TEL AVIV UNIVERSITY אוניברסיטת תל-אביב

המחלקה להנדסת תעשייה וניהול
מבוא למדעי הנתונים (Introduction to Data Science)
סמסטר א' תשע"ח

Assignment 1

Introduction:
In this assignment, you will be using data of taxi rides in Chicago, during the 2nd half of December, 2016.
The purpose of the assignment is introduction to the basics of R, while practicing two key components of the CRISP-DM model you have learned: pre-processing and data exploration.

Guidelines:
Points will be taken off for not adhering to the following guidelines:

- ☒ Questions regarding the exercise will be answered in the forum ONLY.
- ☒ Use the code file from the Moodle. **Do not remove comments from it!**
- ☒ In most cases, there is more than one way to solve a specific question. However, the code efficiency will affect your grade.
- ☒ Points appear next to each question in brackets and provide an indication of the question's difficulty or complexity (so if you are stuck try to move on).
- ☒ Your **final grade** will also be affected by the readability of your code. Use meaningful names for variables, add documentation and notes. Also, make sure you answer all the verbal questions as a comment in the script.
- ☒ Use Google – R is an open-source language. A lot of information should be available online.

Submission:
Deadline: 5.12.17
Teams: Teams of two students.
Submission: R file uploaded to Moodle by **one** of the team members. Make sure to insert all ID's at the top of your code file.
Score: 0-100. A script that will fail execution will receive 0 points.

Useful Tips:

- ☒ You will likely need to install packages along the way. In order to do so execute `install.packages` (look it up on google).
- ☒ If you want to know more about a certain function – execute the function with "?" before it, e.g.: `?nameOfFunction` At the bottom of the help page that will open you can usually find very good examples which you can copy and execute yourself and play around with. If the function's package is not installed – you will not see a help page about it, in which case you are better off searching google for it.

The data set will be in the moodle.
You can find general info about the data set in the following link:
https://www.kaggle.com/chicago/chicago-taxi-rides-2016

GOOD LUCK!!!

TEL AVIV UNIVERSITY אוניברסיטת תל-אביב
המחלקה להנדסת תעשייה וניהול

מבוא למדעי הנתונים (Introduction to Data Science)
סמסטר א' תשע"ח

1. Loading the data and setting an environment (8):

1.a. (0) Download and extract the data from Moodle into a local folder designated for this assignment.

1.b. (2) Set your working directory to be your assignment folder for easy access. From now on, if needed, do not use the full path, but only the name of the file within this path.

1.c. (3) Import the CSV file " chicago_taxi_data.csv " into R and save it by the name **data**. Notice the data in the file has row numbers that are redundant (so pay attention to the function arguments).

1.d. (3) Make sure the data was loaded properly by showing the first few rows of the data.

2. Dataset preparation (23):

2.a. (3) Sample 10000 rows from the dataset without replacement. This file will be our dataset throughout the exercise. Before you sample, set your random seed to be 1.

2.b. (3) We will not use any of geographical columns (pickup/ dropoff - longtitude/ latitude). Delete these columns.

2.c. (3) Show the names and the data type of all the features.

2.d. (5) The Column pickup_census_tract has only NA values. Create a verification check for this claim.
Delete the column pickup_census_tract and dropoff_census_tract from the dataset.
Could we have known in advanced that this column is problematic? Tip: use your answer the previous question.

2.e. (5) What's your opinion about the current type of the column 'company'? Is it adequate?
If yes, explain why. It not, explain why not and change the type of this column and other similar columns.

2.f. (4) Create a summary statistics of the dataset (using one-line command). What is the difference between the output for the numerical columns and the non-numeric columns?

3. Missing values (18):

3.a. (3) Calculate the percentage of rows with at least one missing value (NA).

3.b. (3) Delete all rows with more than 1 missing value (NA)

3.c. (4) Create a histogram of a categorical column (to your choice). Explain the findings of the chart. Pay attention that some histogram functions work only with numerical values, so a transformation is needed.

TEL AVIV UNIVERSITY אוניברסיטת תל-אביב
המחלקה להנדסת תעשייה וניהול

מבוא למדעי הנתונים (Introduction to Data Science)
סמסטר א' תשע"ח

3.d. (8) Choose and implement the best way to deal with the missing values in the dataset, in your opinion (according to your previous findings). As for the columns: [trip_seconds, trip_miles, trip_total] , deal with 0's (zeros) as if they were NA's .
Pay attention - you can decide to delete specific rows or columns, while impute some other remaining missing values. Explain all of your choices.

4. Data normalization (11):

4.a. (4) Make a Q-Q plot for each of the following columns: [trip_seconds, trip_miles, trip_total]. Explain what we can learn from a qq plot about the distribution of the data.

4.b. (7) According to the Q-Q plots ,do we need to normalize these features? Which normalization function should we use for each feature, if any?
For each feature, in case you decided to normalize it, create a new normalized column of the feature (e.g. norm.trip_seconds).

5. Outlier detection (19):

5.a. (5) Create a boxplot of the normalized trip_miles column (or the original column in case you chose not to normalize). Remove the column's outliers from the data based on the box plot. Hint: use the boxplot object.

5.b. (4) Implement a min-max transformation on the normalized columns of [trip_seconds, trip_miles, trip_total] (or the original columns in case you chose not to normalize). Create new column with the transformed data (e.g. minmax.trip_seconds)

5.c. (10) Using the 3 columns you created, you will use a hierarchical-clustering method, followed by density-based method.
First, use hierarchical-clustering method to evaluate the probability of each instance to be an outlier. Exclude all instances with 0.75 chance or higher. Hint: use "DMwR" package.
Then, using LOF, pick k=10 and remove all instances that their LOF score is above 1.4. Hint: use "Rlof" package.

6. Exploration and visualization (21):

6.a. (6) Create a correlation matrix of all the relevant numerical features. In addition, Display a correlation plot for this matrix. Write 3 business insights we can learn from the correlation matrix.

6.b. (15) Create 5 different statistical outputs based on the dataset. Visualize at least 3 of them. Add an explanation. Try to be creative.
Examples:

1. A bar chart that displays the average and median amount of trip_total, for each payment_type.
2. Density plots of trip_second – one for each day.