### **Project Report:**

### **New York City Yellow Cab: Processing Large Datasets**

### **Introduction**

The NYC Yellow Taxi dataset provides comprehensive trip records, including pickup and drop-off locations, timestamps, fare amounts, and passenger counts. This project aims to analyze patterns, trends, and insights from this data to improve urban transportation planning and enhance taxi services. 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**.

**Objectives**

* Understand trip distribution across different times and locations.
* Analyze fare trends based on distance and time.
* Identify peak hours and seasonal variations.
* Evaluate passenger demand and supply imbalance.
* Assess the impact of external factors like weather and events.

1. **Data Acquisition:**

The dataset used for this analysis is obtained from the NYC Taxi & Limousine Commission (TLC). The key attributes include:

* **Pickup & Drop-off Datetime**: Timestamp of ride start and end.
* **Pickup & Drop-off Location IDs**: Identifiers for trip locations.
* **Trip Distance**: Distance covered in miles.
* **Fare Amount**: Cost of the trip.
* **Passenger Count**: Number of passengers.
* **Payment Type**: Cash, credit, or other payment methods.

2. **Data Preprocessing:**

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 and removed 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.
* **Handling outliers:** Filtering out unrealistic fare amounts and trip distances.
* **Feature Engineering**: Creating new variables such as time of day, day of the week.

2. **EDA and Insights:**

**a. Fare Amount Distribution:** A comprehensive examination of fare components, including tips, tolls, and surcharges, revealed that the average trip fare was $28.5. The total fare amount exhibited a strong positive correlation with both trip distance and tip amount. The data distribution was positively skewed, with most rides falling within the lower fare range.

**b. Trip Distance:** A histogram analyzing trip distances indicated a positively skewed distribution, suggesting that most trips covered shorter distances rather than longer ones.

**c.** **Revenue Breakdown:** An analysis of revenue components, including fare amount, tip amount, total amount, extra, and surcharges revealed that the majority of income comes from the fare amount, followed by tips and congestion surcharges, respectively.

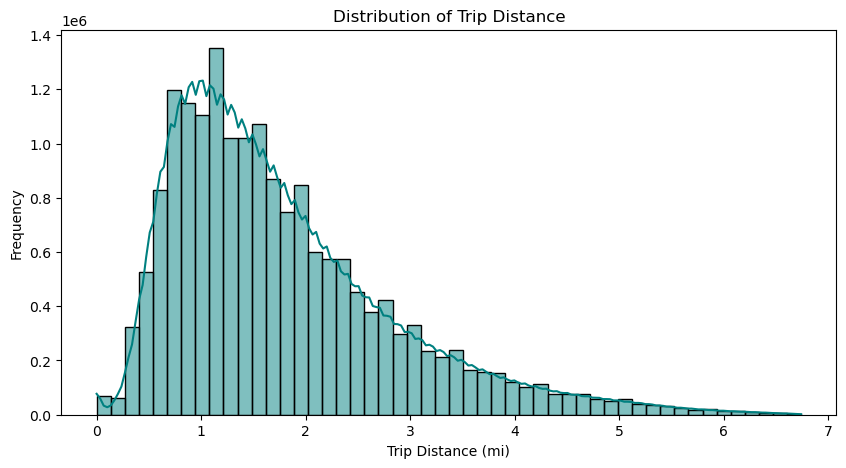
**d. High-Demand Locations:** By analyzing coded location data, the ten most popular pickup spots in New York City were identified. Areas near major transit hubs, such as Upper East Side South, Upper East Side North, Midtown East and Midtown Center, recorded the highest number of pickups. Bar charts were used to illustrate trip density around these locations. Notably, the most frequent drop-off locations mirrored the top pickup points.

