Analyzing Uber Data Using GCP, Python, and Looker Studio

Akshit Anand

1. Introduction

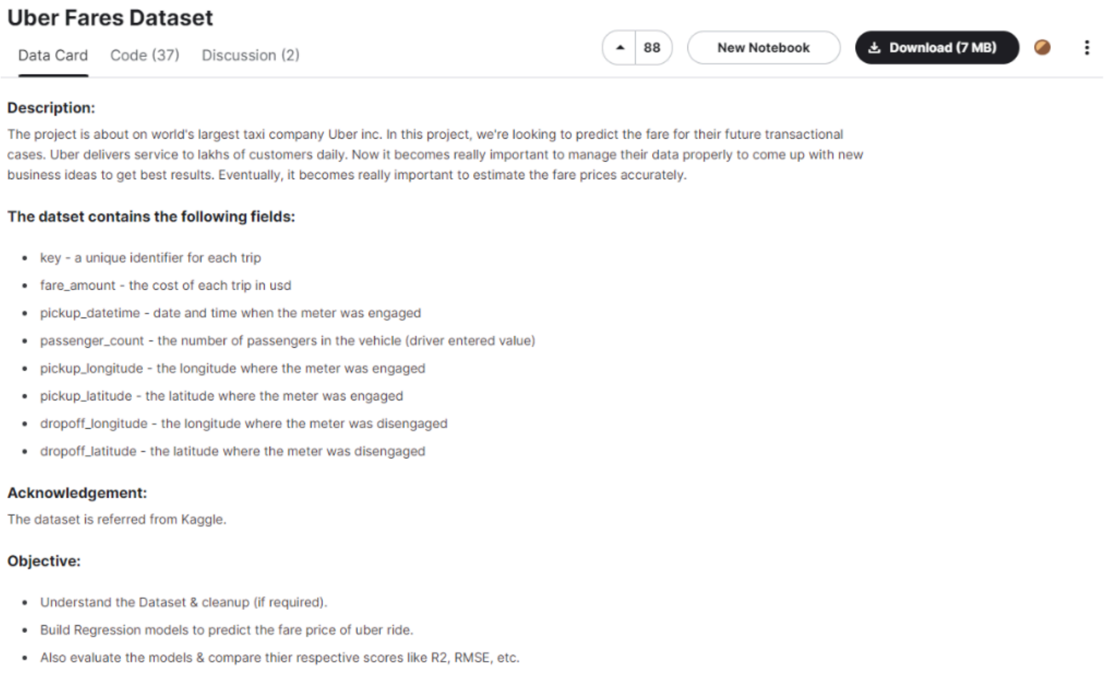
1.1 Objective

The objective of this project was to utilize cloud technology and data analytics to analyze Uber's public datasets. By leveraging Google Cloud Platform (GCP), Python, Mage for ETL, BigQuery, and Looker Studio, the project aimed to uncover patterns and insights in Uber data that could support strategic decision-making.

1.2 Data Overview

The Uber dataset consists of millions of Uber pickups across New York City over a specific period. Key features include:

- **Date/Time of Pickup**: Indicates when each ride occurred.
- **Location of Pickup**: Provides coordinates for pickup locations.
- **Ride Details**: Additional attributes relevant to each trip.



Description:

The project is about on world's largest taxi company Uber inc. In this project, we're looking to predict the fare for their future transactional cases. Uber delivers service to lakhs of customers daily. Now it becomes really important to manage their data properly to come up with new business ideas to get best results. Eventually, it becomes really important to estimate the fare prices accurately.

The datset contains the following fields:

- key - a unique identifier for each trip
- fare_amount - the cost of each trip in usd
- pickup_datetime - date and time when the meter was engaged
- passenger_count - the number of passengers in the vehicle (driver entered value)
- pickup_longitude - the longitude where the meter was engaged
- pickup_latitude - the latitude where the meter was engaged
- dropoff_longitude - the longitude where the meter was disengaged
- dropoff_latitude - the latitude where the meter was disengaged

Acknowledgement:

The dataset is referred from Kaggle.

