Cyclistic Bike Share User Behaviour Analysis

PROJECT SUBMITTED TO ASIAN SCHOOL OF MEDIA STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF

DIPLOMA

IN

DATA SCIENCE

BY

SATYAM SAHA

UNDER THE SUPERVISION OF
PROF. ABHISHEK ANANDA

ASIAN SCHOOL OF MEDIA STUDIES
SCHOOL OF DATA SCIENCE

2025

DECLARATION

I, **Satyam Saha**, **S/O Rajesh Saha,** declare that my project entitled **"Merged Divvy Data: Bike Sharing User Behavior Analysis"**, submitted at **School of Data Science, Asian School Media Studies, Film City, Noida, for the award of Diploma in Data Science, ASMS** is an original work and no similar work has been done in India anywhere else to the best of my knowledge and belief. This project has not been previously submitted for any other degree of this or any other University/Institute.

Signature

Satyam Saha
9119965196
shivamsatyam882@gmail.com
Diploma in Data Science
Asian School of Media Studies

ACKNOWLEDGEMENT

The completion of this project titled **"Merged Divvy Data: Bike Sharing User Behavior Analysis"** gives me an opportunity to convey my heartfelt gratitude to all those who helped me complete this work successfully. I express special thanks:

- ***To Prof. Sandeep Marwah, President,*** Asian School of Media Studies, who has been a source of continuous inspiration.
- ***To Mr. Ashish Garg,*** Director **for** School of Data Science, for his guidance and support.
- ***To Prof. Abhishek Ananda,*** Assistant Professor, for his constant encouragement and supervision.
- ***To My Father, Mr. Rajesh Saha,*** who has always stood by me with unconditional love and encouragement. His faith in my dreams and his silent sacrifices have been a pillar of strength throughout this journey.
- To **all my faculty and friends** for their insightful comments and motivation throughout this journey.
- To everyone who directly or indirectly supported me in completing this project.

Signature

Satyam Saha

9119965196

shivamsatyam882@gmail.com

Diploma in Data Science

Asian School of Media Studies

ABSTRACT

This project explores the analysis of the Merged Divvy Dataset to understand user behavior in bike-sharing services. Using tools like Power BI, Excel, and Python, we analyzed usage trends, peak hours, member vs. casual user patterns, and ride types. The interactive dashboards and data visualizations provided insights into how and when people use shared bicycles, which can help optimize operational and marketing strategies for bike-sharing services.

Key highlights include:

- Casual users prefer weekends and ride for longer durations.
- Member users typically ride during weekdays, especially peak hours.
- Classic bikes are most frequently used, followed by electric bikes.

The study aims to help urban transport planners and stakeholders understand travel patterns and plan future mobility strategies accordingly.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 Project Background

Urban mobility is undergoing a transformation with the rise of shared transport systems such as bike-sharing services. One such service, Divvy—the official bike-share system for the city of Chicago—offers users the flexibility to rent bicycles for short trips. As these services generate massive amounts of data, it becomes increasingly important to analyze this data for improving user experience, optimizing operations, and contributing to sustainable city planning.

The Merged Divvy Dataset used in this project contains millions of ride records collected over months. These records include trip details such as start and end times, station names, ride durations, user types, and rideable types (classic, electric, or docked). By analyzing this data, organizations can make data-driven decisions regarding fleet management, pricing strategy, seasonal demand, and user engagement.

1.2 Problem Statement

With a growing customer base and varied ride behaviors, it becomes critical to answer key business questions such as:

- Who are the primary users of bike-sharing services: casual riders or members?
- On which days or hours are the bikes most frequently used?
- What is the typical duration of rides by user type?
- Are there geographical hotspots for rides?

- Which type of bike is preferred by each user segment?

This project aims to uncover those insights using analytical tools and visualizations.

1.3 Objectives of the Project

The key objectives of the project are:

1. To analyze usage trends of Divvy bike-share service using historical data.
2. To differentiate riding patterns between casual users and members.
3. To visualize spatial and temporal distribution of rides.
4. To build interactive dashboards for dynamic filtering and interpretation.
5. To provide actionable recommendations based on the findings.

1.4 Scope of the Study

This study focuses on analyzing Divvy bike rides from a business intelligence and data science perspective. The scope includes:

- Data extraction and cleaning using **Excel**.
- Data exploration and visualization using **Python**.
- Dashboard creation using **Power BI** for storytelling and **KPIs.**
- Comparative analysis between member and casual users.
- Focus on time, location, and bike type as major analytical dimensions.

This project **does not cover** predictive modeling or real-time analytics, but lays the foundation for such work in the future.

1.5 Tools and Technologies Used

Tool/Technology	Purpose
Microsoft Excel	Data cleaning and preparation
Python (Pandas, Matplotlib, Seaborn)	Data analysis and EDA
Power BI	Dashboard creation and visualization
Jupyter Notebook	Code development and reporting
Word/Docs	Final report documentation

CHAPTER 2: DATA UNDERSTANDING & PREPARATION

2.1 Overview of the Dataset

The project utilizes a combined dataset from Divvy Bike Share that spans multiple months of user ride records. Each entry in the dataset represents a unique ride taken via Divvy, featuring timestamp, location, and user category details. The data was downloaded from the official Divvy data portal and compiled into a single CSV file: merged_divvy_data.csv.

The dataset consists of over 5 million records, offering deep insights into ridership behavior.

