Uber Data Analysis

ARYAN DINESH JAMSUTKAR
VIRAJ VIVEK NAIK
VANDIT VIVEK NAIK

July 3, 2025

Contents

Abstract

This report presents an analysis of Uber ride data using Python-based data science tools. The dataset includes time-stamped records of Uber pickups in New York City. The project involves data preprocessing, visualization, and extracting trends in usage patterns to draw meaningful insights.

Additional analyses focus on identifying peak demand hours, high-traffic pickup zones, and patterns based on days of the week and months. Advanced techniques such as correlation analysis and time series decomposition are also applied to better understand user behavior. The findings provide valuable inputs for operational improvements, driver scheduling, and customer targeting strategies. Furthermore, the visualizations created using tools like Matplotlib and Seaborn make the trends more interpretable. This project showcases the real-world application of data science in solving urban mobility challenges. Overall, the study emphasizes the importance of data-driven decision-making in the transportation industry.

Chapter 1

Introduction

The aim of this project is to explore and analyze Uber ride data to identify key patterns in customer behavior, temporal trends, and usage hotspots. Using data science techniques, we uncover trends in pickup frequency, peak hours, and geographical patterns.

This analysis not only helps in understanding customer demand but also assists in optimizing driver allocation and improving service efficiency. By examining seasonal variations and weekly fluctuations, the project provides insights into operational challenges. Additionally, the data allows us to observe the impact of external factors like holidays and weather conditions on ride frequency. The use of Python libraries such as Pandas, NumPy, Matplotlib, and Seaborn enables robust data handling and visualization. This comprehensive study serves as a foundation for making data-driven business decisions and enhancing urban mobility solutions through predictive insights.

Chapter 2

Libraries and Tools Used

The analysis was conducted using the following Python libraries:

- **pandas** – for data manipulation
- **numpy** – for numerical operations
- **matplotlib** – for plotting graphs
- **seaborn** – for statistical visualizations
- **os** – for interacting with the operating system (e.g., file handling)
- **glob** – for pattern matching and file path retrieval

Chapter 3

Dataset Description

The dataset contains Uber pickup data, including timestamps and locations. Fields typically include:

- Date/Time : This field records the exact timestamp when a pickup occurred. It includes the year, month, day, hour, minute, and second, which allows for detailed temporal analysis. Using this field, we can derive additional features such as day of the week, time of day (morning, afternoon, evening), and detect patterns across different time intervals.

- Latitude and Longitude : These fields represent the geographical coordinates where each Uber ride was initiated. Latitude indicates the north-south position, while longitude represents the east-west position. These values help in plotting ride locations on maps, clustering high-demand areas, and identifying geographical hotspots or zones with higher activity.

- Base (Dispatch center or company ID) This field identifies the base station or company associated with the ride request. It helps in understanding which dispatch centers are most active and how ride requests are distributed across different operational units. This can also be useful for comparing performance and ride volumes between different bases.

Chapter 4

Data Preprocessing

- **Converted timestamps into datetime objects**: The original dataset contains timestamps in string format, which are not suitable for time-based analysis. Therefore, these timestamps were converted into Python's datetime objects to enable efficient extraction and manipulation of date and time components. This conversion is crucial for performing operations like filtering rides by hour, grouping by day of the week, or identifying peak hours.

- **Extracted day, hour, weekday, and month**: Once the timestamps were converted, various time-based features were extracted such as:

  - **Day**: Helps identify which days experience higher demand.

  - **Hour**: Useful for recognizing hourly usage patterns and peak times.

  - **Weekday**: Helps distinguish between weekday and weekend usage trends.

  - **Month**: Useful for tracking seasonal variation and long-term changes in demand.

  These derived features form the basis for most of the temporal analysis in the project.

- **Removed missing or incorrect data entries**: Data quality is vital for accurate analysis. Rows with missing values (such as null timestamps or coordinates) and invalid entries (e.g., latitude and longitude outside NYC range) were removed. This step ensures the integrity and reliability of the dataset, and helps avoid errors during visualization and statistical analysis.

