OPTIMIZING RIDE HAILING : A DEEP DIVE INTO TAXI HAILING DATA FROM NEW YORK

TEAM PIED PIPERS

BHAVIK | NAMRATHA | SMEET | SOURAB | YUTI

TABLE OF CONTENTS

ABSTRACT

- Uber/Lyft, a rideshare application from San Francisco, has transformed global urban commuting with its decentralized model leveraging local drivers and vehicle owners.
- This study compares Uber/Lyft's data-driven approach to New York's traditional taxis.
- The comparison aims to enhance operational strategies like driver allocation and route optimization for Uber/Lyft, leading to cost reductions, improved service quality, and a passenger-focused experience.

PROJECT OBJECTIVE

- Get general insights and analytics about taxi data from New York.
- Compare the performance of new-gen ride-hailing services vs Traditional Yellow Taxis.
- Compare the performance between different new-gen ride-hailing apps.
- An Interface for customers to gauge the cheaper/more reliable option from a certain source to a certain destination.

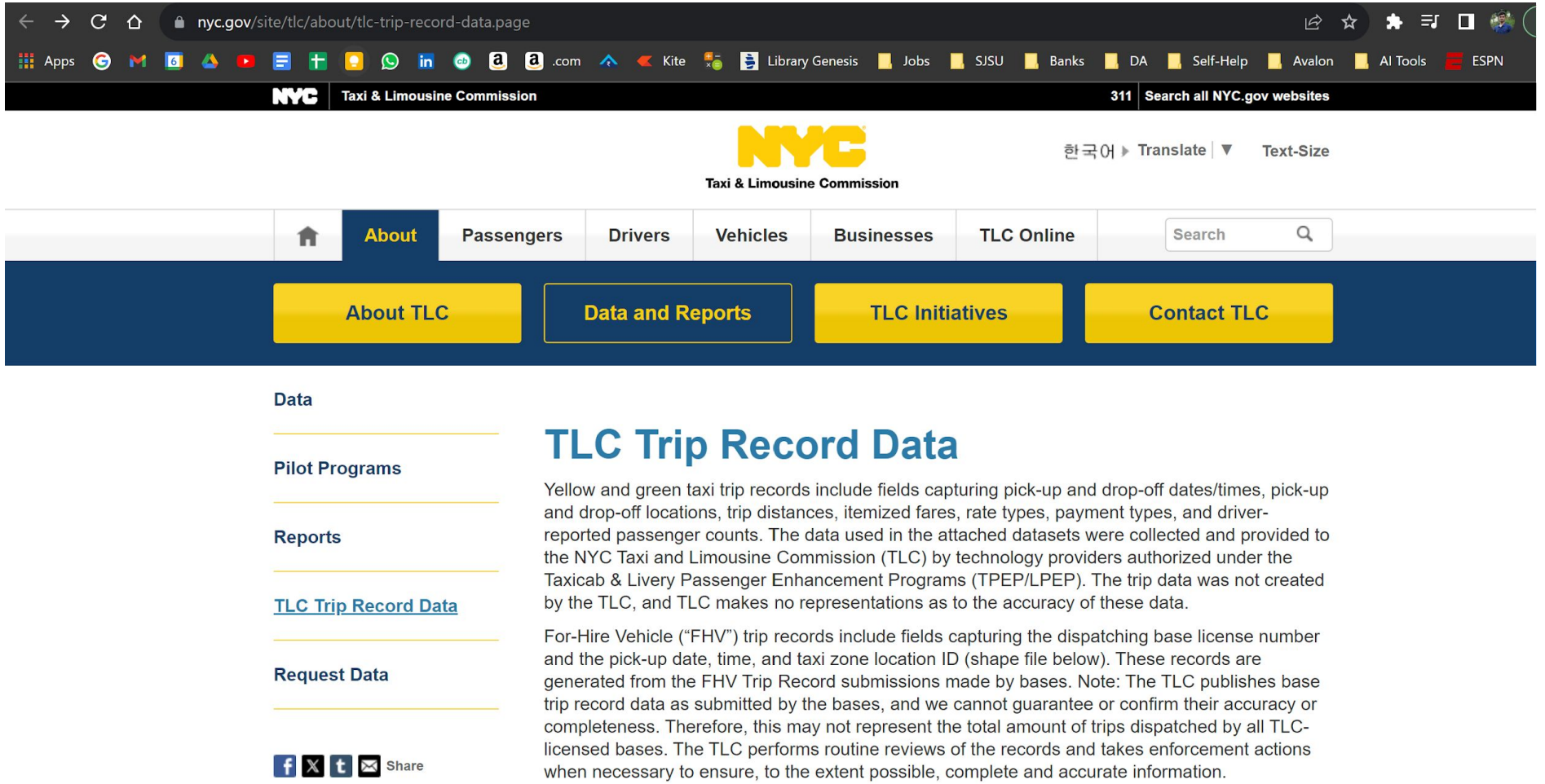
WHAT DOES THE DATA LOOK LIKE?

- Taxi Data for the Q3 of year 2023
- Data from TLC (Taxi & Limo Commission) of New York for Yellow Taxis
- Parquet File Format
- About 25 GB of Data

TOOLS AND TECHNIQUES

- **Data Ingestion** - Mage
- **EDA**- Google Colab
- **Analysis**- BigQuery
- **Visualizations**- Power BI
- **Website Development** - C#
- **Single Sign On** - Google OAuth SSO
- **Website Hosting** - IIS Localhost

