Analyzing Uber Data Using GCP, Python, and Looker Studio: A Comprehensive Study

- Parth Madaan, Palak Sahu

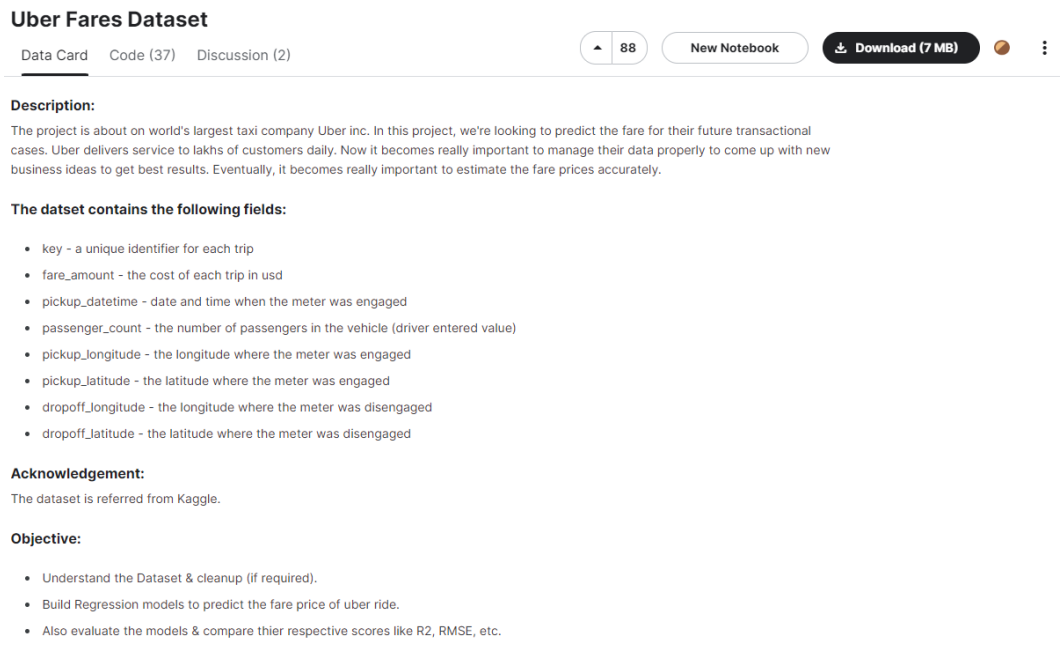
1. Introduction

1.1 Objective of the Project

The goal of this project was to leverage cloud technology and data analytics to explore and gain insights from Uber's public datasets. By using Google Cloud Platform (GCP), Python, Mage Data Pipeline Tool, BigQuery, and Looker Studio, we sought to reveal hidden patterns and trends to facilitate strategic decision-making.

1.2 Data Overview

The Uber dataset consists of millions of Uber pickups in New York City over a certain period. The data features several variables such as date/time, the location of the pickup, and other ride-related details.



Uber Fares Dataset

Data Card | Code (37) | Discussion (2) | 88 | New Notebook | Download (7 MB)

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:

