Academic year 2019-2020

PROJECT 1 - Exploratory data analysis

High-dimensional data analysis

Baudenne Céline s154198
Girineza Guy s144377
Payan Palomo Fatima s190047

Master civil Engineering in Data science
Master in Mathematical Sciences

A. Presentation of the data

The data set chosen for this project is the result of an investigation about the placebo effect done by an Harvard psychologist. It contains physiological measures about 75 female room attendants working in seven different hotels. Among the participants, 41 were randomly selected and were explained that their work satisfies the recommendations for a good and active lifestyle. The 34 other women belonged to the control group and were told nothing. The measures have been collected in two steps; the subjects had first to fill a questionnaire and were taken one after the others to collect their physiological measures, then four weeks later the same measures have been taken again.
The data set is composed of 17 variables in total among which one binary indicator specifying if the subject has been informed or not and a variable indicating their age; the other variables are described in the following table:

Variables	Meaning
Wt - Wt2	Weight
BMI - BMI2	Body mass index
Fat - Fat2	Percentage of body fat
WHR - WHR2	Waist-to-hip ratio
Syst - Syst2	Systolic blood pressure
Diast - Diast2	Diastolic blood pressure
Fat.Ft2	Difference between Fat and Fat2
BMI.BM2	Difference between BMI and BMI2
Dst.Ds2	Difference between Diast and Diast2

The principal question of interest that can be addressed with this data set is to see if there is actually an observable difference in the measures between the women that have been informed about the benefits their work could have on their health and the women that were not aware.

B. Missing data

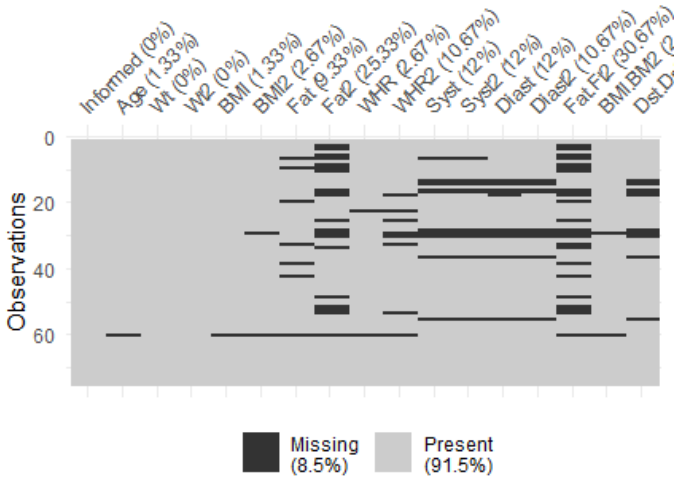


Figure 1 – Visualisation of missing values among the data.

It can be observed on Fig. 1 that for some subjects, there is no value for any of the variables concerning the blood pressure, namely *Syst*, *Syst2*, *Diast* and *Diast2*. We tried to see if this missingness could be related to the age of the subjects. Indeed, it can be seen on the two graphs of Fig. 2 that missing values of the variable *Syst* seem to be mostly found for women of advanced age. To statistically check if the

difference in the means is actually significant, a t-test has been performed from which we can conclude the rejection of the null hypothesis of no difference between the means ($p - value = .02255 < .05$). The same conclusion can be drawn for the three other variables [1]. We can derive from this analysis that the missing values likely follow the MAR mechanism as the missingness seems to be conditioned on an other observed data.

