NYC Taxi Fare Analysis 2023

Jiwoo Jeong (jj4252)
Xintong Li (xl5733)
Alexander Pegot-Ogier (ap9283)
Omni jyothi Gudiguntla (og2148)

Dataset

Source: We use the dataset from the NYC Taxi and Limousine Commision (TLC) about taxi fares in New York City for the year 2023, obtained from the official TLC data page. Additional datasets: NYC 2023 Weather Dataset (Open Meteo Weather API), NYC Uber Rides (NYC Open Data - Rideshare dataset)

Description: This dataset provides valuable insights into taxi fare distribution across different zones, distribution across hours, components of fares, and trip distances.

Objective: The primary objective of this project is to analyze the patterns in NYC taxi fares throughout 2023, identifying key factors that influence fare amounts. By assessing the impact of each variable, we aim to provide taxi drivers insights that would help them optimize their service strategy.

Some major features of the dataset include:

Feature Name	Definition
Trip Distance	The elapsed trip distance in miles reported by the taximeter
PULocationID	TLC Taxi Zone in which the taximeter was engaged
DOLocationID	TLC Taxi Zone in which the taximeter was disengaged
Fare Amount	The time-and-distance fare calculated by the meter
Payment Type	A numeric code signifying how the passenger paid for the trip.

Research Questions

Primary Question

What are some factors influencing NYC taxi fares, and how do factors such as time, passenger count, weather conditions, pickup locations contribute to fare variability?

Sub Questions

1. Can we identify temporal patterns in taxi fares based on the time of the day, day of the week, and month of the year throughout 2023?
2. How does having multiple passengers impact the total fare? Additionally, if relevant, what other factors influence passenger count?
3. How are weather conditions associated with ride frequency and fares?
4. Does the pickup location influence the total fare? If so, which pickup locations tend to generate the highest fares?
5. Can we determine the most impactful variables that are associated with ride frequency and fares?

Potential Questions

Given specific weather conditions at a particular location, is a customer more likely to choose a taxi or an Uber ride?

Data Analysis & Possible Methods

Data Handling (Pandas, Pyspark)

- **Data processing** - Merge the 12 parquet files into one and read it in as a csv file.
- **Handling Missing Values**

Visualizations (Matplotlib, Seaborn)

- **Line Plot -** Identify the temporal patterns in taxi fares over time.
- **Bar Plot** - Visualize to understand the relationship between specific weather conditions and fare amounts.
- **Heatmap** - Display the distribution of average taxi fares & pickup rates across different pickup zones in NYC.

Feature Importance

- **Multiple Linear Regression** - Use multiple linear regression to assess the impact of various factors on taxi fares by estimating the contribution of each predictor.
- **Tree-based Methods -** Explore algorithms such as Random Forest and XGBoost to evaluate feature importance to obtain insights into the relative importance of each variable in estimating taxi fares.

Statistical Testing:

- **Hypothesis Testing -** Conduct hypothesis testing to determine statistical significance across factors affecting fares.