Chapter 5

Exploratory Data Analysis

1. Uber pickups by hour

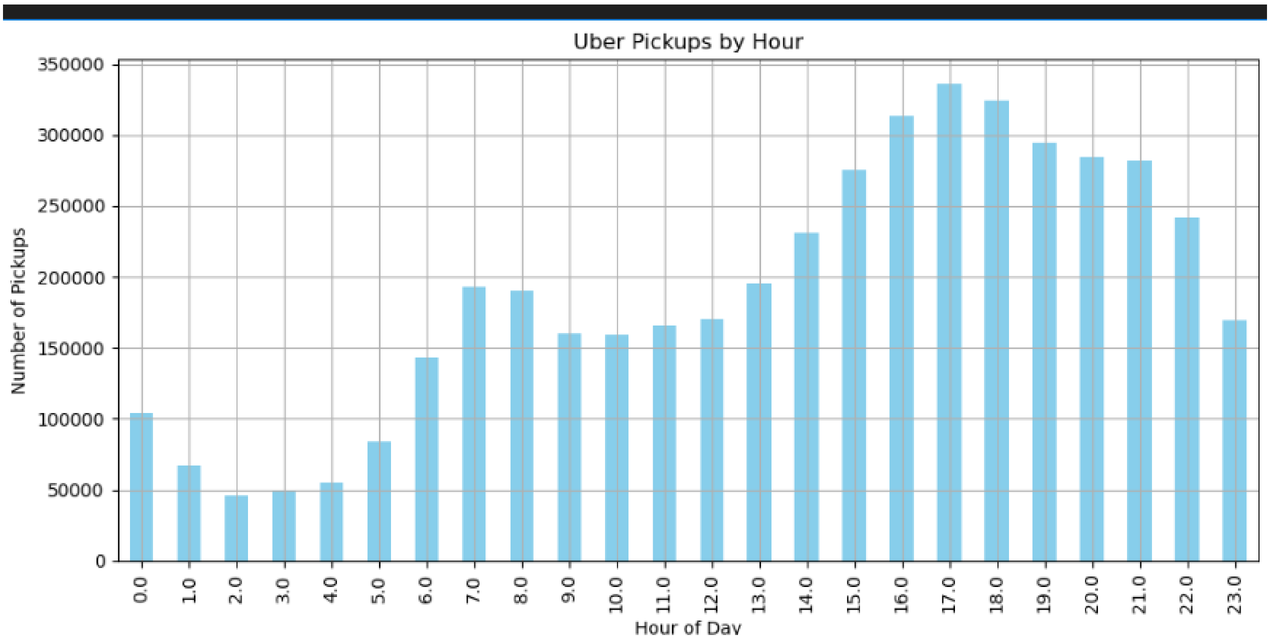

Figure 5.1: Uber Pickups by Hour

2. Weekday Pickups

