Uber Data Analysis

A Detailed Exploratory Study on Ride Patterns

Harsh Gangavane

Abhishek Debaje

July 2025

Contents

Abstract

This project presents an exploratory data analysis of Uber ride data to uncover usage patterns, peak travel times, and trip trends. Using Python and various data science libraries, we clean, visualize, and derive key insights from the dataset collected from Uber rides in New York City.

1 Introduction

Uber has revolutionized urban transportation. With millions of daily rides, analyzing this data can help understand commuter behavior and improve service efficiency. This project focuses on identifying temporal and spatial trends, such as busiest hours, days, and areas with high demand.

2 Libraries and Tools Used

The analysis was performed using:

- **Python 3.11** – Core programming language

- **Pandas** – Data manipulation and analysis

- **NumPy** – Numerical operations

- **Matplotlib & Seaborn** – Data visualization

- **Jupyter Notebook** – Interactive coding environment

3 Dataset Description

The dataset used is publicly available on Kaggle: Uber NYC Data. It contains information about Uber pickups including:

- Date and Time of pickup

- Latitude and Longitude

- Base code of the vehicle

It spans several months and helps analyze trends in time and location dimensions.

4 Data Preprocessing

Before analysis, several steps were taken to clean and prepare the data:

- Converted timestamp strings to Python datetime format

- Removed rows with null or invalid values

- Extracted useful features such as hour, weekday, and month

Listing 1: Datetime conversion example

```python
import pandas as pd

df = pd.read_csv("uber_data.csv")
df['Date/Time'] = pd.to_datetime(df['Date/Time'])
df['Hour'] = df['Date/Time'].dt.hour
df['Weekday'] = df['Date/Time'].dt.weekday
```

5   Exploratory Data Analysis

The exploratory analysis revealed:

- **Most active hours:** 5 PM – 7 PM

- **Busiest days:** Fridays and Saturdays

- **Popular locations:** Midtown Manhattan, Lower Manhattan

trip_frequency.png

Figure 1: Trip frequency by hour of day

Listing 2: Plotting trip frequency

```
import seaborn as sns
import matplotlib.pyplot as plt
```

```
sns.countplot(x='Hour', data=df)
plt.title("Trip Frequency by Hour")
plt.xlabel("Hour")
plt.ylabel("Number of Trips")
plt.show()
```

6   Key Findings

1. Evening hours see the highest ride volumes

2. Weekends are significantly busier than weekdays

3. Geographic hotspots correlate with business and tourist areas

4. Monthly trends indicate seasonal variations in ride demand

These insights can help Uber improve surge pricing, fleet availability, and targeted marketing strategies.

7 Conclusion

This project effectively demonstrated how data analysis can be applied to real-world ride data. By cleaning, exploring, and visualizing Uber trip data, we extracted actionable insights. In the future, further analysis could involve clustering or predicting demand using machine learning models.

References

[1] Kaggle. Uber Pickups in New York City.
https://www.kaggle.com/datasets/fivethirtyeight/uber-pickups-in-new-york-city

[2] Wes McKinney. *Python for Data Analysis*. O'Reilly Media.

[3] Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science Engineering.

Appendix: Sample Code Snippets

Listing 3: Basic data summary

```python
# Display basic structure of dataset
print(df.info())


# Describe statistics of numerical columns
print(df.describe())
```

Listing 4: Heatmap by day and hour

```python
import numpy as np
heatmap_data = df.groupby(['Weekday','Hour']).size().unstack()
sns.heatmap(heatmap_data, cmap='YlGnBu')
plt.title("Heatmap of Rides by Day and Hour")
plt.show()
```