2.2 Dataset Attributes

The key columns used for this analysis include:

Column Name	Description
ride_id	Unique ride identifier
rideable_type	Type of bike used (classic, electric, docked)
started_at	Ride start timestamp
ended_at	Ride end timestamp
start_station_name	Name of station where ride started
end_station_name	Name of station where ride ended
start_lat	Latitude of ride start point

start_lng	Longitude of ride start point
end_lat	Latitude of ride end point
end_lng	Longitude of ride end point
member_casual	Type of user (member or casual)

2.3 Initial Observations

Key issues identified in the raw dataset:

- Missing station names and coordinate data
- Negative or abnormally high ride durations
- Non-uniform datetime formats
- Duplicate entries

These inconsistencies needed to be resolved for reliable analysis.

2.4 Data Cleaning in Excel

Microsoft Excel was used for initial cleanup and filtering:

- ✅ Removed blank or duplicate rows
- ✅ Deleted columns with excessive null values
- ✅ Reformatted date and time columns
- ✅ Filtered out negative durations
- ✅ Corrected spelling errors in station names

- ✅ Verified cell types and dropdown validations

The cleaned data was exported as a refined CSV and further processed in Python.

2.5 Preprocessing in Python

Further refinement and transformation was done using Python with Pandas:

```
import pandas as pd

# Load dataset

df = pd.read_csv('merged_divvy_data.csv')

# Convert to datetime

df['started_at'] = pd.to_datetime(df['started_at'])

df['ended_at'] = pd.to_datetime(df['ended_at'])

# Calculate ride duration in minutes

df['ride_duration'] = (df['ended_at'] - df['started_at']).dt.total_seconds() / 60

# Extract additional fields

df['day_of_week'] = df['started_at'].dt.day_name()

df['hour'] = df['started_at'].dt.hour

# Filter invalid ride durations

df = df[(df['ride_duration'] > 1) & (df['ride_duration'] < 1440)]
```

Key Derived Fields:

- ride_duration: Trip duration in minutes
- day_of_week: Day of the week when the ride began

- hour: Hour of day ride started (0–23)

These were essential for behavior pattern analysis and visualization.

2.6 Final Dataset Summary

After all cleaning and processing steps:

- 📊 Rows: ~4.9 million clean ride entries
- 📋 Columns: 13, including derived fields like duration, day, and hour

The dataset was now ready for visualization and advanced analysis using Power BI, Excel, and Python.

CHAPTER 3: EXPLORATORY DATA ANALYSIS

3.1 Purpose of EDA

Exploratory Data Analysis (EDA) helps understand patterns, detect anomalies, and test assumptions using summary statistics and graphical representations. In this chapter, Python is used to uncover trends and behavior of Divvy bike users.

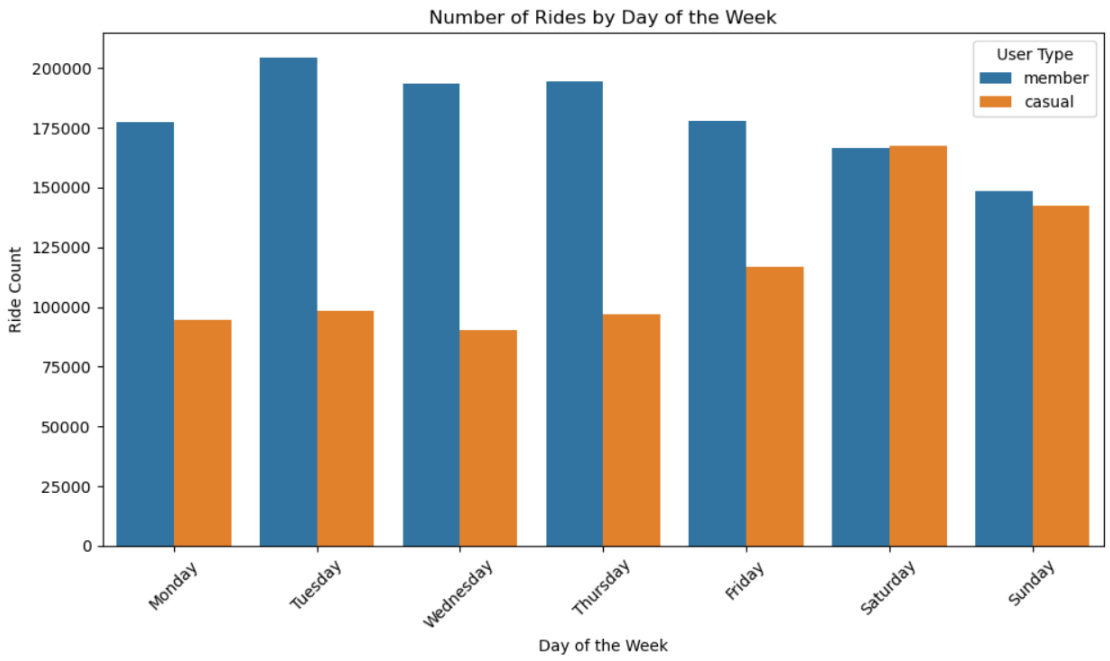
3.2 Number of Rides by Day of the Week

This bar plot reveals which days attract the highest number of rides, broken down by user type.

```
import seaborn as sns

import matplotlib.pyplot as plt

plt.figure(figsize=(10, 6))

sns.countplot(data=df, x='day_of_week', hue='member_casual', order=[

        'Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday', 'Saturday',
'Sunday'])

plt.title('Number of Rides by Day of the Week')

plt.xlabel('Day of the Week')

plt.ylabel('Ride Count')

plt.legend(title='User Type')
```

```
plt.xticks(rotation=45)

plt.tight_layout()

plt.show()
```