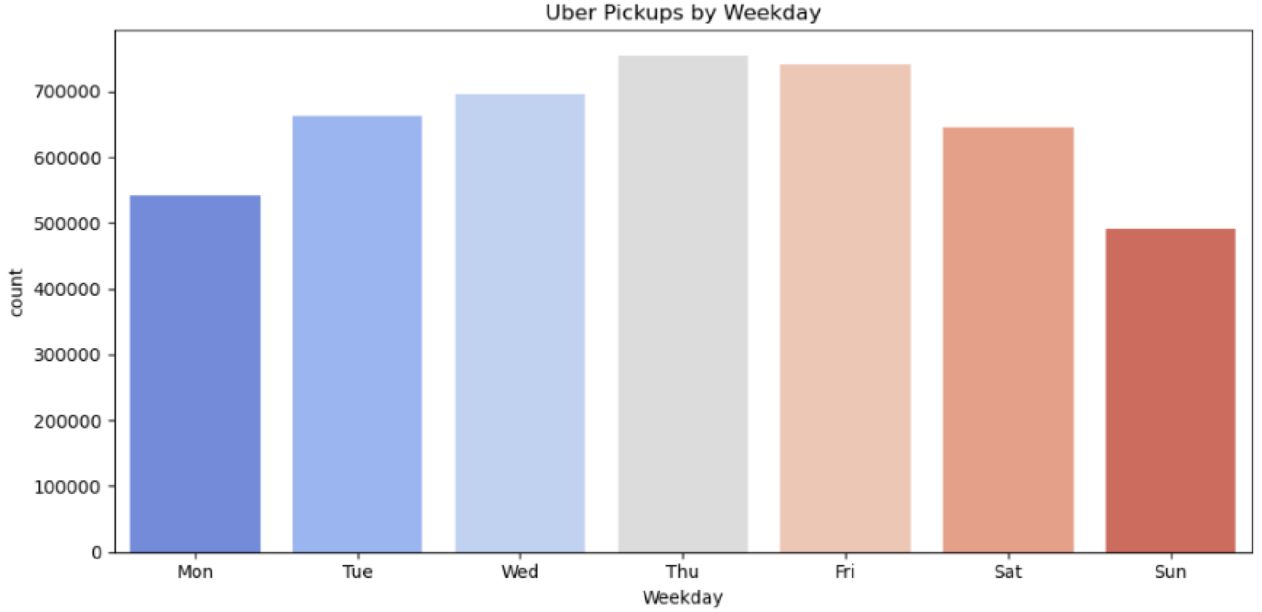

Figure 5.2: Uber pickups by weekday

3. Uber Pickups by Month

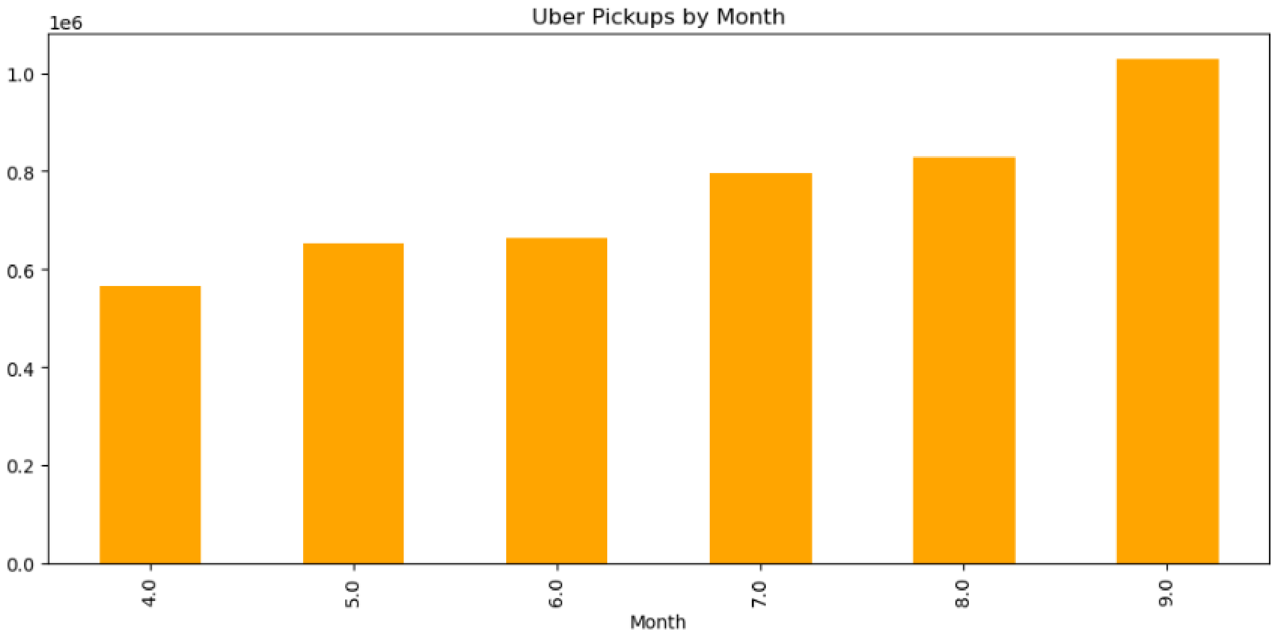

Figure 5.3: Uber Pickups by Month

