Information Management Project

NYC TAXI COMPANY

Group 8

Rochan, Tanvi, Snehal, Muskaan, Pranav

AGENDA

01 REAL-WORLD COMPANY

02 TRANSACTION MANAGEMENT APPLICATIONS

03 DATA MODEL

04 DATA LAKE

05 ANALYSIS PATTERNS

06 REFLECTIONS

NYC TAXI COMPANY

DATASET

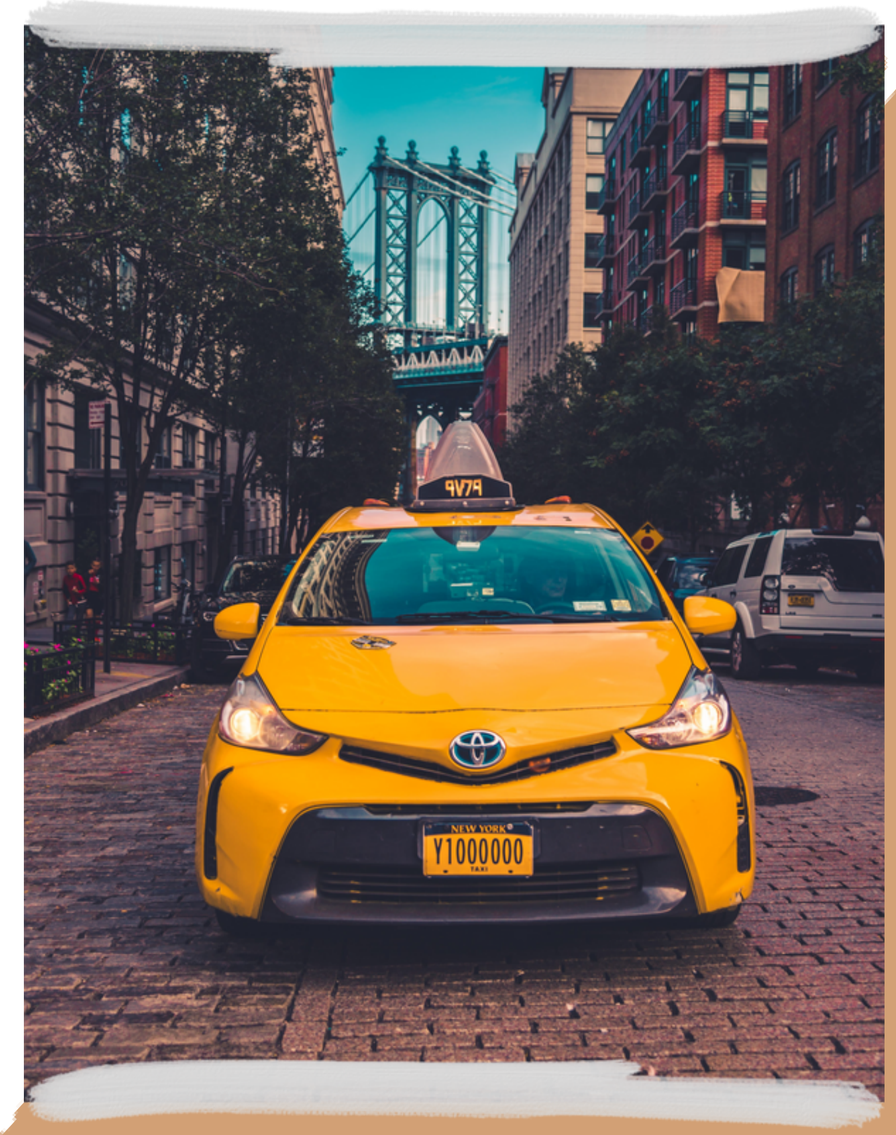
- 50 million+ entries
- Transactions, trips, distance, passenger count, etc.

