Taxi Rides Transactional Dataset

Name : Aakanksha Pednekar

Date : 30-01-2026 to 05-02-2026

INDEX

INTRODUCTION

This study focuses on the analysis of a Taxi Rides Transactional Dataset which contains nearly 181 million taxi trip records, including operational, financial, transactional and geographical information.

The main objective of this study is to explore the dataset and derive meaningful insights that help in understanding travel patterns, passenger behaviour , and revenue trends and to improve services.

The analysis was done in seven days using multiple tools such as Python, SQL, Microsoft Excel, and Power BI.

Problem Statement

The Taxi Rides dataset contains a very large amount of data collected over years. Because the dataset is so big, complex and contains null values , it cannot be analysed directly in its raw form.

The data needs to be cleaned, organized, and prepared before useful information can be obtained.

The main challenge of this , is to turn the raw dataset into a clean and well-structured format that can use for analysis.

The goals of this study are to understand trip duration and distance patterns , study fare and revenue patterns, understand time-based demands, explore payment and tipping behaviour, and identify areas with high pickup and drop-off activity.

Dataset Description

The Taxi Rides Transactional Dataset contains nearly 181 million rows and twenty-three columns, and its total size is about 13.3 GB.

- **Each row represents a single completed taxi trip.**
- **The dataset includes different types of data such as categorical, numerical, time-based, and geographical information.**
- **Categorical columns include trip ID, taxi ID, payment type, and company name.**
- **Numerical columns contain details such as trip duration, trip distance, fare amount, tips, tolls, extra charges, and total trip cost.**
- **Time-based columns record the start and end time of trip.**
- **Geographical columns store pickup and drop-off locations using census tracts, community areas, and latitude-longitude coordinates.**

Since the dataset is extremely large, it was not possible to upload and analyse the entire dataset at once. Therefore, a sample dataset containing 10,48,575 rows and 23 columns was selected to use for analysis.

Tools and Technologies

- **Python (Pandas, NumPy) – Data loading, cleaning, and manipulation.**
- **Matplotlib and Seaborn – Data visualization and plotting.**
- **SQL Workbench – Query-based analysis and subqueries.**
- **Microsoft Excel – Pivot tables for summarization and Visualisations.**
- **Power BI – Dashboard creation.**

Exploratory Data Analysis

Stages Performed during EDA :

- Imported important libraries like Pandas , NumPy , Matplotlib and seaborn .

- I have loaded the CSV file into a Pandas Data Frame.[df]

- Used df.shape to check the number of rows and columns.

- Converted column to standardize form , lowercase and replaced spaces with underscores.

- Using df.dtypes , displayed data types of all columns.

- Then, checked Missing Values in each column , and also counted duplicate rows in the dataset. As, there are no such duplicate rows.

Data Type Adjustment

- Removed commas and converted trip_seconds to numeric from object.

- Converted fare, tips, tolls, extras, and trip_total into numeric format as it contains ($) and was in object type .

- Converted trip_start_timestamp and trip_end_timestamp into datetime.

- Converted pickup centroid latitude and longitude , dropoff centroid latitude and longitude , pickup census_tract, dropoff_census_tract , pickup_community_area , dropoff community area to string/object type as it works as Categorial Column.

- Then, Verified Updated Data Types after conversions.

- **Missing Values Handling -**
  1.Filled null taxi_id values.
  2. Filled missing trip_end_timestamp using trip_start_timestamp.

  3. Calculated the trip duration in seconds by subtracting the trip start time from the trip end time.

4.Also, where the distance is 0 miles and the trip duration is still missing , have set trip_seconds to 0.

5. Vice Versa done for trip_miles , where trip_seconds = 0 and miles are missing, have set trip miles = 0 , and for other places have put the mode value of miles.

- **Creating new columns**
  Created new columns such as trip_minutes, trip_hour, trip_day, trip_month, and trip_year ,distance_category and trip_duration_group .

