Automatidata Project: NYC Taxi Trip Data Analysis

Milestone 2 - Preliminary Data Inspection and Key Variable Identification

OVERVIEW

This report summarizes the initial inspection and analysis of the New York City Taxi and Limousine Commission (NYC TLC) taxi trip data. The primary objective was to understand the dataset's structure, identify potential data quality issues, and determine key variables for building a predictive model. This preliminary analysis, conducted as part of Milestone 1, lays the groundwork for subsequent exploratory data analysis (EDA) and the development of targeted predictive models.

PROJECT STATUS

- The project is in the **initial data inspection and analysis phase (Milestone 2).**
- The dataset has been successfully loaded and inspected using Python and pandas.
- Initial data quality issues and potential key variables have been identified.

NEXT STEPS

- **Clean Data:** Address negative values, zeros, and outliers. Convert datetime columns.
- **Conduct EDA:** Explore variable relationships and visualize patterns.
- **Engineer Features:** Create new predictive features.
- **Build Models:** Develop and evaluate fare/total amount models.
- **Communicate:** Present findings to NYC TLC.

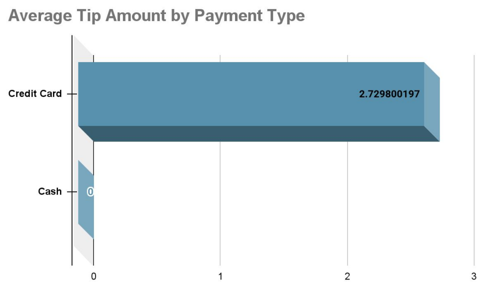
KEY INSIGHTS

Data Integrity: Dataset is complete (no nulls), but requires cleaning:

- Negative monetary values (fare, total) indicate errors.
- Zero values in `passenger_count` and `trip_distance` need investigation.
- Outliers in `trip_distance` and `RatecodeID` to be addressed.
- `tpep_pickup_datetime` and `tpep_dropoff_datetime` require datetime conversion.

Payment & Vendor Analysis:

- Credit card payments (type 1) yield significantly higher average tips than cash (type 2).
- Vendor IDs (1 & 2) show comparable average total amounts.
- Zero values in passenger count require further investigation.



Predictive Variable Identification:

- `trip_distance` is a strong predictor for fare and total amount.
- `payment_type` offers insights into payment behavior and potential tip prediction.