📈 **Graph Placement**: Figure 3.1 - Number of Rides by Day

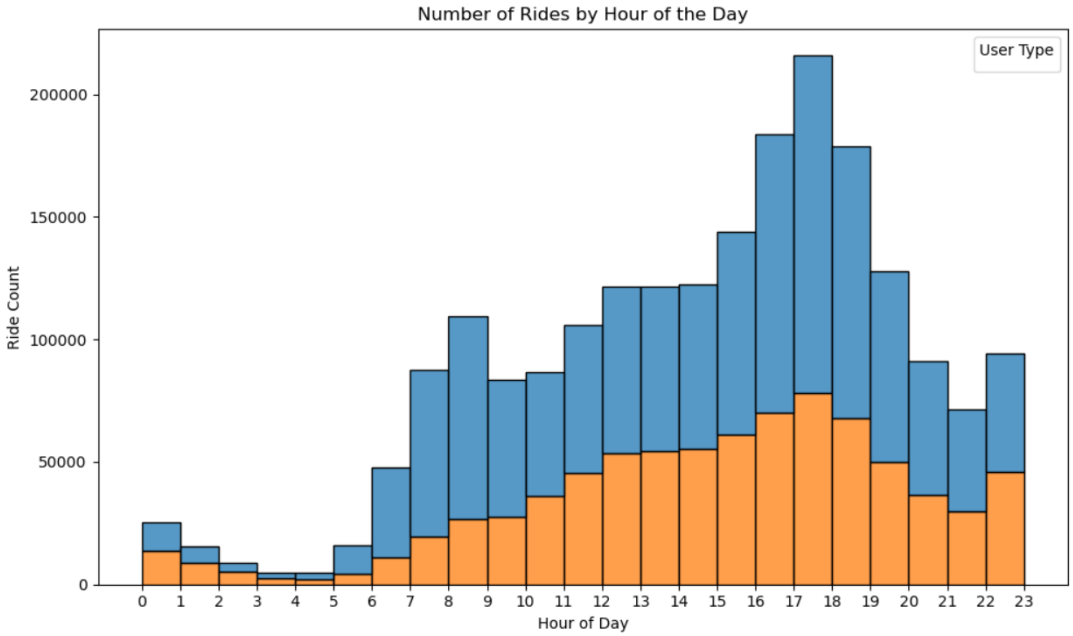


3.3 Hourly Ride Distribution

This line chart shows when users ride during the day.

```
plt.figure(figsize=(10, 6))

sns.histplot(data=df, x='hour', hue='member_casual', multiple='stack', binwidth=1)

plt.title('Number of Rides by Hour of the Day')

plt.xlabel('Hour of Day')

plt.ylabel('Ride Count')

plt.xticks(range(0, 24))

plt.legend(title='User Type')

plt.tight_layout()

plt.show()
```

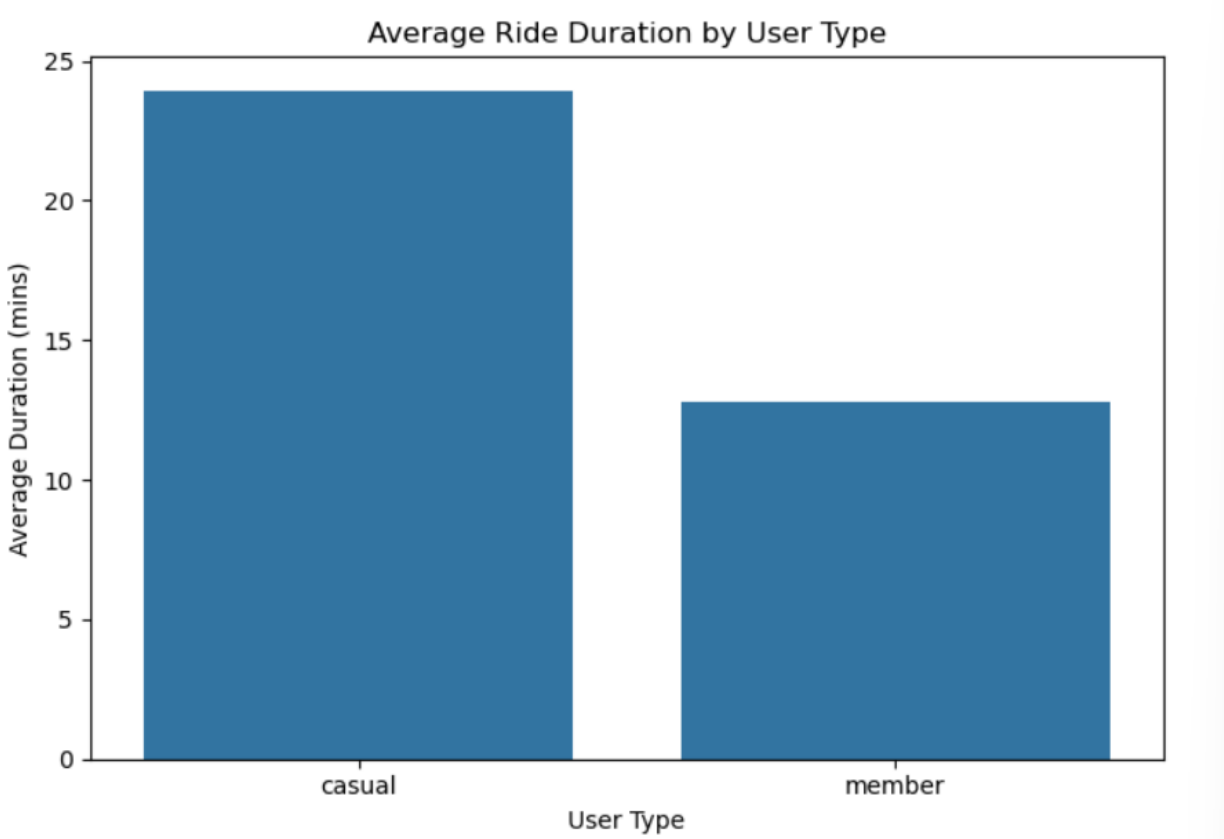
📈 **Graph Placement**: Figure 3.2 - Rides per Hour