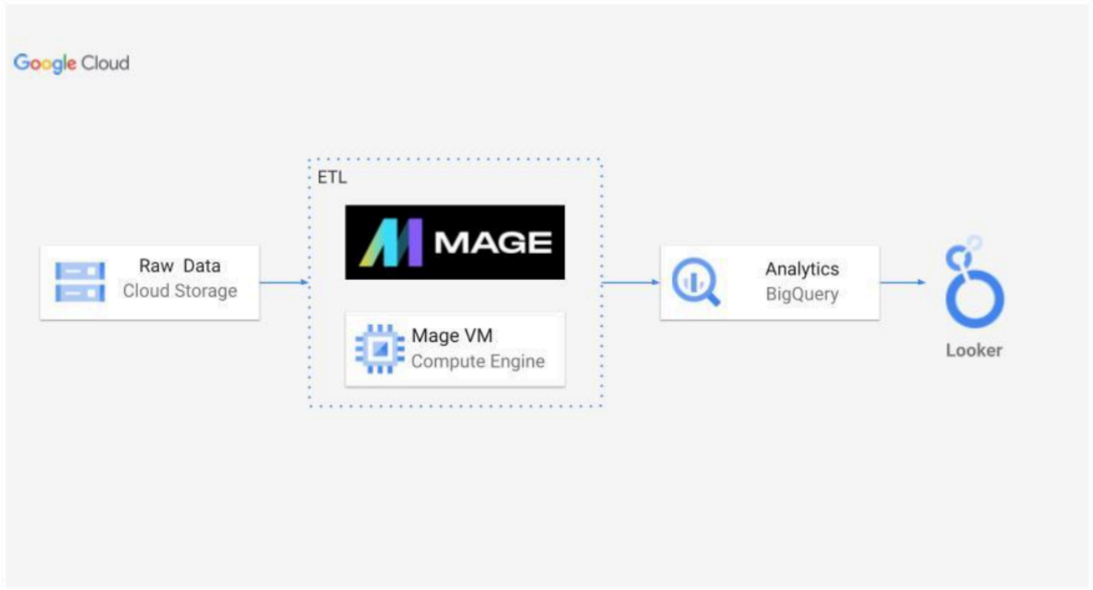
Objective:

- Understand the Dataset & cleanup (if required).
- Build Regression models to predict the fare price of uber ride.
- Also evaluate the models & compare thier respective scores like R2, RMSE, etc.

1.3 Technologies Used

This project used several cloud-based tools and technologies:

- **Google Cloud Storage**: For durable and scalable storage of the large dataset.
- **Python**: For data manipulation, pre-processing, and initial analysis.
- **Mage (ETL Tool)**: To extract, transform, and load data into BigQuery.
- **BigQuery**: For data warehousing, SQL querying, and storage.
- **Looker Studio**: To visualize data insights interactively.



