Introduction to Statistics Project 1

BMI Survey

Project Report

Wojciech Ciok s182125

Contents

Chapter 1

Introduction

This paper was written by Wojciech Ciok student number s182125. It is a result of the first mandatory assignment under the course 02402 Introduction to Statistics E18.

This project focuses on overweight in Denmark. Overweight has an undeniable impact on health, happiness of people and economy. In a consumer world it is a very common problem. It is extremely easy, and often cheapest, to get some kind of fast food whenever you want it. As it leads to different kinds of diseases it also leads to deterioration of financial wealth. The costs of health insurance and treatment keep increasing. It is also worth noting that the weight loss industry was worth $64billion in 2014 worldwide. Which means that for many people it is important to loose weight but also shows that the scale of the problem is massive.

One way of measuring overweight is calculating BMI score. This is how it is defined:

$$BMI = weight/height^2 \tag{1.1}$$

where weight is expressed in kilograms and height in meters. One cannot forget that there are other factors which might be relevant like gender. Here is the assessment depending on the BMI score:

BMI score	**Assessment**
Less than 18.5	The person is underweight
Between 18.5 and 25	The person's weight is normal
Between 25 and 30	The person is moderately overweight
Between 30 and 35	The person is severely overweight (Obesity Class I)
Between 35 and 40	The person is severely overweight (Severe Obesity Class II)
Above 40	The person is severely overweight (Extremely severe obesity Class III)

Table 1.1: Categories of BMI scores

A BMI survey was conducted where number of people was ask questions about their height, weight, habits and so on. The survey resulted in a data set which will be the subject of analysis in this project (more about the data in the next chapter). Such a survey might provide a lot of useful information and answer important questions and this is the goal of this project.

Chapter 2

Descriptive Analysis

In this chapter, in order to give the reader reasonable knowledge about data, I will give a description and an overview of the data. I will try to explain how the data set is built, what information it consists of and give a summary of the data.

2.1 Variables

The data might be represented as a table where each column is a different variable and each row is an observation. There are 5 columns and 145 rows. There are no missing values.
To explain the values which the variables might have I include a table with explanation of the values. Note that in the dataset values of the attribute fastfood were scaled to times per year.

height	measured in centimeters (cm)
weight	measured in kilograms (kg)
gender	0: Female 1: Male
urbanity	1: Outside urban areas 2: City with less than 10,000 inhabitants 3: City with 10,000 to 49,999 inhabitants 4: City with 50,000 to 99,999 inhabitants 5: City with over 100,000 inhabitants
fastfood	1: Never 2: Less than 1 time per year 3: 1-11 times per year 4: 1-3 times per month 5: 1-2 times per week 6: 3-4 times per week 7: 5-6 times per week 8: Every day

Table 2.1: Explanation of the values

Here is the information if the variables are categorical or quantitative.

height	quantitative
weight	quantitative
gender	categorical
urbanity	categorical
fastfood	categorical

Table 2.2: Variables categorical/quantitative

2.2 Distributions of the BMI Scores

At this point we will calculate and look at the BMI scores obtained from the data.
To make initial observations I prepared a density histogram of the BMI scores.