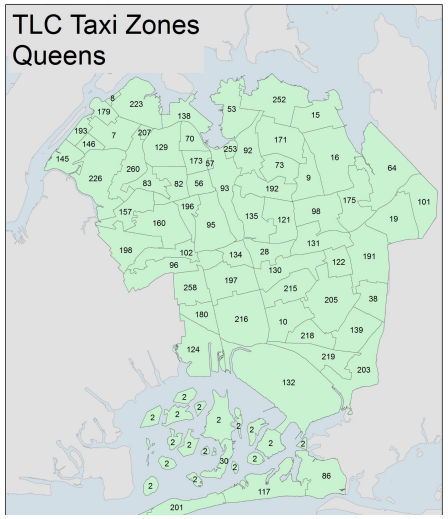
DATA SOURCE

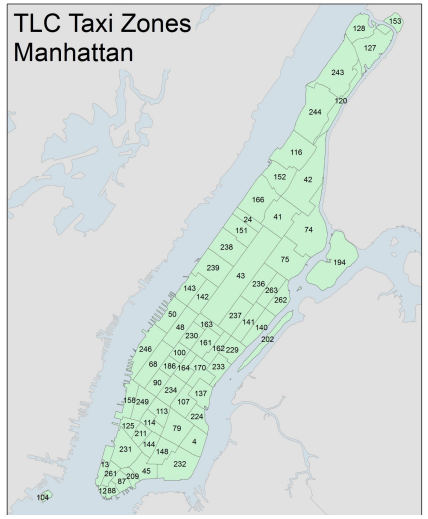

nyc.gov/site/tlc/about/tlc-trip-record-data.page
Apps
Kite
Library Genesis
Jobs
SJSU
Banks
DA
Self-Help
Avalon
AI Tools
ESPN
NYC
Taxi & Limousine Commission
311
Search all NYC.gov websites
NYC
Taxi & Limousine Commission
한국어
Translate
Text-Size
About
Passengers
Drivers
Vehicles
Businesses
TLC Online
Search
About TLC
Data and Reports
TLC Initiatives
Contact TLC
Data
Pilot Programs
Reports
TLC Trip Record Data
Request Data
Share
TLC Trip Record Data
Yellow and green taxi trip records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. The data used in the attached datasets were collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers authorized under the Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP). The trip data was not created by the TLC, and TLC makes no representations as to the accuracy of these data.
For-Hire Vehicle ("FHV") trip records include fields capturing the dispatching base license number and the pick-up date, time, and taxi zone location ID (shape file below). These records are generated from the FHV Trip Record submissions made by bases. Note: The TLC publishes base trip record data as submitted by the bases, and we cannot guarantee or confirm their accuracy or completeness. Therefore, this may not represent the total amount of trips dispatched by all TLC-licensed bases. The TLC performs routine reviews of the records and takes enforcement actions when necessary to ensure, to the extent possible, complete and accurate information.

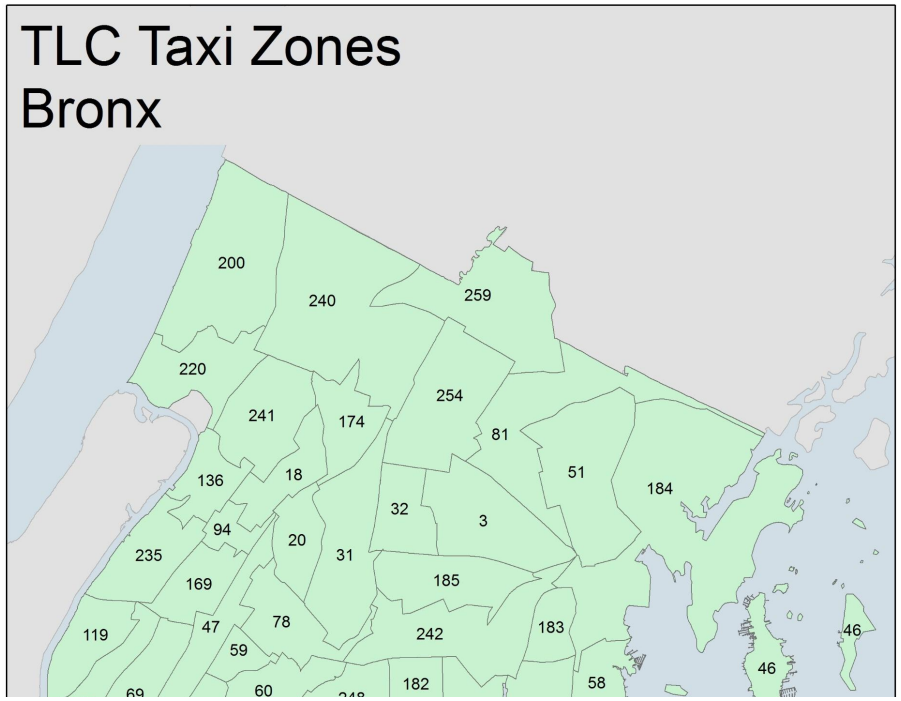
Field Name	Description
Hvfhs_license_num	The TLC license number of the HVFHS base or business As of September 2019, the HVFHS licensees are the following: • HV0002: Juno • HV0003: Uber • HV0004: Via • HV0005: Lyft
Dispatching_base_num	The TLC Base License Number of the base that dispatched the trip
Pickup_datetime	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
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)
trip_miles	total miles for passenger trip
trip_time	total time in seconds for passenger trip
base_passenger_fare	base passenger fare before tolls, tips, taxes, and fees

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)
shared_request_flag	Did the passenger agree to a shared/pooled ride, regardless of whether they were matched? (Y/N)
shared_match_flag	Did the passenger share the vehicle with another passenger who booked separately at any point during the trip? (Y/N)

access_a_ride_flag	Was the trip administered on behalf of the Metropolitan Transportation Authority (MTA)? (Y/N)
wav_request_flag	Did the passenger request a wheelchair-accessible vehicle (WAV)? (Y/N)
wav_match_flag	Did the trip occur in a wheelchair-accessible vehicle (WAV)? (Y/N)