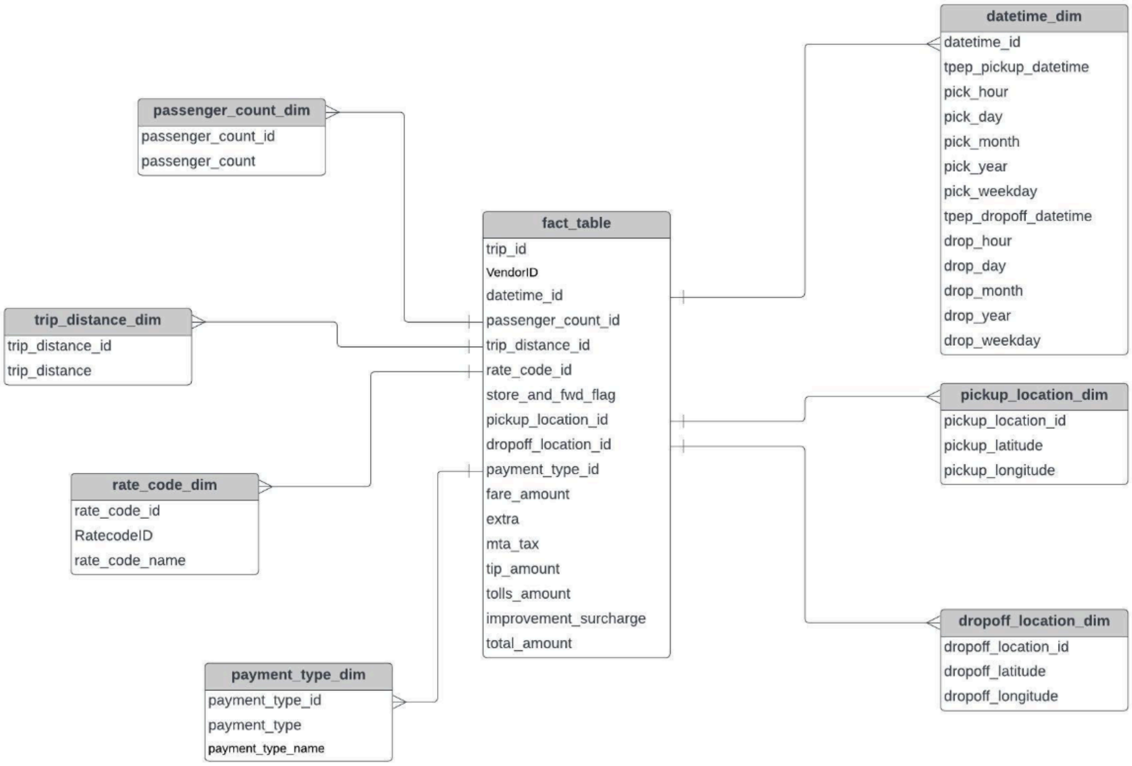
2. Data Modeling

2.1 Schema Design

A **star schema** was designed to structure the data efficiently, facilitating fast query performance:

- **Fact Table**: Contains core information about each ride, including timestamps, pickup locations, and trip details.
- **Dimension Tables**: Separate tables for time, location, and ride characteristics, which allow for easier filtering and aggregations.
- **Screenshot Suggestion**: Include a **screenshot of the schema diagram** (e.g., from Lucidchart or another modeling tool) here. Place it below this section to visually depict

the data structure used in the project.

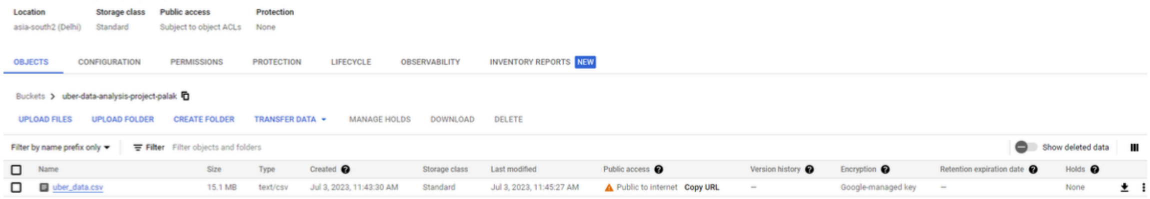


3. Data Ingestion and Storage

3.1 Uploading Data to Google Cloud Storage

The initial dataset (in CSV format) was uploaded to **Google Cloud Storage**. This storage service provides reliability, durability, and scalability, making it suitable for handling large datasets efficiently.

- **Screenshot Suggestion**: Add a **screenshot showing the successful upload of the data to Google Cloud Storage**. This screenshot should display the dataset in GCP, confirming its availability for further processing.



Location	Storage class	Public access	Protection
asia-south2 (Delhi)	Standard	Subject to object ACLs	None

Buckets > uber-data-analysis-project-palak

☐	Name	Size	Type	Created	Storage class	Last modified	Public access	Version history	Encryption	Retention expiration date	Holds
☐	uber_data.csv	15.1 MB	text/csv	Jul 3, 2023, 11:43:30 AM	Standard	Jul 3, 2023, 11:45:27 AM	Public to internet	—	Google-managed key	—	None

A screenshot that the data is successfully uploaded to GCP storage.