MANAGEMENT STRATEGY

- Offensive
  - Competitive position, Flexible, MVOT

REASON FOR SELECTION

- Plentiful information
- Useful for other cities
- Learning experience



Dataset: *https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page*

TRANSACTION MANAGEMENT APPLICATIONS

OPERATIONS

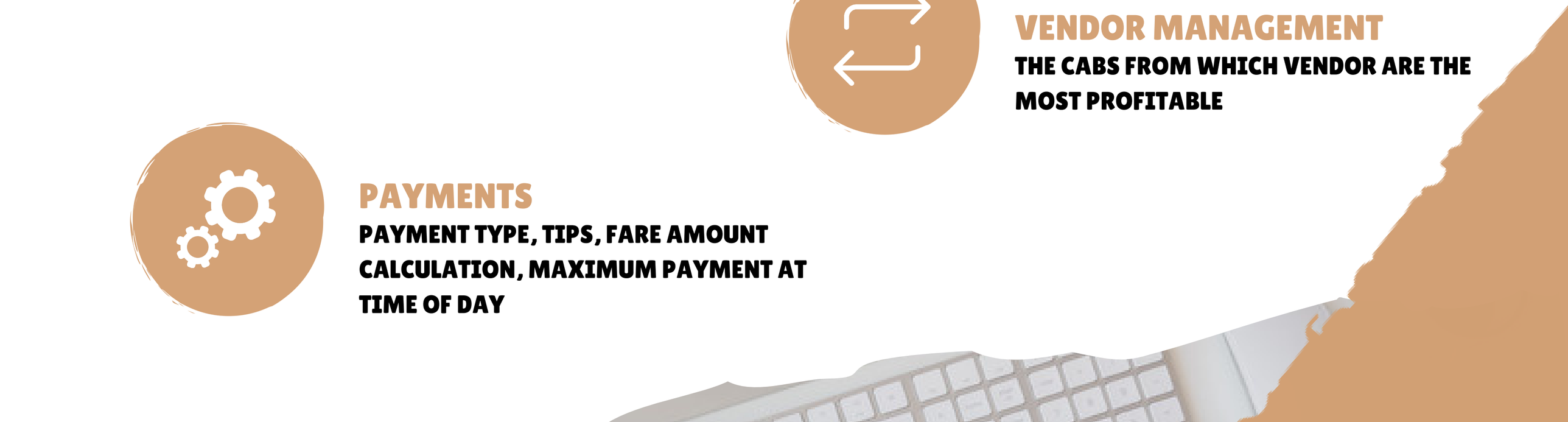
OPTIMAL ROUTES, PICK-UP DECISIONS, INVENTORY

