# 1st homework assignment

## Task 1 - cleaning data (2 points)

Work with the **customer_behaviour** dataset.

```
load("customer_behaviour.RData")
```

The dataset has 4 columns, each row represents an individual customer: *money_spent* describes the average amount of money customer spends during one visit, *age* is self-explanatory, *web_visits* describes how many times a month customer checks out the shop website, *mail_ads* describes how many advertisement emails the customer gets monthly, *shop_visits* described how many times the customer visits a shop in person a month. Explore each variable and **delete** any rows which have mistakes in them. **Do not fix the mistakes, delete whole rows.**

| number of rows in the cleaned dataset |
|---|
| 481 |

## Task 2 - descriptive statistics (3 points)

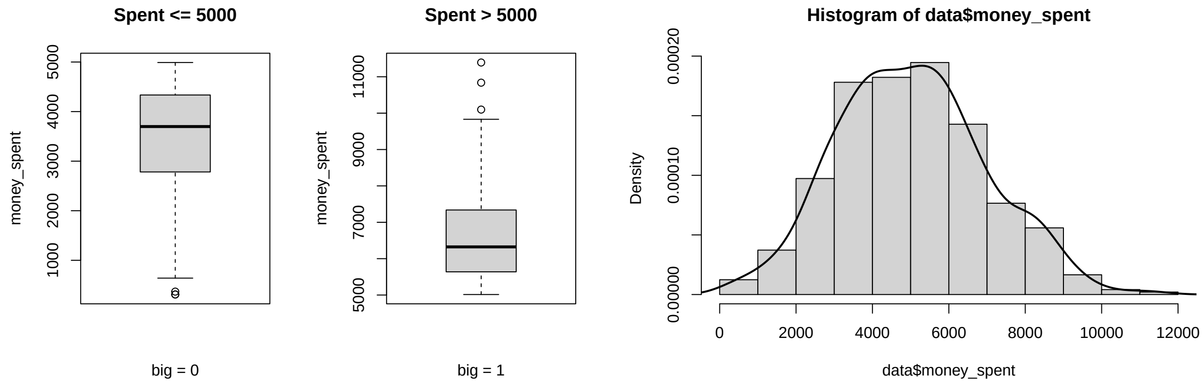
Work with the cleaned dataset from the previous month **customer_behaviour2**.

```
load("customer_behaviour2.RData")
```

Firstly, create a new variable called `big` where each value equals either 1 (if the person spent more money than 5000 USD), or 0 (if he spent less or equal):

```
data$big = as.numeric(data$money_spent > 5000)
```

Plot two boxplots of the variable *money_spent* into one figure: the first one for observations with the value of *big* equal to 0, the second one for observations with the value of *big* equal to 1. Then create a histogram for the variable *money_spent* together with its kernel density estimation.



Finally, compute following numerical characteristics of the variable *age*:

| mean | median | $1^{st}$ quartile | $3^{rd}$ quartile | interquartile range | variance |
|---|---|---|---|---|---|
| 54.78882 | 55 | 40 | 68.5 | 28.5 | 344.7146 |

Choose one appropriate measure of location and one appropriate measure of variability for the *money_spent* variable. Input the name of the measure into the following table. Briefly explain why you chose these measures.

| measure of location | measure of variability |
|---|---|
| median | interquartile range |
| It is less influenced by extremes in the dataset. And as the dataset seems to be skewed, that's why I think it is a good measure of location. | I think it is a good measure of variablity, because the dataset seems to be skewed. And just like median, the IQR is less influcenced by extremes in the dataset. |

## Task 3 - correlation (2 points)

Compute the correlation matrix of the data from the previous task (excluding the *money_spent* and *big* variables) and the sum of all its diagonal elements. Explain the result of the sum:

| Sum of diagonal elements | Explanation |
|---|---|
| 4 | The correlation matrix is a square matrix with correlation coefficient (between -1 and 1) between all variables. The value on the diagonal is a correlation of the variable with itself, which is always 1. So the sum tells us the number of variables in the correlation matrix, but it otherwise does not give us any useful information. |

Compute and **interpret** correlation coeficients between following variables:

| Variables | Results | Interpretation |
|---|---|---|
| `example` | 0 | The correlation is zero, which means... |
| `money_spent`, `age` | -0.6133549 | Moderately high negative correlation. This can be interpreted such that the older people were, the less money they spent. |
| `money_spent`, `web_visits` | 0.3725449 | Moderate possitive correlation. This indicates that people who visited the eshop website more often also spent more money. |

Interpretation in the form "correlation coefficient is 0.8 which means the correlation is high" will not be accepted.