4. ETL Pipeline

4.1 Deploying Mage on Google Compute Engine

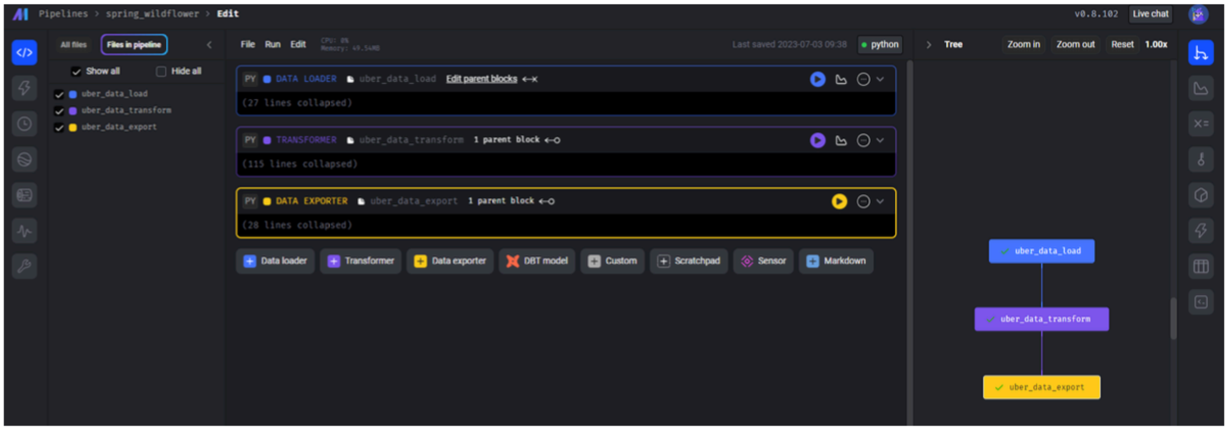
Mage, an ETL tool, was deployed on **Google Compute Engine** to handle data extraction, transformation, and loading processes. This setup allowed us to efficiently manage large data volumes and automate ETL tasks.

4.2 ETL Process Using Mage

Mage was configured to:

1. **Extract** the data from Google Cloud Storage.
2. **Transform** the data according to the pre-designed schema, including data cleansing steps.
3. **Load** the transformed data into BigQuery for further analysis.

- **Screenshot Suggestion**: Include a **screenshot of Mage deployed on Compute Engine** or a **screenshot of an ETL job in progress** to show the data pipeline setup.



5. Data Cleansing in BigQuery

5.1 Data Import to BigQuery

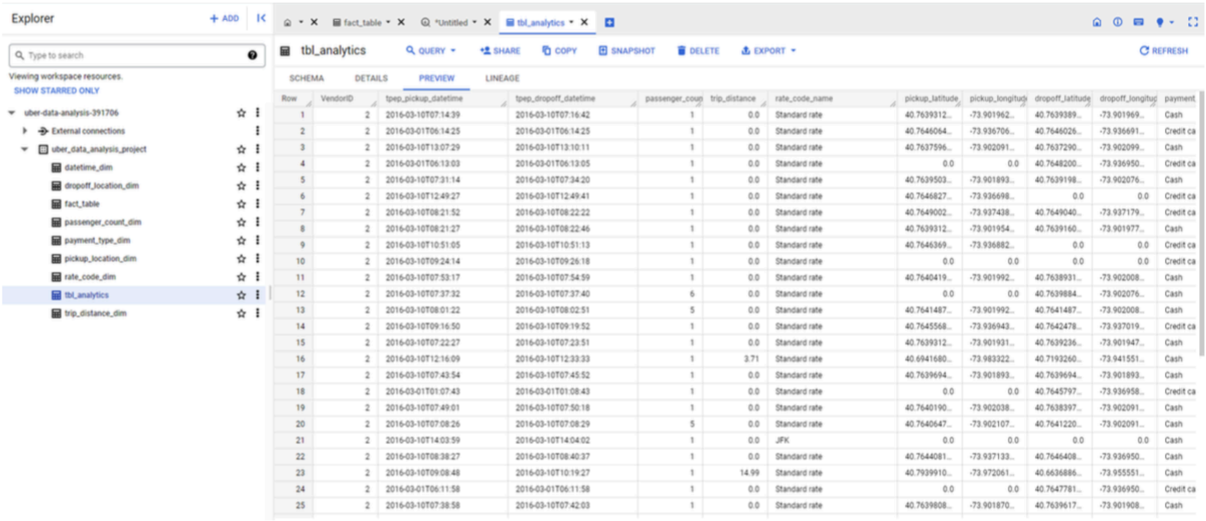
Once the ETL process was complete, the data was imported into BigQuery. This allowed us to utilize SQL-based querying to manage and analyze the data quickly and effectively.

5.2 Data Cleansing Operations

We performed several cleansing operations in BigQuery to ensure data quality:

- **Handling Missing Values**: Rows with missing data were removed or imputed to ensure complete records.
- **Removing Duplicates**: Duplicated rows were identified and removed to prevent double-counting.

- **Standardizing Formats**: Consistent formatting was applied to ensure data uniformity across all entries.