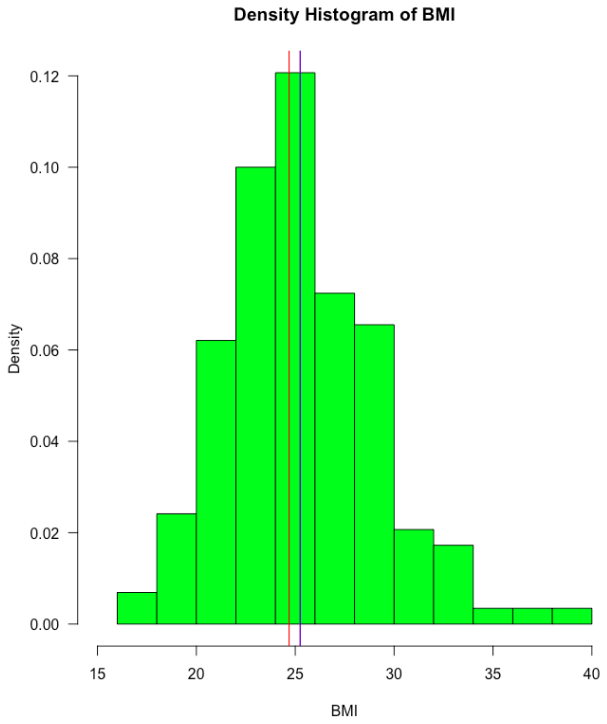


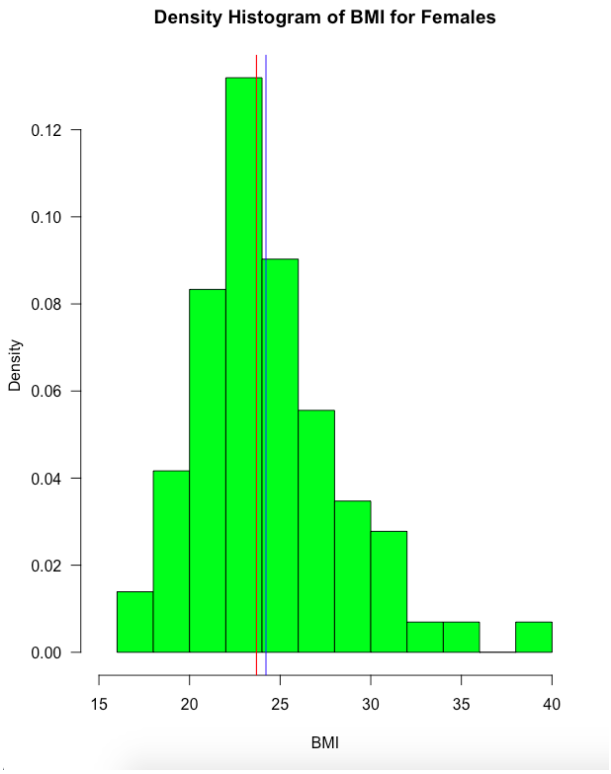
Figure 2.1: BMI density histogram

First of all the density is not symmetrical, it is right skewed meaning its right tail is longer. This means that there is a small number of scores which are relatively big, there are more extreme results in overweight than underweight. We can also see median marked with red color and mean with blue, they are very close to each other and are equal to around 25. Although the right tail is longer than the left one, they cover almost exactly the same areas.
The lowest BMI score is around 16, it is reasonable as BMI score cannot be negative.

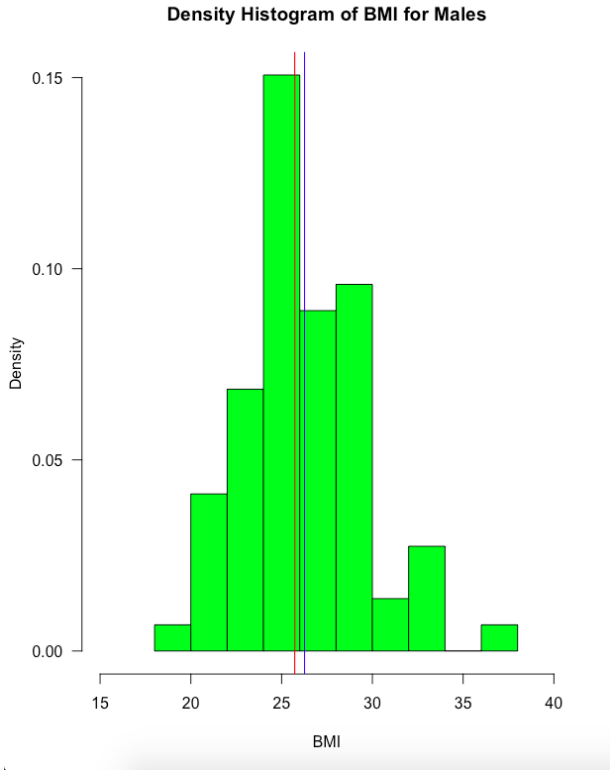
In the next three subsections we will focus on comparing scores based on gender.

2.3 Distributions of the BMI Scores by Gender

First of all we will take a look at histograms similar to that in the previous section but now with a distinction of gender. Once again medians marked with red color and means with blue.



(a) Women



(b) Men

Figure 2.2: Distributions of the BMI Scores by Gender

On women's histogram we can see most popular answers are between 20 and 25. The right tail seems to be longer. The range of scores is around 15 to 40. The mean and the median are just below 25.
In males' case we can see main mass concentration in range 25 to 30 with most popular answers definitely in 25 area. After point 30 we can see a clear drop. Mean and median values are just above 25.
When comparing the histograms we can make an observation that men generally have bigger BMI scores. Although men are more obese women have a bigger range of scores, we can see more extreme results in both directions in their case.