## Task 4 - PCA (3 points)

Use PCA on the dataset from the previous task (**customer_behaviour2**, excluding variables *money_spent* and *big*). Use as little components as possible to capture at least 80 % of data variance.

| Number of components used |
|---|
| 2 |

State which variable has the most influence on each component.

| - | Component 1 | Component 2 | Component 3 | Component 4 |
|---|---|---|---|---|
| most impactful variable | web_visits | age | mail_ads | shop_visits |

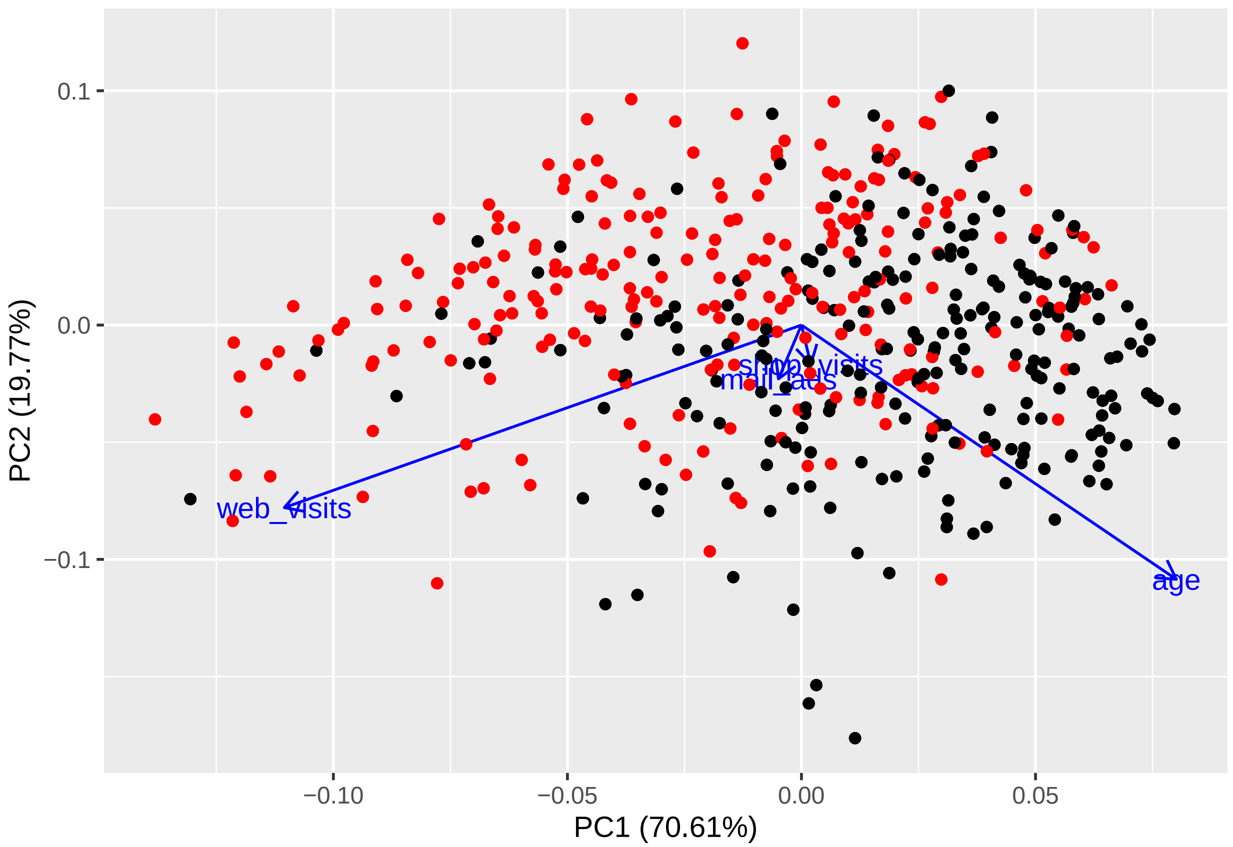
Create a scatter plot of data points using the first two components. Plot the points in different colours depending on the value of the *big* variable. What is your evaluation of the final plot? Can you decipher from the plot which variable(s) seems best at separating *big* shoppers from the customers who spend less?

```
#Either graph will do. The autoplot has loadings which nicely visualize
#the main variables that seperate the spenders (age and web_visits)
short_data.pca <- prcomp(short_data, center = T, scale = F)
pca_scores = predict(short_data.pca, short_data)
df = data.frame(PC1 = pca_scores[, 1], PC2 = pca_scores[, 2], big = data$big)
#library(ggplot2)
#ggplot(df, aes(x = PC1, y = PC2))+
#    geom_point(color = ifelse(data$big == 1, "red", "black")) +
#    geom_smooth(method = "lm", formula = y ~ x, data = df[df$big == 0,],
#    se = F, color = "black") +
#    geom_smooth(method = "lm", formula = y ~ x, data = df[df$big == 1,],
#    se = F, color = "red")

library(ggfortify)
```

```
## Loading required package: ggplot2
```

```
autoplot(short_data.pca, data = df, shape = T,
         colour = ifelse(data$big == 1, 'red', 'black'), loadings = TRUE,
         loadings.label = TRUE, loadings.color = "blue",
         loadings.label.color = "blue") +
    geom_point(color = ifelse(data$big == 1, "red", "black"))
```



| Scatter plot evaluation | Which variable(s) best separates heavy spenders |
|---|---|
| I can see that both light and heavy spenders do mix quite a bit around center of the graph, but there are regions dominated mostly by either a heavy or light spenders. So it seems that there is a difference between them. | Age and web_visits. From the graph it seems that the older people were, the less money they spent. I can also see that the more people visited the eshop website, the more money they spent. So age had a negative impact on the money spent while visitation count had a possitive impact. Both conclusions seem to fit the results from task 3. |