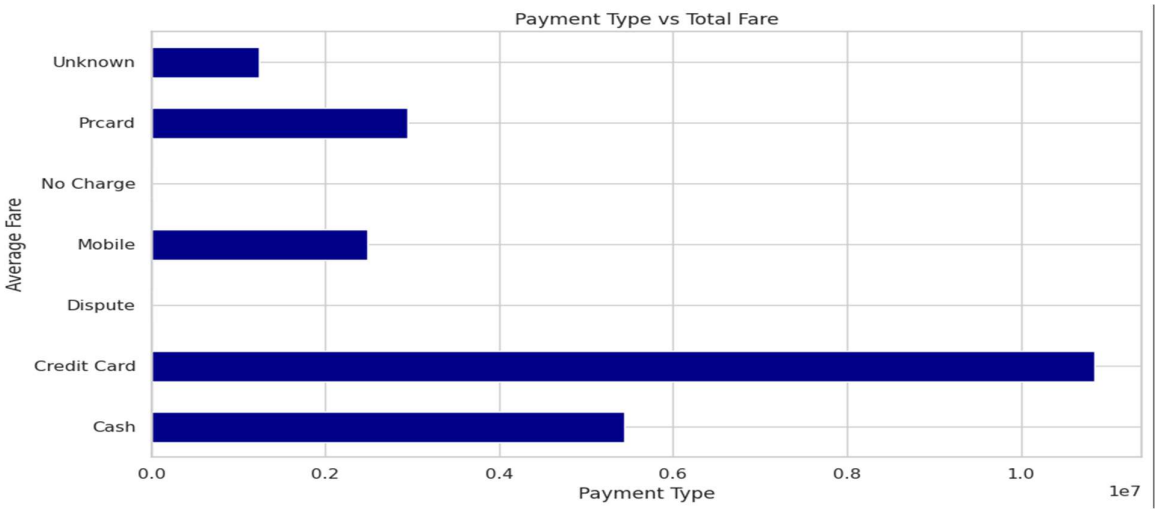
- After doing all the above steps , again checked the shape of the dataset and datatypes and null value counts.

- When I got all the outputs correct with all the conditions, saved the file in CSV format.

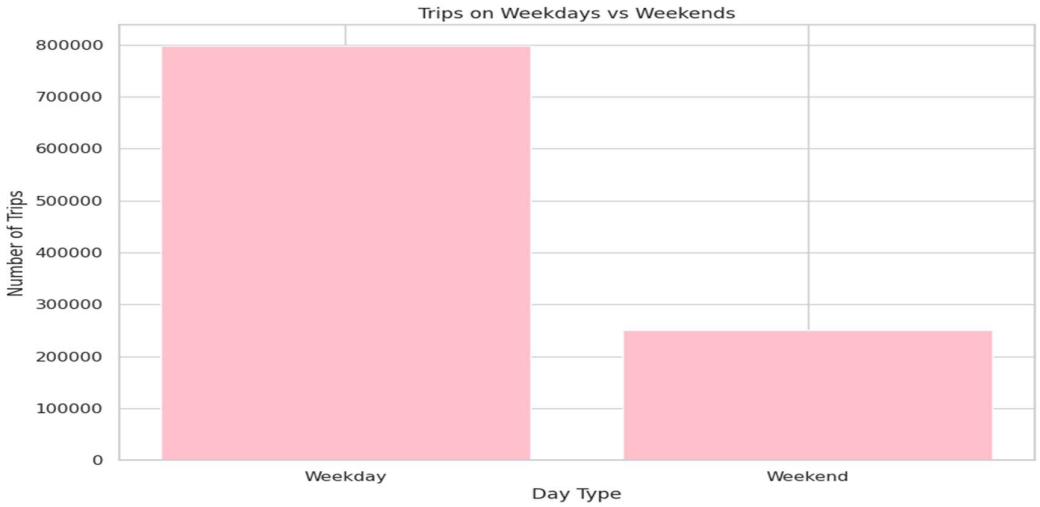
- Then, I have moved further for Visualisations Using libraries like Matplotlib and seaborn.

Following Chart shows Payment Distribution across Rides:

- Card payments hold the largest share of total fares, making them the primary source of revenue.
- Cash transactions follow as the second- largest contributor, while Prcards and mobile payments add smaller portions.
- Unspecified, disputed, and no-charge payments make up only a minimal share, suggesting they used rarely for payments.



Trip demands on weekdays vs weekends :



-Most trips occur on weekdays, showing much higher travel activity compared to weekends.

- Weekend trips are noticeably lower, indicating reduced demand outside the regular workweek.