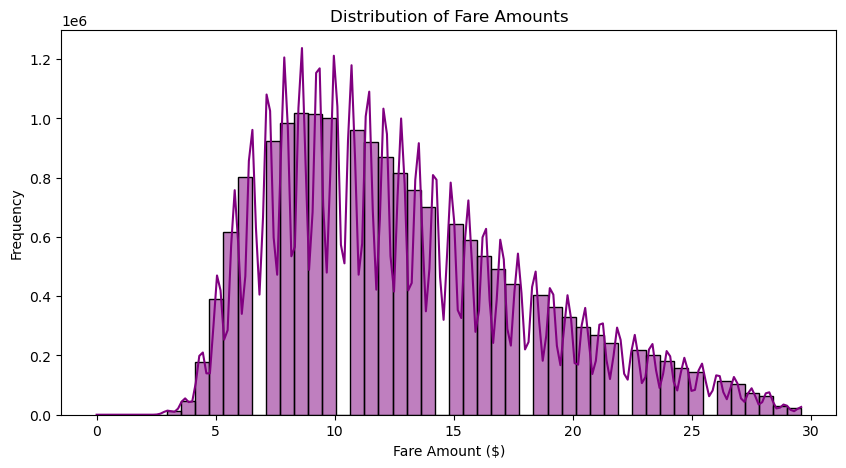
**e. Trip Distribution:** Analyzed trip frequency by hour, day, and month. Peak hours were observed during morning (8 AM - 10 AM) and evening (5 PM - 8 PM) rush hours. The busiest months were identified as March, May, and October, while July, August, and September saw notably fewer rides. Wednesdays and Thursdays appear to be the most popular days for taxi use, while Sundays tend to be slower.

**3. Visualizations**

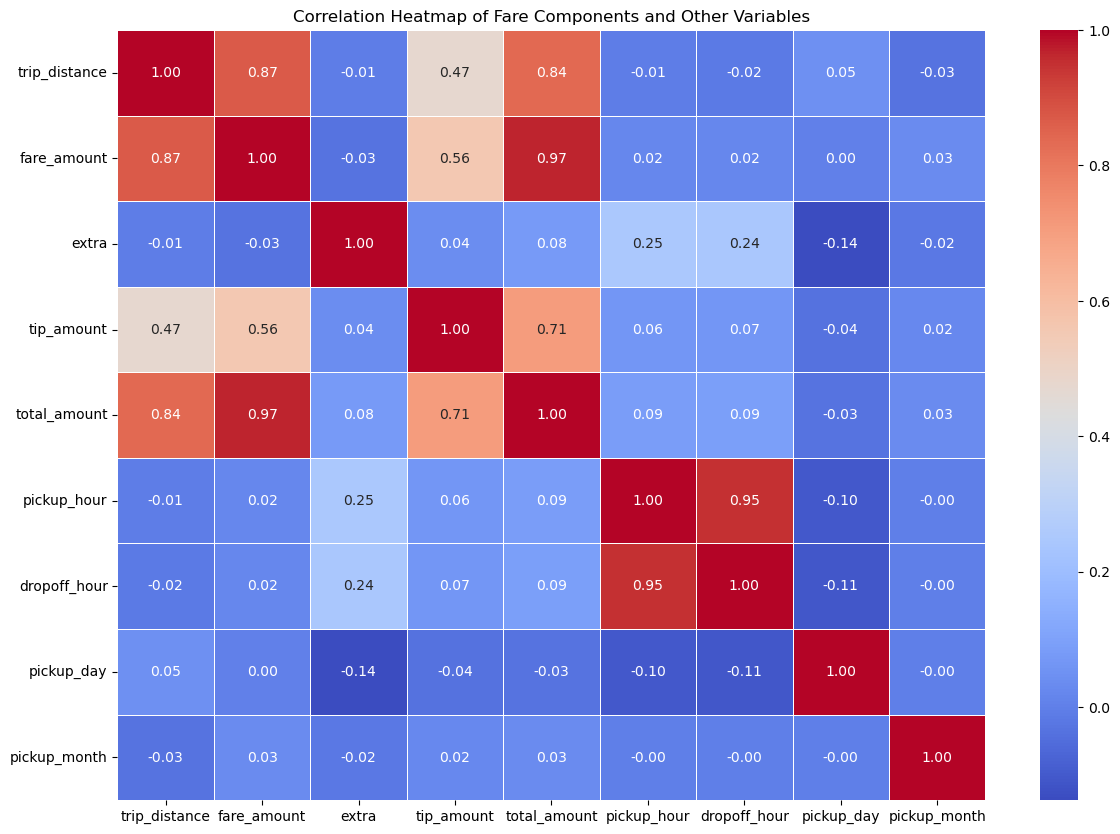
A variety of visualizations were created to present the analysis clearly:

