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. (12 points) In the following, you will have to analyse a dataset to perform tasks that data analysts working for a insurance company perform regularly: identifying the risky customers. In the dataset, 'insurance.csv', there is some information about the insured people: (i) age, (ii) gender, (iii) BMI (body mass index), (iv) number of children, (v) whether the person is a smoker, (vi) the region within US where the person lives, (vii) charges, i.e. the amount of money the insurance company had to pay for different medical costs of the person. You need to perform the following steps:

   - Determine the region with the highest total amount of charges!
   - Create a pivot table indexed with gender and number of children variables, and focusing on the charges. Which combination of gender and number of children has the highest average charge value?
   - Do smokers or non-smokers have higher average charges? Check whether the difference is statistically significant using a t-test!
   - Using groupby(), calculate the correlation between age and charges for each region in the dataset separately. In which region do you find the highest correlation?
   - In general, a BMI value between 18.5 and and 24.9 is considered healthy. Create a new column in the data, which is True (or 1) when the BMI in that row is in the healthy range, and False (or 0) otherwise. Are there more people with healthy or unhealthy BMI in the dataset?
   - Using the column created in the previous task, calculate the average value of charges for people with healthy and unhealthy BMI separately. Can you confirm that the healthy BMI range results in lower charges? Use a t-test to confirm your finding.

2. (13 points) In this exercise you have to work with the data about customers and their purchasing behaviour. The main outcome of interest in the data is the variable 'Response', indicating whether the customer reacted positively or not to the most recent marketing campaign. Using the data provided, you need to form some ideas on what variables are related to the customer's response. 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 this dataset, both graphical and non-graphical. As the first step, import and then combine the files using the single shared column, 'ID'.

   - 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 'Response' column and in general the spending patterns of customers. The created visuals should address at least the following issues : (i) relationship between income of the customer and response to campaigns (in the past and in the most recent case); (ii) spending (amount and purchase) across countries; (iii) relationship between income and total amount/purchase; (iv) relationship between total amount/purchase and response to current and past campaigns.

- Descriptive statistics: You are free to explore the data with any of the tools we used in the course to understand the relationship between Response and general spending of the customers and other variables. You need to at least address the issues mentioned in the visual analysis part. Note that, as 'Response' is not considered a numeric variable, as it only 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 continue working with the Insurance dataset, and prepare it for further modeling through the following steps. You have to solve these tasks in sequence, so each task needs to be performed on the output data of the previous task.

    - Remove outliers: (i) for the column 'charges', remove the top 2% of values, and (ii) for column 'bmi', remove the top 2% and the bottom 2%
    - Create dummy variables (one-hot encoding) for the columns 'region'.
    - Create a categorical version of 'bmi' column with four categories and corresponding labels: (i) below 18.5, 'Underweight', (ii) between 18.5 and 24.9, 'Healthy', (iii) between 24.5 and 30, 'Overweight', and (iv) above 30, 'Obese'.
    - Create a simple version of the column 'children', which is 0, if the original value is 0, and 1 otherwise (i.e. an indicator specifying whether a person has any children or not at all).
    - Scale the columns 'charges' and 'age' using the StandardScaler transformation.