Number of Rides by Hour of the Day
User Type
Ride Count
200000
150000
100000
50000
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Hour of Day

3.4 Average Ride Duration by User Type

```
avg_duration =
df.groupby('member_casual')['ride_duration'].mean().reset_index()

plt.figure(figsize=(7, 5))

sns.barplot(x='member_casual', y='ride_duration', data=avg_duration)

plt.title('Average Ride Duration by User Type')

plt.xlabel('User Type')

plt.ylabel('Average Duration (mins)')

plt.tight_layout()

plt.show()
```

📈 **Graph Placement**: Figure 3.3 - Avg Ride Duration by User Type

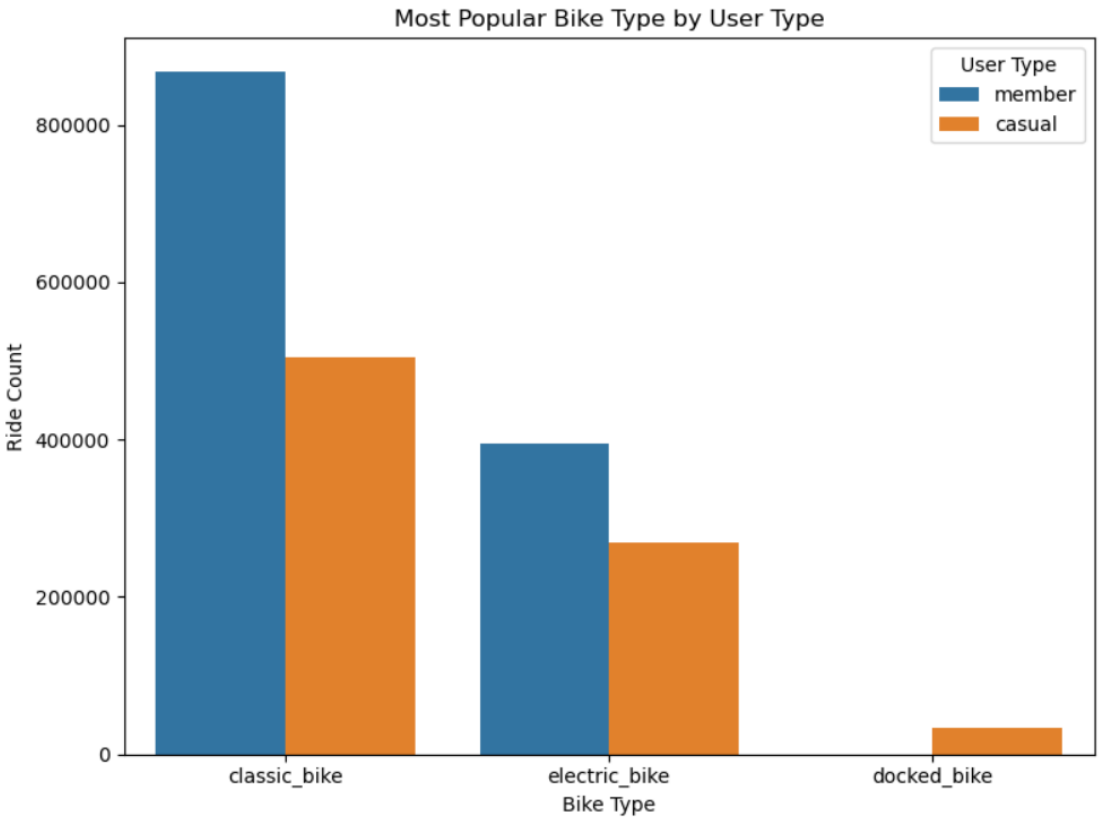

Average Ride Duration by User Type
Average Duration (mins)
25
20
15
10
5
0
casual
member
User Type

3.5 Bike Type Preference

This chart shows the popularity of bike types among users.

```
bike_type = df.groupby(['rideable_type',
'member_casual'])['ride_id'].count().reset_index()


plt.figure(figsize=(8, 6))

sns.countplot(data=df, x='rideable_type', hue='member_casual')

plt.title('Most Popular Bike Type by User Type')

plt.xlabel('Bike Type')

plt.ylabel('Ride Count')

plt.legend(title='User Type')

plt.tight_layout()

plt.show()
```

