Assignment 2 for "Business Analytics I" course

The assignment should be done individually. The exercises worth 35 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. (13 points) In this exercise you have to work with the data about customers and their purchasing behaviour. Using the data provided, you need to form some ideas on the relationship of variables. In the first file 'marketing_demographics.csv', you find basic information about the customers: (i) education level, (ii) marital status, (iii) yearly income, (iv) country, (v) age, and (vi) number of children. In the second file 'marketing_business.csv', you can find information about the customer's interaction with the company: (i) total amount of money spent on items, (ii) total number of purchases, also separately for purchases performed online and in the physical store, (iii) the number of times the customer accepted some campaign offers in the past, (iv) the number of times the customer visited the company website in the month before the most recent campaign, (v) the number of times the customer made a complaint in the past, and (vi) the customer's response to the most recent campaign.

   You have to perform some descriptive analysis tasks on these datasets. As the first step, import and then combine the files using the single shared column, 'ID'. In each question, unless stated otherwise, use this complete dataset to calculate the answer (so if I ask you to filter, it applies only to that specific question, and you do not use this filtered data in the next question).

   - Calculate the average of 'Total_Amount' and 'Total_Purchase' for each Education category. Do you find the same category to have the highest average for both amount and purchase?
   - Filter your data for countries from which there is at least 200 customers. From which country in this filtered dataset you can find the most complaints (column Complaint)?
   - Calculate a new column that is 0, if the customer has 0 children, and 1 otherwise (so if the customer has at least 1 child). Do customers with or without children have higher average Income? Check whether the difference is statistically significant using a t-test!
   - The company believes that their main target group is the customers aged between 18 and 45 years. To check whether it is reflected in the sales, check whether the average number of Web_Purchase is higher for the customers in the target age group than for other customers. Do you have the same result when you use Store_Purchase? (Hint: you can create a new column that is 1 when the customer is in the target age group and 0 otherwise).
   - Calculate the correlation between the columns Total_Purchase, Age, and Income. Which of the other two variables seem to be more related to Total_Purchase? Based on this and the previous question, do you think it would be useful for the company to focus a lot of efforts on age-based segmentation/marketing?

2. (12 points) In this exercise, you will have to work with a telecom churn dataset (churn-bigml-80.csv). The data includes information of customer activity data, along with a churn label specifying whether a customer canceled the subscription (`https://www.kaggle.com/mnassrib/telecom-churn-datasets?select=churn-bigml-80.csv`). You need to write the code to answer the following questions.

   - Visualization: create 6 plots of your choice based on the data and explain what information you gain from them; the plots can be histograms, boxplots etc. You need to create some univariate and multivariate plots, and focus mainly on the 'Churn' column. The created visuals should address at least the following issues : (i) relationship between Churn and having international or voice mail plan in the subscription; (ii) total minutes/calls/charge during different parts of the day and Churn (you may also want to create a column that has these values for the whole day, i.e., the sum of day, eve and night minutes/calls/charges); (iii) relationship between international calls/charge and Churn; (iv) relationship between Customer service calls and Churn.

- Descriptive statistics: You are free to explore the data with any of the tools we used in the course to understand the relationship between Churn and other data available about customers. You need to at least address the issues mentioned in the visual analysis part. Note that, as 'Churn' has two possible values, you want to use either cross-tabulation (to compare with other categorical variables) or groupby and aggregation (to compare with numeric variables).

3. (10 points) In this exercise, you will have to work with the World Happiness Report dataset ('happiness.csv'; when importing, use sep = ';'). The main Score asks respondents to think of a ladder with the best possible life for them being a 10 and the worst possible life being a 0 and to rate their own current lives on that scale. The other columns estimate the extent to which each of six factors – economic production, social support, life expectancy, freedom, absence of corruption, and generosity – contribute to making life evaluations higher in each country. In the following you have to perform some data preparation tasks; this time, in each task use the data that you obtain after performing the previous steps (so in the end you will obtain a dataframe that is modified according to all the specifications)

   - Remove outliers: (i) for the column 'Healthy life expectancy', remove the top 3% of values, and (ii) for column 'Perceptions of corruption', remove the bottom 2%
   - Handle missing values: (i) remove countries (rows) where there are 3 or more missing values; (ii) fill in the rest of the missing values with the mean value of their column
   - Create a categorical version of 'GDP per capita' column with four categories and corresponding labels (keep also the original column): (i) below 0.58, 'Low', (ii) between 0.58 and 0.96, 'Average', (iii) between 0.96 and 1.23, 'High', and (iv) above 1.23, 'Very High'.
   - Scale all the numeric columns using the StandardScaler transformation.