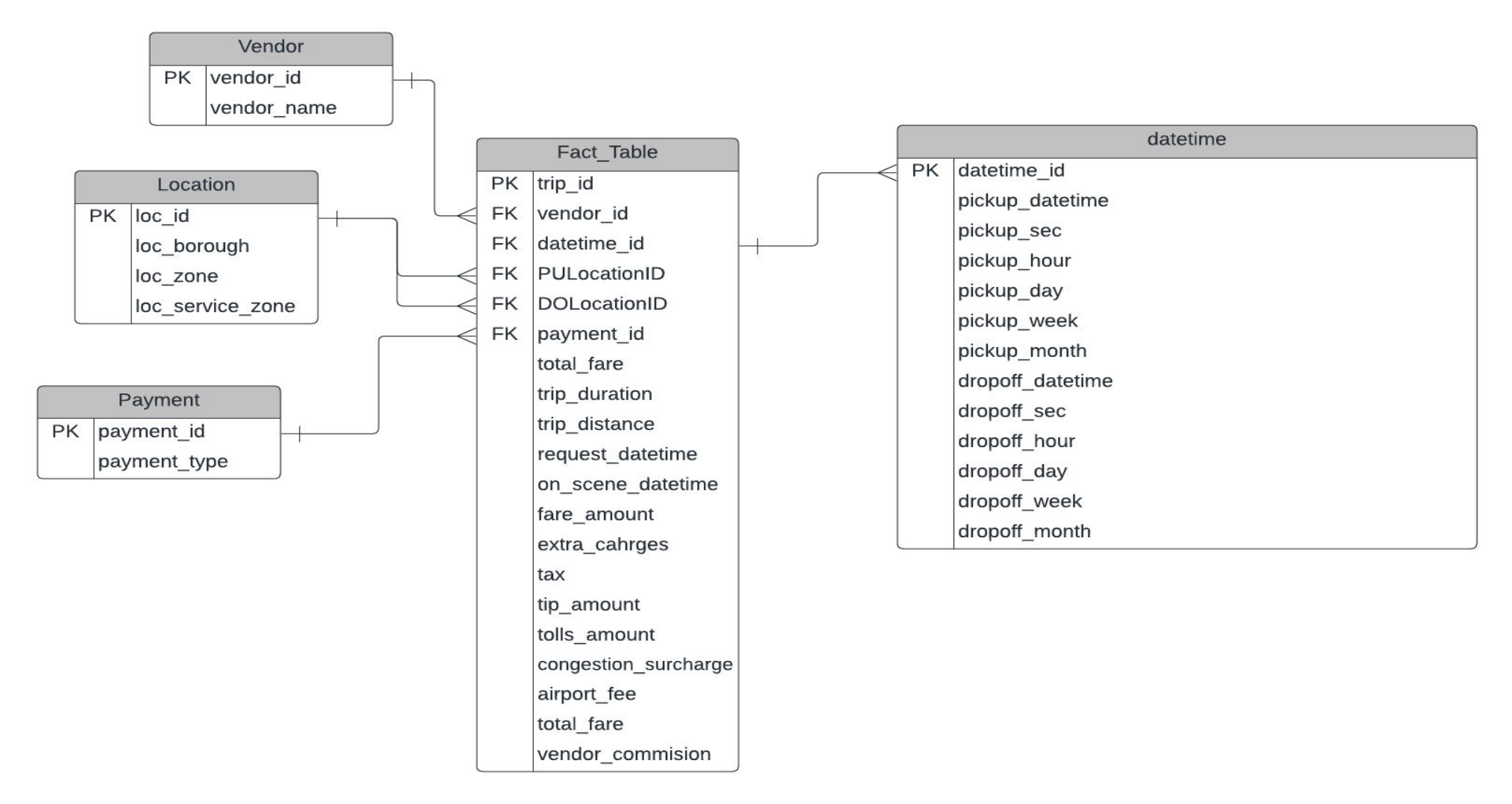




	A	B	C	D
1	LocationID	Borough	Zone	service_zone
2	1	EWR	Newark Airport	EWR
3	2	Queens	Jamaica Bay	Boro Zone
4	3	Bronx	Allerton/Pelham Gardens	Boro Zone
5	4	Manhattan	Alphabet City	Yellow Zone
6	5	Staten Island	Arden Heights	Boro Zone
7	6	Staten Island	Arrochar/Fort Wadsworth	Boro Zone
8	7	Queens	Astoria	Boro Zone
9	8	Queens	Astoria Park	Boro Zone
10	9	Queens	Auburndale	Boro Zone
11	10	Queens	Baisley Park	Boro Zone
12	11	Brooklyn	Bath Beach	Boro Zone
13	12	Manhattan	Battery Park	Yellow Zone
14	13	Manhattan	Battery Park City	Yellow Zone
15	14	Brooklyn	Bay Ridge	Boro Zone
16	15	Queens	Bay Terrace/Fort Totten	Boro Zone
17	16	Queens	Bayside	Boro Zone
18	17	Brooklyn	Bedford	Boro Zone
19	18	Bronx	Bedford Park	Boro Zone
20	19	Queens	Bellerose	Boro Zone
21	20	Bronx	Belmont	Boro Zone
22	21	Brooklyn	Bensonhurst East	Boro Zone
23	22	Brooklyn	Bensonhurst West	Boro Zone
24	23	Staten Island	Bloomfield/Emerson Hill	Boro Zone
25	24	Manhattan	Bloomingdale	Yellow Zone