SQL ANALYSIS

The goal of the SQL analysis was to examine the cleaned Taxi Rides dataset using SQL .

The dataset was exported as a CSV file, imported into SQL Workbench .

Due to the large size of the dataset (over one million rows), MySQL Workbench experienced performance lagging and was able to load only a subset of approximately 62,000 records.

SQL-based analysis was performed on this sample. Then, checked to ensure all fields were correctly loaded with suitable data types.

Core SQL commands such as SELECT, WHERE, GROUP BY, ORDER BY, and LIMIT were used to explore and organize the data.

Aggregate functions like COUNT, SUM, AVG, MIN, and MAX helped summarize important numerical fields including trip duration, distance, fare, and total cost.

Subqueries were included for deeper analysis, such as identifying trips with fares above the overall average, finding taxis with higher-than-average trip volumes, and highlighting high-revenue rides. These techniques allowed for more detailed and meaningful understanding of the dataset.

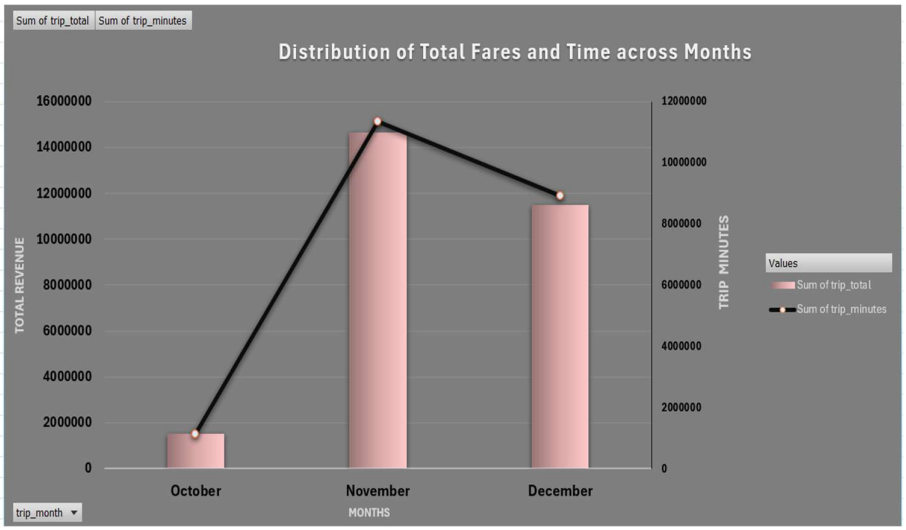
Excel

The Excel analysis based on exploring the Taxi Rides dataset using Pivot Tables and charts to summarize large size of data and identify patterns.

After importing the cleaned dataset into Microsoft Excel, Pivot Tables were created to organize the data.

A. Distribution of Total Fare And Trip Duration across Months:

Sum of trip_total	Sum of trip_minutes
1507623.4	1138416
14680424.62	11337448
11501471.17	8917030
27689519.19	**21392894**



- November has the highest total fares and the longest total trip time, showing it was the busiest month overall.

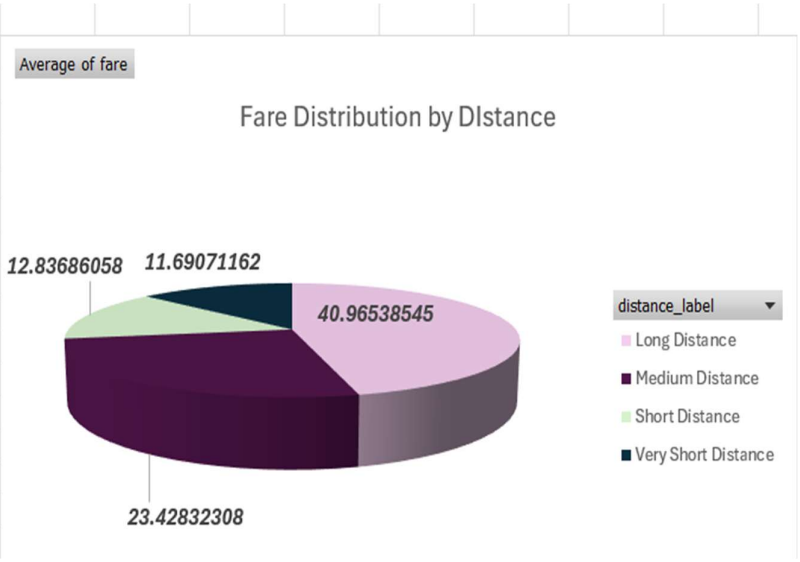
- The average fare increases from October to November and then decreases in December, indicating fewer rides.

