Assignment 1 for "Business Analytics I" course

The assignment should be done individually, the deadline is the 14th of February, 23:59. The exercises worth 20 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 *full_group.csv* (more details about the data `https://www.kaggle.com/imdevskp/corona-virus-report?select=full_grouped.csv`). Note I added three columns, 'year', 'month' and 'day', extracted from the Date column, that may be useful in some of the tasks. The dataset contains information about the number of confirmed, recovered and death cases related to COVID 19 in the initial months. The dataset has one row for each day and country combination. The columns with name starting 'New' record the numbers for that specific day, while other columns provide cumulative values up to that day. In the tasks, you have to select, summarize and compare information in the dataset.

   - Use either the original date column or the new year-month-day columns to determine what is the earliest and latest date that appears in the dataset.
   - Select the data for the month April. What is the total number of deaths in this month?
   - Create a new dataframe with the two columns CountryRegion and WHO_Region. Remove the duplicates so that each combination of CountryRegion and WHO_Region appears only once. What is the number of different CountryRegion that appears in the dataset? Which WHO_Region has the most CountryRegion associated to it?
   - Select the data for the months March and April, and for the country US. Was the average number of daily new deaths higher in March or April in US?

2. (4 points) In this exercise, you will have to analyze a dataset (AB_NYC_2019.csv) that we worked with in the lecture. The data includes information about hosts, geographical availability, and different metrics available from Airbnb places in New York City (`https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data`). You need to write the code to answer the following questions to understand the data further.

   - It is an important task to understand the difference between different neighbourhood groups. In order to do this, identify the neighbourhood group with (i) the highest average price, and (ii) the highest total number of reviews.
   - Next, analyse the data from the perspective of room type and identify which room type has (i) the highest average price, and (ii) the highest average number of reviews.

3. (8 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 (features), along with a churn label specifying whether a customer canceled the subscription (more information at `https://www.kaggle.com/mnassrib/telecom-churn-datasets?select=churn-bigml-80.csv`). You need to write the code to answer the following questions.

   - Focusing on day, eve and night calls as considered in the data, which part of the day has the highest total number of calls? Is the part of day with the highest total charge is the same?
   - What is the maximum amount of customer service calls that appear in the dataset, and how many individual customers had exactly that number of service calls?
   - Choose one categorical column which has less than 5 possible values, and add new dummy columns to the dataset for the selected column.

- Create a new column as the sum of the total day, eve and night call columns. Considering this new total daily call value, who had in average higher number of calls: customers with or without international plan?