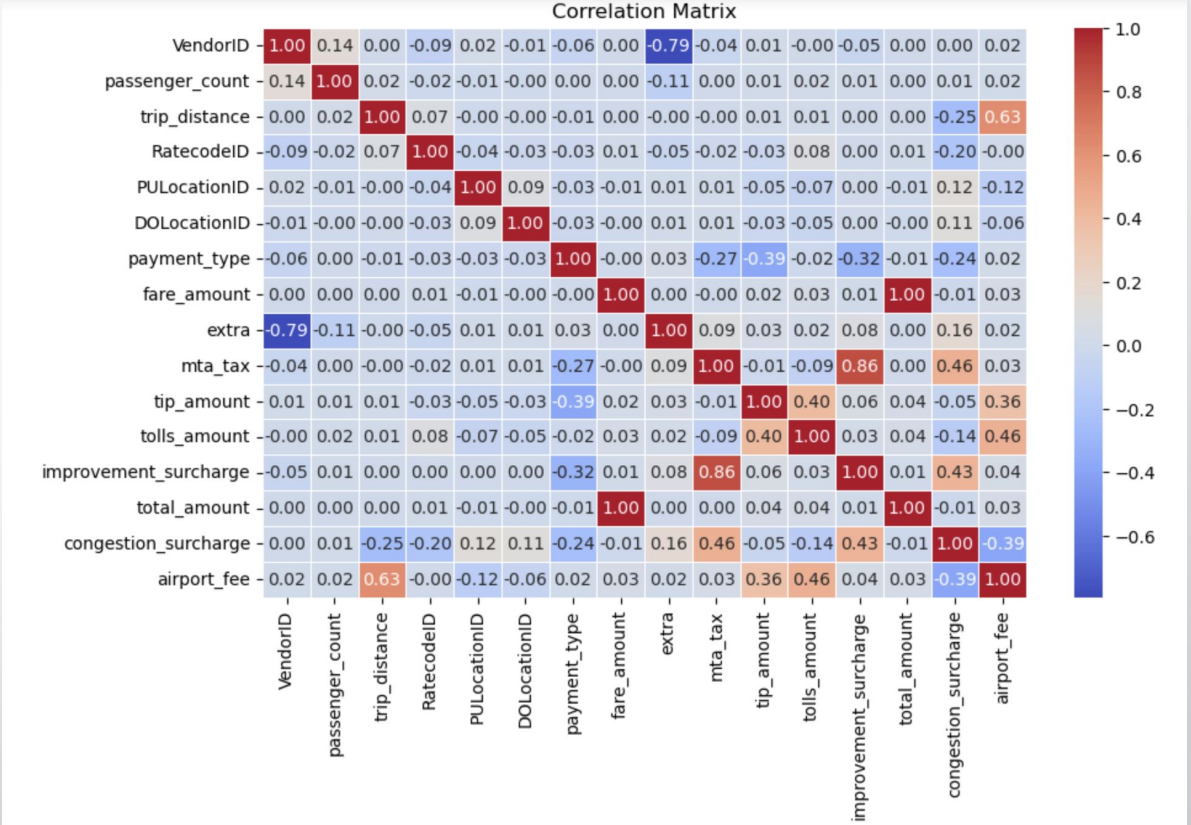
taxi+_zone_lookup

ER-DIAGRAM


Vendor
PK vendor_id
vendor_name
Location
PK loc_id
loc_borough
loc_zone
loc_service_zone
Payment
PK payment_id
payment_type
Fact_Table
PK trip_id
FK vendor_id
FK datetime_id
FK PULocationID
FK DOLocationID
FK payment_id
total_fare
trip_duration
trip_distance
request_datetime
on_scene_datetime
fare_amount
extra_cahrges
tax
tip_amount
tolls_amount
congestion_surcharge
airport_fee
total_fare
vendor_commision
datetime
PK datetime_id
pickup_datetime
pickup_sec
pickup_hour
pickup_day
pickup_week
pickup_month
dropoff_datetime
dropoff_sec
dropoff_hour
dropoff_day
dropoff_week
dropoff_month

EXPLORATORY DATA ANALYSIS

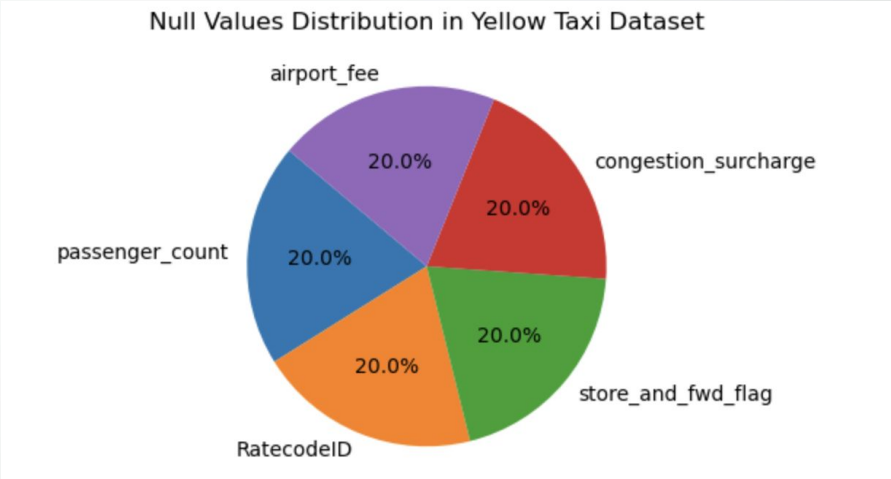
- positive correlation between "airport_fee" and "tolls_amount"
- strong positive correlation between "trip_distance" and "airport_fee"
- positive correlation between "mta_tax" and "Improved_surcharge"
- strong negative correlation between "tip_amount" and "payment_type"

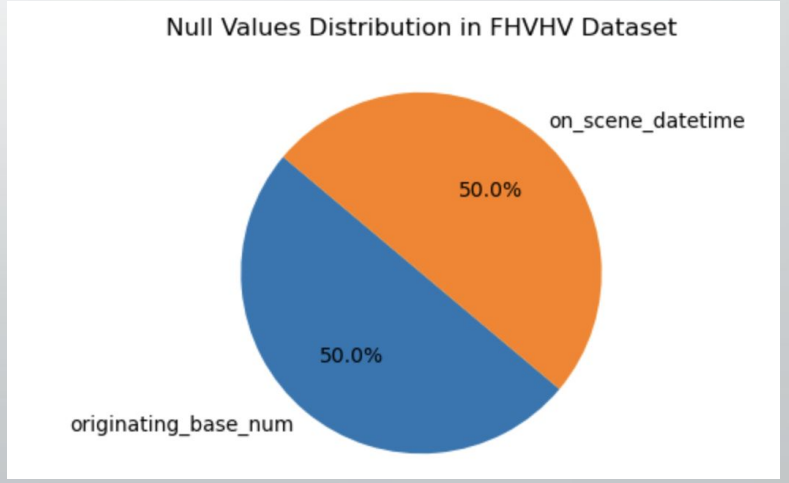


EXPLORATORY DATA ANALYSIS

Yellow Taxi dataset:
Columns such as airport_fee, congestion_surcharge, store_and_fwd_flag, RatecodeID, and passenger_count each contribute around 20% of nulls.

FHVHV vehicle dataset:
Originating_base_num and on_scene_datetime exhibit approximately 50% null values, indicating potential data quality issues.