4. Uber Pickups per Day

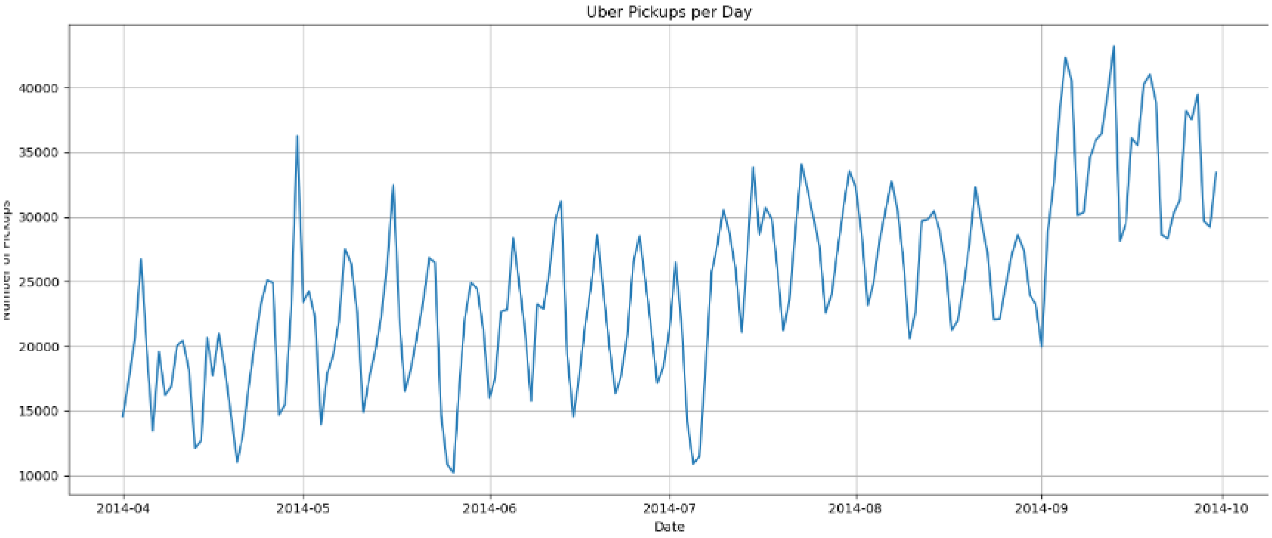

Figure 5.4: Uber Pickups per Day

5. Daily Pickups with Weekly Trends

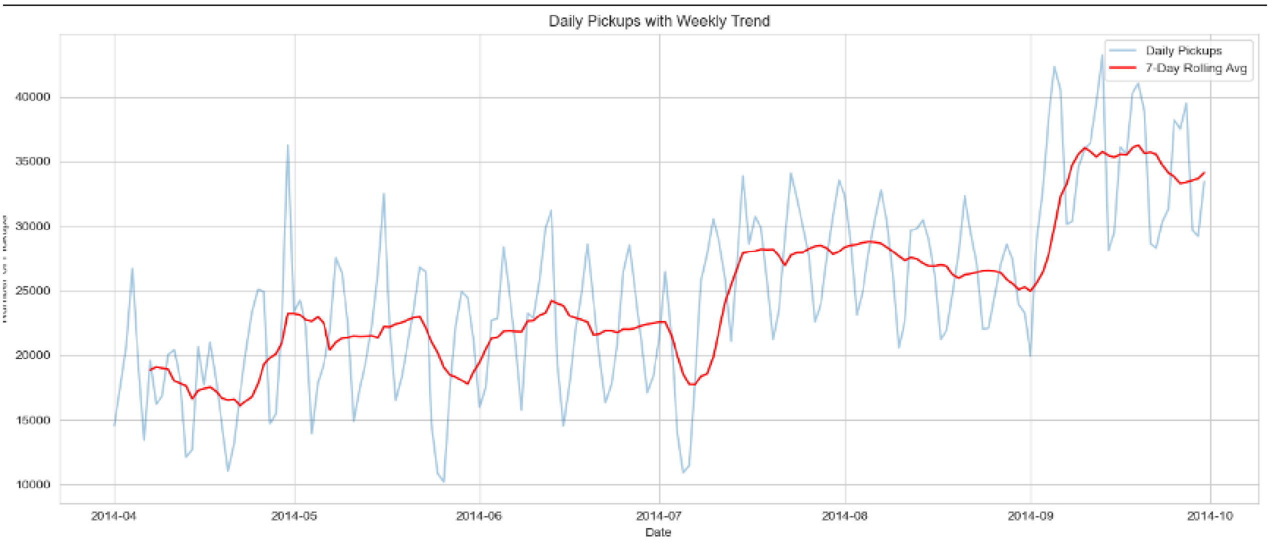

