PROJECT MILESTONE 2

PACE: Plan Stage

- How can I best prepare to understand and organize the provided information?

First of all, I need to frame the problem, meaning equip myself with the necessary tools to understand and organize the provided information. Then, I will inspect the data. Begin by exploring the dataset and consider reviewing the Data Dictionary. One can prepare to understand the information by reading the taxi cab data fields and understanding the impact of each one. Reviewing the fact sheet could also provide helpful background information. However, the primary goal is to get the data into Python, inspect it, and provide stakeholders with initial observations. The next step would be to learn more about the data and check for any anomalies.

- What follow-along and self-review codebooks will help me perform this work?

Data Dictionary because it provides comprehensive information about each variable in the dataset.

- What are some additional activities a resourceful learner would perform before starting to code?

Data Cleaning and Preparation Codebook Documents the steps taken to clean and prepare the data for analysis. - Data cleaning techniques used (e.g., handling missing values, outlier detection, normalization) - Transformations applied (e.g., feature engineering, scaling) - Reasons for specific cleaning or preparation step

PACE: Analyze Stage

- Will the available information be sufficient to achieve the goal based on my intuition and the analysis of the variables?

After careful inspecting of the data, I found a number of outliers and incorrect negative values for the 'total_amount' variable. From the total number of 22.699 records, only 14 records have negative values of the 'total_amount' variable, which is acceptable for further analysis. From the 'trip_distance' variable, there are 11.304 outliers, which is almost half of the total records. Therefore, I think that I will need other dataset with much less outliers.

- How would I build summary dataframe statistics and assess the min and max range of the data?

By using the describe() function on the dataset.

- Do the averages of any of the data variables look unusual? Can I describe the interval data?

Most of the data variables have averages greater than 3 standard deviations. The interval data is usually between 25% and 75%.

PACE: Construct Stage

The Construct stage does not apply to this workflow. The PACE framework can be adapted to fit the specific requirements of any project.

PACE: Execute Stage

- Given my current knowledge of the data, what would I initially recommend to the stakeholders to investigate further prior to performing exploratory data analysis?

  I will ask for different data source with much less outliers.

- What data initially presents as containing anomalies?

  All the variables with enormous deviations in their values are containing anomalies which could affect the further analysis.

- What additional types of data could strengthen this dataset?

  An additional type of data could be idle time of the taxi driver without customers, or off-ride.