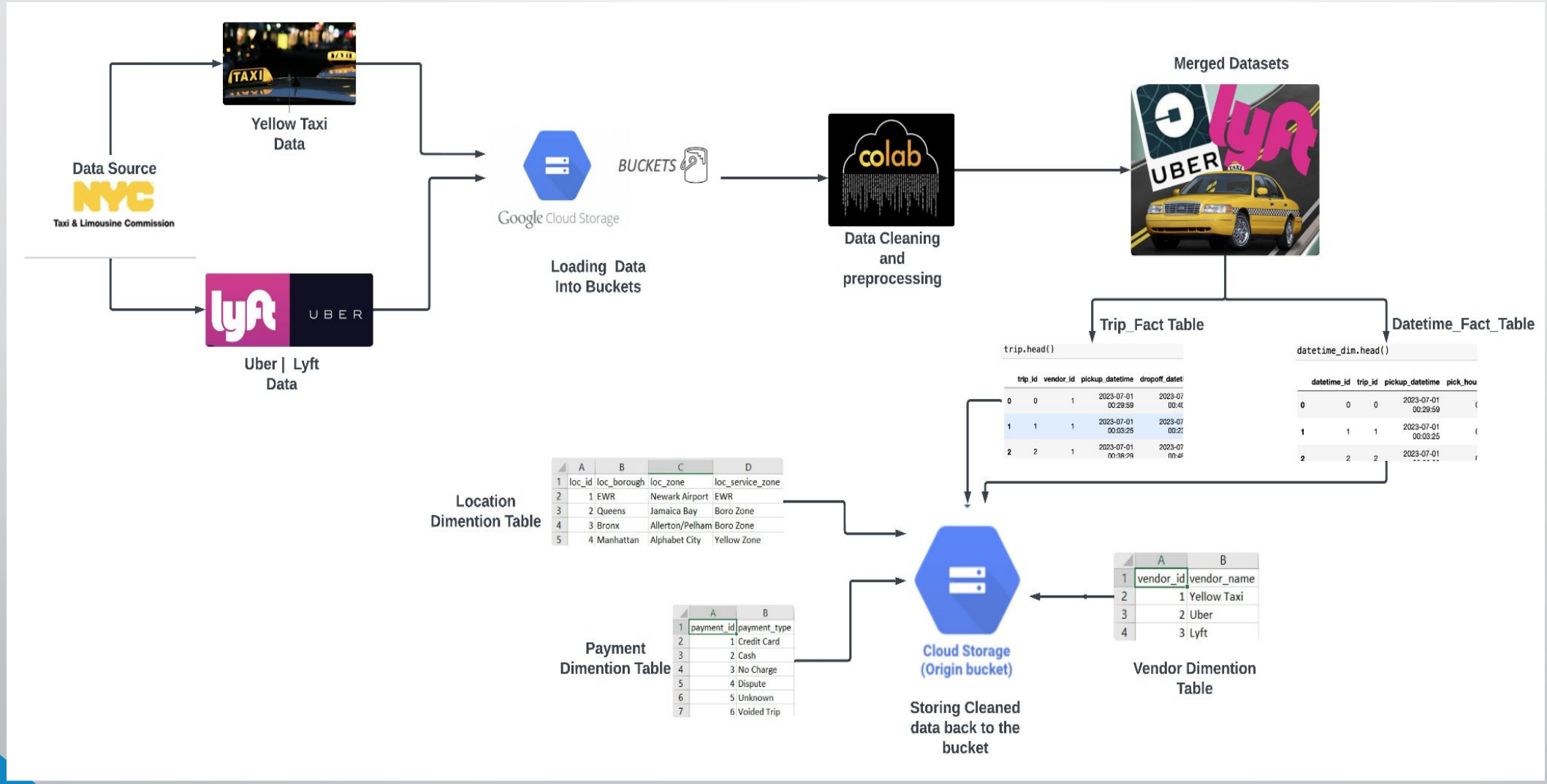




DATA CLEANING

- Standardizing Column Names
- Handling Missing Values
- Consolidating and Cleaning Categorical Values
- Deriving Relevant Attributes
- Creating Composite Features

DATA FLOW

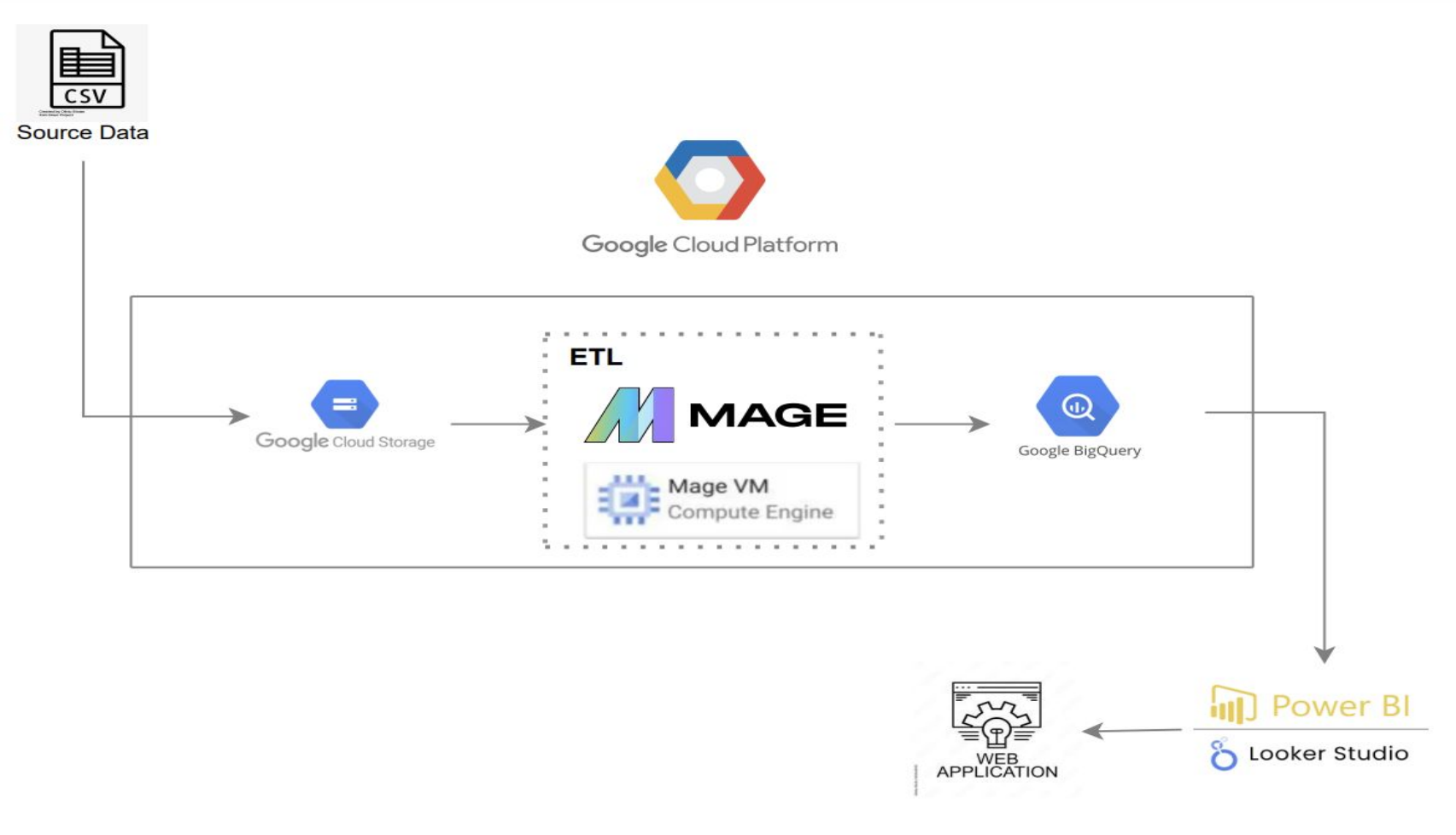


loc_id	loc_borough	loc_zone	loc_service_zone
1	EWR	Newark Airport	EWR
2	Queens	Jamaica Bay	Boro Zone
3	Bronx	Allerton/Pelham	Boro Zone
4	Manhattan	Alphabet City	Yellow Zone

