Assignment 1 for "Business Analytics I" course

The assignment should be done individually, the deadline is February 16. The exercises worth 25 points in total. You have to submit a Jupyter notebook file with the code that you used to solve the different tasks; use comments in the notebook to discuss the results and explain the output of your code.

1. (8 points) In this task, you have to work with a dataset in the file *students.csv*. The data provides some information about students, including: (i) gender, (ii) the educational background of the student's parents, (iii) whether the student has participated in test preparation courses, and (iv) the test scores in math, reading and writing. By making use of the basic data analysis tools introduced in the course, answer the following questions:

   - Are there more male or female students included in the dataset?
   - In which subject did the students perform best overall (i.e. what is the subject with the highest average score)?
   - How many students can you find with perfect results (100 score in all three subjects)?
   - Can we say that participating in the test preparation course improves the results in math (i.e. is the average math score higher for those who completed the test preparation course)?

2. (7 points) In this exercise, you will have to analyze a dataset about some employees of a company (*salary.csv*). The data includes information about (i) years of experience of the employee, (ii) the age of the employee, and (iii) the salary of the employee. You need to write the code to answer the following questions to transform and understand the data further.

   - There are some missing values in the data. In the experience_years column, replace all missing values by 3, and in the age column replace all the missing values by the mean of the column.
   - Check the correlation between the columns of the data. Based on the results, what is your opinion, does experience or age impact salary more?
   - Create a categorical version of the salary column with 3 groups: (i) values between 0 and 55000, (ii) values between 55000 and 90000, (iii) values between 90000 and 125000. Label the groups ['Low', 'Medium', 'High']. Calculate the average age of employees who have High salary.

3. (10 points) In this exercise, you will have to work with a dataset about movies (*movies.csv*). In the dataset, you can find information about: (i) movie title, (ii) short overview of the movie, (iii) the original language, (iv) number of votes given by people who have seen the movie, and (v) the average of the votes for the movie (a vote is between 0 and 10). You need to write the code to answer the following questions.

   - Evaluations that are based on too few votes cannot be completely trusted. For this reason, start by removing movies that has less than 100 votes (vote_count). Perform the following tasks on this filtered dataset.
   - What is the 3rd most frequent original language in the data?
   - How many Finnish movies are in the data (original_language is fi), and which is the highest rated one?
   - Calculate the average of votes separately for all French (original_language is fr) and Spanish (original_language is es) movies. Which country has higher average value?
   - Create a categorical version of the vote_average column with 4 groups: (i) values between 0 and 2.5, (ii) values between 2.5 and 5, (iii) values between 5 and 7.5, and (iv) values between 7.5 and 10. Label the groups ['Bad', 'OK', 'Good', 'Excellent']. Which of the four created groups has the highest number of movies?