VENDOR MANAGEMENT

THE CABS FROM WHICH VENDOR ARE THE MOST PROFITABLE

PAYMENTS

PAYMENT TYPE, TIPS, FARE AMOUNT CALCULATION, MAXIMUM PAYMENT AT TIME OF DAY

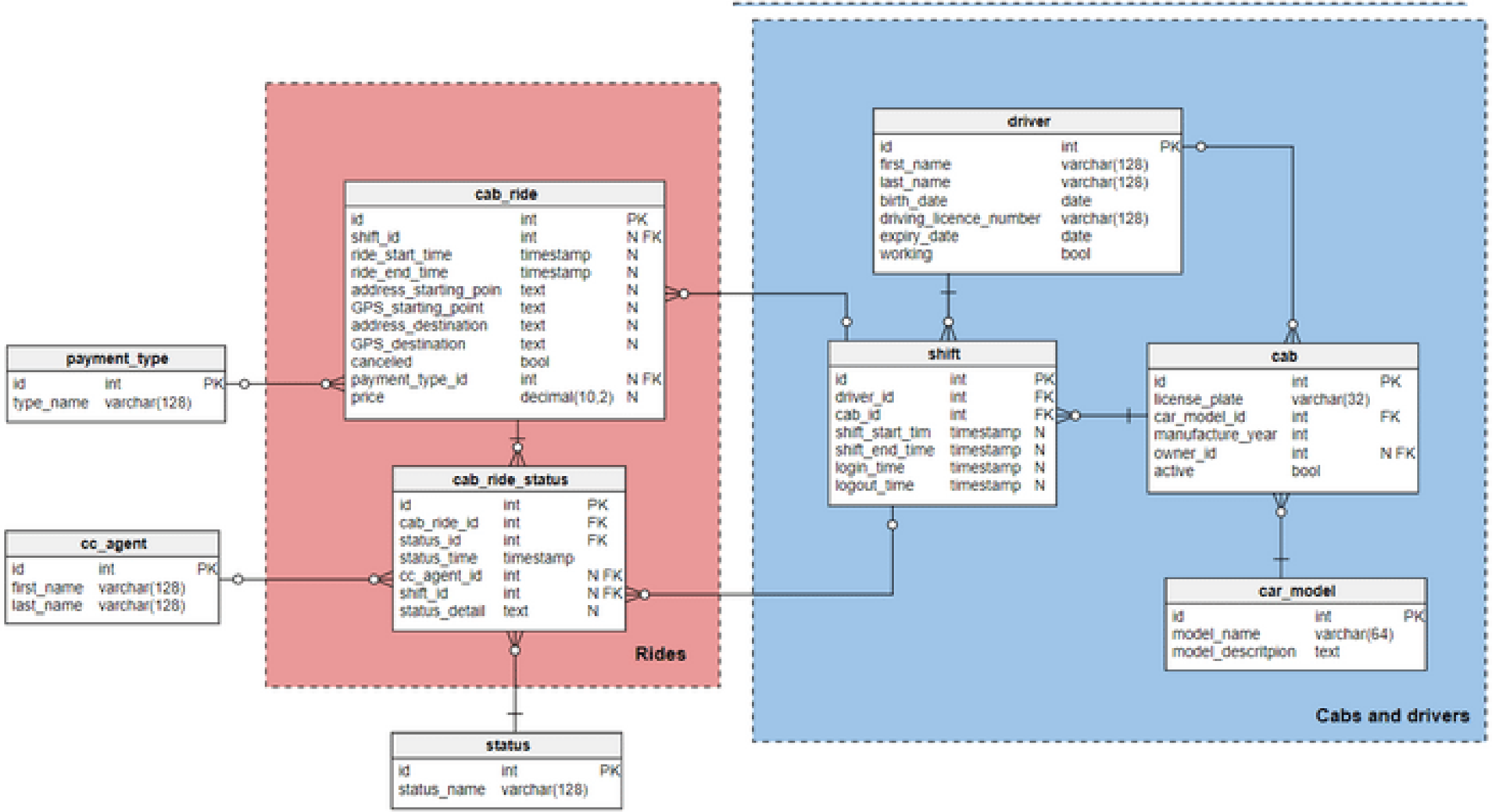


DATA MODEL

CONCEPTUAL MODEL

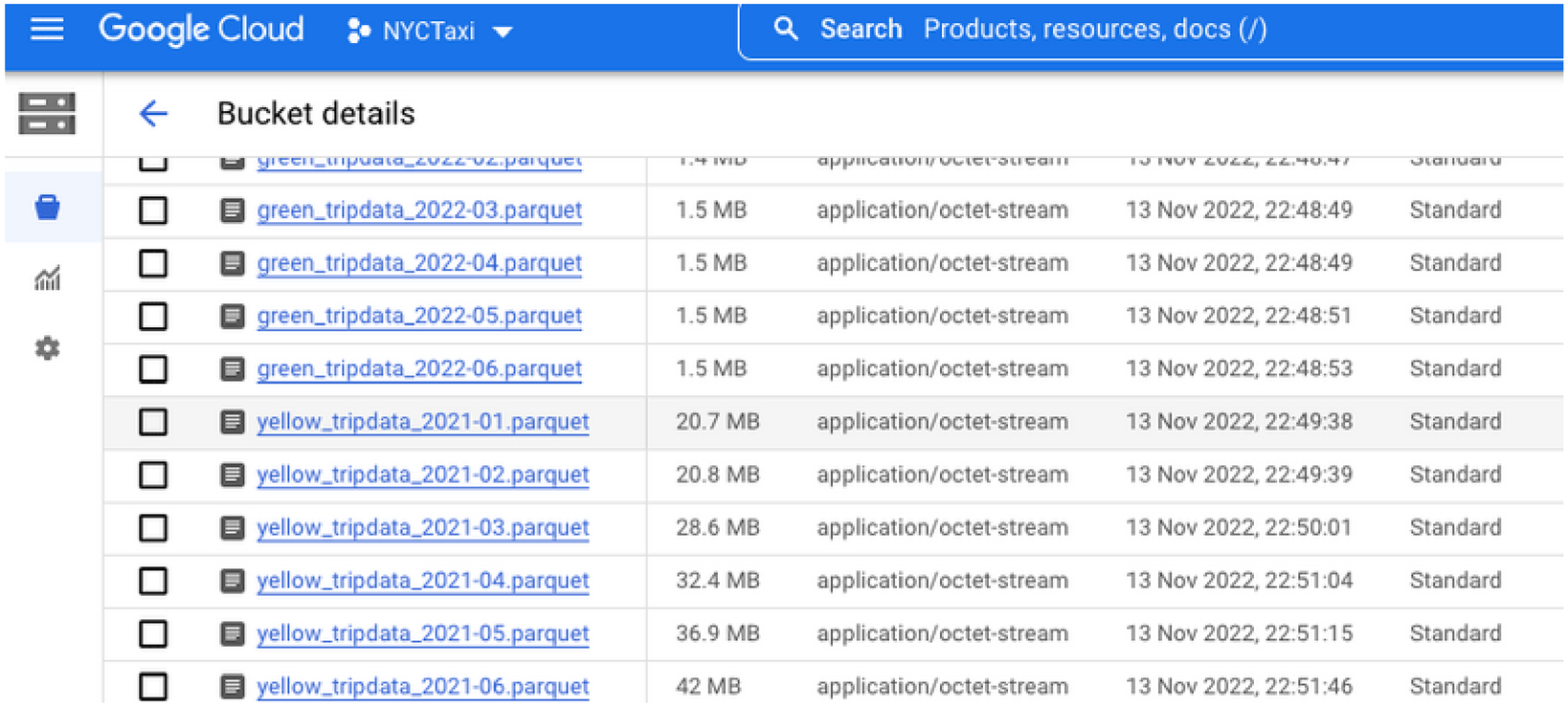
- We got data about customer rides including pickup time, distance, location, etc.
- Data is processed to analyze the rides and patterns in the data.
- Based on patterns, we can decide where to put drivers and cabs for the maximum demand filled.
- We can also decide how much to supply to locations.
- We can also decide what time of day and what day of the week is the most lucrative, and where.