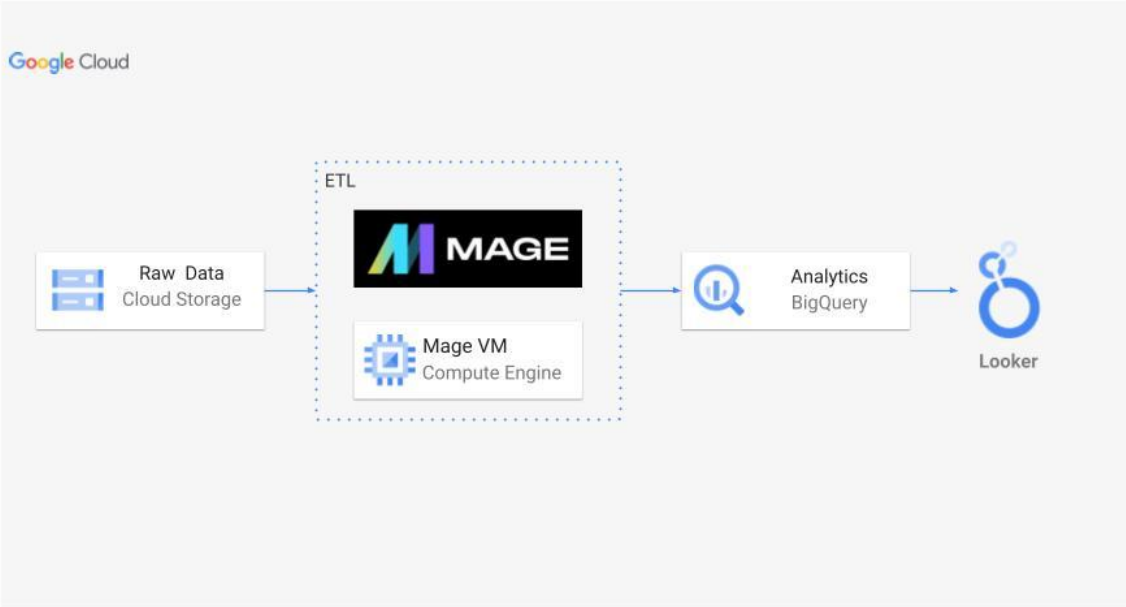
The dataset is referred from Kaggle.

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 utilized several technologies including Google Cloud Storage, Python for data manipulation, Mage for ETL processes, BigQuery for storage and querying, and Looker Studio for data visualization.



