DATA 603 Platforms for Big Data Processing

Uber Trip Data Analysis

Final Project Delivery Report

Group 5
Havish Manikya Vakkalanka
Rohith Kumar Koutike
Sai Akhil Sadula

Table of Contents

Introduction

Project Significance: This Uber Trip Data Analysis provides insights into urban travel patterns, highlighting the importance of data in enhancing transportation policies and strategies.

Uber Overview: As a leading ride-sharing company, Uber's extensive data offers a unique perspective on urban mobility, crucial for understanding modern transportation trends.

Project Goals: The aim is to analyze trip patterns and trends, contributing to the efficiency of ride-sharing services and overall urban transportation.

Methodology Overview: Utilizing advanced data analytics tools, the project dissects Uber trip data to reveal key insights into rider behavior and peak travel times.

Introduction

- In New York City, all taxi vehicles are managed by TLC (Taxi and Limousine Commission) established in 1971.

- TLC regulates New York City's Medallion (Yellow) taxi cabs, for-hire vehicles (community-based liveries, black cars, and luxury limousines), commuter vans, and paratransit vehicles.

- Over 200,000 (2 Lakhs) TLC licensed vehicles complete approximately 1,000,000 (1 Million) trips each day.

- High-volume-for-hire vehicle bases(HVFH) are companies that dispatch 10,000+ trips per day.

- We have selected UBER for our analysis which is also an HVFH company.

About Uber

- It is founded in 2009
- Uber used in 70 countries
- 131 million users and 5 million drivers
- 23 million trips done everyday all over World

Project Objective and Goals

- **Analyze Urban Travel Patterns:** Uncover the trends and behaviors in urban travel using Uber's comprehensive trip data.
- **Enhance Ride-Sharing Efficiency:** Utilize insights from the data to propose improvements for ride-sharing services, focusing on efficiency and customer satisfaction.
- **Predictive Trend Analysis:** Employ predictive models to forecast future transportation trends and rider preferences.
- **Data-Driven Decision Making:** Provide actionable recommendations for urban transportation planning and policy-making based on analyzed data.

Problem Statement

- Using the uber data, to develops a time series and financial analysis analysis and uses Apache Spark in Databricks to produce accurate projections in the years 2021 and 2022.
- Since the data we have in Parquet file format Jupyter makes it difficult to handle files with big amounts of data, the system should use Apache Spark and Databricks to produce interactive dashboards and visualizations to make Time and Financial analysis.
- The system should be able to analyze past data to identify patterns and trends and provide recommendations to help Uber drivers make informed decisions.
- The system must be scalable and easy to use, with the ability to process large volumes of data quickly and efficiently.

About Dataset

- Data Source: We were able to obtain data in the form of Parquet files from NYC Taxi & Limousine commission.
  - https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page - 2021 & 2022
- We considered data for 2 years
- In total there were up to 23 million records of data in which each month consists of upto millions of rows or trips

Schema

Name	Description
Hvfhs_license_num	The TLC license number of the HVFHS base or business As of September 2019, the HVFHS licensees are the following: • HV0002: Juno • HV0003: Uber • HV0004: Via • HV0005: Lyft
Dispatching_base_num	The TLC Base License Number of the base that dispatched the trip
originating_base_num	base number of the base that received the original trip request
request_datetime	date/time when passenger requested to be picked up
on_scene_datetime	date/time when driver arrived at the pick-up location (Accessible Vehicles-only)
Pickup_datetime	The date and time of the trip pick-up
DropOff_datetime	The date and time of the trip drop-off
PULocationID	TLC Taxi Zone in which the trip began
DOLocationID	TLC Taxi Zone in which the trip ended