Row	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_coun	trip_distance	rate_code_name	pickup_latitude	pickup_longitud	dropoff_latitude	dropoff_longitud	payment
1	2	2016-03-10T07:14:39	2016-03-10T07:16:42	1	0.0	Standard rate	40.7639312…	-73.901962…	40.7639389…	-73.901969…	Cash
2	2	2016-03-01T06:14:25	2016-03-01T06:14:25	1	0.0	Standard rate	40.7646064…	-73.936706…	40.7646026…	-73.936691…	Credit ca
3	2	2016-03-10T13:07:29	2016-03-10T13:10:11	1	0.0	Standard rate	40.7637596…	-73.902091…	40.7637290…	-73.902099…	Cash
4	2	2016-03-01T06:13:03	2016-03-01T06:13:05	1	0.0	Standard rate	0.0	0.0	40.7648200…	-73.936950…	Credit ca
5	2	2016-03-10T07:31:14	2016-03-10T07:34:20	1	0.0	Standard rate	40.7639503…	-73.901893…	40.7639198…	-73.902076…	Cash
6	2	2016-03-10T12:49:27	2016-03-10T12:49:41	1	0.0	Standard rate	40.7646827…	-73.936698…	0.0	0.0	Credit ca
7	2	2016-03-10T08:21:52	2016-03-10T08:22:22	1	0.0	Standard rate	40.7649002…	-73.937438…	40.7649040…	-73.937179…	Credit ca
8	2	2016-03-10T08:21:27	2016-03-10T08:22:46	1	0.0	Standard rate	40.7639312…	-73.901954…	40.7639160…	-73.901977…	Cash
9	2	2016-03-10T10:51:05	2016-03-10T10:51:13	1	0.0	Standard rate	40.7646369…	-73.936882…	0.0	0.0	Credit ca
10	2	2016-03-10T09:24:14	2016-03-10T09:26:18	1	0.0	Standard rate	0.0	0.0	0.0	0.0	Credit ca
11	2	2016-03-10T07:53:17	2016-03-10T07:54:59	1	0.0	Standard rate	40.7640419…	-73.901992…	40.7638931…	-73.902008…	Cash
12	2	2016-03-10T07:37:32	2016-03-10T07:37:40	6	0.0	Standard rate	0.0	0.0	40.7639884…	-73.902076…	Cash
13	2	2016-03-10T08:01:22	2016-03-10T08:02:51	5	0.0	Standard rate	40.7641487…	-73.901992…	40.7641487…	-73.902008…	Cash
14	2	2016-03-10T09:16:50	2016-03-10T09:19:52	1	0.0	Standard rate	40.7645568…	-73.936943…	40.7642478…	-73.937019…	Credit ca
15	2	2016-03-10T07:22:27	2016-03-10T07:23:51	1	0.0	Standard rate	40.7639312…	-73.901931…	40.7639236…	-73.901947…	Cash
16	2	2016-03-10T12:16:09	2016-03-10T12:33:33	1	3.71	Standard rate	40.6941680…	-73.983322…	40.7193260…	-73.941551…	Cash
17	2	2016-03-10T07:43:54	2016-03-10T07:45:52	1	0.0	Standard rate	40.7639694…	-73.901893…	40.7639694…	-73.901893…	Cash
18	2	2016-03-01T01:07:43	2016-03-01T01:08:43	1	0.0	Standard rate	0.0	0.0	40.7645797…	-73.936958…	Credit ca
19	2	2016-03-10T07:49:01	2016-03-10T07:50:18	1	0.0	Standard rate	40.7640190…	-73.902038…	40.7638397…	-73.902091…	Cash
20	2	2016-03-10T07:08:26	2016-03-10T07:08:29	5	0.0	Standard rate	40.7640647…	-73.902107…	40.7641220…	-73.902091…	Cash
21	2	2016-03-10T14:03:59	2016-03-10T14:04:02	1	0.0	JFK	0.0	0.0	0.0	0.0	Cash
22	2	2016-03-10T08:38:27	2016-03-10T08:40:37	1	0.0	Standard rate	40.7644081…	-73.937133…	40.7646408…	-73.936950…	Cash
23	2	2016-03-10T09:08:48	2016-03-10T10:19:27	1	14.99	Standard rate	40.7939910…	-73.972061…	40.6636886…	-73.955551…	Cash
24	2	2016-03-01T06:11:58	2016-03-01T06:11:58	1	0.0	Standard rate	0.0	0.0	40.7647781…	-73.936950…	Credit ca
25	2	2016-03-10T07:38:58	2016-03-10T07:42:03	1	0.0	Standard rate	40.7639808…	-73.901870…	40.7639617…	-73.901908…	Cash

