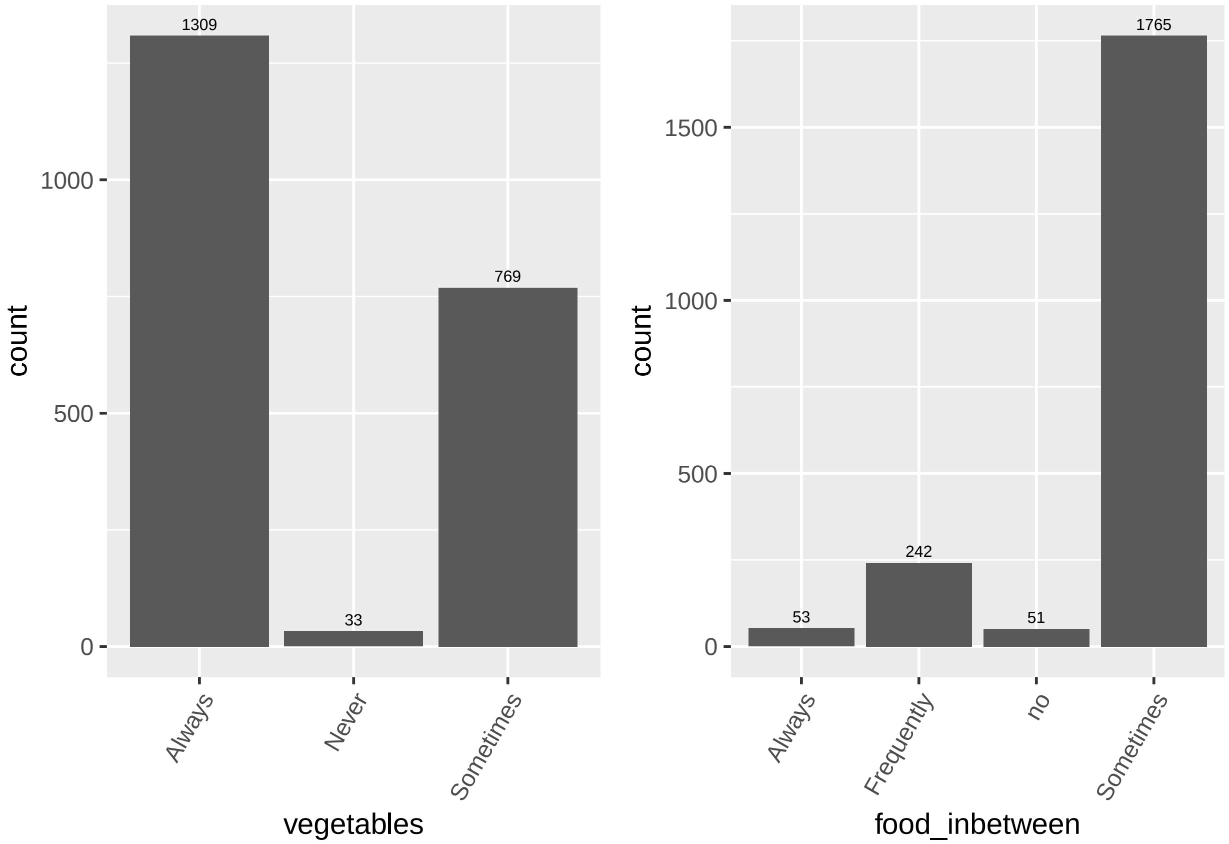
```r
plot_4 = ggplot(data = obesity, aes(x = family_history)) +
    geom_bar(aes(y = ..count.., group = 1)) +
    theme(axis.text.x = element_text(angle = 60,
        hjust = 1)) + geom_text(stat = "count",
    aes(label = ..count..), vjust = -0.5, size = 2.2)

plot_5 = ggplot(data = obesity, aes(x = vegetables)) +
    geom_bar(aes(y = ..count.., group = 1)) +
    theme(axis.text.x = element_text(angle = 60,
        hjust = 1)) + geom_text(stat = "count",
    aes(label = ..count..), vjust = -0.5, size = 2.2)

plot_6 = ggplot(data = obesity, aes(x = food_inbetween)) +
    geom_bar(aes(y = ..count.., group = 1)) +
    theme(axis.text.x = element_text(angle = 60,
        hjust = 1)) + geom_text(stat = "count",
    aes(label = ..count..), vjust = -0.5, size = 2.2)



# Arranging them two-by-two :

grid.arrange(plot_1, plot_2, ncol = 2)
```



```r
grid.arrange(plot_3, plot_4, ncol = 2)
```



```r
grid.arrange(plot_5, plot_6, ncol = 2)
```


count
Always
Never
Sometimes
vegetables
count
Always
Frequently
no
Sometimes
food_inbetween