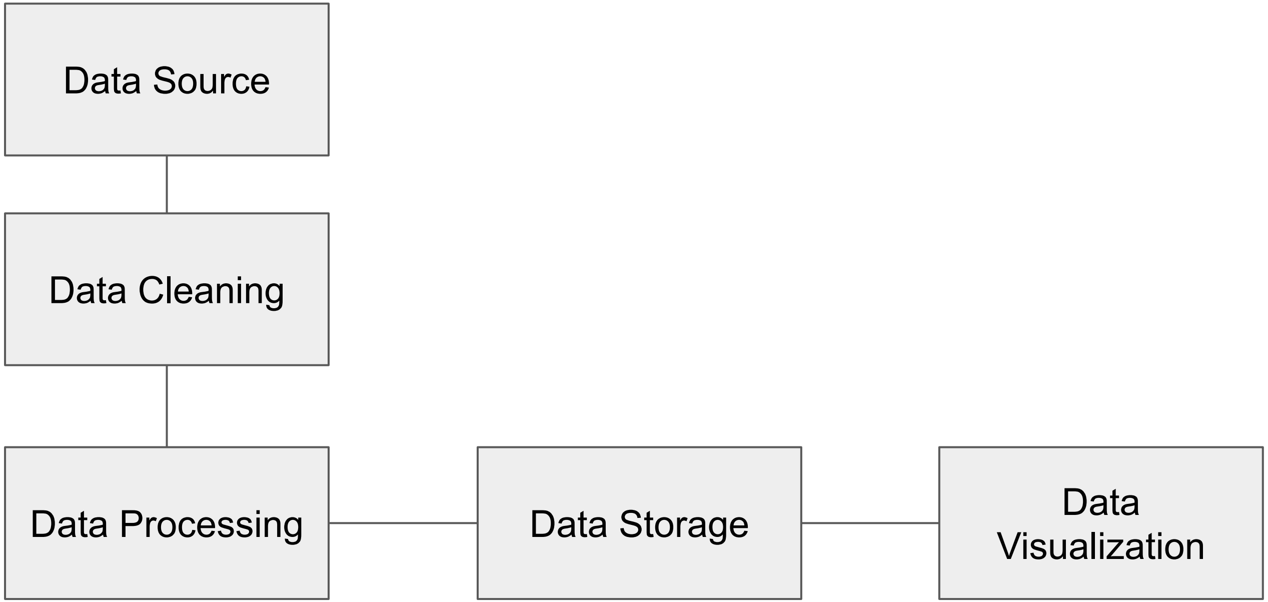
Schema

Name	Description
trip_miles	total miles for passenger trip
trip_time	Total airtime in minutes
base_passenger_fare	base passenger fare before tolls, tips, taxes, and fees
tolls	total amount of all tolls paid in trip
bcf	total amount collected in trip for Black Car Fund
sales_tax	total amount collected in trip for NYS sales tax
congestion_surcharge	total amount collected in trip for NYS congestion surcharge
airport_fee	$2.50 for both drop off and pick up at LaGuardia, Newark, and John F. Kennedy airports
tips	total amount of tips received from passenger
driver_pay	total driver pay (not including tolls or tips and net of commission, surcharges, or taxes)