2.4 Box Plots of the BMI Scores by Gender

Another way of looking at the BMI data is by drawing a box plot.

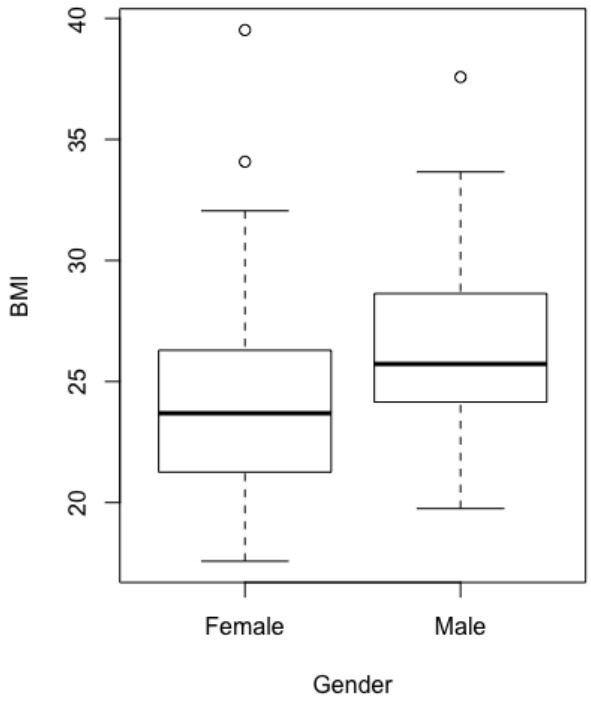


Figure 2.3: Box Plots of the BMI Scores by Gender

The plot confirms that men have generally bigger BMI scores. There are three suspected outliers, all with "too high" BMI. Two of them on the female part and one on the male part. We can also clearly see that in the men box the median is a not in the middle.

2.5 Summary Statistics of BMI with Gender Distinction

While the graphs provided a better overall trends in the data in this section we will take a look at exact values which we couldn't precisely read from the graphs.

	Number of obs.	Sample mean	Sample Variance	Sample st. dev.	Lower quartile	Median	Upper quartile
Everyone	145	25.25	14.69	3.83	22.59	24.69	27.64
Women	72	24.22	16.42	4.05	21.26	23.69	26.29
Men	73	26.27	11.07	3.33	24.15	25.73	28.63

Table 2.3: Summary Statistics of BMI with Gender Distinction

In the table above we can see for the first time the number of scores is divided almost in half between genders with 72 scores from women and 73 from men. Again, with mean, median, Q1 and Q3 we can say that men have bigger BMI scores. Women data has more variance which is with accordance with previous observations.

Chapter 3

Statistical Analysis

In this section we will perform a simple statistical analysis of the BMI for men and women.

3.1 Confidence Intervals and hypothesis tests

To make the data more "normal" we are going to work on log transformed BMI scores.

3.1.1 Statistical Model

To check if the logBMI scores may be assumed to be normal distributed we will take a look at a Q-Q plot.

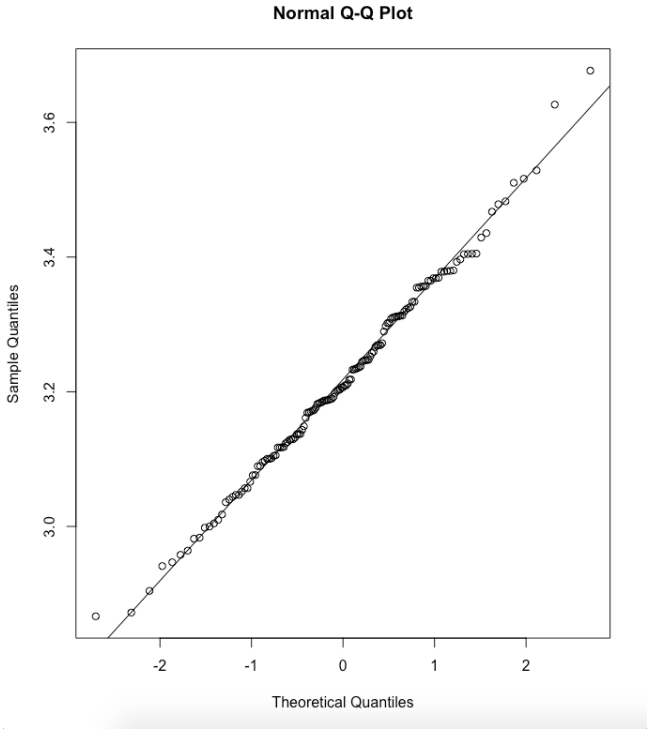


Figure 3.1: Log-transformed BMI Q-Q plot

The data seems to be normally distributed.