* **Histograms:** Represented the distribution of trip distances and fare amounts

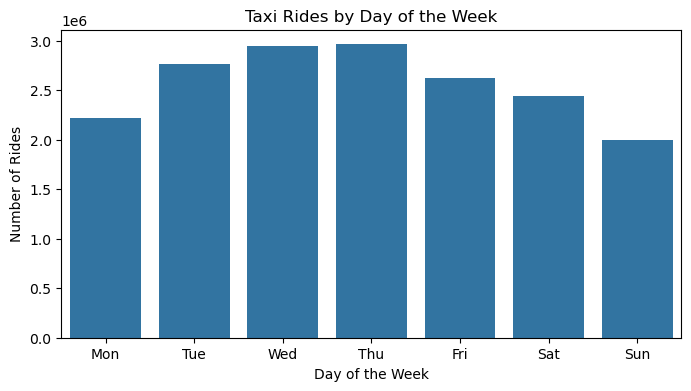


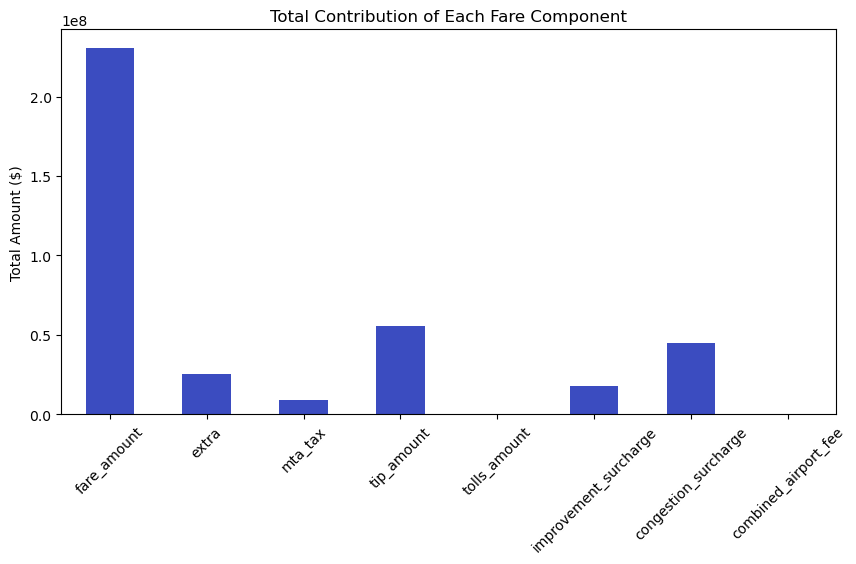
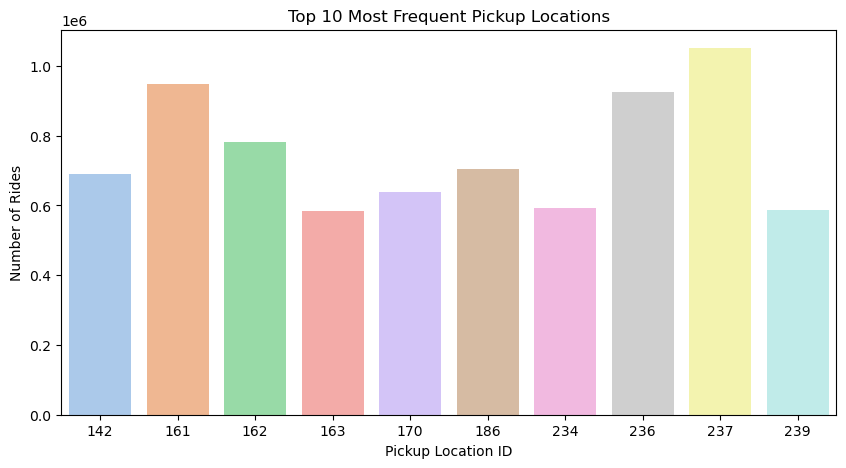


* **Heatmaps:** Displayed the correlation between fare components and other ride-related factors.