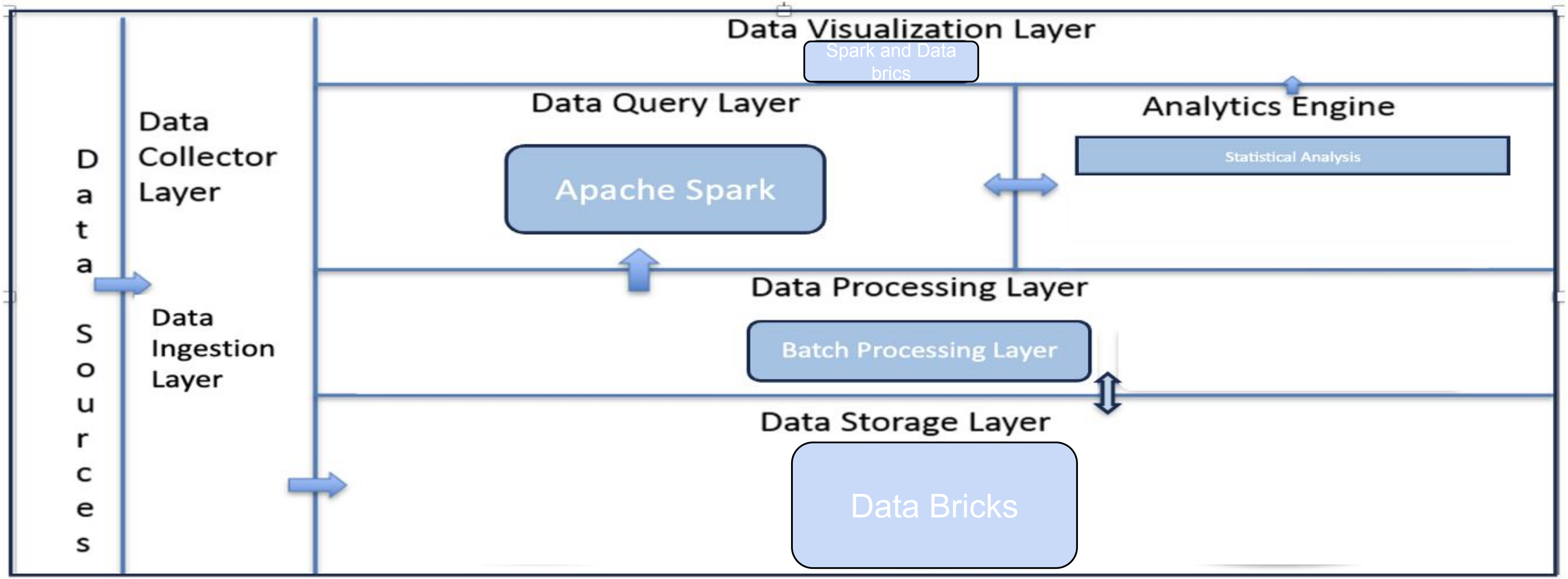
Challenges in Dataset

- **Handling Large Data Volume:** The 7.2GB parquet data presented challenges in loading and processing due to its size and complexity.
- **Database Integration Issues:** Initial attempts to integrate the data with MongoDB and HDFS were hindered by compatibility and performance constraints.
- **Utilization of Databricks:** To overcome these challenges, we employed Databricks, which facilitated efficient storage and analysis of the large dataset.
- **Data Management and Processing:** The need for advanced data management strategies was highlighted to effectively handle and analyze the extensive dataset.

Workflow


Data Source
Data Cleaning
Data Processing
Data Storage
Data Visualization

Stack Diagram



Sample Dataset

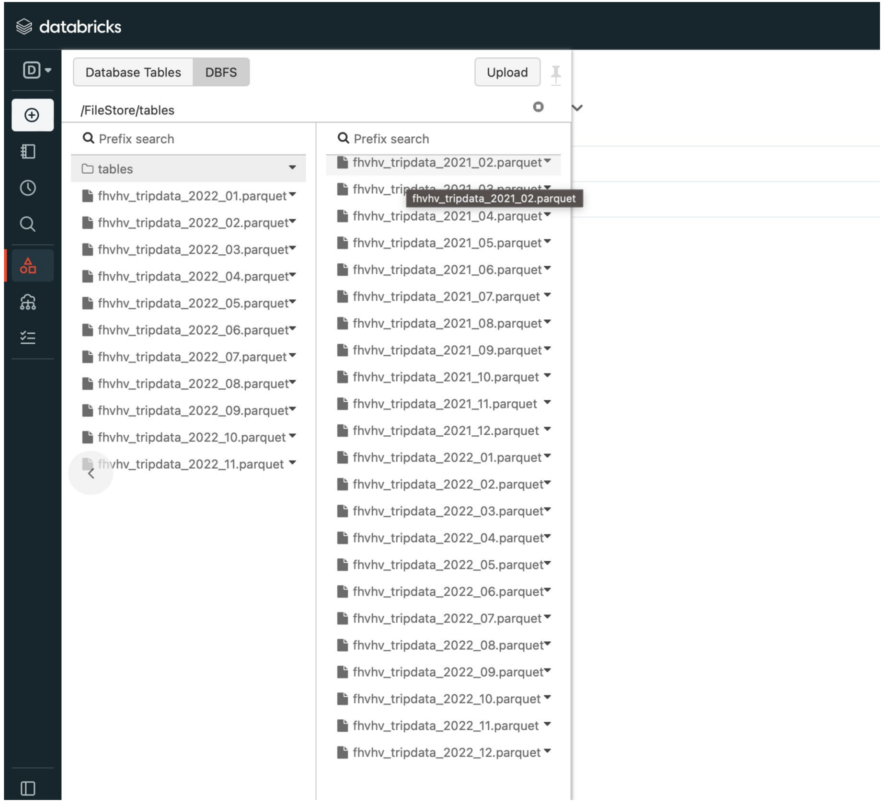
	hvfhs_license_num	dispatching_base_num	originating_base_num	request_datetime	on_scene_datetime	pickup_datetime	dropoff_datetime	PULocatio
0	HV0003	B02682	B02682	2021-01-01 00:28:09	2021-01-01 00:31:42	2021-01-01 00:33:44	2021-01-01 00:49:07	
1	HV0003	B02682	B02682	2021-01-01 00:45:56	2021-01-01 00:55:19	2021-01-01 00:55:19	2021-01-01 01:18:21	
2	HV0003	B02764	B02764	2021-01-01 00:21:15	2021-01-01 00:22:41	2021-01-01 00:23:56	2021-01-01 00:38:05	
3	HV0003	B02764	B02764	2021-01-01 00:39:12	2021-01-01 00:42:37	2021-01-01 00:42:51	2021-01-01 00:45:50	
4	HV0003	B02764	B02764	2021-01-01 00:46:11	2021-01-01 00:47:17	2021-01-01 00:48:14	2021-01-01 01:08:42	
...	...	...	...	...	...	...	...	
11908463	HV0003	B02765	B02765	2021-01-31 23:13:51	2021-01-31 23:25:03	2021-01-31 23:25:40	2021-01-31 23:40:10	
11908464	HV0003	B02872	B02872	2021-01-31 23:23:56	2021-01-31 23:29:03	2021-01-31 23:29:31	2021-01-31 23:47:44	
11908465	HV0003	B02872	B02872	2021-01-31 23:42:53	2021-01-31 23:49:23	2021-01-31 23:49:32	2021-02-01 00:04:36	
11908466	HV0003	B02764	B02764	2021-01-31 23:04:32	2021-01-31 23:09:13	2021-01-31 23:09:29	2021-01-31 23:27:46	
11908467	HV0003	B02764	B02764	2021-01-31 23:22:20	2021-01-31 23:28:33	2021-01-31 23:28:33	2021-01-31 23:56:36	