We will assume the statistical model to be:

$$X_i \tilde{} N(\mu, \sigma^2) \text{ and i.i.d., where } i = 1, ..., 145. \tag{3.1}$$

We will estimate real-life parameters of the distribution with sample mean $\bar{x}$=3.217641 and sample standard deviation s=0.1488778.

Since the original data is not normal distributed it is also important to mention the Central Limit Theorem (CLI) which, since our sample size is big enough, will make the means of the sample points converge to a normal distribution.

3.1.2 Confidence Intervals

Our sample mean $\bar{x}$ is only an estimation. In order to give more reliable information about the mean we will compute 95% confidence interval (CI). In simple words we will calculate a range and we will say that we are 95% sure that the mean belongs to this range. Here is the formula:

$$\bar{x} \pm t_{0.975} \frac{s}{\sqrt{n}} \tag{3.2}$$

where n is number of scores, s is sample standard diviation and $t_{0.975}$ is 0.975th percentile of n-1 fold Student's t distribution.
After a quick calculation we obtain the following interval:

[3.193203, 3.242078]

Now, by simply applying exponential function to the range we can obtain the CI for median of the original not transformed BMI scores:

[24.36635, 25.58684]

3.1.3 Hypothesis Test

We will perform hypothesis test to check whether the mean log-transformed BMI score is different from log(25) (the median BMI score is different from 25). This can be done by testing the following hypothesis:

$$H_0 : \mu_{logBMI} = log(25) \tag{3.3}$$

$$H_1 : \mu_{logBMI} \neq log(25) \tag{3.4}$$

I will assume conidence level $\alpha = 0.5$. In order to test the hypothesis we need the observed test-statistic t_{obs}.

$$t_{obs} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{3.5}$$

μ_0 is value against which we test, so $\mu_0 = log(25)$.
The last thing we will need is p-value.

$$\text{p-value} = 2 \cdot P(T > |t_{obs}|) \tag{3.6}$$

T is a random variable following Student's t distribution with n-1 degrees of freedom (144 degrees in our case).
After plugging in the values we obtain:

$$t_{obs} = -0.09991274$$
$$\text{p-value} = 0.9205526$$

The p-value is bigger then α meaning we do not have a strong evidence against hypothesis H_0 and we accept it. We cannot conclude that half of the population is overweight.

3.1.4 Statistical Models by Gender

Women

We will assume the statistical model to be:

$$W_i \tilde{} N(\mu, \sigma^2) \text{ and i.i.d., where } i = 1, ..., 72. \tag{3.7}$$

We will estimate the parameters with sample mean $\bar{x}_{women}$=3.174097 and sample standard deviation s_{women}=0.1598877.

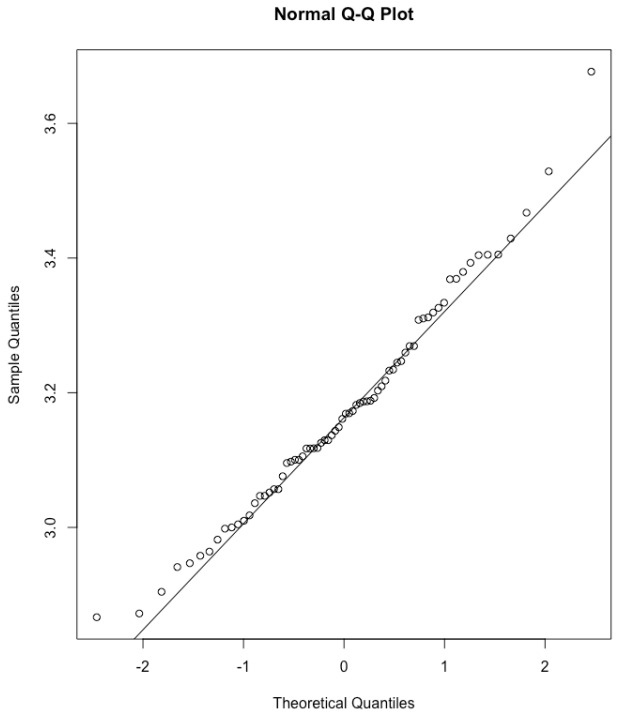


Figure 3.2: Q-Q Plot for Women logBMI

The plot gives a reasonable confidence in the assumed model.