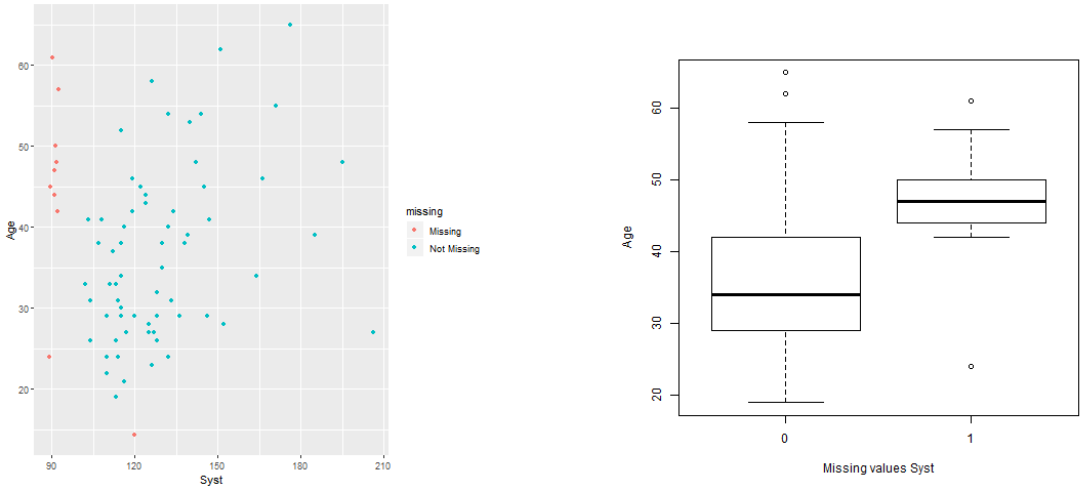


Figure 2 – Visualisation of the missingness of variable Syst with respect to the age.

Concerning the variable *WHR2* (waist-to-hip ratio), the visual inspection of the data and the comparison of the means with respect to the variable *Age* and *Wt* showed no difference between the two groups (missing and not missing). We might thus assume that the MCAR mechanism holds for this variable.

Regarding the variable *Fat* and *Fat2*, again the visual inspection of the data and the comparison of the means with respect to the variable *Age* and *Wt* do not show some links between the missing values and the other observed data. However, the high proportion of missing values of the variable *Fat2* could be explained by a problem in the measurement device. The percentage of body fat is computed according to an equation based on the weight and the impedance which were measured using a scale. It could be that for some reason the device failed at measuring the impedance or at computing some results during the second phase of measurements.

For the treatment of the missing values in the following exploratory analysis, given the configuration of the data set and the fact that the main goal of the measurements are to observe a possible different evolution between the two phase of measurement in the two different groups, we decided that no imputation techniques could be used as it could weakened the relationships between the variables and their evolution. For most of the analysis, the cell-deletion technique will be applied as most of them consist in pair-wise comparisons.

[1] It should be noted that the normality assumption does not hold for some groups