11908468 rows × 24 columns

Tools Implemented

- In our original proposal, we wanted to use jupyter notebook for data cleaning and processing.
- But most of the cleaning was done in DataBricks and Apache Spark.
- There was not necessity of utilizing the services of jupyter notebook.

Data feeding to Databricks



- Data consisting 23 million records was imported in DataBricks.
- For accurate analysis of recent Uber patterns, data of year 2021-2022 was stored.
- Data was imported in .parquet format and stored in the database.
- Data, specific to the requirement was added to the collection.

Data cleansing using DataBricks

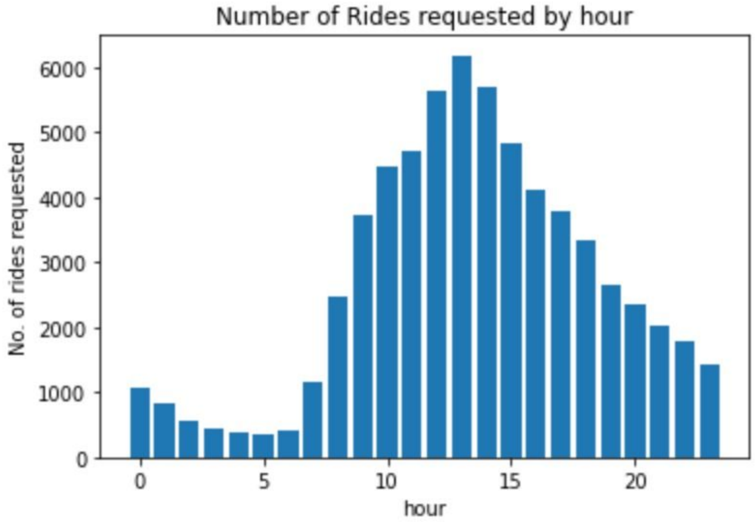
- Data cleansing was performed using in-built aggregate functions present in Databricks.

- Queries were run to remove null values, remove duplicates and replace irrelevant values in the pipeline.

- The cleaned data set was then stored in Pyspark Dataframe containing the combined data of all 12 months

EDA - Rides requested by hour

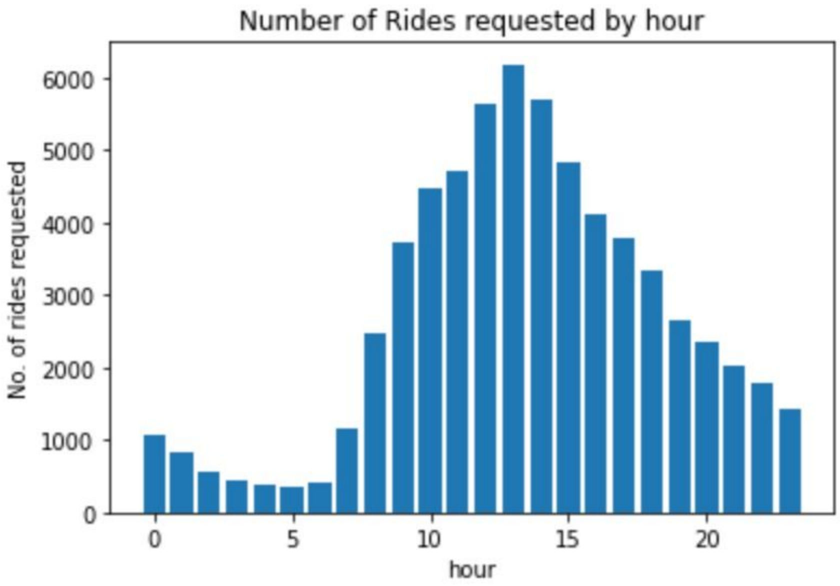
pickup_hour	avg_trip_duration
12	1322.0442397977608
22	1183.3990903922684
1	1084.404052443385
13	1303.4444979919679
6	1160.1767676767677
16	1343.4551937247445
3	1183.5644444444445
20	1170.7235668789808
5	1183.7664835164835
19	1209.151106111736
15	1317.7156398104266
17	1314.6259925886714
9	1290.034966887417
4	1179.5340050377833
8	1373.143855322647
23	1183.6850828729282
7	1303.4048913043478
10	1322.7816014394962



