New York City TLC Collaboration Project Preliminary Data Summary

Executive Summary

Tasks Completed

A preliminary analysis was performed on the provided taxi data to identify trends, anomalies, and potential areas for future analysis. We engaged in data exploration, cleaning, feature engineering, visualization, and initial statistical analysis. We assessed the available variables and verified no missing values while identifying various datatypes present. Data variable ranges were comprehensively assessed, identifying outliers and potential errors.

Findings

Upon analysis, we identified unusual patterns and potential anomalies in the data:

1. Short trip durations with high total amounts: Certain records show high total fare amounts for trips with very short durations. This anomaly requires further investigation to confirm if it is due to data errors, large tips, or other reasons.
2. Zero trip distance with high total amounts: Some trips reported zero distance traveled but still registered high total fare amounts. These records need further examination to determine the cause.
3. Negative total amount: Some records showed negative values for the total fare amount, which is counterintuitive. These instances might represent refunds or data errors, necessitating further inquiry.
4. RateCodeID: Some trips with high fares, regardless of distance or duration, were associated with RateCodeID 5 (Negotiated fare). However, further understanding of this variable and its influence on fare calculation is required.

These findings warrant further investigation before proceeding with predictive modeling.

Recommendations

Based on our initial analysis, we recommend the following steps:

1. Investigate anomalous data: To ensure the validity of our predictive model, we need to understand the causes of the identified anomalies. This involves further investigation of the records with high total amounts for short trips, trips with zero distance, and negative total amounts.

2. Consider including additional data: Enhancing the dataset with external factors like weather, traffic conditions, socio-economic variables, special events data, and more granular time data can provide a more holistic understanding of taxi demand patterns, which will be useful for predictive modeling.

3. Perform detailed exploratory data analysis (EDA): Before building predictive models, we must explore the data thoroughly, identify trends, understand the distribution of variables, and discover underlying patterns. This will also involve a deeper examination of the identified anomalies.

4. Data preparation for predictive modeling: Prepare the data for predictive modeling, which includes handling outliers, feature selection, and potential feature engineering based on findings from the EDA.

5. Predictive modeling: Upon completion of data preparation, we can proceed with predictive modeling. This will involve selecting the appropriate machine learning models, training, and testing these models, followed by model selection and validation.

In conclusion, the initial data assessment has laid a robust foundation for further analysis and predictive model development. The unusual patterns identified will drive our investigative efforts in the next phase, which will be instrumental in ensuring the success of this project.