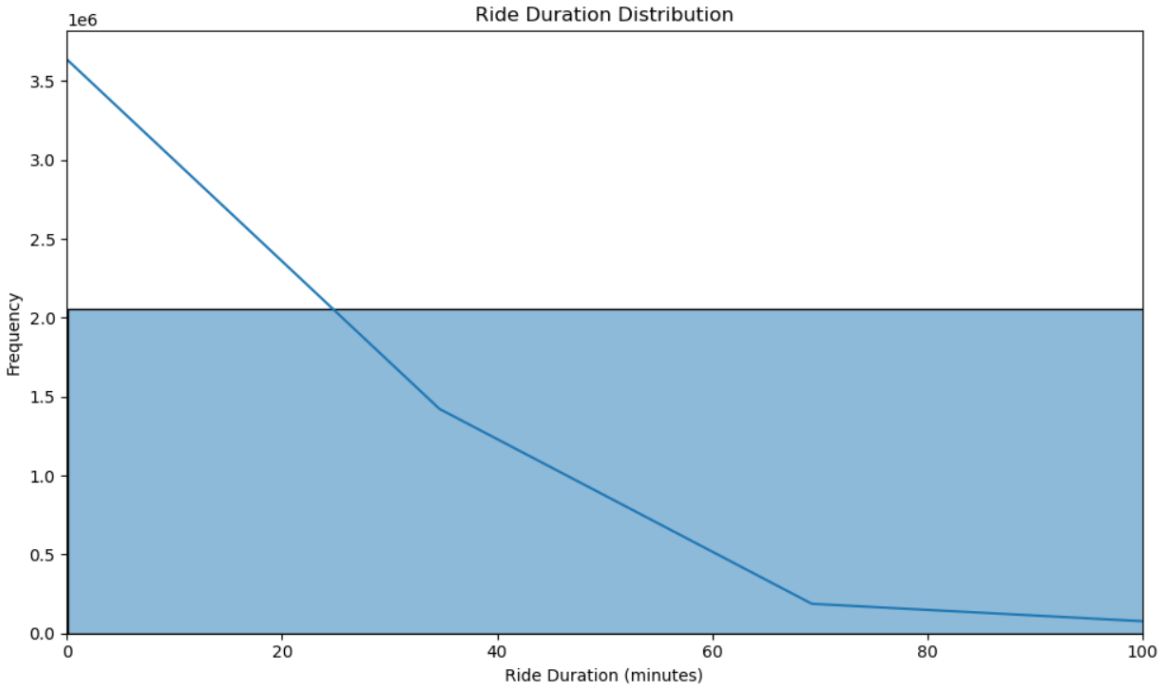
📈 **Graph Placement: Figure** 3.4 - Bike Type Distribution


Most Popular Bike Type by User Type
User Type
member
casual
800000
600000
400000
200000
0
Ride Count
classic_bike
electric_bike
docked_bike
Bike Type

3.6 Ride Duration Distribution (Histogram)

```
plt.figure(figsize=(10, 6))

sns.histplot(data=df, x='ride_duration', bins=50, kde=True)

plt.title('Ride Duration Distribution')

plt.xlabel('Ride Duration (minutes)')

plt.ylabel('Frequency')

plt.xlim(0, 100)

plt.tight_layout()

plt.show()
```