Figure 5.5: Daily Pickups with Weekly Trends

6. Uber Pickups Location Scatter

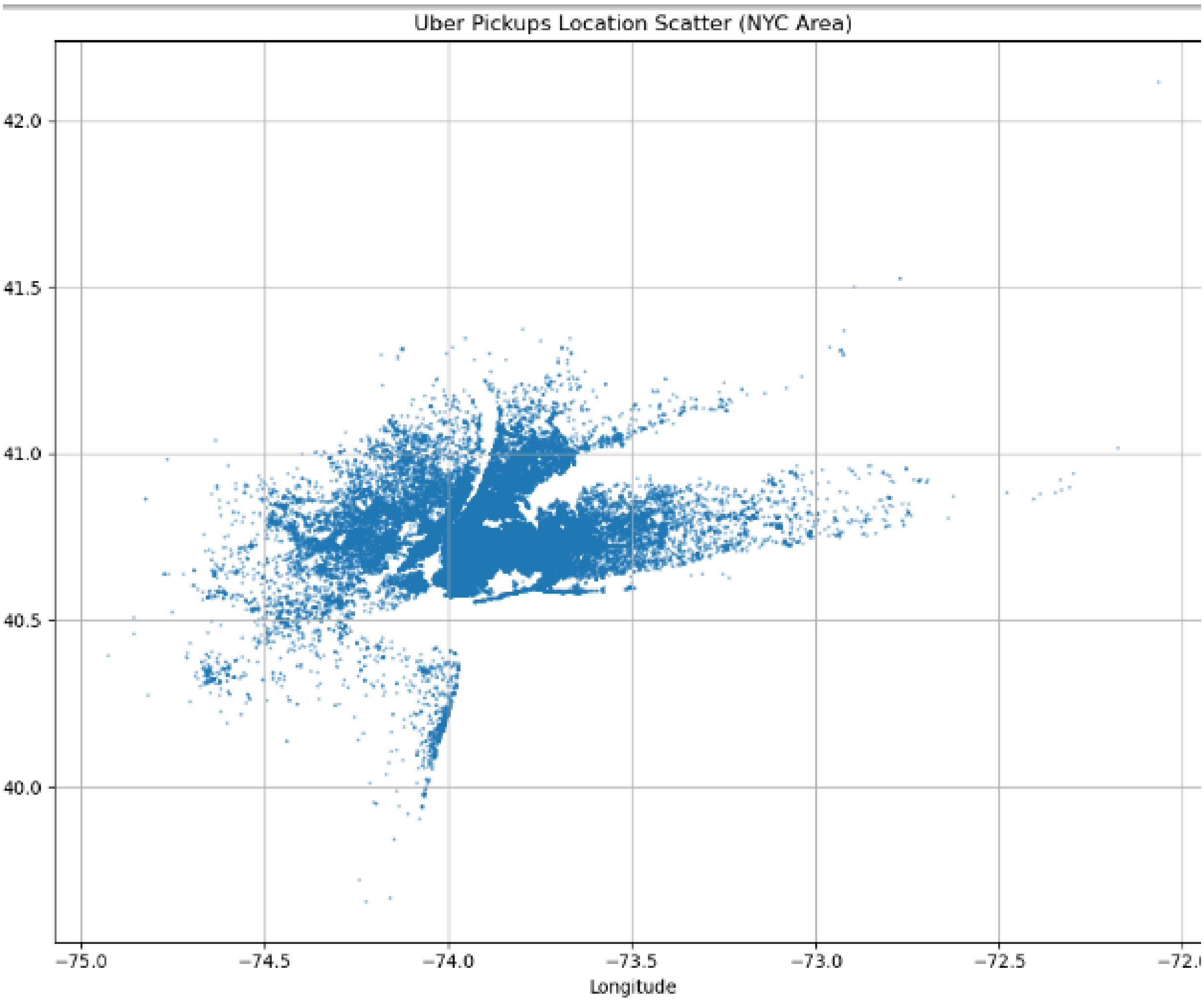

Figure 5.6: Uber Pickups Location Scatter (NYC)