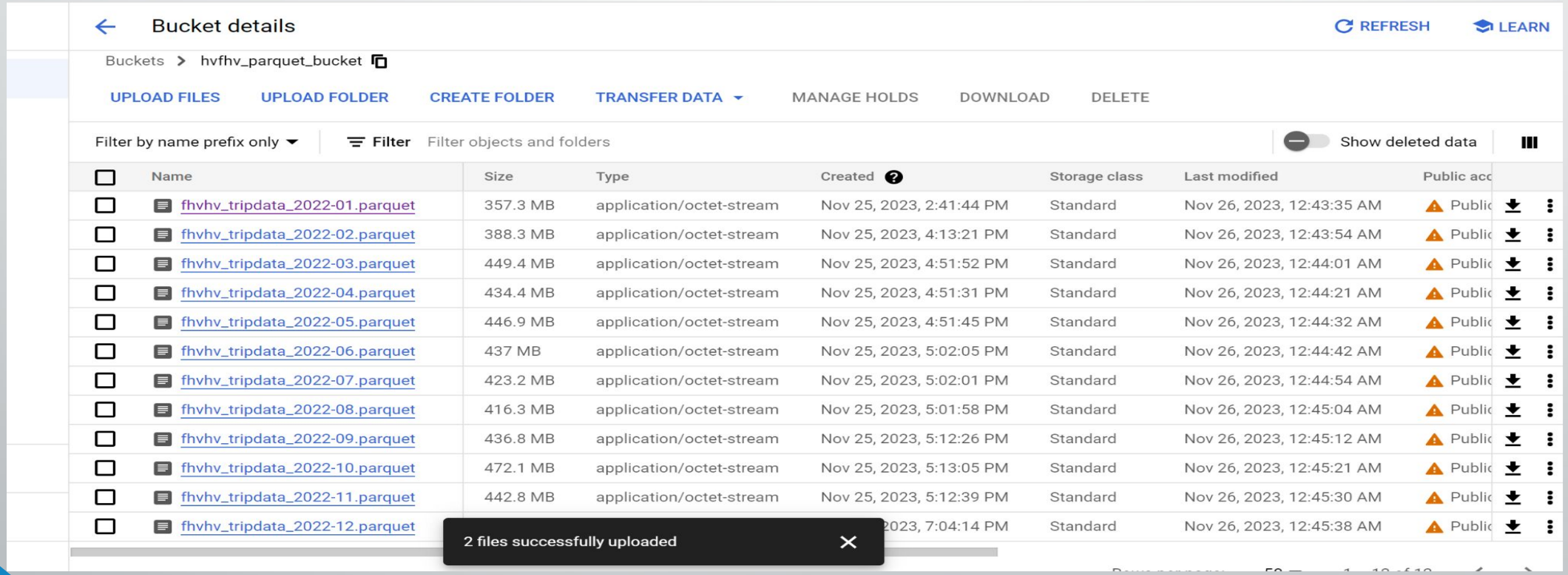
payment_id	payment_type
1	Credit Card
2	Cash
3	No Charge
4	Dispute
5	Unknown
6	Voided Trip

vendor_id	vendor_name
1	Yellow Taxi
2	Uber
3	Lyft

ARCHITECTURE

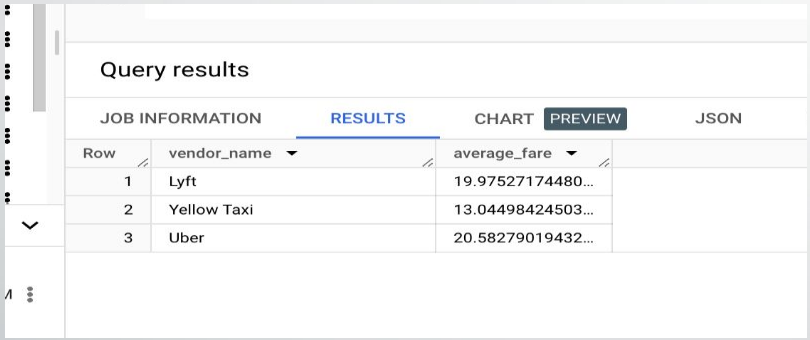


DATA LOADING AND TRANSFORMATION



BIGQUERY ANALYTICS

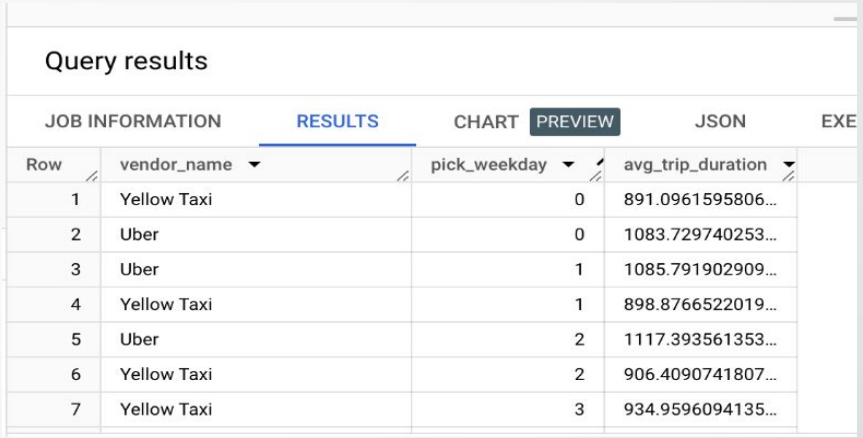
- The Analytics on Comparing the average fare amounts for Uber, Lyft and Yellow Taxi.
- Uber earns the highest avg_fare.



Query results

JOB INFORMATION | RESULTS | CHART PREVIEW | JSON

Row	vendor_name	average_fare
1	Lyft	19.97527174480...
2	Yellow Taxi	13.04498424503...
3	Uber	20.58279019432...



Query results

JOB INFORMATION | RESULTS | CHART PREVIEW | JSON | EXE

Row	vendor_name	pick_weekday	avg_trip_duration
1	Yellow Taxi	0	891.0961595806...
2	Uber	0	1083.729740253...
3	Uber	1	1085.791902909...
4	Yellow Taxi	1	898.8766522019...
5	Uber	2	1117.393561353...
6	Yellow Taxi	2	906.4090741807...
7	Yellow Taxi	3	934.9596094135...