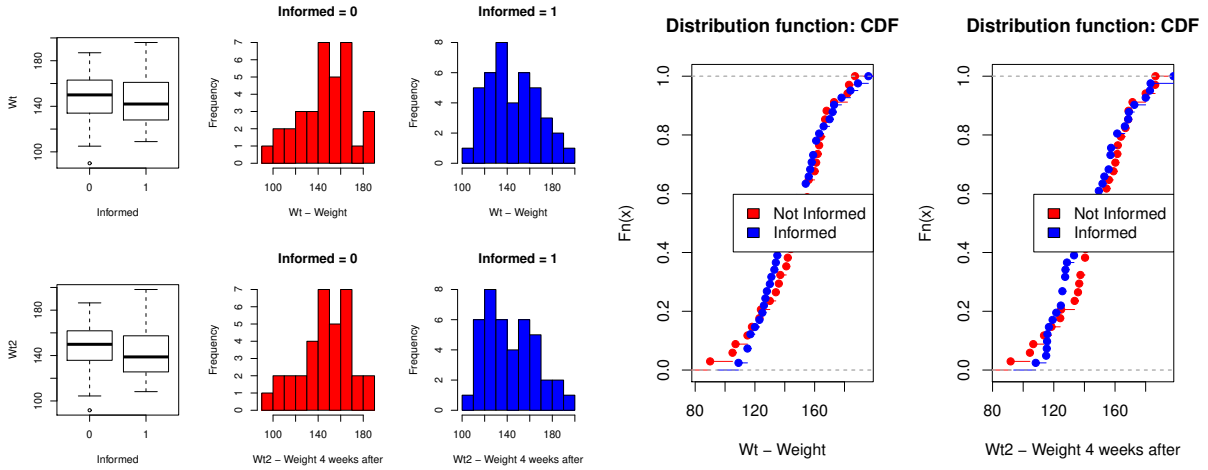
C. Exploratory analysis

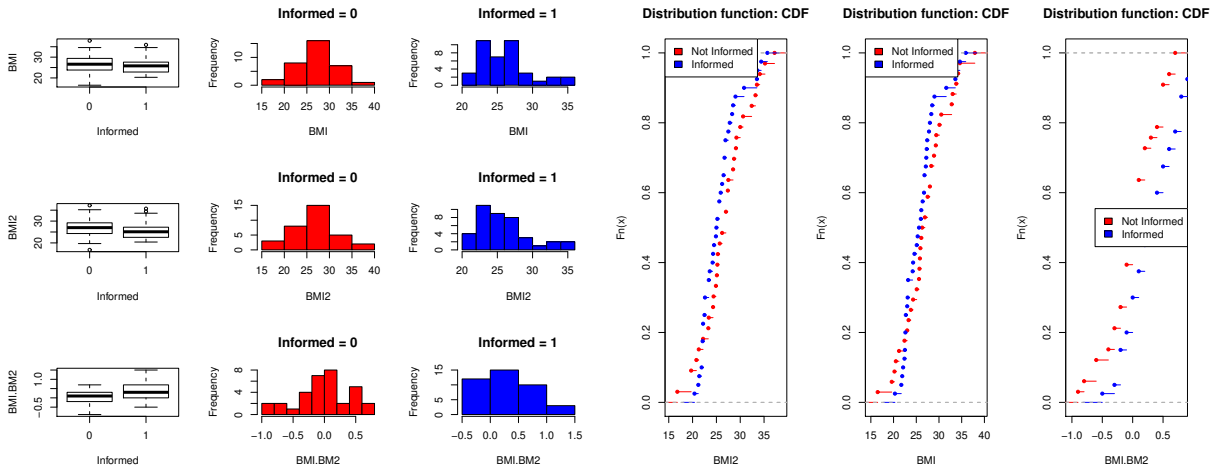
1. Statistical and graphical summary of the data.

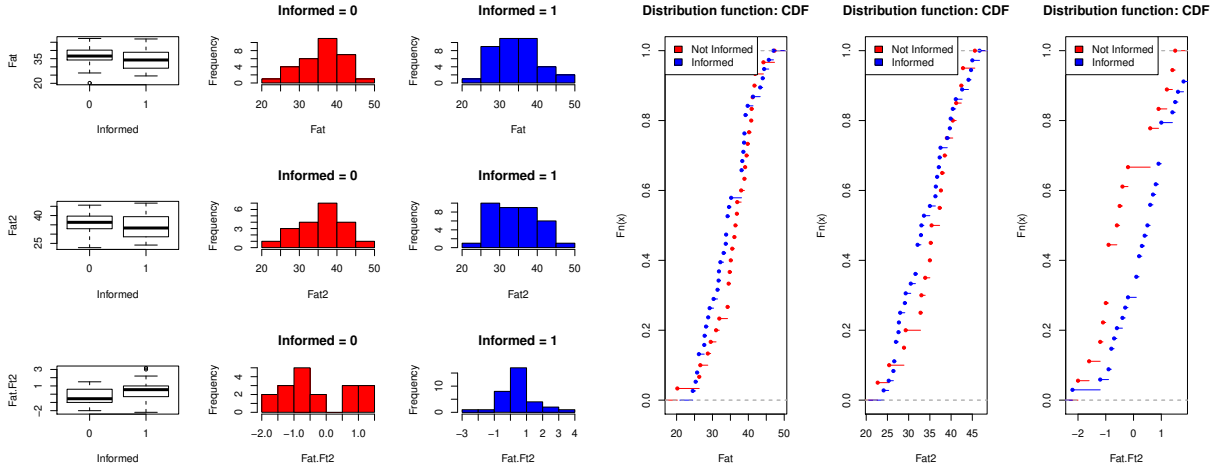
The following table contains the statistical summary of the data.

