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 with data from a bike sharing platform ('bike_data.csv'). Your job is to build a predictive model to estimate demand for bicycles at different locations using the following information (one row in the data is for one day of the year):

   - season: the season of the observaton (1: winter, 2: spring, 3: summer, 4: fall)
   - month : the year of the month (1 to 12, from January to December)
   - holiday : whether the day is holiday or not (extracted from [Web Link])
   - day: day of the month (1-31)
   - weekday : day of the week (1 to 7)
   - workingday : 1, if day is neither weekend nor holiday is 1, and 0 otherwise
   - weather: weather conditions (1: clear, partly cloudy; 2: cloudy, light rain; 3: thunderstorm, light snow)
   - temp : temperature in Celsius
   - hum: humidity
   - wind: wind speed
   - registered: total number of registered users checking the platform app within the day
   - cnt: count of total rental bikes during the day

   You need to perform the following tasks:

   - Exploratory data analysis: try to understand the different variables in the data. Identify the variables, based on exploratory data analysis methods, that you think have an effect on the count of rental bikes needed. As part of this exploratory analysis, create visualizations that show the relationship between 'cnt' 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 'cnt' varies across months, days, working days, and holidays), calculate correlation of the variables.
   - Develop a regression model that the company can use to predict the count of total rental bikes. Start with all the variables included in the data file, then follow the process suggested in the lectures to remove variables as long as you still find the model performance acceptable.
   - By looking at the coefficients of your final model, would you say that, in general, the company will need more bikes: (i) on working days or non-working days; (ii) on a day when temperature is 20 or on a day when temperature is 25?

2. (15 points) In this assignment, your task is to create a classification model that can predict whether a client of a bank will positively respond to a marketing campaign and invest some amount of money. The data is in the file 'bank.csv', and contains the following information about the client:

   - age: age of the client in years
   - job: type of job of the client, 11 possible categories
   - marital: marital status, 3 possible categories
   - education: highest level of education, 6 possible categories

- balance: average yearly balance
- housing: whether the client has housing loan (1 or 0)
- loan: whether the client has personal loan (1 or 0)
- day: day of month when the client was contacted the last time about the campaign
- month: month when the client was contacted the last time about the campaign, values 1-12
- duration: the duration of the last contact by phone, in seconds
- contact_count: number of contacts performed during this campaign for the client
- previous: number of contacts performed in a previous campaign for the client
- outcome: whether the client has invested in the product advertised in the current campaign

You have to perform the following tasks:

- Perform one-hot encoding on the categorical columns job, marital and education
- Check the histograms of the columns balance, duration and contact_count. If you think there are outliers in the data, remove them.
- Build a logistic regression classification model with 'outcome' column as the target, and using all other variables as predictors. Divide the data set into training (75 %) and test set (25 %), use random_state = 0, and follow the process of building a classification model as discussed in the course. (Hint: if you encounter a warning, you can set the parameter max_iter = 1000 within LogisticRegression()).
- 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 'outcome', i.e. for positive and negative class? If not, do you think it is a problem? How many false negatives do you find, i.e. clients who would invest in the advertised product but the model predicts that they would not?

3. (10 points) In this exercise you have to work with the data in the file 'patients.csv', that contains some measurements about patients, who experienced angina, which can typically be a symptom of coronary artery disease. You can find the following variables in the data:

- age: age of the patient
- gender: gender of the patient (0 - female, 1 - male)
- pain: intensity of the chest pain (integer value, 0-3)
- blood_pressure: blood pressure of the patient
- cholesterol: cholesterol in blood, mg/dl
- blood_sugar: indicating whether the blood sugar level is normal or not (1 when it is above 120 mg/dl, and 0 otherwise)
- heart_rate: maximum heart rate
- exercise: whether the chest pain was induced by some physical exercise or not (1 or 0)
- outcome: 1 for patients with heart attack, and 0 for patients who did not have heart attack

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

- Scale all the variables.
- 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)?
- Can you identify some variables that clearly have different average values for the clusters?
- Perform k-means clustering now with k=2 (if your chosen value was 2 in the analysis already, you can continue with the previously created clusters). Compare the created two clusters to the 'outcome' column. Do the created clusters separate patients who had heart attack from the other patients, or are the two clusters a mix of not healthy (have heart attack) and healthy (no heart attack) patients?