- Identify the busiest days of the week. (0 = Monday, 1 = Tuesday…)

BIGQUERY ANALYTICS

- Identify and compare popular pickup and drop-off locations for all services.
- Uber is very popular in most of the main cities in the NYC

Query results

JOB INFORMATION | RESULTS | CHART PREVIEW | JSON | EXECUTION DET

Row	vendor_name	loc_borough	trip_count
1	Uber	Manhattan	9290025
2	Uber	Brooklyn	5773696
3	Yellow Taxi	Manhattan	4943083
4	Uber	Queens	4100961
5	Lyft	Manhattan	3168155
6	Uber	Bronx	2803114
7	Lyft	Brooklyn	2486339

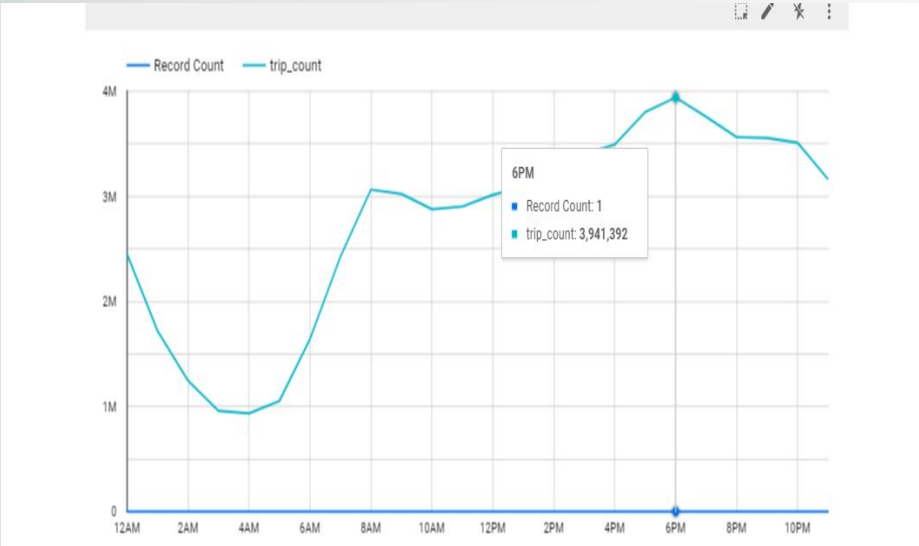
JOB INFORMATION | RESULTS | CHART PREVIEW

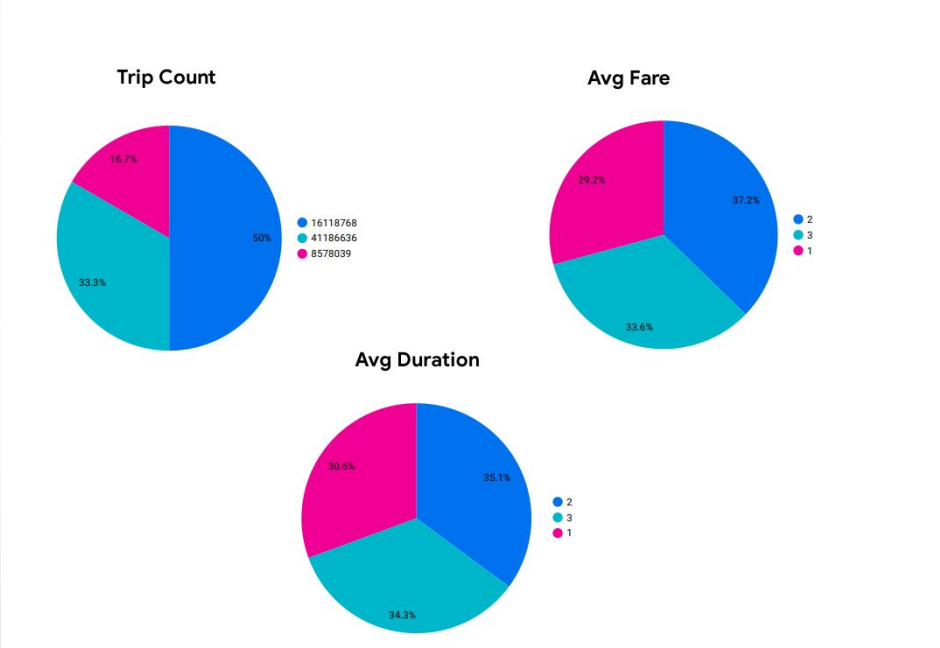
Row	vendor_name	avg_net_profit
1	Yellow Taxi	4.30924852324...
2	Uber	9.119872842311...
3	Lyft	9.409626982092...

- Calculate the net profit for each trip by subtracting driver pay and fees from the total fare.
- Here, lyft is leading in avg_net_profit