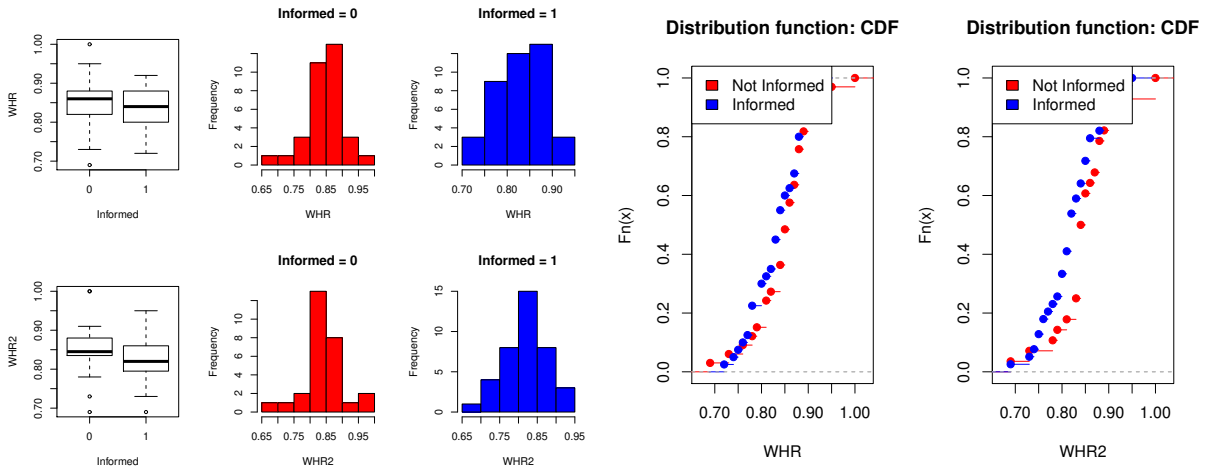
Variable	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Age	19.0	29.0	38.0	37.72	45.0	65.0
Wt	90.0	130.0	145.0	146.1	161.5	196.0
Wt2	91.8	127.4	143.0	145.1	161.4	198.2
BMI	16.5	23.12	26.1	26.46	28.3	37.9
BMI2	16.8	23.3	25.6	26.22	28.5	37.2
Fat	20.2	31.3	35.35	35.31	39.27	47.3
Fat2	22.7	29.3	35.1	34.82	39.25	46.7
WHR	0.69	0.81	0.85	0.8436	0.88	1
WHR2	0.69	0.805	0.84	0.836	0.875	1
Syst	102.0	114.2	124.5	129.2	137.5	206.0
Syst2	83.0	112.0	117.5	121.7	126.0	190.0
Diast	56.0	69.0	75.0	78.68	83.75	131.0
Diast2	48.0	67.5	74.0	74.78	78.5	124.0
Fat.Ft2	-2.2	-0.725	0.25	0.2269	1	3.2
BMI.BM2	-0.9	-0.1	0.2	0.2068	0.5	1.5
Dst.Ds2	-0.15	-3.75	2.0	3.894	7.750	55.0

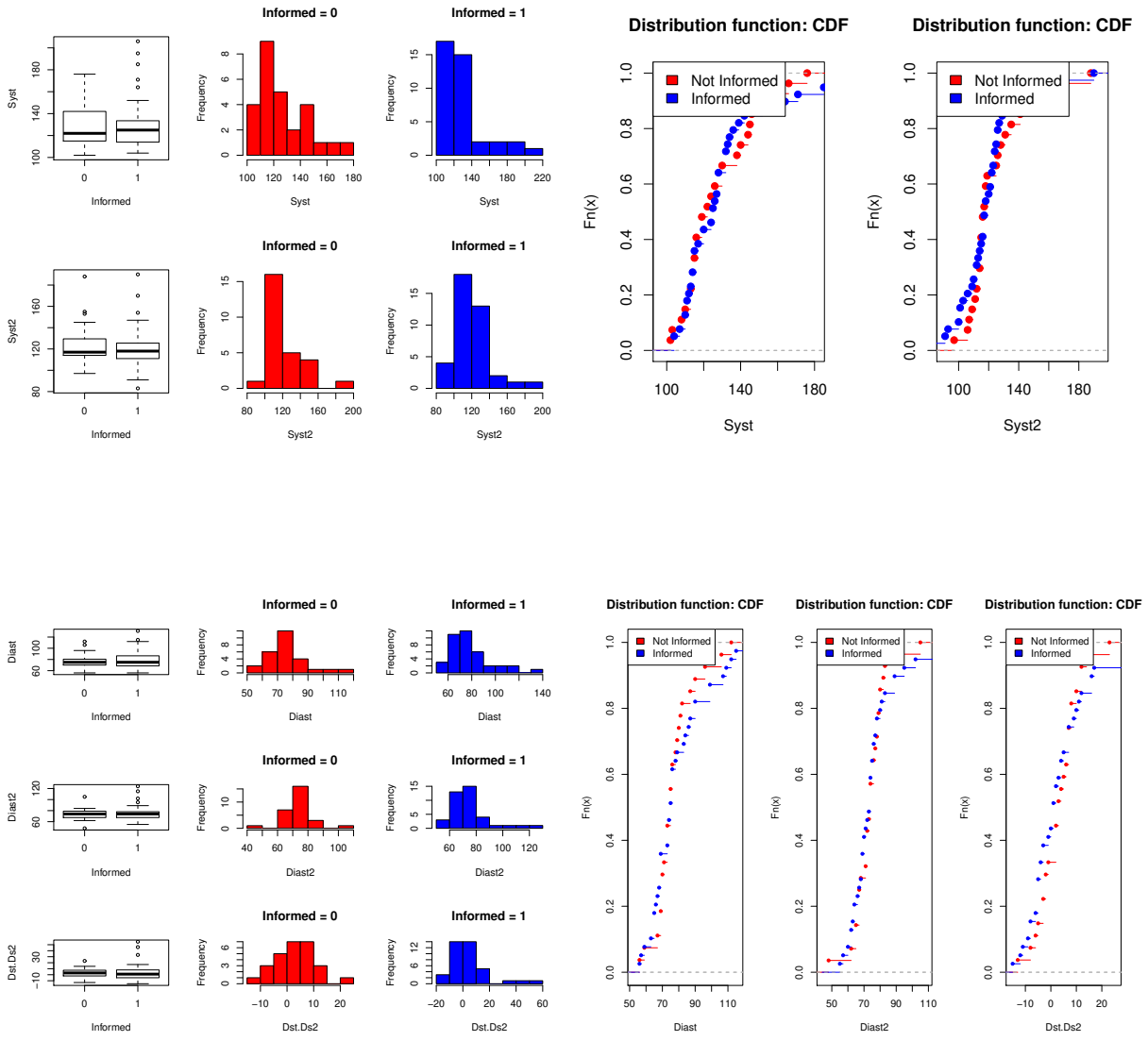
The following graphs give a general representation of the data. We chose to observe the data from the point of view of the difference between the informed and not informed groups. A table containing information about those split variables can be found in the annex 1.




