**NYC Yellow Cab Data Analysis Report**

**Project Overview**  
The **New York City Yellow Cab: Processing Large Datasets** project aims to analyze and derive valuable insights from the vast dataset of NYC Yellow Cab trips. Using authoritative sources like the NYC Taxi & Limousine Commission (TLC), the dataset covers millions of taxi trips from 2023, including trip details such as fare amounts, pickup/drop-off locations, and passenger counts.

The goal is to understand key operational patterns in NYC’s taxi services in 2023, including **high-demand locations**, **peak travel hours**, **fare trends**, and **revenue breakdowns**. The insights from this analysis will help optimize taxi operations and support decision-making for urban mobility improvements.

**1. Data Acquisition and Preprocessing**

The first step involved gathering raw data from **NYC TLC**. The dataset was provided in **Parquet file format**, which contains the following fields:

* **tpep\_pickup\_datetime** and **tpep\_dropoff\_datetime** – Timestamp data for when a trip started and ended.
* **fare\_amount**, **tip\_amount**, **total\_amount**, etc. – Different fare components.
* **pickup\_location** and **dropoff\_location** –data for understanding popular pickup and drop-off locations.

**Data Cleaning & Transformation**

The data was cleaned and preprocessed as follows:

* **Handling Missing Values**: Missing values in certain columns were filled with appropriate methods, such as using the **mean** for numeric columns.
* **Duplicates Removal**: Duplicate rows were identified to ensure data accuracy.
* **Datetime Conversion**: The **pickup** and **dropoff** timestamps were converted into datetime formats, and additional features like **hour of the day** were extracted for time-based analysis.

**2. Extract-Transform-Load (ETL) Pipeline**

To streamline the data intake process, an **ETL pipeline** was built. The pipeline automates the extraction of the dataset, performs transformations (such as cleaning, filtering, and aggregations), and loads the data into a **relational database** for optimized storage and querying.

**Key Components:**

* **Extract**: Raw data is acquired from **NYC TLC** and stored temporarily.
* **Transform**: The data is cleaned, transformed, and enhanced (e.g., extracting hourly data).
* **Load**: The processed data is loaded into a structured database that can be easily queried for analysis.

This automated pipeline helps improve the scalability and efficiency of handling large datasets.

**3. Data Analysis & Insights**

The core of this project focused on uncovering key insights from the NYC Yellow Cab data. Below are some major findings:

**a. High-Demand Locations**

Using coded location data, the **top 10 pickup locations** in New York City were identified. Areas near major transportation hubs like **Upper East Side South, Upper East Side North** and **Midtown Center** had the highest pickup counts. **Bar charts** were created to visualize the **density of trips** around these locations. The top drop-off locations were the same as the top pick up destinations.

**b. Peak Travel Hours**

By analyzing the **pickup hour**, we identified that the **morning rush hour** (8-9 AM) and **evening rush hour** (5-7 PM), peaking at **6 PM** are when the majority of trips are taken. This insight helps taxi services allocate resources more efficiently during these peak times.

The peak months were recognized as **March, May and October** with rides being strikingly low in the months of July, August and September. Perhaps the **walkability** during these times of the year results in people opting to walk instead of taking short duration rides.

**Wednesdays** and **Thursdays** seem to be the most popular days for the taxi service while Sundays seem to be a bit slow.

**c. Fare Distribution**

A detailed analysis of the **fare amounts**, including components such as **tip**, **tolls**, and **surcharges**, showed that the average fare for trips was $28.5. The total amount of fare was rightly, positively correlated with trip distance as well as the tip amount. The distribution of the data was positively skewed with majority of the rides being accounted for on the lower side of the fares.

**d. Revenue Breakdown**

A breakdown of revenue across various components (e.g., **fare\_amount**, **tip\_amount**, **total\_amount**) highlighted the highest proportion of income is derived from fare amount, followed by tip amount and congestion surcharge amounts respectively.

**e. Trip Distance**

A histogram of data related to trip distances revealed that the distribution of the data was positively skewed which signals to majority of the trips being accounted for in shorter distances than in longer distances.

**4. Performance Optimization**

Given the size of the dataset, performance optimization strategies were essential for efficient analysis. Key optimizations include:

* **Using pandas optimized functions** for aggregations to speed up processing.
* **Efficient memory handling** for working with large datasets, including using data types as in **.parquet** format to reduce memory usage.

**5. Visualizations**

Several visualizations were created to present the analysis effectively:

* **Heatmaps**: Used for visualizing **correlation between fare components** and other ride related components.
* **Line Plots**: Showed **trends in fare amounts** over time (monthly and hourly), as well as **peak hours** for trips.
* **Bar Charts:** Showed **weekly** **trends**, **popular drop-off** and **pick-up** destinations.
* **Histogram:** Showed distribution of **trip distance** and **fare amounts**.

**6. Conclusions & Recommendations**

Based on the analysis, the following recommendations were made:

* **Resource Allocation**: Focus on facilitating with more resources during **rush hours** (morning and evening) to better handle high demand.
* **Targeted Promotions**: Consider implementing **promotions** or **discounts** for trips to less congested areas to spread demand.
* **Revenue Opportunities**: Evaluate opportunities to **increase fares** during peak congestion periods and **promote tipping** in high-value trips (such as to/from airports).

**Conclusion**

This project demonstrated the power of **big data** processing and **analytics** in understanding **NYC Yellow Cab operations**. By uncovering key trends and patterns, this analysis can drive **data-driven decisions** that improve taxi operations, enhance customer experience, and optimize revenue generation.

A graph of a number of blue lines

AI-generated content may be incorrect.

A graph of a trip distance

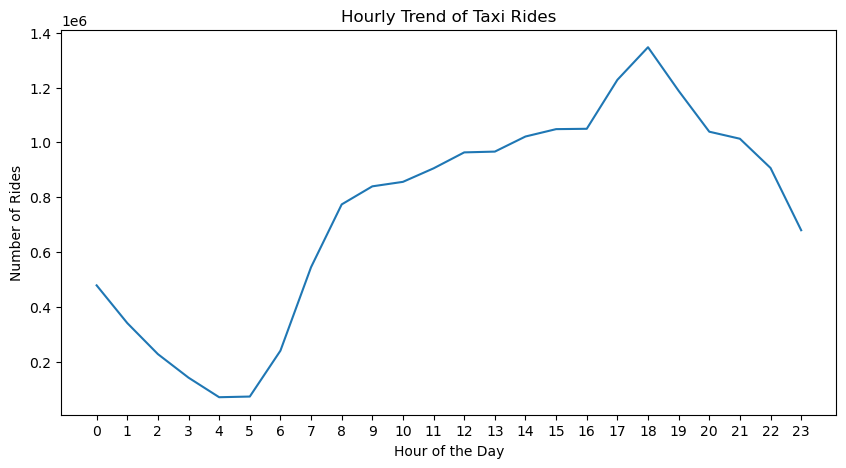
AI-generated content may be incorrect.

A graph of a taxi ride

AI-generated content may be incorrect.

A graph of blue bars

AI-generated content may be incorrect.



A graph of blue bars

AI-generated content may be incorrect.

A graph of blue bars

AI-generated content may be incorrect.

A graph with blue squares

AI-generated content may be incorrect.

A screenshot of a computer screen

AI-generated content may be incorrect.