B. Fare Distribution by Distance Categories :

- Longest distance rides usually cost more on average, while very short and short trips are much cheaper.

- Medium-length trips sit in the middle range, with fares increase in fare as trip distance increases.

Row Labels	Average of fare
Long Distance	40.96538545
Medium Distance	23.42832308
Short Distance	12.83686058
Very Short Distance	11.69071162
Grand Total	**21.89285253**

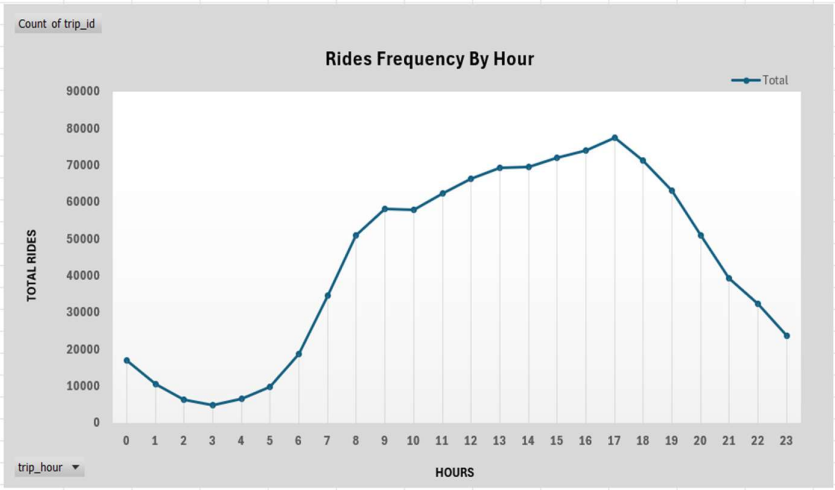


- Overall, the chart shows a clear relationship between distance and fare, as longer trips generally result in higher charges.

C. Ride Frequency by Hour :

- Ride activity is lowest during the late-night and early-morning hours and begins to rise steadily after 6 AM.

Row Labels	Count of trip_id
0	17051
1	10503
2	6371
3	4967
4	6504
5	9888
6	18694
7	34547
8	50990
9	58117
10	58088
11	62464
12	66337
13	69466
14	69630
15	72145
16	74143
17	77619
18	71279
19	63143
20	50933
21	39446
22	32388
23	23862
Grand Total	**1048575**



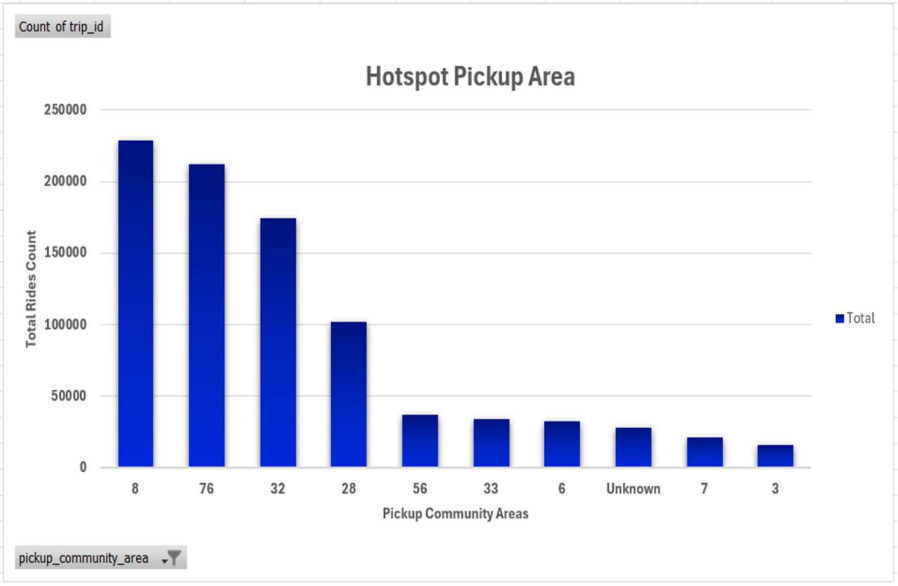
- The number of rides increases throughout the day and reaches its peak during the late afternoon to early evening hours.