DATA VISUALIZATIONS

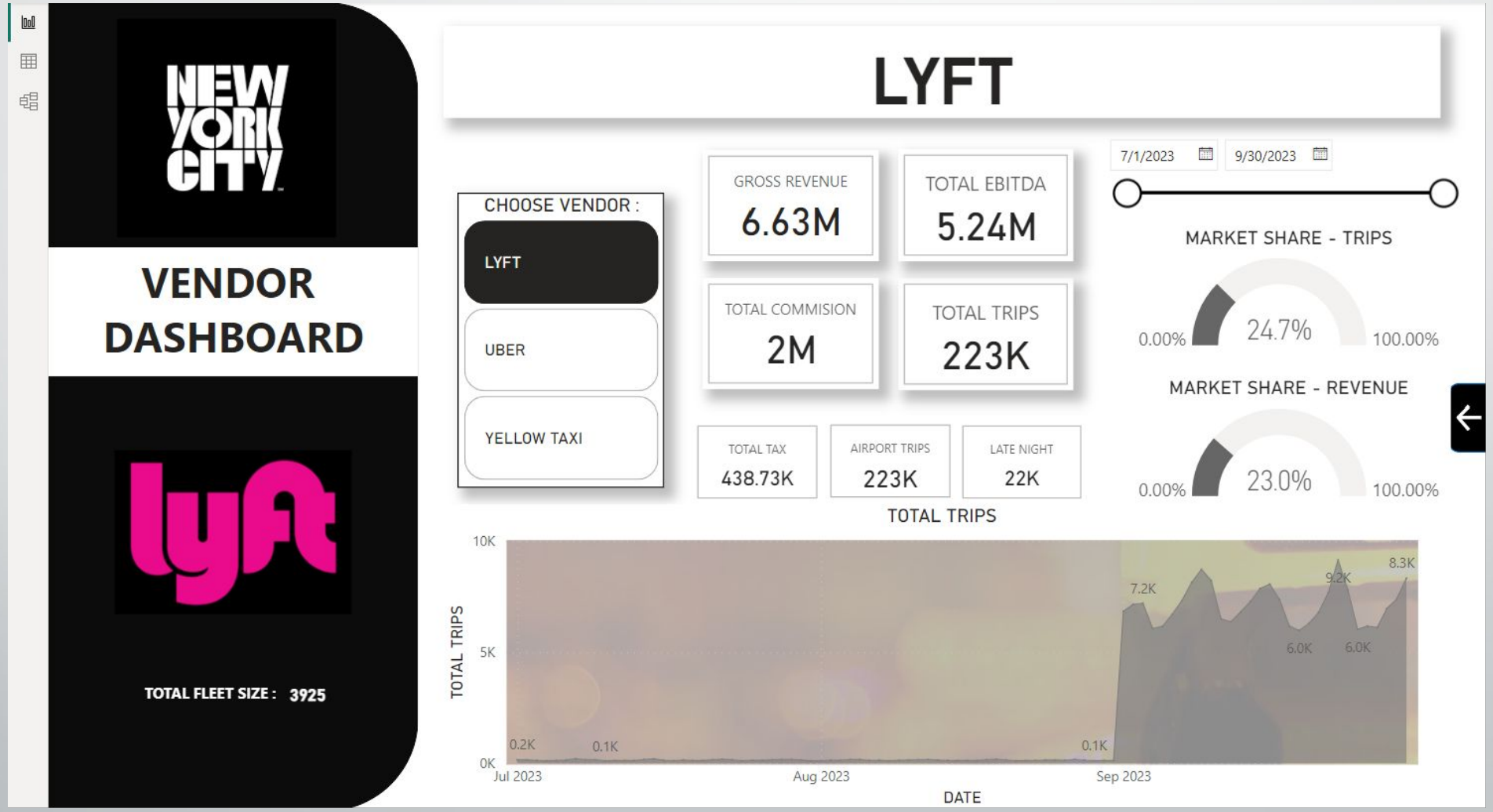
- Trip Count for every hour in a single day



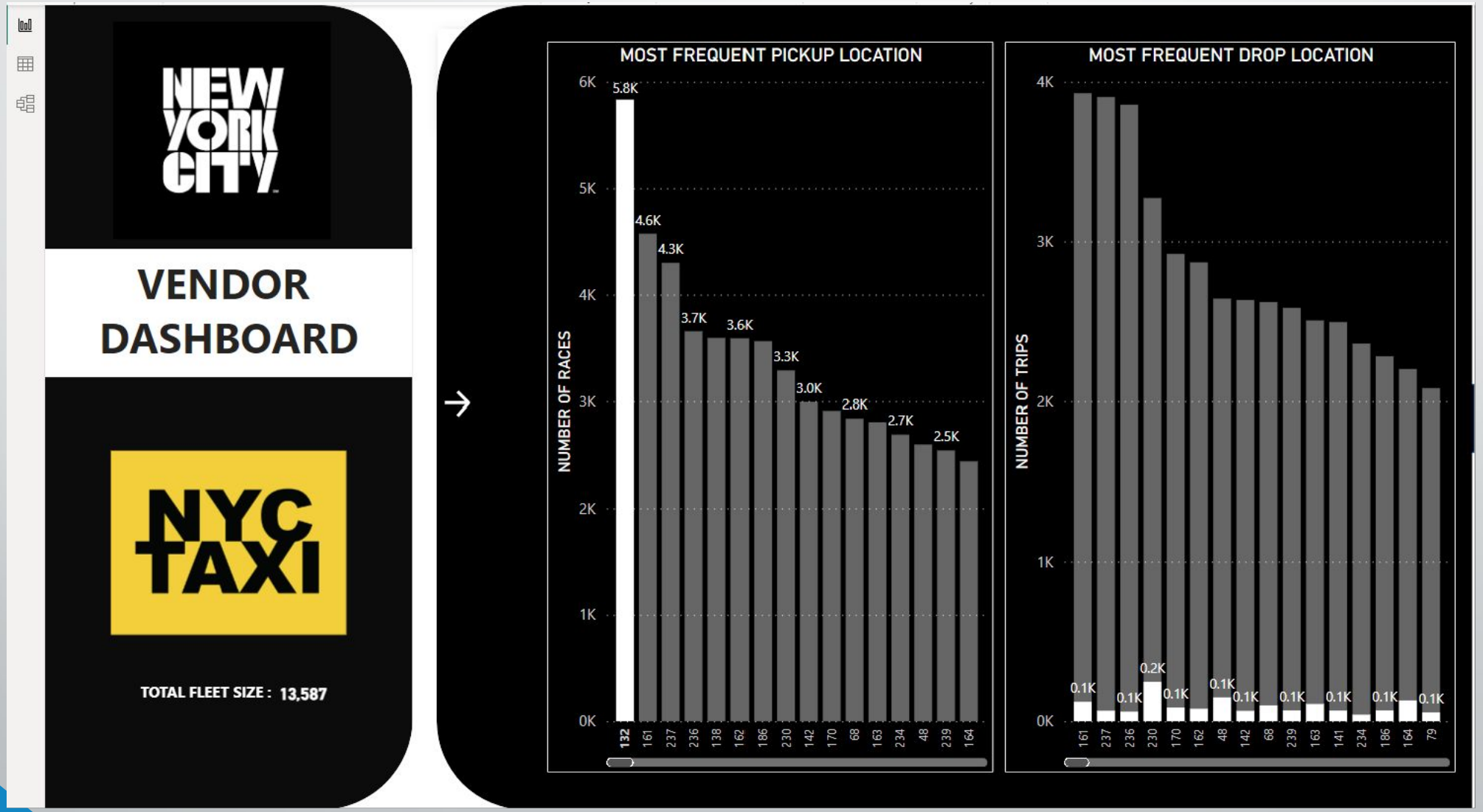


- Comparison of trip count, avg fare and duration for all three vendors
- **Results -** Uber Generates more average revenue, as the number of count and avg fare is higher than its competitors.

POWER BI DASHBOARD



POWER BI DASHBOARD



LIVE DEMO

FUTURE WORK

- Real-time data pipeline
- Make a replica of data on the cloud and deploy it to another location as a backup in case of a data loss.
- Create Mobile app to handle dashboard real time.

CONCLUSION

- Data driven decision making for vendors to better analyse each trips.
- Proficient time management by analysing ETA of drivers and trip duration for any hour during the day. leading to cost reductions, improved service quality, and a passenger-focused experience.

REFERENCES

Data Source: https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

THANK YOU!