Informed = 0
Informed = 1
Distribution function: CDF
Not Informed
Informed
BMI
BMI2
BMI.BM2
Frequency
Fn(x)
Informed


Informed = 0
Informed = 1
Distribution function: CDF
Not Informed
Informed
Fat
Fat2
Fat.Ft2
Frequency
Fn(x)
Informed


Informed = 0
Informed = 1
Distribution function: CDF
Not Informed
Informed
WHR
WHR2
Frequency
Fn(x)
Informed



2. Correlation structure

The figure 3 represents visually the correlation matrix of the quantitative variables. It can be observed that the variables *Wt*, *BMI* and *Fat* are highly positively correlated. This is quite logical as the body-mass index and the percentage of body fat are computed namely based on the weight. The same is true for the variables *Syst* and *Diast* as there are strongly related. Also, obviously the variables measured after four weeks are strongly correlated with the variables measured the first time.

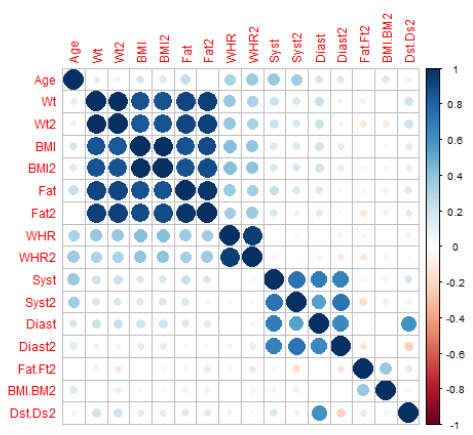


Figure 3 – Visualisation of the correlation matrix of the quantitative data.

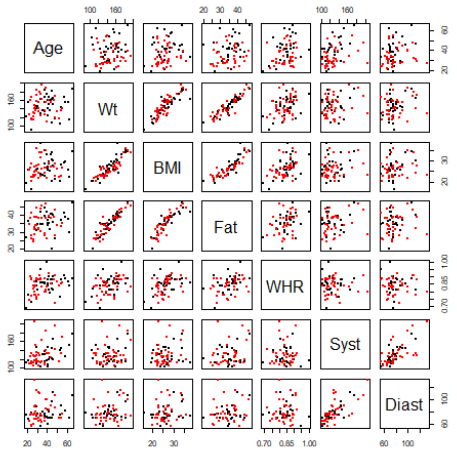


Figure 4 – Scatter plot matrix.