- After the evening peak, ride frequency gradually declines as the night progresses.

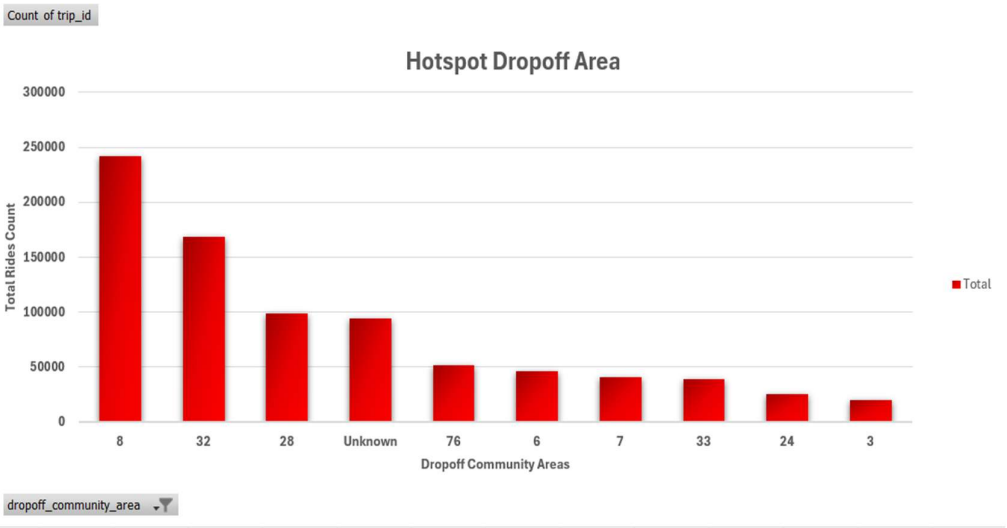
D. Hotspot Pickup and Dropoff Areas –

- Community Areas 8 and 32 main centers of taxi activity, showing consistently high volumes for both trip origins and destinations.

Row Labels	Count of trip_id
8	228784
76	212260
32	174724
28	102236
56	36983
33	34309
6	32342
Unknown	27707
7	21283
3	16254
Grand Total	**886882**



Row Labels	Count of trip_id
8	241863
32	168146
28	98702
Unknown	93981
76	51569
6	46323
7	41325
33	39202
24	25498
3	20513
Grand Total	**827122**



- Some locations experience noticeably more pickups than drop-offs, while others show the opposite pattern, indicating different roles these areas play in trip behaviour.