Men

We will assume random variables to follow normal distribution:

$$M_i \tilde{} N(\mu, \sigma^2) \text{ and i.i.d., where } i = 1, ..., 73. \tag{3.8}$$

We will estimate the parameters with sample mean $\bar{x}_{men}$=3.2605881 and sample standard deviation s_{men}=0.1239114.

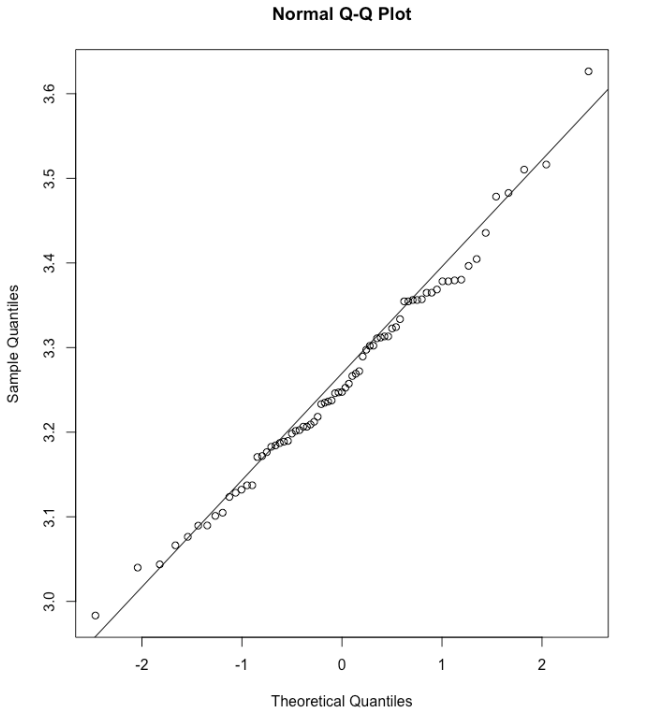


Figure 3.3: Q-Q Plot for Men logBMI

The plot gives a reasonable confidence in the assumed model.

3.1.5 Confidence Intervals by Gender

Since the method was explained in section 3.1.2, in this section I will only present the results of 95% confidence intervals for the median BMI score of women and men.

	Lower bound of CI	Upper bound of CI
Women	23.02372	24.82047
Men	25.32209	26.82940

Table 3.1: CI of median BMI by Gender

3.1.6 Hypothesis Test by Gender

To see if there is a difference between men and women BMI we will perform a Welch two-sample t-test statistic. I will assume conidence level $\alpha = 0.5$. Here is the hypothesis:

$$\delta = \mu_{women} - \mu_{men} \tag{3.9}$$

$$H_0 : \delta = 0 \tag{3.10}$$

In our case the Welch two-sample t-test statistic is:

$$t_{obs} = \frac{(\bar{x}_{women} - \bar{x}_{men}) - 0}{\sqrt{s^2_{women}/72 + s^2_{men}/73}} \tag{3.11}$$

The p-value is once again:

$$\text{p-value} = 2 \cdot P(T > |t_{obs}|) \tag{3.12}$$

In this case random variable T follows a t-distribution with degrees of freedom, where

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} \tag{3.13}$$

For our considerations $s_1 = s_{women}$, $s_2 = s_{men}$, $n_1 = 72$, $n_2 = 73$.
After plugging in the results we obtain the following results:

$$t_{obs} = -3.637453$$
$$v = 133.7501$$
$$\text{p-value} = 0.0003919648$$

Comparing the p-value to α we can state that there is a strong evidence against the hypothesis H_0. In other words we expect there is a difference between the BMI of women and men.

3.1.7 Comment to Section 3.1.6

The calculations in the previous section were not necessary. Confidence intervals for men and women BMI does not overlap, this fact itself mean that the two groups are significantly different.