The figure 4 is an other way of representing the relationship between the variables two by two. We can observed the same finding as in figure 3. The high positive correlation between the variables *Wt*, *BMI* and *Fat* is clearly observable. Here, it can also be seen that the variable *WHR* and the variable *BMI* seem to be slightly positively correlated. This is also reasonable as it is excepted that the body-mass index computed based on the weight and the height has some relation with the waist-to-hip ratio.

3. Impact of the qualitative variable on the quantitative variables.

The graphical summary of the data already shows the impact of the qualitative variable on the other variables, here we chose to look to some variables more in depth.

Both graphs of Fig. 5 represent the variable *Age* with respect to the two groups of the qualitative variable *Informed* namely the informed group and the control group. A discrepancy can be observed between the two groups. In fact, women in the informed group seem to have in average a younger age than women in the control group.

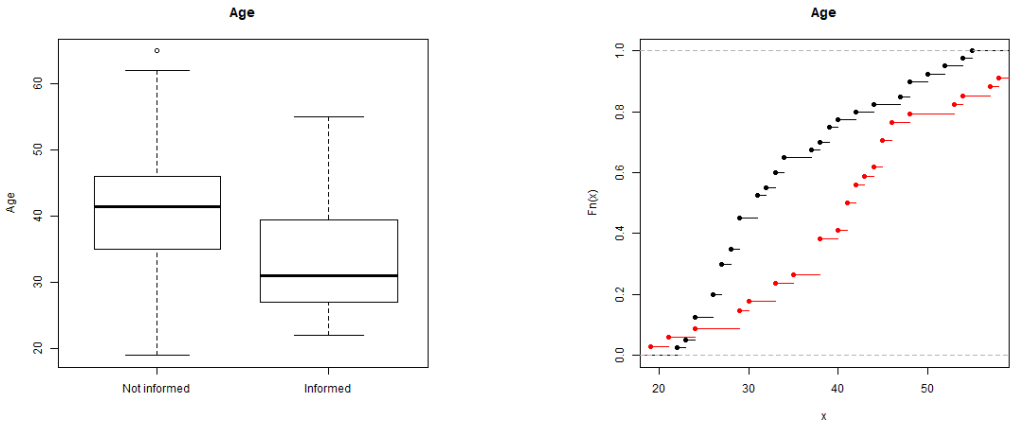


Figure 5 – Impact of the qualitative variable on the variable Age *(Black = Informed)*.

The figure 6 display the impact on the variable *BMI.BM2* which is the difference between the body-mass index at the two measurement moments with respect to the qualitative variable *Informed*. This enable to see if there is a difference in the evolution of the variable measured at four weeks interval. The results show a slightly bigger and positive difference for the informed women compared to the women of

the control group. When doing the t-test we obtained a p-value equal to .00248 which implies a rejection of the null hypothesis of no difference in the means[2]. The difference can thus be considered as quite significant.

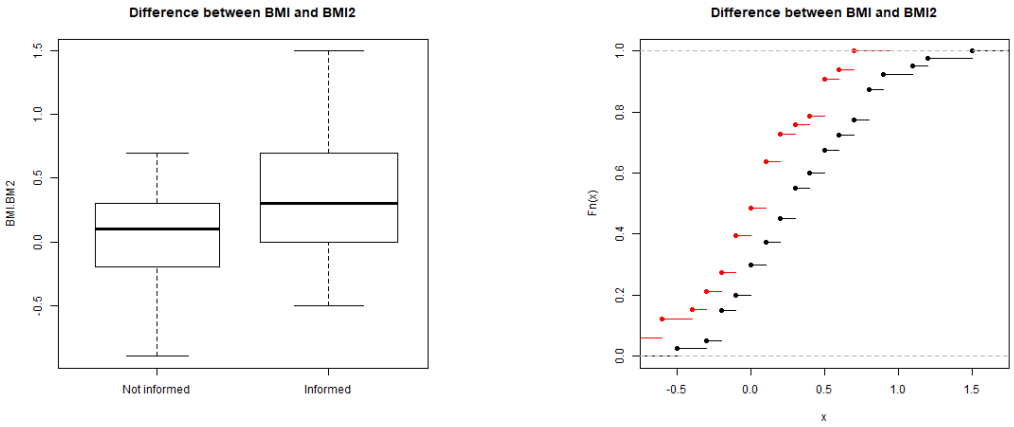


Figure 6 – Impact of the qualitative variable on the variable BMI.BM2 *(Black = Informed)*.