6. Data Visualization with Looker Studio

6.1 Visualization Process

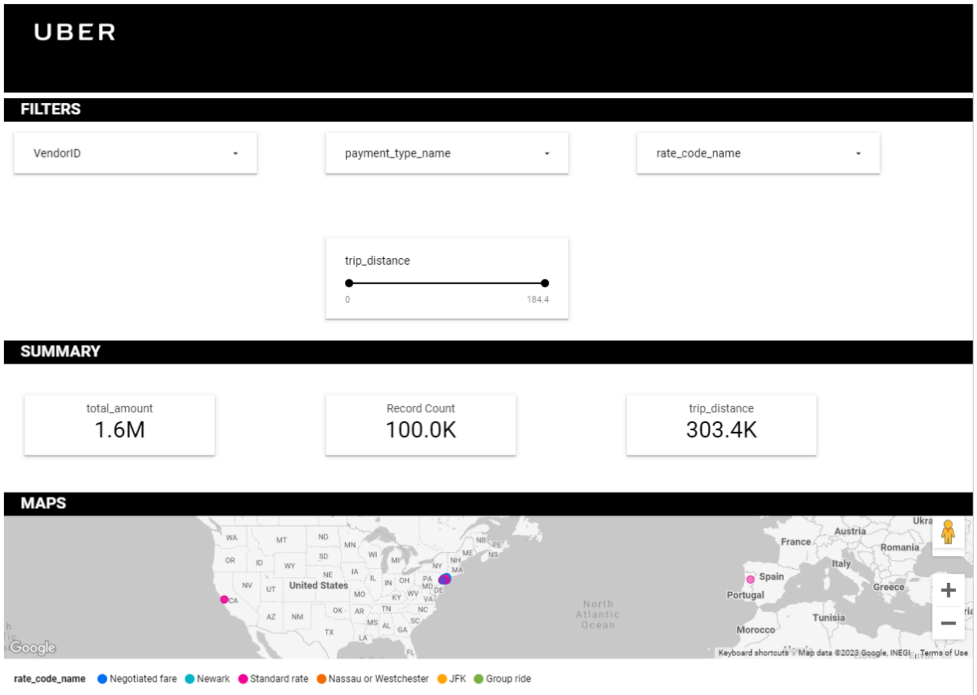
Using **Looker Studio**, we created a series of interactive dashboards to visualize key insights and trends. Looker Studio's drag-and-drop interface enabled easy customization, and we designed the following visualizations:

- **Heatmaps** of pickup locations to identify high-demand areas.
- **Time-based trend graphs** showing peak hours for Uber pickups.

6.2 Insights from Visualizations

The visualizations provided a comprehensive view of Uber data patterns, including:

- **Peak Hours**: High demand during specific hours of the day.
- **Hotspot Locations**: High-demand areas were visible in the heatmaps, revealing popular pickup locations within New York City.
- **Screenshot Suggestion**: Add **screenshots of key Looker Studio dashboards** such as heatmaps and time-series graphs here. Place them below this section to illustrate insights visually.



total_amount 1.6M

Record Count 100.0K

trip_distance 303.4K

rate_code_name ● Negotiated fare ● Newark ● Standard rate ● Nassau or Westchester ● JFK ● Group ride

7. Conclusion

7.1 Summary of the Project

This project demonstrated the effectiveness of cloud-based data analytics tools in handling large datasets. By combining GCP, BigQuery, and Looker Studio, we successfully processed, cleansed, and visualized Uber data, gaining valuable insights.

7.2 Key Insights

Our analysis uncovered essential insights, such as:

- **Peak Ride Times**: Certain hours showed higher ride frequencies, useful for optimizing driver allocation.
- **High-Demand Areas**: Identified through location-based heatmaps, helping pinpoint areas where Uber services are frequently required.

7.3 Future Scope and Enhancements

Potential extensions of this project could include:

- **Incorporating Additional Data**: Using data sources such as weather or local events to uncover more nuanced patterns.
- **Machine Learning for Predictive Analysis**: Applying machine learning algorithms to predict demand in specific locations or times for strategic planning.