3.2 Correlation

In this section we put focus on pair-wise correlation of three attributes: weight, fastfood and bmi. We will calculate correlation coefficients and plot scatter plots. In every case we use this general formula for sample correlation coefficient of sets x and y:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{s_{xy}}{s_x \cdot s_y} \tag{3.14}$$

bmi vs weight

$r_1 = 0.828261$

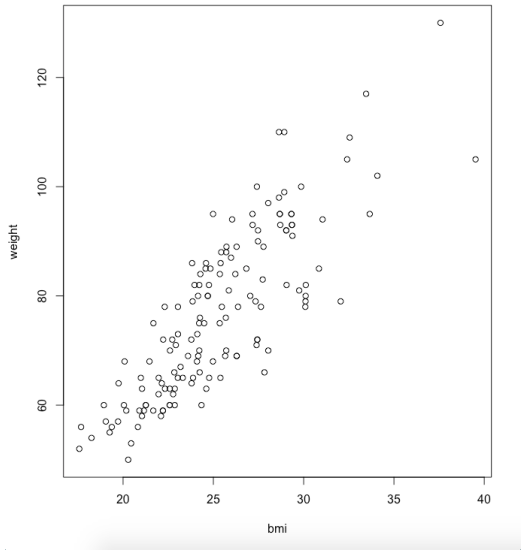


Figure 3.4: Scatter plot bmi vs weight

We can see a clear positive correlation. This was expected as weight is in the definition of BMI.

bmi vs fastfood

$r_2 = 0.1531578$

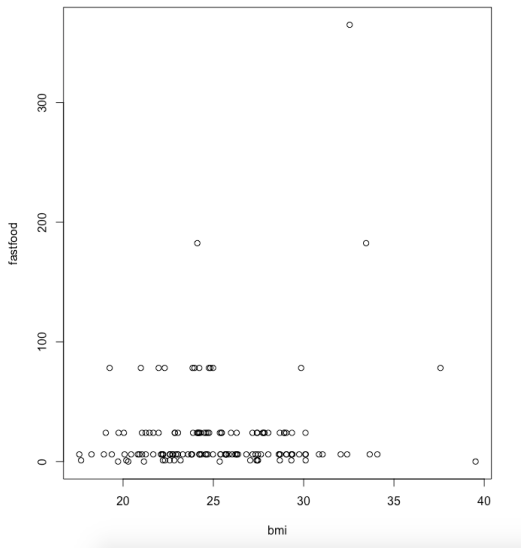


Figure 3.5: Scatter plot bmi vs fastfood

There is a big group of people who eat fastfood once or not at all per year and are still obese. Moreover we can also observe points with bmi in norm and 78.2 fastfood meals per year. Both facts and low correlation coefficient suggest which I did not expect that there is little correlation between eating fast food and being obese.

weight vs fastfood

$r_3 = 0.2793223$

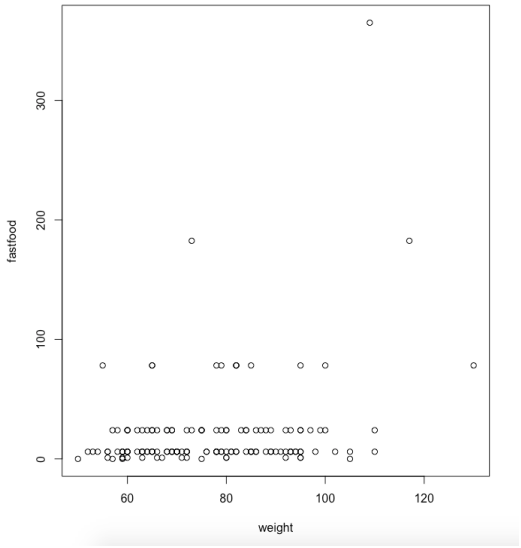


Figure 3.6: Scatter plot weight vs fastfood

Since there is a strong correlation between bmi and weight it is not surprising to see a similar result to bmi vs fastfood.

Chapter 4

Conclusions

After performed analysis I have a much better understanding of the data. The data set and the attributes especially BMI scores have been presented in numerous ways. I learned a lot about basic properties of the attributes. The project gave me a better grasp over mathematical formulas and allowed me to use them in practise. I performed two hypothesis tests and analyzed the results. Also, I made a basic correlation analysis between chosen attributes. All this gave me a solid ground and confidence with the data and prepared for the next part of the project.