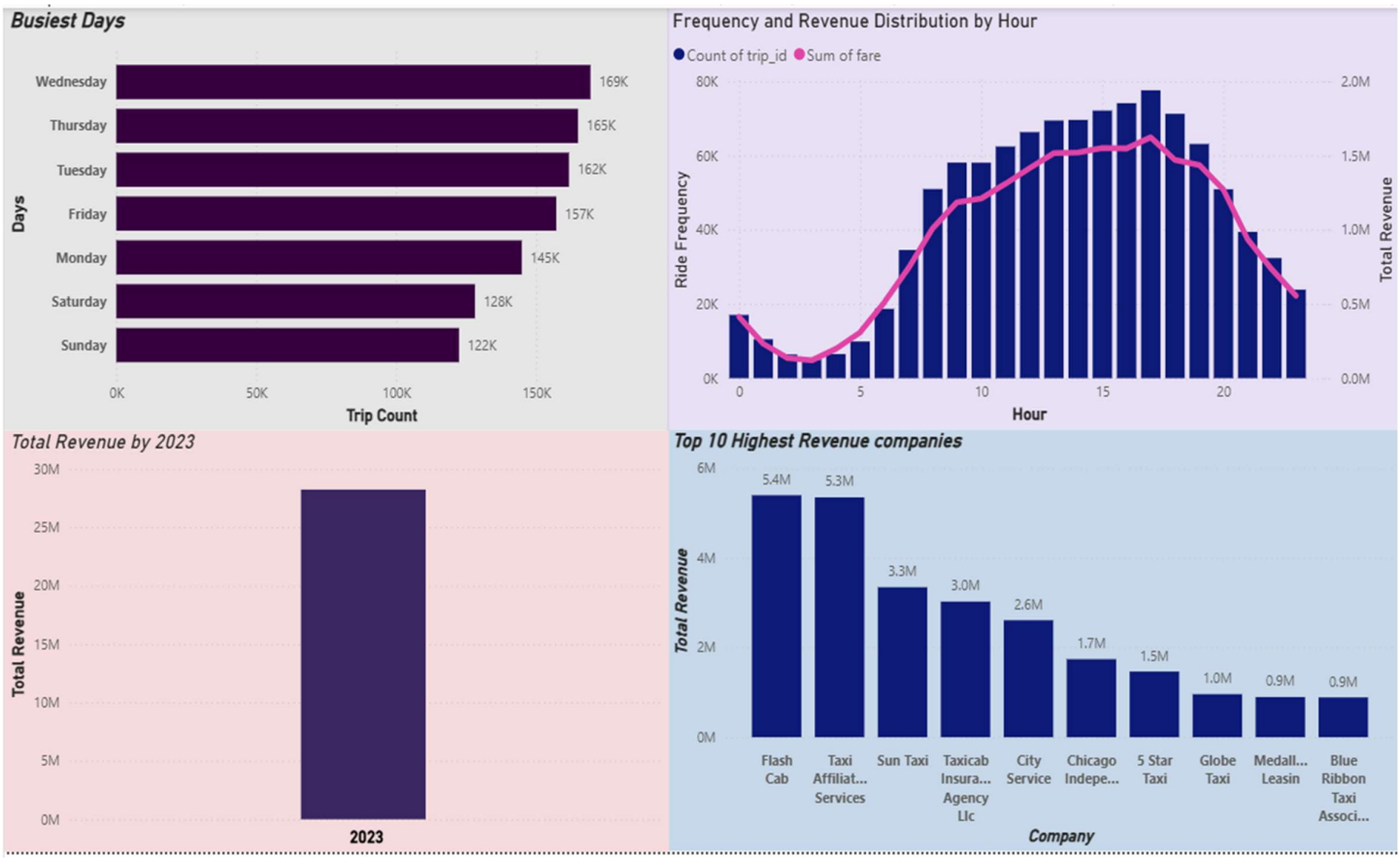
- some noticeable records come under an “Unknown” category, missing location data that may need closer examination before further analysis

Power BI Analysis

The purpose of using Power BI in this study is to design meaningful dashboards that display important insights from the Taxi Rides Transactional Dataset in a clear and easy-to-understand format.

The cleaned dataset was first imported into Power BI from a CSV file. After loading the data, columns were reviewed to ensure correctness.

Existing measures such as total trips, total fare, and total trip amount were used, and additional calculations like minimum tip, maximum tip, and target tip were created using DAX formulas.