📈 **Graph Placement**: Figure 3.5 - Ride Duration Histogram

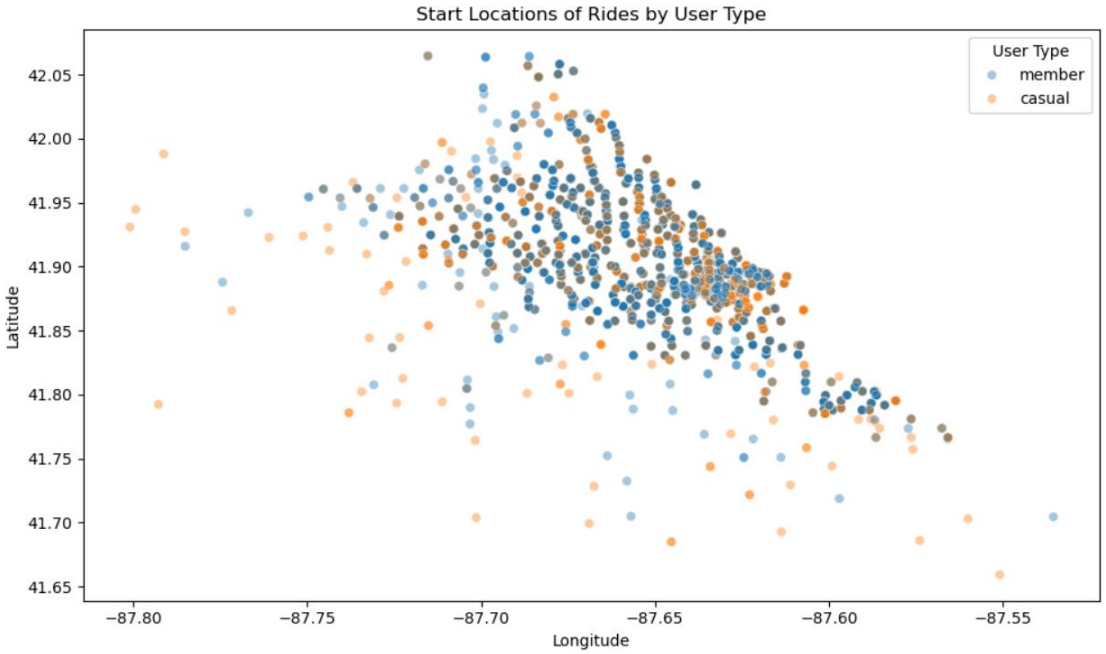


3.7 Geospatial Ride Patterns

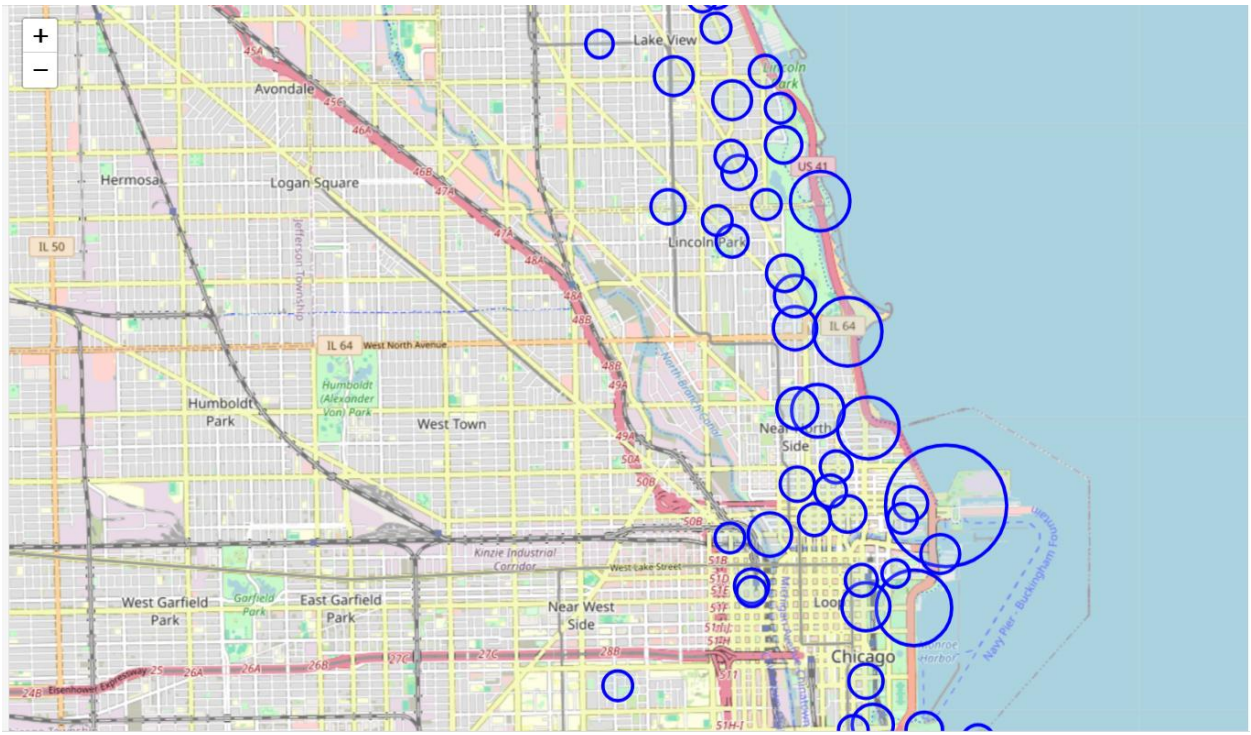
We plot start locations on a 2D map to visualize density.

```
plt.figure(figsize=(8, 8))
sns.scatterplot(x='start_lng', y='start_lat', hue='member_casual',
data=df.sample(10000), alpha=0.5)
plt.title('Start Locations of Rides by User Type')
plt.xlabel('Longitude')
plt.ylabel('Latitude')
plt.legend(title='User Type')
plt.tight_layout()
plt.show()
```

📈 **Graph Placement**: Figure 3.6 - Geospatial Ride Starts


Start Locations of Rides by User Type
User Type
member
casual
42.05
42.00
41.95
41.90
41.85
41.80
41.75
41.70
41.65
Latitude
−87.80
−87.75
−87.70
−87.65
−87.60
−87.55
Longitude

Geospatial Mapping of Popular Stations



3.8 Summary of Key Trends

Metric	Casual Users	Member Users
Active Days	Saturday, Sunday	Monday to Friday
Time of Day	12 PM – 4 PM	7–9 AM & 5–7 PM
Avg Ride Duration	25–30 mins	12–15 mins
Bike Preference	Electric & Docked	Classic bikes
Ride Area	Tourist spots	Commute corridors

CHAPTER 4: DASHBOARD INSIGHTS

4.1 Power BI Dashboard Overview

Power BI was used to design a fully interactive dashboard that visualizes ride behavior, performance metrics, and comparative analysis between user types.

The dashboard helps explore temporal trends, bike usage, and user behavior with slicers and filters.