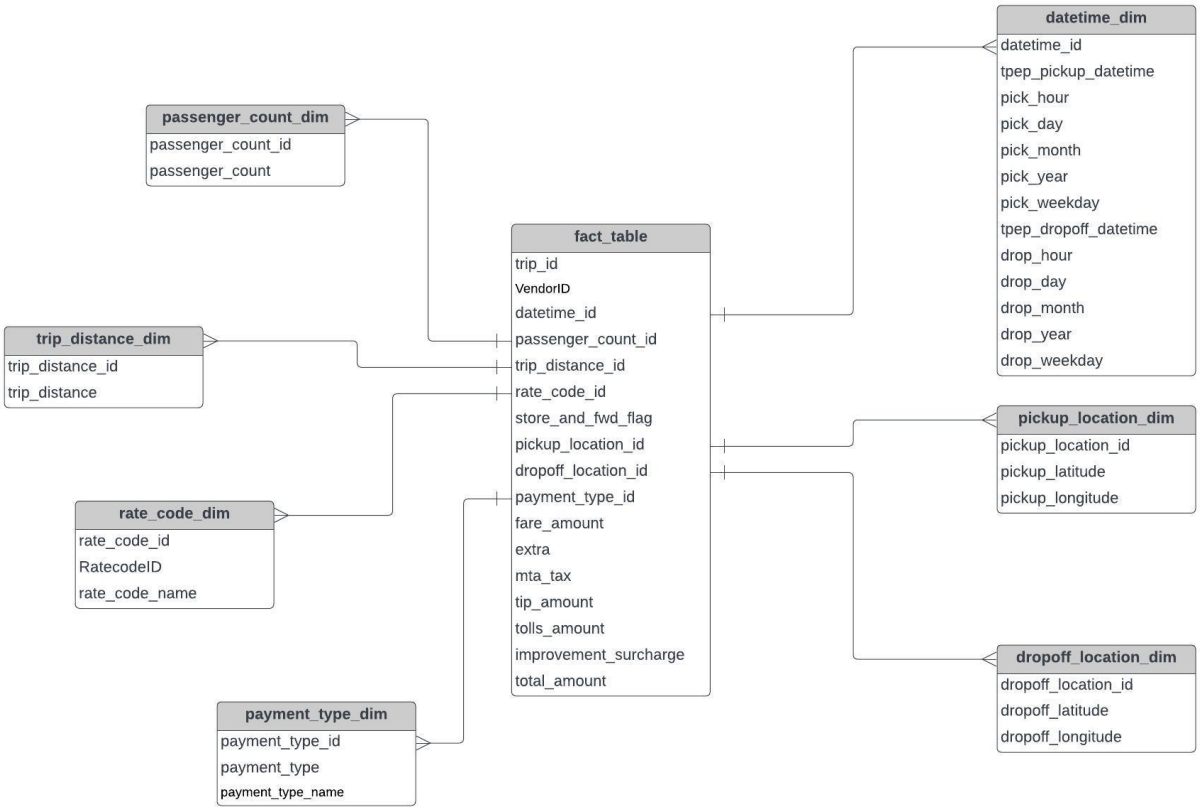
1.4 Project Overview

We began by designing a data model, followed by uploading data to GCP. We then processed the data using an ETL pipeline before performing data cleansing in BigQuery. Lastly, we visualized the results in Looker Studio.

2. Data Modeling

2.1 Designing the Schema

We modeled the data in Lucidchart, creating a fact table for the rides and dimension tables for time, location, and ride characteristics. This star schema allows efficient querying and aggregation.



2.2 Implementing the Schema

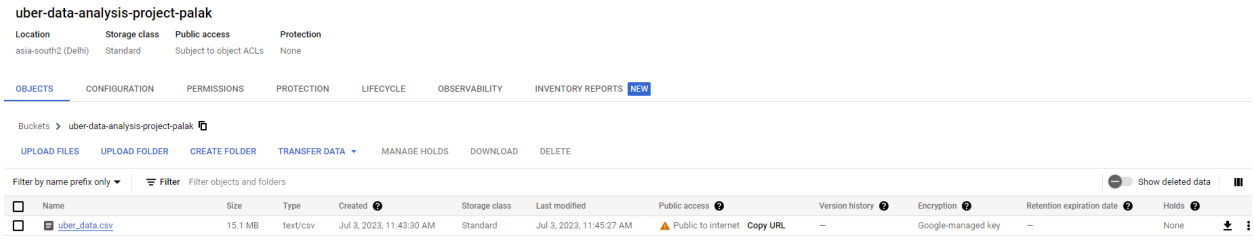
After uploading our data to GCP, we implemented our designed schema using Python and GCP tools, aligning the structure of our data with the Lucidchart model.

3. Data Ingestion and Storage

3.1 Uploading Data to Google Cloud Storage

We uploaded our CSV file to GCP's cloud storage. GCP provides durable and scalable object storage, making it an ideal choice for our large dataset.

3.2 CSV Data in Google Cloud Storage



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

4. ETL Pipeline

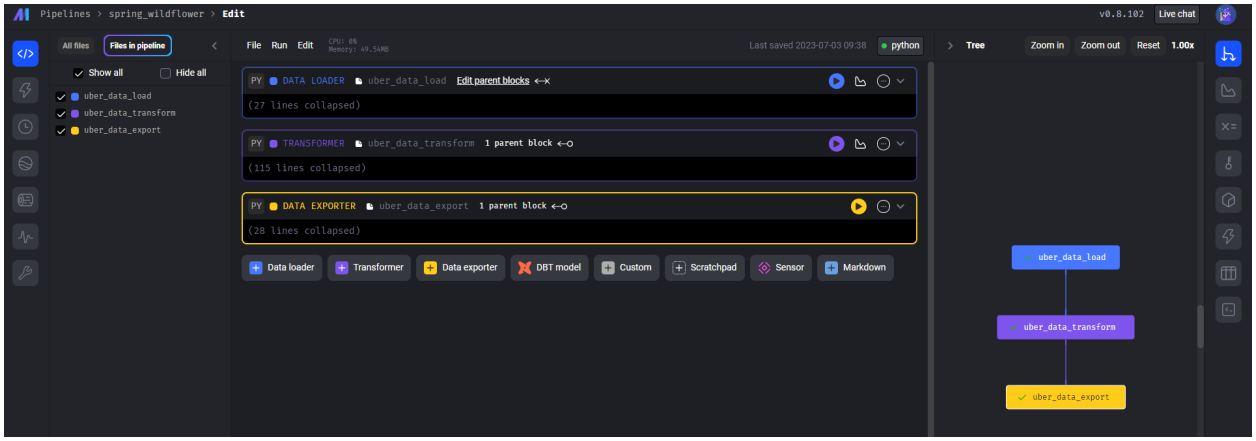
4.1 Deploying Mage on Compute Engine API

