Exploratory Data Analysis of Taxi Company Data (Milestone 3)

Executive summary report

Prepared by **PhD Aleksandar Osmanli**

Project Overview

The taxi company has asked to build a regression model that predicts taxi cab ride fares. In this milestone, the data needs to be analyzed, explored, cleaned and structured prior to any modeling.

Key Insights

The Problem: After running initial exploratory data analysis (EDA) on a sample of the data provided by the taxi company, it is clear that some of the data will prove an obstacle for accurate ride fare prediction. Namely, trips that have a total cost entered, but a total distance of "0." At this point, the analysis indicates these to be anomalies or outliers that need to be factored into the algorithm or removed completely.

Proposed solution: After analysis, it is recommended to remove outliers with a total distanced recorded of "0".

Keys to success

- Ensuring with the taxi company that the sample provided is an accurate reflection of their data as a whole.
- Plan for handling other outliers, such as low trip distance paired with high costs.

Details

As a result of the conducted exploratory data analysis, trip distance and total amount were considered as key variables to depict a taxi cab ride. The provided scatter plot shows the relationship between the two variables. This scatter plot was created in Tableau to enhance the provided visualization.

Alt Text: Graph displaying the taxi company data plotting variables for total distance and total amount.

Next Steps

- Determine any unusual data points that could pose a problem for future analysis in predicting trip fares.
  - For example, locations that have longer durations.
- Determine the variables that have the largest impact on trip fares.
- Filter down to consider the most relevant variables for running regression, statistical analysis, and parameter tuning.