4. Outliers

It is very important to highlight that all the statistical techniques (boxplot, z-Score, Mahalanobis distance) used in this part are applied under the underlying assumption of normality. Some of the attributes do not really follow the shape of a normal distribution. This is why any conclusion should take into account this observation[3].

In order to analyse the outliers of this study, we first compute the box-plots of all attributes. Figure 7 display only the variables for which the number of outlying observations was higher.

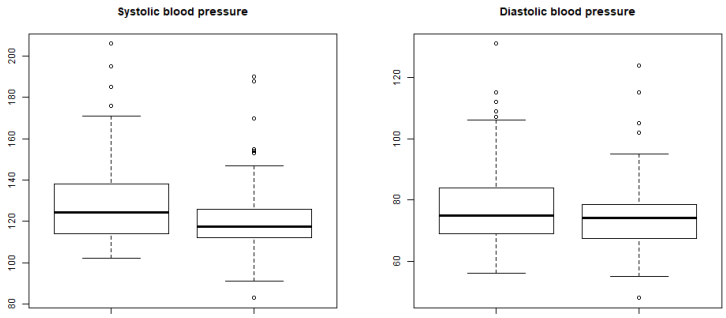


Figure 7 – Graphs of the boxplot for first data gathering (left side) and second data gathering (right side).

In order to see if we can confirm that those observations are real outliers, we computed and plotted their z-score.

[2]It can be observed on the graph in the data summary that the normality assumption seems more or less to hold here.
[3]Graphs showing the distribution's shape of the variables studied in the following can be found in annex 2

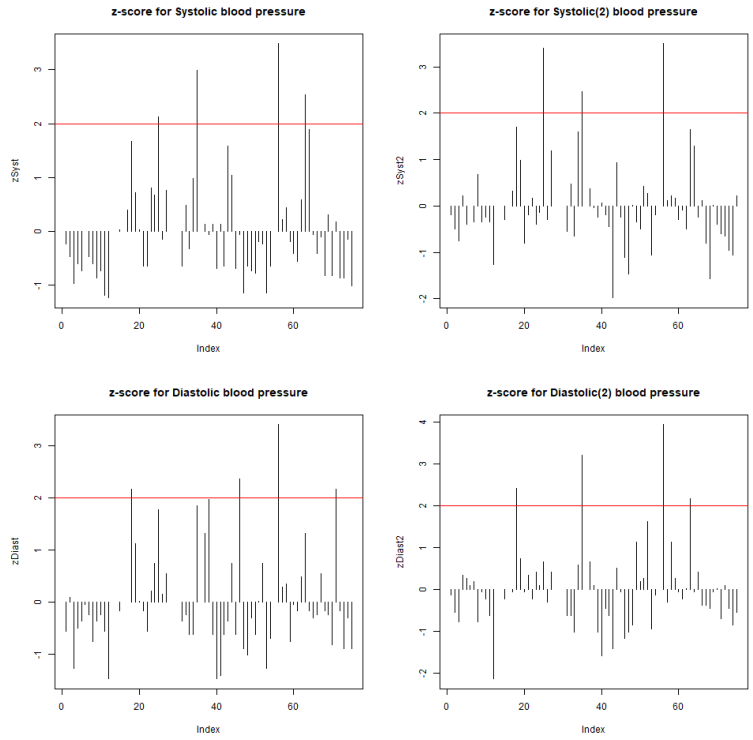


Figure 8 – Graphs of the z-score for the variables Syst, Syst2, Diast and Diast2.

As you can see, in all the four plots one observations (the 56^{th}) seems really to come out as being suspicious. When looking in more details into the data, we saw that this observations had actually the maximum values for the four variables. Let's compare this with the outliers emerging when analysing the Mahalanobis distance.

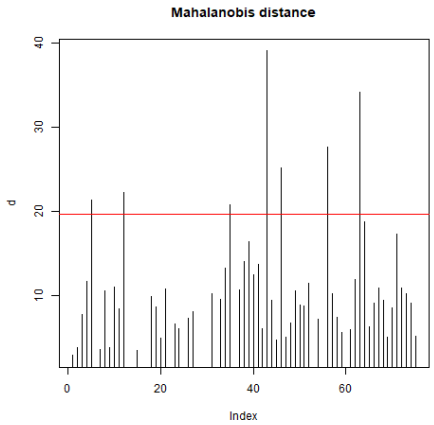


Figure 9 – Graphs of the Mahalanobis distances.