7. Uber Pickups by Base

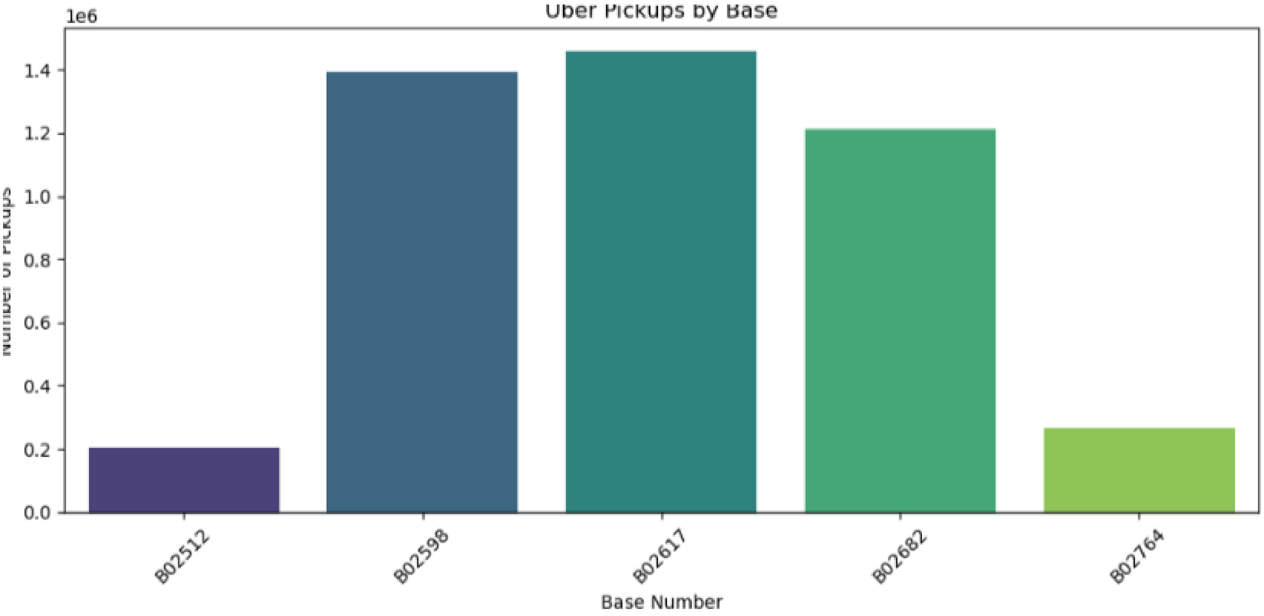

Figure 5.7: Uber Pickups by Base

8. Uber Pickups Density Heatmap

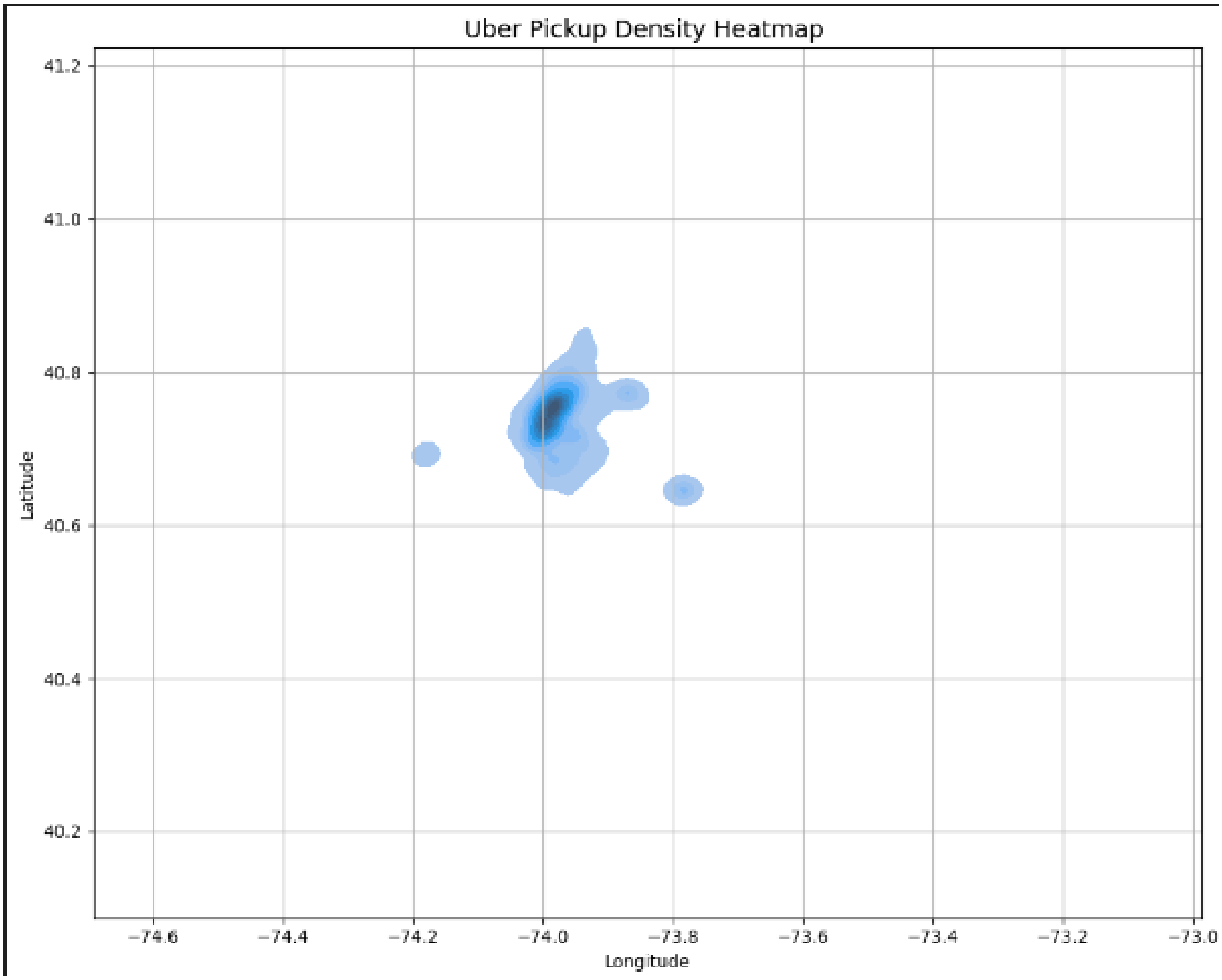

Figure 5.8: Uber Pickups Density Heatmap

9. Heatmap of Uber pickups by Hour and Weekday

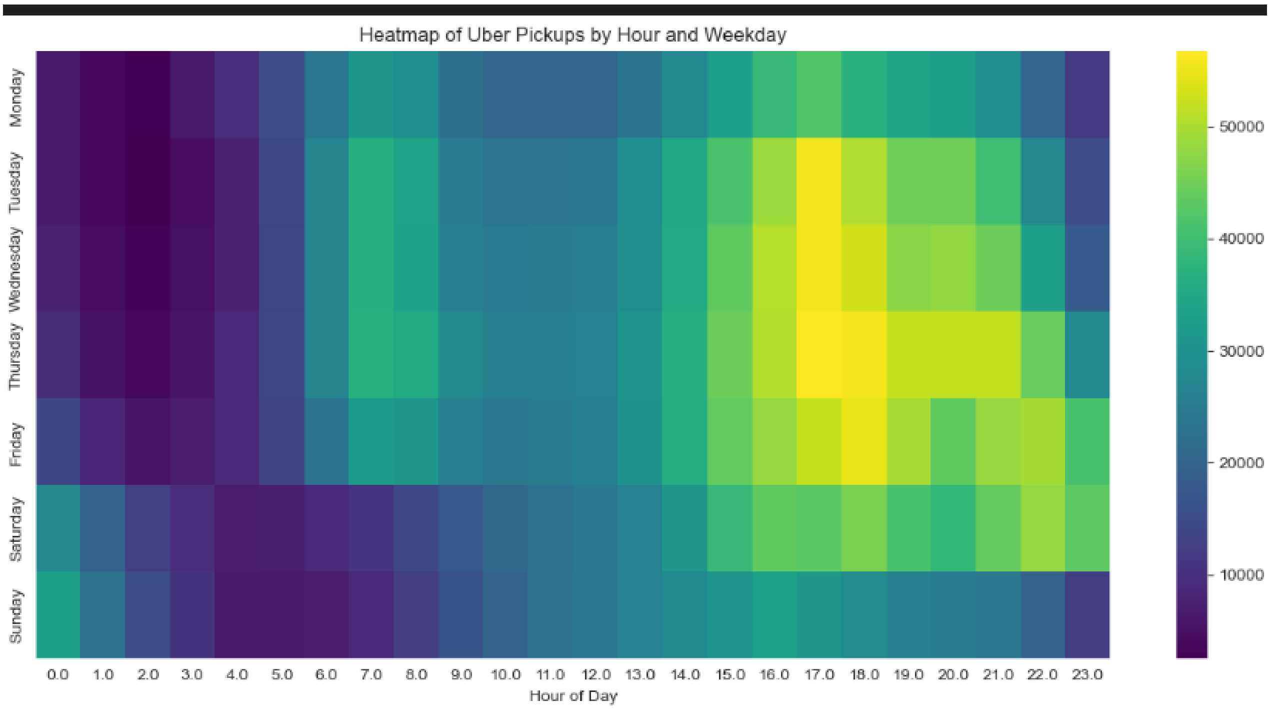

Figure 5.9: Uber Pickups Density Heatmap

Chapter 6

Key Findings

- Peak Usage Periods: Uber experiences the highest demand during the late afternoon and early night hours, indicating strong usage for evening commutes and nighttime activities.

- Busiest Days: Activity significantly increases on Fridays and Saturdays, suggesting that users frequently rely on Uber for weekend outings and social events.

- Geographic Trends: Heatmap analysis reveals that Manhattan is the primary hotspot for Uber rides, reflecting dense population, tourism, and a high concentration of businesses and entertainment venues.

Chapter 7

Conclusion

This project offers valuable insights into **urban mobility** by analyzing Uber usage patterns. The findings highlight temporal and geographic trends in ride demand, such as peak hours, busy days, and high-activity zones.

Such data-driven understanding can assist **Uber** and **local authorities** in:

- Optimizing transportation services
- Managing fleet and driver resources efficiently
- Addressing traffic congestion in hotspot areas

Overall, the analysis contributes to better planning and smarter urban transportation strategies.

References

- Uber dataset: `https://www.kaggle.com/datasets`
- Python documentation: `https://docs.python.org/3/`

Appendix: Sample Code Snippets

Listing 7.1: Loading and parsing the dataset

```python
import pandas as pd

df = pd.read_csv("uber.csv")
df[’Date/Time’] = pd.to_datetime(df[’Date/Time’])
df[’day’] = df[’Date/Time’].dt.day
df[’weekday’] = df[’Date/Time’].dt.weekday
df[’hour’] = df[’Date/Time’].dt.hour
```