Seminar 4

1. Load dataset *market.csv* into RStudio. Your task is to study and illustrate the dependencies of working market segment proportions across the EU. You should also study differences across states of the EU.

   (a) Transform data from the long format to the short format. Use column `state` (indicates state) as key variable for each row. Use values in column `market` (indicates market segment) to identify new columns. Use values from column `value` (indicates percent of people working in the given segment).

   (b) Create a correlation matrix using build-in R function. Try different options for correlation matrix visualization using `corrplot`.

   (c) Create a single scatter plot for `Industry` and `Manufacturing`. Create a scatter plot matrix for all segments.

   (d) Use PCA to analyze the data. Use built-in R function `prcomp` and `autoplot` function from package `ggfortify`.

   (e) FOR VOLUNTEERS: Use PCA to analyze the data without using the built-in R function `prcomp`.

2. Work with the dataset `customer_behaviour.RData` describing the customers behaviour (money spent during some time period, their age, number of web and shop visits and number of mail ads).

   (a) Create a new column called `big` containing value 2 if `money_spent` variable is greater than 5000 USD and value 3 if `money_spent` is smaller than 5000 USD.

   (b) Perform PCA using the built-in R function (remember scaling Your data), investigate the summary and store the summary object into a special variable.

   (c) Plot the cumulative proportion of the explained variance (see structure of the summary object and the previous examples). How many principal components are needed for explaining at least 90%?

   (d) State which variable has the most influence on each component.

   (e) Visualize the first two principal components (using the `autoplot()` function), use `colour` input argument to color the data points by two colors according to the `big` variable.

3. **FOR VOLUNTEERS: Using PCA for image pattern recognition:** Use R library `jpeg` to load figures 001.jpg - 009.jpg from study materials. Transform data from matrix to vector and save them into a data frame.

   (a) Images were loaded as a data matrix with 9 rows (observations) and 47 988 columns (variables). Each variable defines a specific pixel of some image. Use PCA to identify common patterns of the data.

   (b) Figure out the minimum number of components so components explained at least 80% variability of the data.

(c) Use the number of components from the previous task to reconstruct the original images.