2021

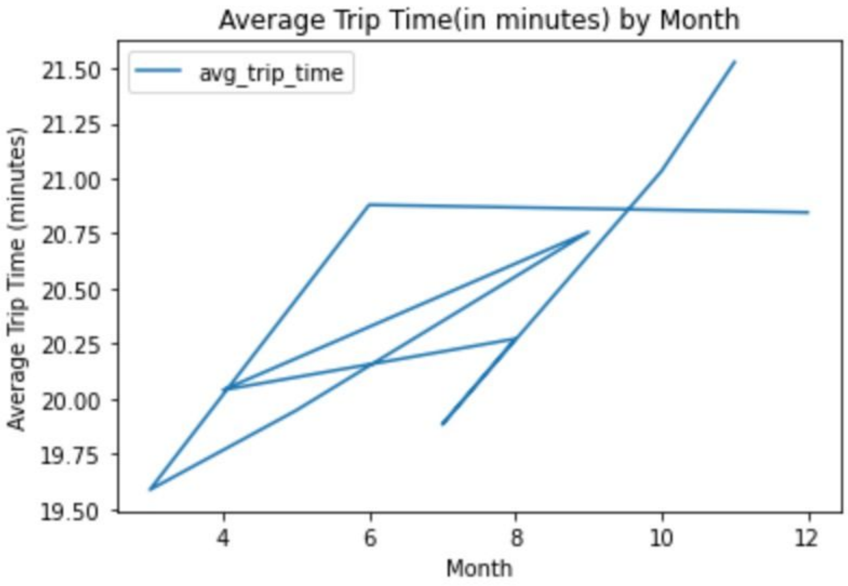
EDA - Rides requeste dby hour

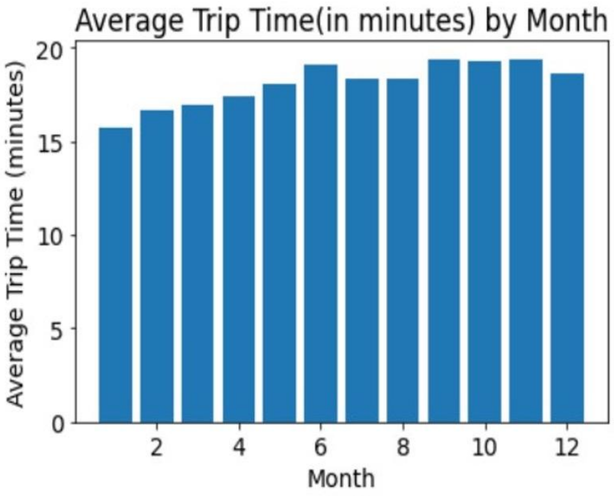
pickup_hour	avg_trip_duration
12	1275.7314731473148
22	1148.7966985230235
1	1032.119266055046
13	1266.134353268428 4
16	1279.5437753971187
6	1264.2215568862275
3	1049.4830769230769
20	1163.93631778058
5	1202.9431818181818
19	1178.2867058195409
15	1285.0228013029316
17	1239.9241744802282
9	1255.0997023809523
4	1158.74609375
8	1295.5418950665623
23	1146.3377823408625
7	1262.549019607843
10	1261.8618541590326