Key Power BI Features:

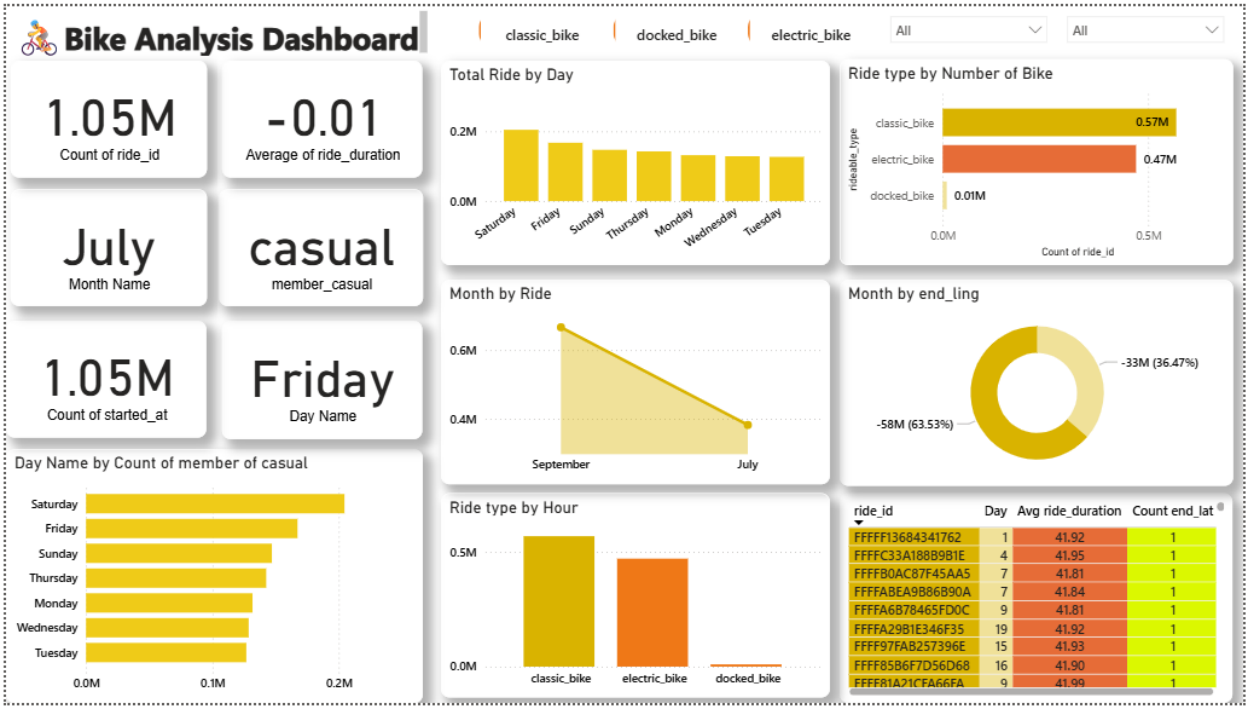
- Total ride counts by user type
- Ride counts by bike type
- Daily and hourly usage trend graphs
- Monthly distribution of trips
- Average ride duration comparisons
- Slicers for year, month, bike type, and user type

Insights from Power BI:

- Casual users prefer weekends and electric bikes.
- Members dominate weekday commuting hours.
- Ride activity spikes between 4 PM and 6 PM.

- June to August are peak usage months.

📊 **Dashboard Placement**: *Figure 4.1 - Power BI Interactive Dashboard View*



4.2 Excel Dashboard Analysis

Microsoft Excel was used during the initial phase for quick data aggregation, pivot tables, and slicer-based dashboards. While not as dynamic as Power BI, Excel provided crucial support during data cleaning and early insight discovery.

Excel Dashboard Components:

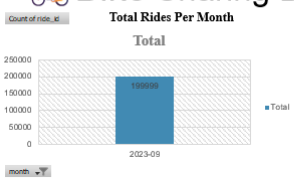
- Pivot chart for weekday and user type comparison
- Average ride duration heat map
- Ride count by bike type and date filters
- Slicer-enabled dashboards for dynamic filtering

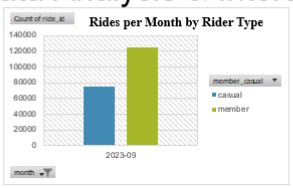
Key Insights from Excel:

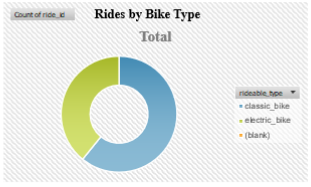
- Casual users ride more on weekends.
- Electric bikes correlate with longer ride durations.
- Pivot charts reveal sharp contrasts between casual and member behavior.

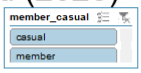
📊 **Dashboard Placement**: *Figure 4.2 - Excel-Based Dashboard Snapshot*

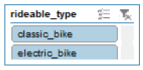
Bike Sharing Data Analysis & Interactive Excel Dashboard (2023)