We deployed Mage on the Google Compute Engine API. This was done to utilize Mage's capabilities in handling large-scale data processing tasks efficiently.

```
mage start uber_data_project
WARNING:traitlets:Message signing is disabled.  This is insecure and not recommended!
WARNING:traitlets:Message signing is disabled.  This is insecure and not recommended!
{
  "environment": "dev",
  "platform": "Linux-5.10.0-23-cloud-amd64-x86_64-with-glibc2.31",
  "project_uuid": "b7e9b05c985c4dcfb66886fbd9b412f4",
  "version": "0.8.102",
  "action": "impression",
  "object": "project"
}
```

4.2 ETL Process Using Mage

Mage extracted the data from the CSV file, transformed the data according to our schema, and loaded the transformed data into BigQuery for further analysis.



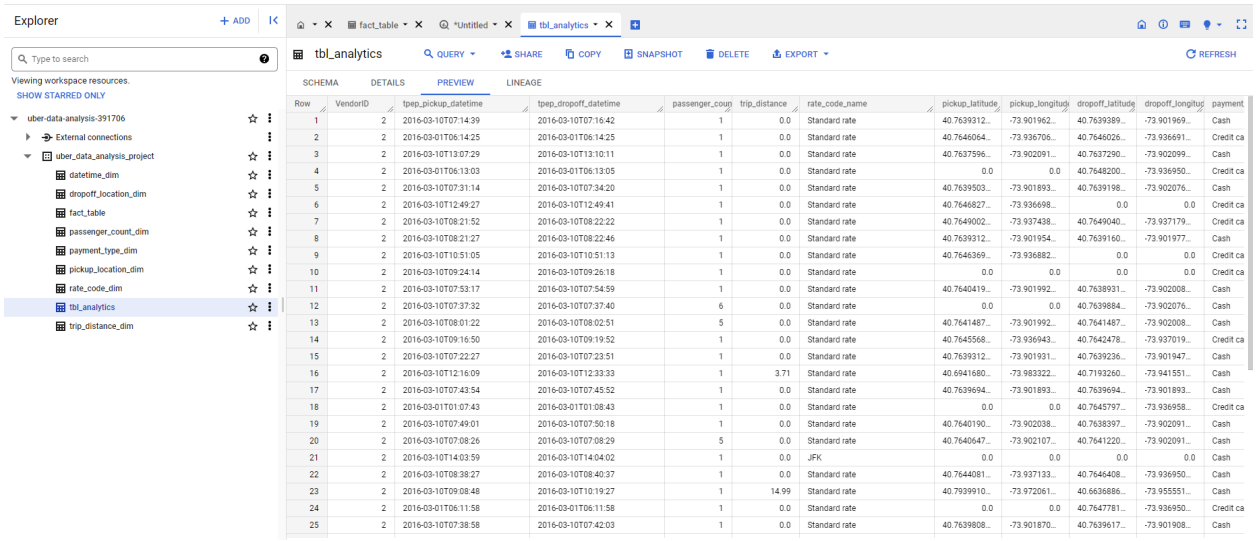
5. Data Cleansing in BigQuery

5.1 Uploading the Data into BigQuery

Once the ETL process was completed, we uploaded our transformed data into BigQuery. BigQuery allows for super-fast SQL queries against append-mostly tables using the processing power of Google's infrastructure.

5.2 Data Cleansing Operations

We performed several data cleansing operations in BigQuery to ensure data consistency and quality. These operations included handling missing data, removing duplicates, and standardizing data formats.