The Mahalanobis distance has been computed with all quantitative variables except the three last variables as well as the variables *Fat* and *Fat2*. Indeed those two variables include a lot of missing values which implies a serious decrease in the size of the data set because of the deletion of the cases containing missing values. We can see on fig. 9 that four observations are above the cutoff of the 95% quartile of the corresponding χ^2 distribution, among those is the 56^{th} observations. Again when looking into the data for those observations, we saw that there values are mostly among the highest for some attributes.

ANNEX

1

	Not Informed ($Informed = 0$)			Informed ($Informed = 1$)		
Variable	mean	sd	average rate of missingness	mean	sd	average rate of missingness
Age	41.88	11.27	0%	34.17	9.35	2.44%
Wt	146.91	23.23	0%	145.48	22.07	0%
Wt2	146.71	23.00	0%	143.70	22.42	0%
BMI	26.95	4.77	0%	26.04	3.76	2.44%
BMI2	26.85	4.77	2.94%	25.69	3.78	2.44%
Fat	36.08	5.82	11.76%	34.7	6.18	7.32%
Fat2	35.71	5.88	41.18%	34.32	6.28	12.2%
WHR	0.85	0.05	2.94%	0.83	0.051	2.44%
WHR2	0.85	0.06	17.65%	0.82	0.05	4.88%
Syst	127.74	18.85	20.59%	130.15	24.24	4.88%
Syst2	123.92	19.03	20.59%	120.15	19.94	4.88%
Diast	77.33	12.38	20.59%	79.61	17.18	4.88%
Diast2	73.92	9.71	17.65%	75.38	14.25	4.88%
Fat.Ft2	-0.28	1.06	47.06%	0.49	1.14	17.07%
BMI.BM2	0.03	0.40	2.94%	0.35	0.45	2.44%
Dst.Ds2	3.40	7.60	20.59%	4.23	14.40	4.88%

2

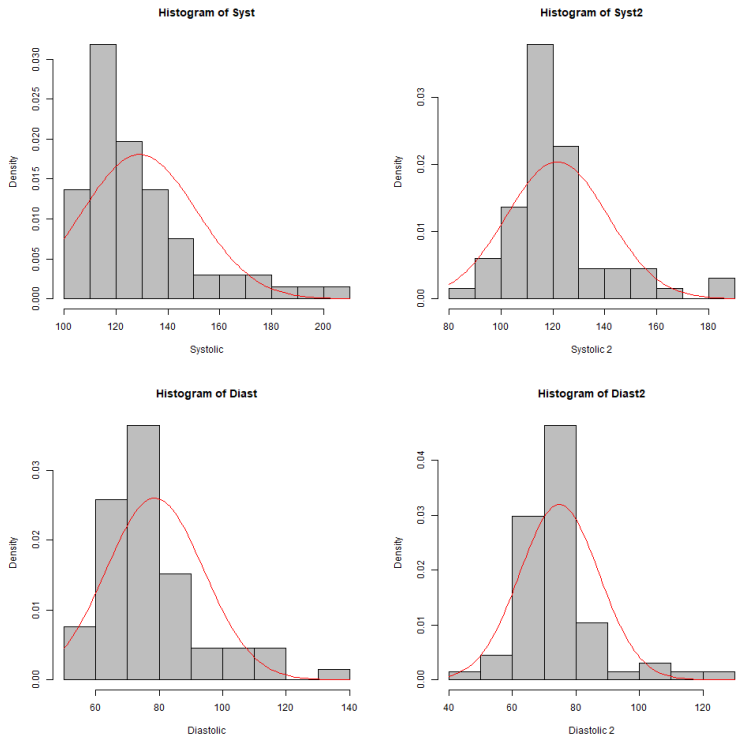


Figure 10 – Histogram of the variables Syst, Syst2, Diast, Diast2

References

[1] Crum, Alia J., and Ellen J. Langer. 2007. *Mind-set matters: Exercise and the placebo effect*. Psychological Science 18, no. 2: 165-171.

[2] *Hotel maids* DASL - Datafiles. Retrieved from https://dasl.datadescription.com/datafile/hotel-maids/.