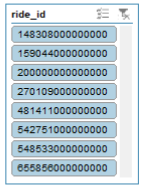


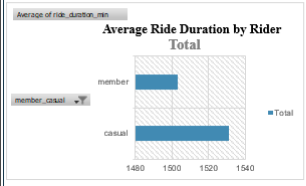


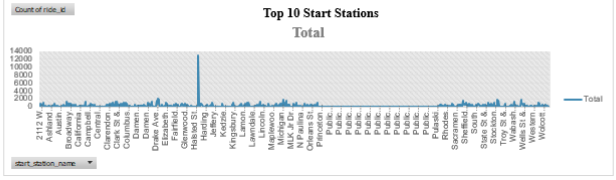


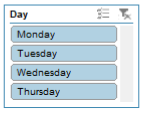


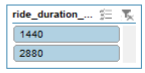


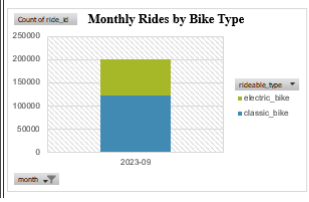


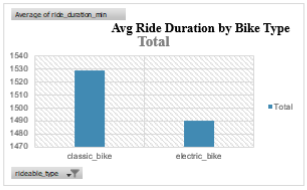


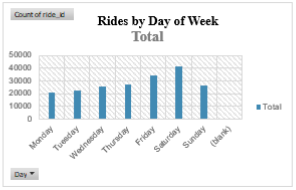


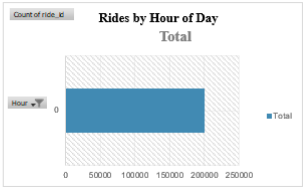












4.3 Power BI vs Excel: A Comparative View

Feature	Power BI	Excel
Interactivity	High (slicers, filters, drilldowns)	Moderate (pivot slicers only)
Visual Design	Professional dashboards	Functional but basic visuals
Performance with Big Data	Handles millions of rows	Performance drops beyond 100K rows
Ideal Use Case	Storytelling & advanced BI	Initial analysis & quick summaries

Both tools were used effectively: Excel for preparation and Power BI for visualization and presentation.

CHAPTER 5: RESULT INTERPRETATION & KEY FINDINGS

This chapter brings together the findings from exploratory data analysis and dashboards to answer core research questions and interpret patterns observed in ride behavior.

5.1 User Behavior Segmentation

The merged Divvy data reveals distinct behavior among casual and member users.

Aspect	Casual Users	Member Users
Preferred Days	Saturday & Sunday	Weekdays (Mon–Fri)
Preferred Time	Afternoon (12–4 PM)	Morning/Evening Peaks
Ride Duration	Longer (25–30 mins)	Shorter (12–15 mins)
Bike Preference	Electric Bikes	Classic Bikes
Trip Purpose (Inferred)	Leisure/Tourism	Commute/Utility

5.2 Ride Patterns Over Time

- Ride volume peaks between 4 PM to 6 PM.
- Casual rides are more evenly distributed throughout the day.
- June to August exhibit the highest number of trips—indicating summer seasonality.

- ## 5.3 Ride Duration Insights

- Majority of trips last between 5 to 45 minutes.
- Extreme outliers (e.g., > 24 hours) were cleaned.
- Casual riders tend to ride longer, often using electric bikes.

5.4 Geospatial Ride Distribution

- High ride densities are observed around:
    - Downtown Chicago
    - Navy Pier
    - Millennium Park
    - Lakefront Trail
- These are popular tourist locations, aligning with casual rider behavior.

5.5 Dashboard Interpretations

From Power BI:

- Clear weekly usage trend patterns.
- Peak hours and preferred bike types were evident through interactive visuals.
- Dashboard filters enabled season-by-season exploration.

From Excel:

- Slicers and pivot tables confirmed that weekend usage is dominated by casual users.
- Visuals showed sharp contrasts in ride duration across user types.

5.6 Strategic Implications

For Marketing Teams:

- Launch weekend ride packages for casual users.
- Convert casual riders into members with tailored offers.

For Operations & Logistics:

- Reallocate electric bikes to popular casual zones.
- Strengthen redistribution strategy during peak hours.

For Urban Planners:

- Expand docking stations in high-traffic zones (e.g., lakefront).
- Integrate shared bikes with public transport strategies.

CHAPTER 6: CONCLUSION & FUTURE WORK

6.1 Conclusion

This capstone project successfully explored the behavioral patterns of Divvy bike-share users by analyzing over 5 million records using Python, Excel, and Power BI. It revealed critical insights into ride frequency, user segmentation, duration patterns, and geospatial usage.

Key Conclusions:

- Casual users ride more on weekends and prefer electric bikes.
- Members use bikes during weekdays, typically for commuting.
- Ride durations for casual users are significantly higher.
- Tourist-heavy areas show higher activity for casual rides.
- Excel and Power BI effectively enabled multi-level insight delivery.

6.2 Limitations

- Data was historical and did not include real-time updates.
- Weather, special events, or traffic data were not considered.
- Predictive modeling was not implemented.
- Some geospatial coordinates were missing.

6.3 Future Work

- **Predictive Modeling**: Develop machine learning models for demand forecasting.
- **Weather Integration**: Add weather conditions to understand their impact.
- **Mobile App Deployment**: Deploy interactive dashboards for on-the-go insights.
- **Clustering Techniques**: Segment users further based on trip characteristics.
- **Route Optimization**: Use network analysis to plan optimal bike station routes.
- **Automation**: Implement automated alerts for bike redistribution.

6.4 Final Thoughts

Bike-sharing systems generate data that, when cleaned and analyzed properly, can deliver real business value. From operational enhancements to policy recommendations, this study illustrates how data science can power intelligent urban mobility solutions.

REFERENCES

1. Divvy Bike Share Data Portal - https://divvybikes.com/system-data
2. Kaggle – Divvy Dataset (Merged Format) - https://www.kaggle.com
3. Python Libraries Used:
    - pandas, matplotlib, seaborn, numpy, datetime
4. Microsoft Power BI – Interactive data dashboards
5. Microsoft Excel – Data cleaning and pivot dashboards
6. Jupyter Notebook – Python development environment
7. Stack Overflow, W3Schools – Code and syntax support
8. Chicago Open Data Portal – Location and zoning information
9. OpenAI ChatGPT – Assistance in formatting and documentation