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

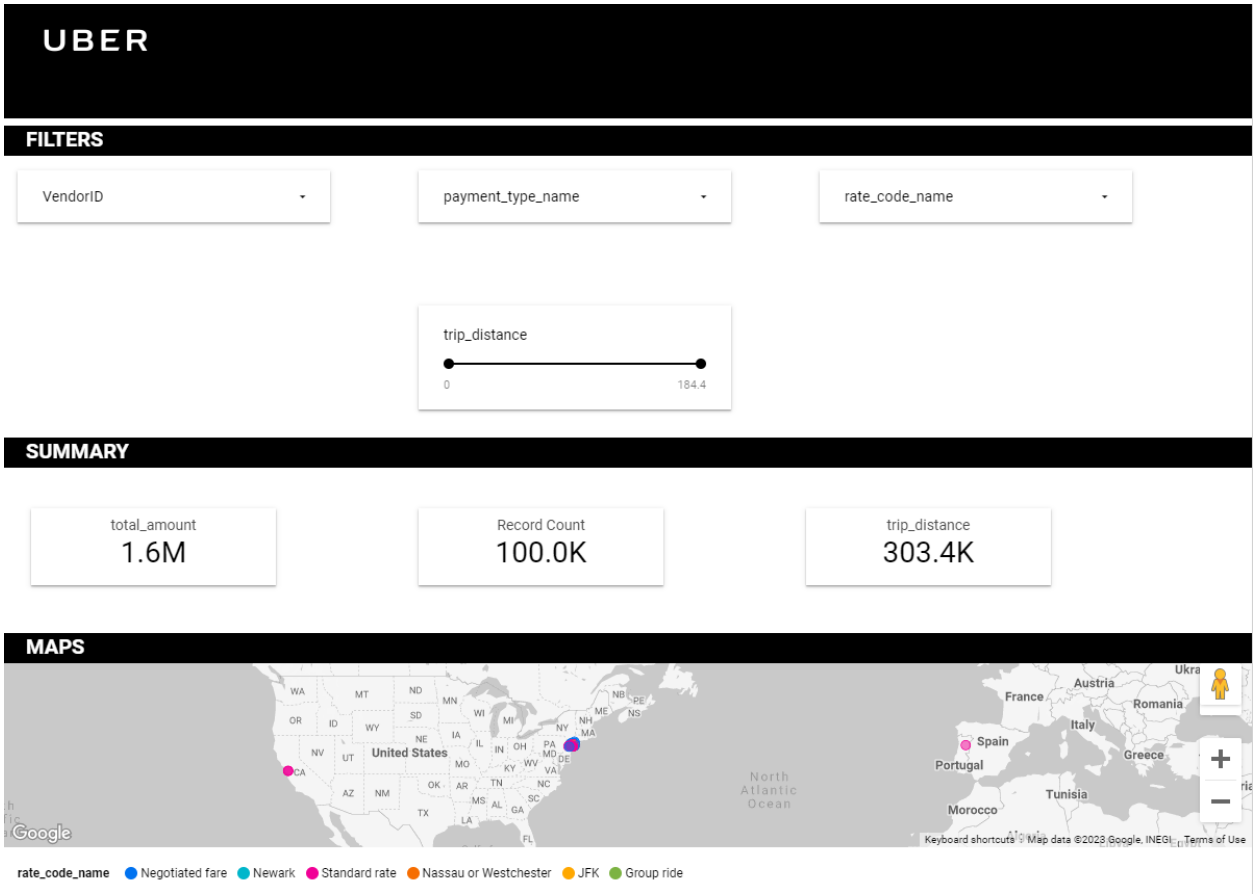
6. Data Visualization with Looker Studio

6.1 Visualization Process

We used Looker Studio to create interactive dashboards and visualizations. Looker Studio's drag-and-drop interface and powerful data modeling features made it easy to create detailed and insightful visualizations.

6.2 Insights from the Visualizations

Through our visualizations, we discovered trends in ride frequency during different times of day and across various locations. For example, a heatmap of pick-up locations revealed certain hotspots within the city.



7. Conclusion

7.1 Summary of the Project

This project demonstrated the power of cloud computing and data analytics tools in analyzing large datasets. Through this process, we uncovered valuable insights from the Uber dataset that could inform business decisions and strategies.

7.2 Insights and Findings

Our analysis revealed key insights, such as peak ride times and high-demand locations, which could potentially assist Uber in optimizing their driver allocation and improving service efficiency.

7.3 Future Scope and Enhancements

Future iterations of this project could incorporate additional data sources, like weather or event data, to uncover more nuanced patterns and trends. Additionally, machine learning algorithms could be utilized for predictive analysis.