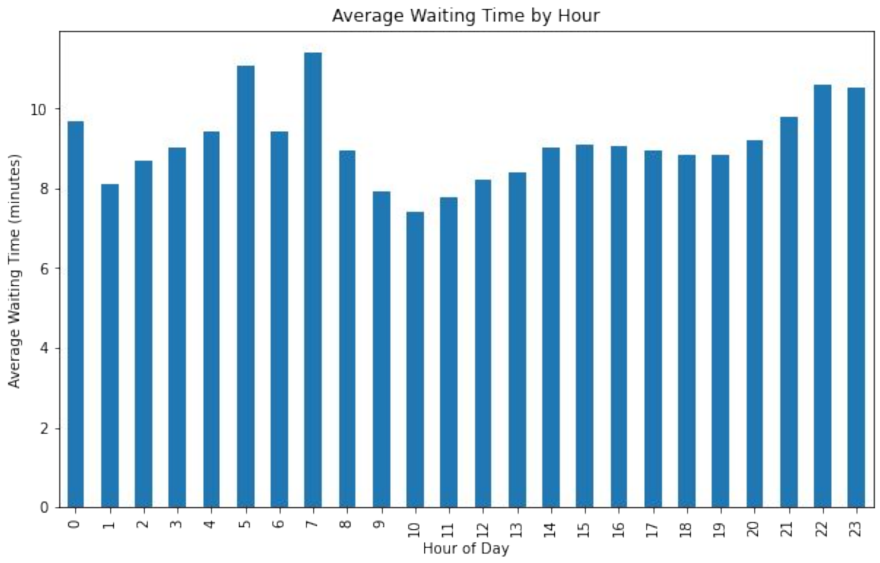
2022

EDA - Trip time over the months

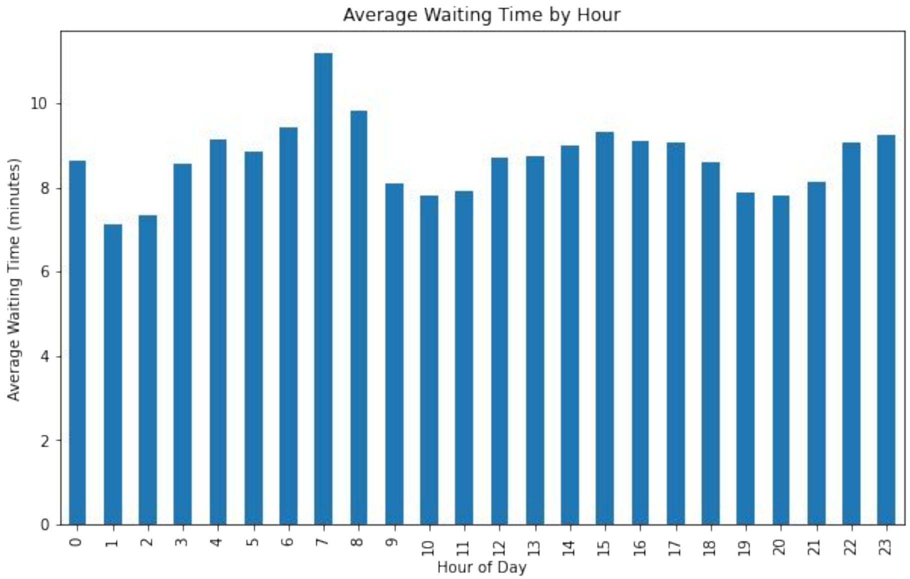




EDA - Average wait time

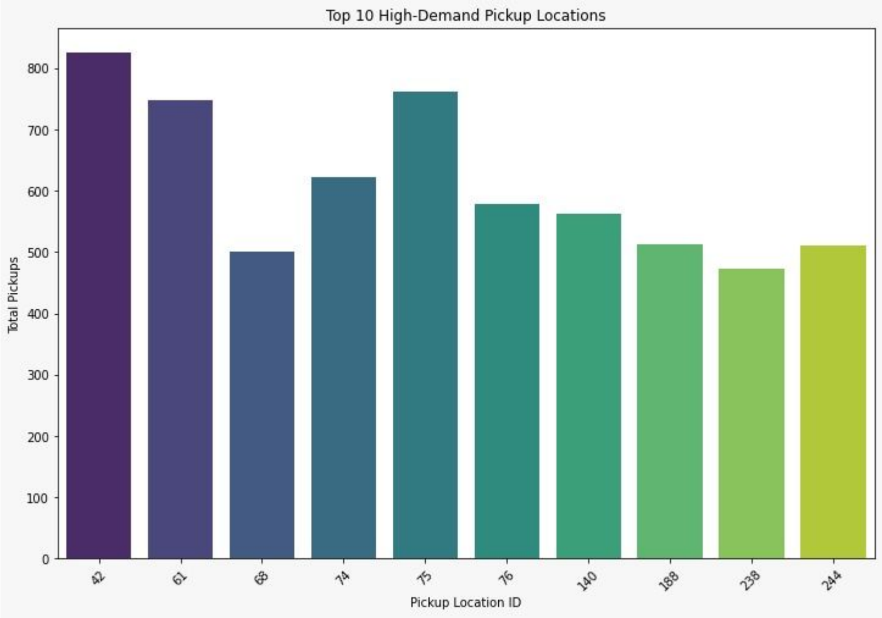


2021

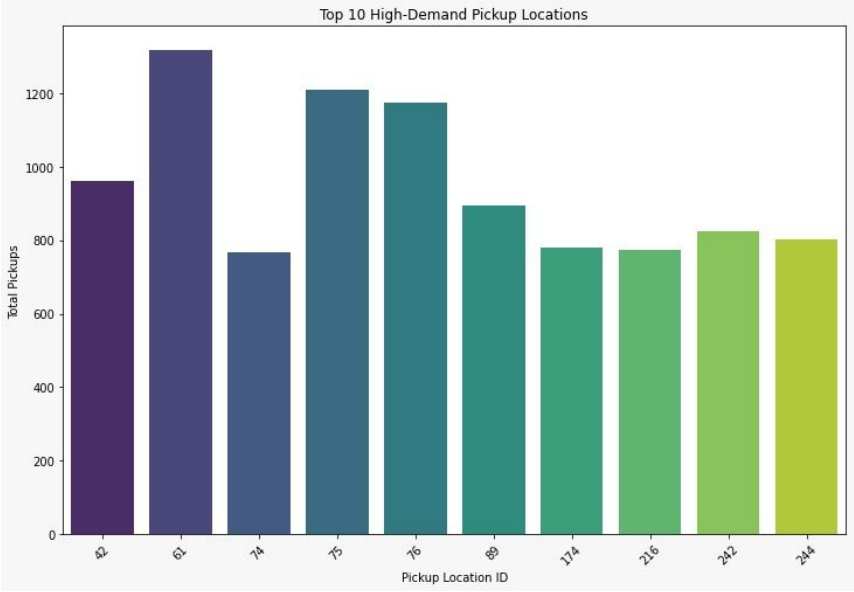


2022

EDA - High Demand Pickup Locations

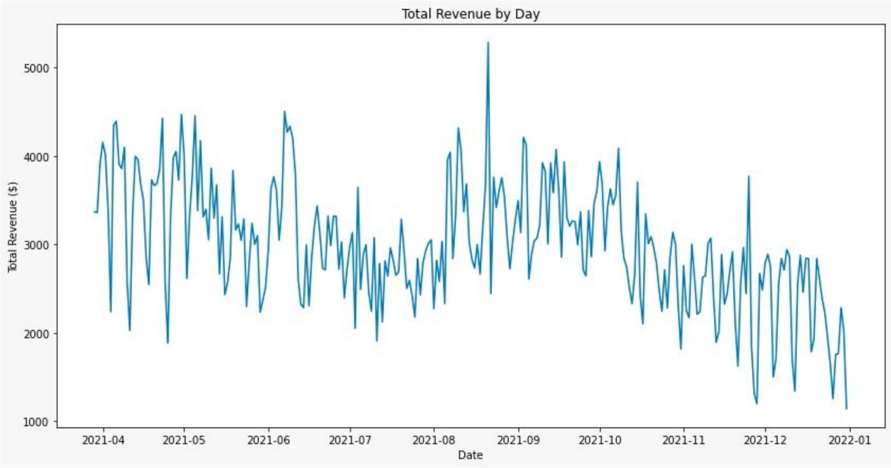


2021

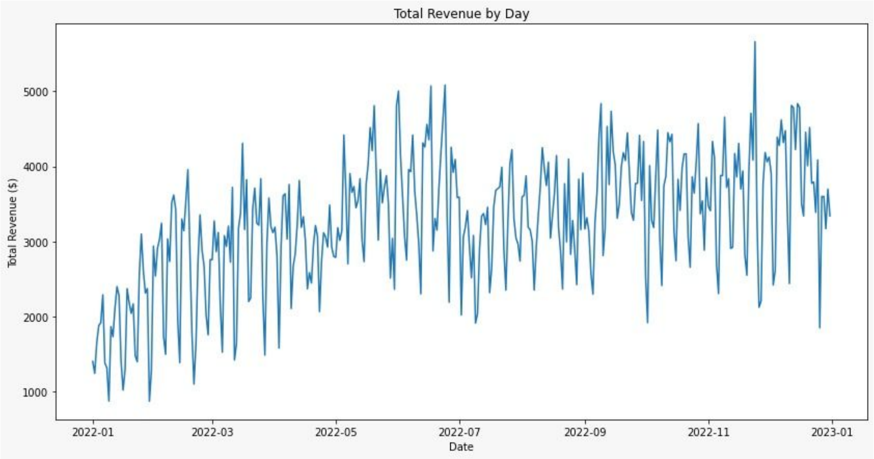


2022

EDA - Total Revenue by Day

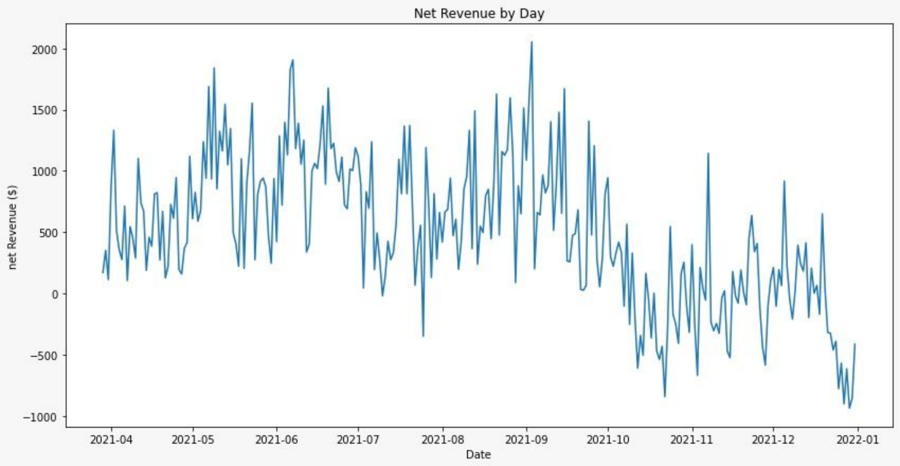


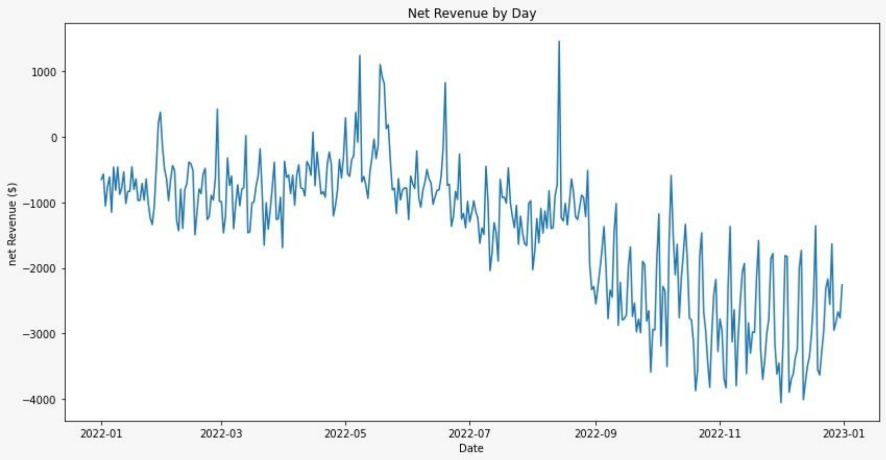
2021



2022

EDA - Net Revenue By Day





2021

2022

Reports and Insights

Trip Time Analysis: Examination of trip times highlighted peak periods for Uber services, aiding in understanding daily transportation flows.
Trip Duration Patterns: Insights into average trip durations helped identify common journey lengths and their implications on service demand.
Wait Time Evaluation: Analysis of rider wait times offered perspectives on service efficiency and areas for improvement.