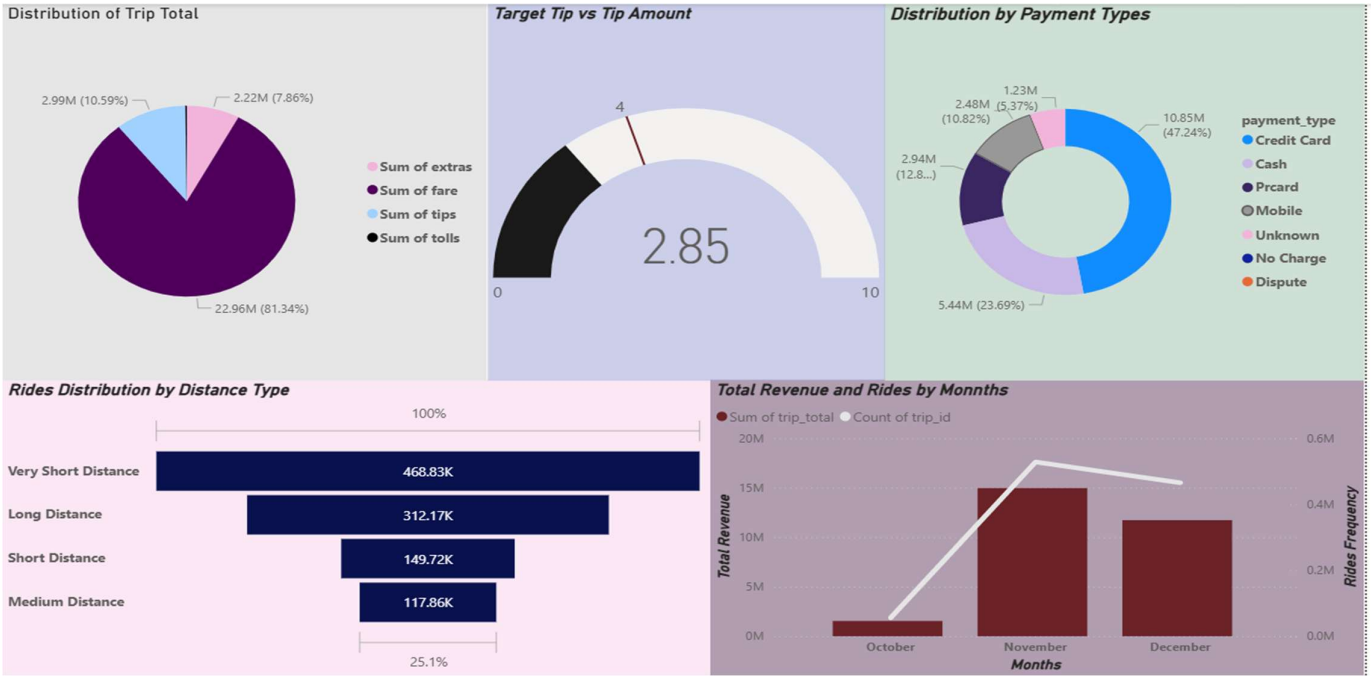


1. The **Busiest Days** visual indicates that ride demand is greater on weekdays than on weekends, indicating higher usage during standard weekdays.

2. The **Hourly Frequency and Revenue Distribution** chart shows that trip volume and earnings rise through the daytime, peak in the late afternoon to early evening, and decrease during nighttime hours.

3. The **Total Revenue** for 2023 chart summarizes the overall income generated throughout the year, offering a high-level financial overview.

4.The **Top 10 Revenue-Generating Companies** chart highlights the taxi operators with the largest contributions to total earnings, enabling performance comparison.



1. The **Distribution of Trip Total** chart indicates that large amount of fares covers most of the total trip cost, while tips, additional charges, and tolls represent smaller portions.

2. The **Target Tip vs Tip Amount** gauge shows that the current average tip falls below the desired benchmark, indicating there should be improvement in services for more tips.

3. The **Distribution by Payment Types** visual indicates that credit cards are the most commonly used payment method, followed by cash and prepaid cards, with fewer transactions made through mobile and other options.

4.The **Rides Distribution by Distance Type** chart shows, very short trips are the most frequent, long trips are most common, and medium and short trips occur less often.

5.The **Total Revenue and Rides by Months** chart shows that both earnings and ride volume rise from October to November and then decrease slightly in December.

Overall Observations

The Taxi Rides Transactional Dataset includes a large volume of records containing details about trips, pricing, time, and locations. An initial review done in Python showed that while the dataset was well structured ,still certain fields had missing values and required adjustments to their data types. After completing data cleaning and preparation, the dataset was ready for detailed analysis.

Exploratory analysis revealed clear patterns in trip length, distance, and fare values. Longer trips and extended travel times Is likely having higher fares. Ride activity is lowest during late-night and early-morning hours, increases steadily throughout the day, and peaks in the late afternoon and evening. Also , study related to payment types showed that card payments are more used than cash or other options.

SQL was particularly effective for aggregating large datasets and identifying top-earning companies and commonly occurring trip types.

Excel Pivot Tables and charts were used to quickly compare trends across months, distance groups, and locations. These summaries made it easier to find patterns and differences.

Power BI dashboards combined all insights into an interactive dashboard , allowing users to explore ride volume, revenue patterns, payment preferences through pictures and filters.

Conclusion

This project shows a full data analysis process using Python, SQL, Excel, and Power BI.

The data was cleaned, prepared, studied, and visualized to find useful insights.

The findings show strong links between trip distance, trip time, and fare amount, along with clear patterns in ride demand at different times of day.

By using multiple tools, the project provides a broad view of taxi operations.