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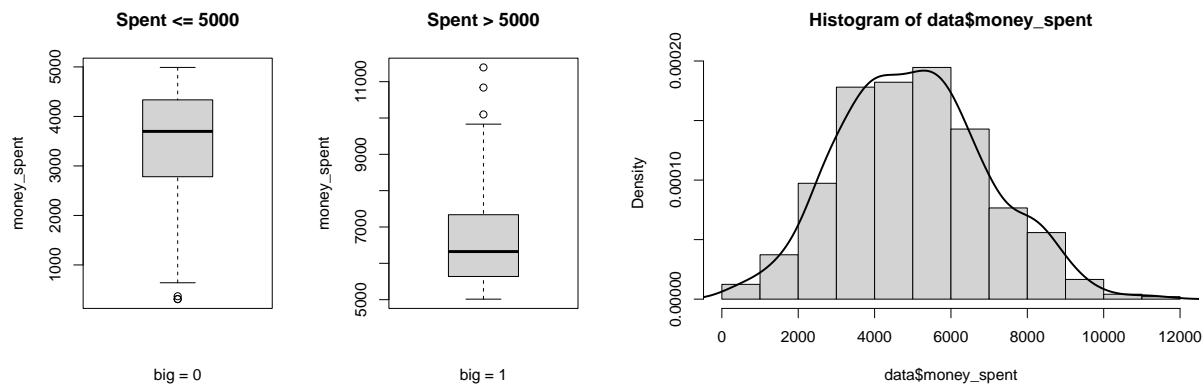
Finally, compute following numerical characteristics of the variable *age*:

				interquartile	
mean	median	1st quartile	3rd quartile	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	It is less influenced by extremes in the dataset (the dataset seems to be skewed)

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	It tells us that the every variable perfectly correlates with itself (and so it tells us the number of variables)

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. The more money_spent increases, the more age decreases.
money_spent, web_visits	0.3725449	Moderate possitive correlation. The more money_spent increases, the more web_visits increases.

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?

```
short_data.pca <- prcomp(short_data, center = T, scale = F)
s <- summary(short_data.pca)

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, color = factor(big))) +
    geom_point()
```
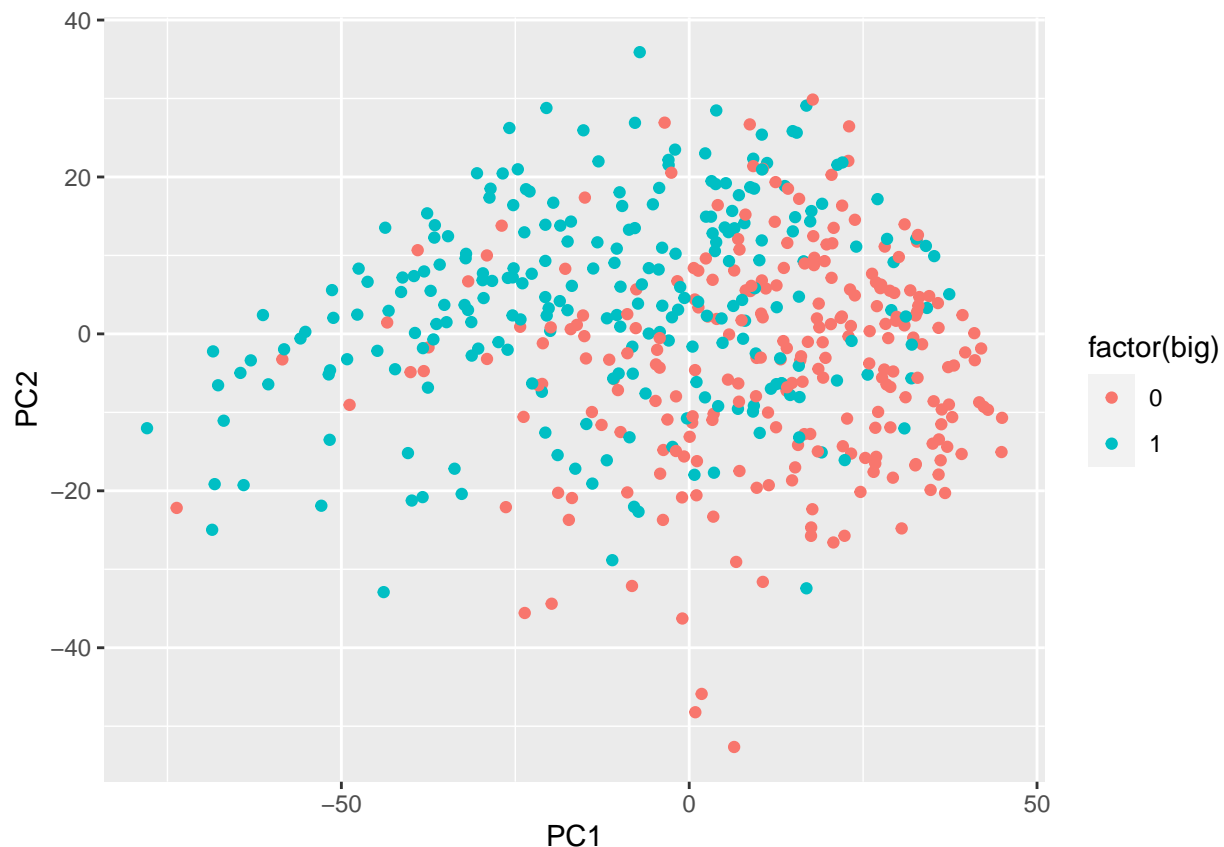

Scatter plot evaluation	Which variable(s) best separates heavy spenders
I can see that both groups (light and heavy spenders) do mix quite a bit in the center. There is a difference between them, but they do mix.	Age and web_visited variables. It seems that heavy spenders visit the eshop more often and are younger while light spenders are older and visit it less.