Pickup Demand Hotspots: Identified key locations with high pickup demands, crucial for strategic planning and resource allocation.
Revenue Analysis: Detailed examination of total and net revenue provided an understanding of the financial aspects of Uber services.

Learnings

- We've gained skills in using Databricks as a NoSQL database for data storage. This experience taught us how to manage flexible data structures and efficiently retrieve and clean data using specialized queries.
- Additionally, our work with Databricks involved Apache Spark for large-scale data processing. We learned to use Spark's capabilities for distributed computing, data manipulation, and running large-scale machine learning models.

Learnings

- Our experience with Apache Spark and Databricks has shown us the ropes of crafting dynamic and interactive data visualizations. We've picked up skills in integrating Tableau with various data sources, constructing informative charts and dashboards, and employing visual storytelling to convey our insights.

- We've also discovered the significance of collaboration and teamwork. Effective communication, coordination, and a shared focus have been key to our success in achieving project objectives.

Future Opportunities and Challenges

- **Scalability with Growing Data:** As data volume increases, scaling analysis tools and infrastructure will be a key challenge.
- **Real-Time Data Analysis:** Future opportunities lie in developing capabilities for real-time data analysis to enhance responsiveness.
- **Integration with Smart City Initiatives:** Collaborating with smart city projects presents opportunities for broader urban planning impacts.
- **Data Privacy and Security:** Ensuring the privacy and security of user data will remain a paramount challenge amidst expanding analysis scopes

References

- *TLC Trip Record Data - TLC*. (n.d.). https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page
- Databricks on AWS. https://docs.databricks.com/sql/index.html
- PySpark Overview — PySpark 3.4.0 documentation. (n.d.). https://spark.apache.org/docs/latest/api/python

Thank You!