Assignment 3 for "Business Analytics I" course

The assignment should be done individually. The exercises worth 40 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. (15 points) In this exercise, you will perform tasks faced by a data analyst working for a supermarket chain ('supermarket.csv', when importing the file, use sep=';'). Your job is to build a predictive model to estimate the purchase amount ('Total' column) for a customer. You have the following information:

   - Branch: branch of the supermarket chain, there is data about three different branches (A, B, and C)
   - Customer type: indicating whether the customer is a Member of the loyalty program or not
   - Gender: the gender of the customer
   - Quantity: number of products purchased by the customer
   - Total: amount of purchase
   - Payment: payment type used by the customer
   - Income: income of the customer
   - Rating: rating given by the customer to the store

   You need to perform the following tasks:

   - Exploratory data analysis: try to understand the different variables in the data. As part of this exploratory analysis, create visualizations that show the relationship between 'Total' and the other variables (create at least 4 plots, you are free to create more if you think it can help in understating the problem), perform aggregation (check how average 'Total' varies across categorical variables).
   - Divide the data into training and test set: the training set should contain datapoints from branches A and B, and the test set from branch C. Develop a regression model using the training set that the company can use to predict the amount of purchase for a customer. Evaluate the performance for the test set.
   - By looking at the coefficients of your final model, would you say that, in general, (i) male or female customers spend more money in the supermarket; (ii) members of the loyalty program or normal customers spend more money?

2. (15 points) In this assignment, your task is to create a classification model that can predict whether a reservation in a hotel will be canceled or not. The data is in the file 'hotel.csv' (when importing, use sep = ';'), and contains the following information:

   - no_of_adults: Number of adults
   - no_of_children: Number of children
   - no_of_weekend_nights: Number of weekend nights (Saturday or Sunday) the guest booked to stay at the hotel
   - no_of_week_nights: Number of week nights (Monday to Friday) the guest booked to stay at the hotel
   - required_car_parking_space: Does the customer require a car parking space? (0 - No, 1- Yes)
   - room_type_reserved: Type of room reserved by the customer.
   - lead_time: Number of days between the date of booking and the arrival date
   - arrival_month: Month of arrival date
   - arrival_date: Day of the month

- repeated_guest: Is the customer a repeated guest? (0 - No, 1- Yes)
- no_of_previous_cancellations: Number of previous bookings that were canceled by the customer prior to the current booking
- no_of_previous_bookings_not_canceled: Number of previous bookings not canceled by the customer prior to the current booking
- no_of_special_requests: Total number of special requests made by the customer (e.g. high floor, view from the room, etc)
- avg_price_per_room: Average price per day of the reservation
- booking_status: Flag indicating if the booking was canceled or not

You have to perform the following tasks:

- Perform one-hot encoding on the categorical columns if any
- Check the histograms of the columns lead_time, and no_of_previous_bookings_not_canceled. If you think there are outliers in the data, remove them.
- Replace any missing values you find.
- Build a logistic regression classification model with 'booking_status' column as the target, and using all other variables as predictors. Divide the data set into training (70 %) and test set (30 %), use random_state = 0, and follow the process of building a classification model as discussed in the course.
- Create the confusion matrix, calculate classification performance measures. What is the accuracy of the model on the test set?
- Does the model perform similarly for the two possible categories of the outcome column, i.e. for hotel visits and cancelations? How many false positives do you find, i.e. guests who would visit the hotel but the model predicts that they would cancel the booking?

3. (10 points) In this exercise you have to work with the data in the file 'iris.csv', that contains 150 records of Iris flowers of three related species under 5 attributes - Petal Length, Petal Width, Sepal Length, Sepal width and Class (Species). In the data there are three different species present, 50 samples each. You have to perform clustering and assess whether clustering is able to distinguish the three different species.

   Your task is to perform K-Means clustering on the dataset; in the model building process, do not use the column 'species'. You need to perform the following steps:

   - Determine the optimal number of clusters using the elbow method, and perform k-means clustering with the chosen value (set random_state = 0).
   - What is the average of each variable in each cluster (the original, not the scaled variables)?
   - Perform k-means clustering now with k=3 (if this was not your selected k value). When you compare the resulting clusters with the original 'species' column, you will find that one of the species is perfectly identified by clustering (i.e., one of the three clusters contains all the datapoints belonging to that species, and no datapoints from other species). Which one is the correctly identified species? When you look at the mean value of the variables, can you identify which variable(s) 'confuse' the clustering model, i.e., which variables you think are responsible for the other two clusters being the mix of the other two species?