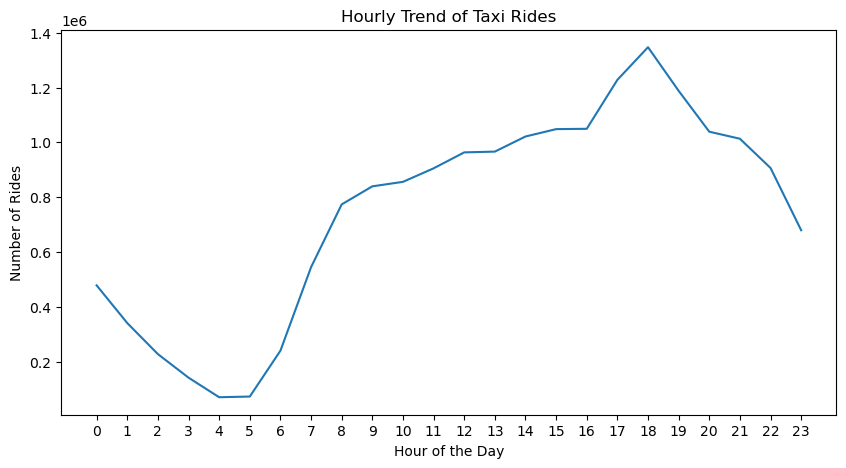


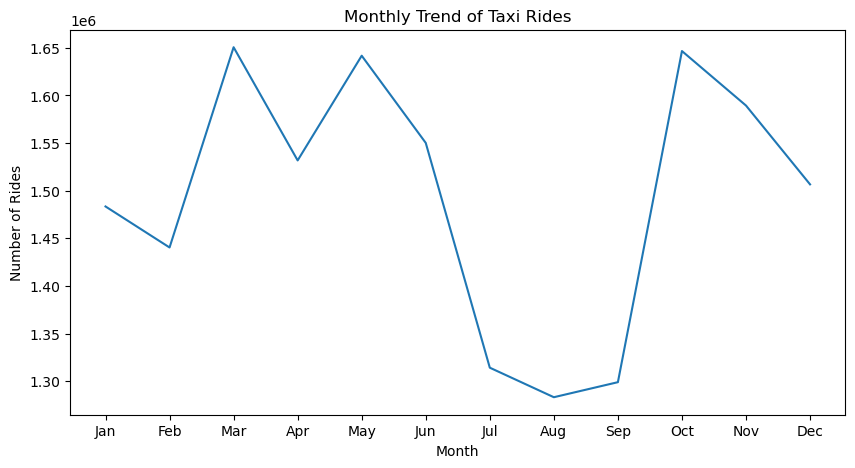
* **Bar Charts:** Depicted weekly trends , the most popular pick-up and drop-off locations and the revenue breakdown.





* **Line Plots:** Illustrated trends in fare amounts over time (both monthly and hourly), highlighting peak trip hours and months.





**Conclusion:**

The analysis of NYC Yellow Taxi data provided valuable insights into ride patterns, fare trends, and peak usage periods. It was observed that March, May, and October had the highest number of rides, while July through September saw a decline, possibly due to favorable weather conditions encouraging walking. Midweek days, particularly Wednesdays and Thursdays, experienced the highest taxi demand, whereas Sundays recorded lower activity. Additionally, fare trends varied across different times of the day, with peak hours aligning with commuting and nightlife periods. The analysis also highlighted popular pick-up and drop-off locations, helping to identify high-demand areas.

**Recommendations:**

* Optimizing Taxi Availability: Since demand fluctuates throughout the year and week, taxi services should adjust fleet distribution accordingly, ensuring higher availability during peak months and busy midweek days.
* Dynamic Pricing Strategies: Implementing fare adjustments based on peak hours and seasonal demand could help maximize revenue while maintaining service efficiency.
* Targeted Promotions: During low-demand periods, discounts or ride-sharing incentives could encourage more taxi usage.
* Enhancing Service in High-Demand Areas: Identifying and focusing on frequently visited locations, such as major transit hubs and entertainment districts, can improve customer satisfaction and operational efficiency.

By leveraging these insights, taxi operators and city planners can enhance service efficiency, improve user experience, and optimize fleet management strategies.