LAYOUT: LOGICAL & PHYSICAL MODELS



DATA LAKE

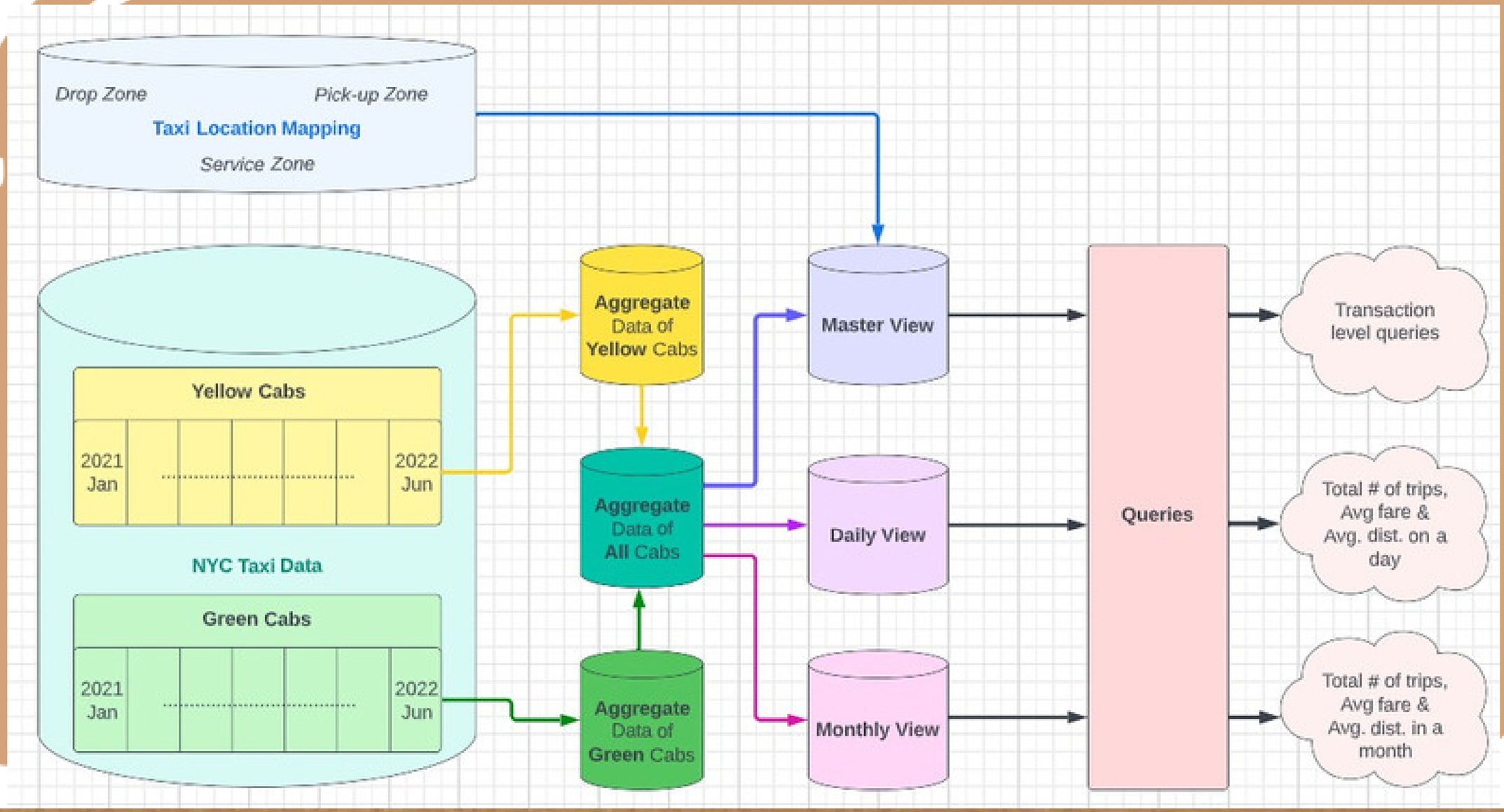
Google Cloud Storage as Data Lake



green_tripdata_2022-03.parquet	1.5 MB	application/octet-stream	13 Nov 2022, 22:48:49	Standard
green_tripdata_2022-04.parquet	1.5 MB	application/octet-stream	13 Nov 2022, 22:48:49	Standard
green_tripdata_2022-05.parquet	1.5 MB	application/octet-stream	13 Nov 2022, 22:48:51	Standard
green_tripdata_2022-06.parquet	1.5 MB	application/octet-stream	13 Nov 2022, 22:48:53	Standard
yellow_tripdata_2021-01.parquet	20.7 MB	application/octet-stream	13 Nov 2022, 22:49:38	Standard
yellow_tripdata_2021-02.parquet	20.8 MB	application/octet-stream	13 Nov 2022, 22:49:39	Standard
yellow_tripdata_2021-03.parquet	28.6 MB	application/octet-stream	13 Nov 2022, 22:50:01	Standard
yellow_tripdata_2021-04.parquet	32.4 MB	application/octet-stream	13 Nov 2022, 22:51:04	Standard
yellow_tripdata_2021-05.parquet	36.9 MB	application/octet-stream	13 Nov 2022, 22:51:15	Standard
yellow_tripdata_2021-06.parquet	42 MB	application/octet-stream	13 Nov 2022, 22:51:46	Standard

ETL Diagram

ETL DIAGRAM



ETL Diagram: *https://lucid.app/documents/view/7186d4ea-04a2-41e3-931b-3c73e85e3810*

ETL

ORIGINAL DATABASE

- Yellow Taxis data (across months)
- Green Taxis data (across months)
- Taxi Location Mapping

AGGREGATION

- The yellow and green taxi data across the months was aggregated into respective data lakes
- Then the data across both yellow and green taxis was aggregated, mapping the locations into a master view

PERIODICAL DATA

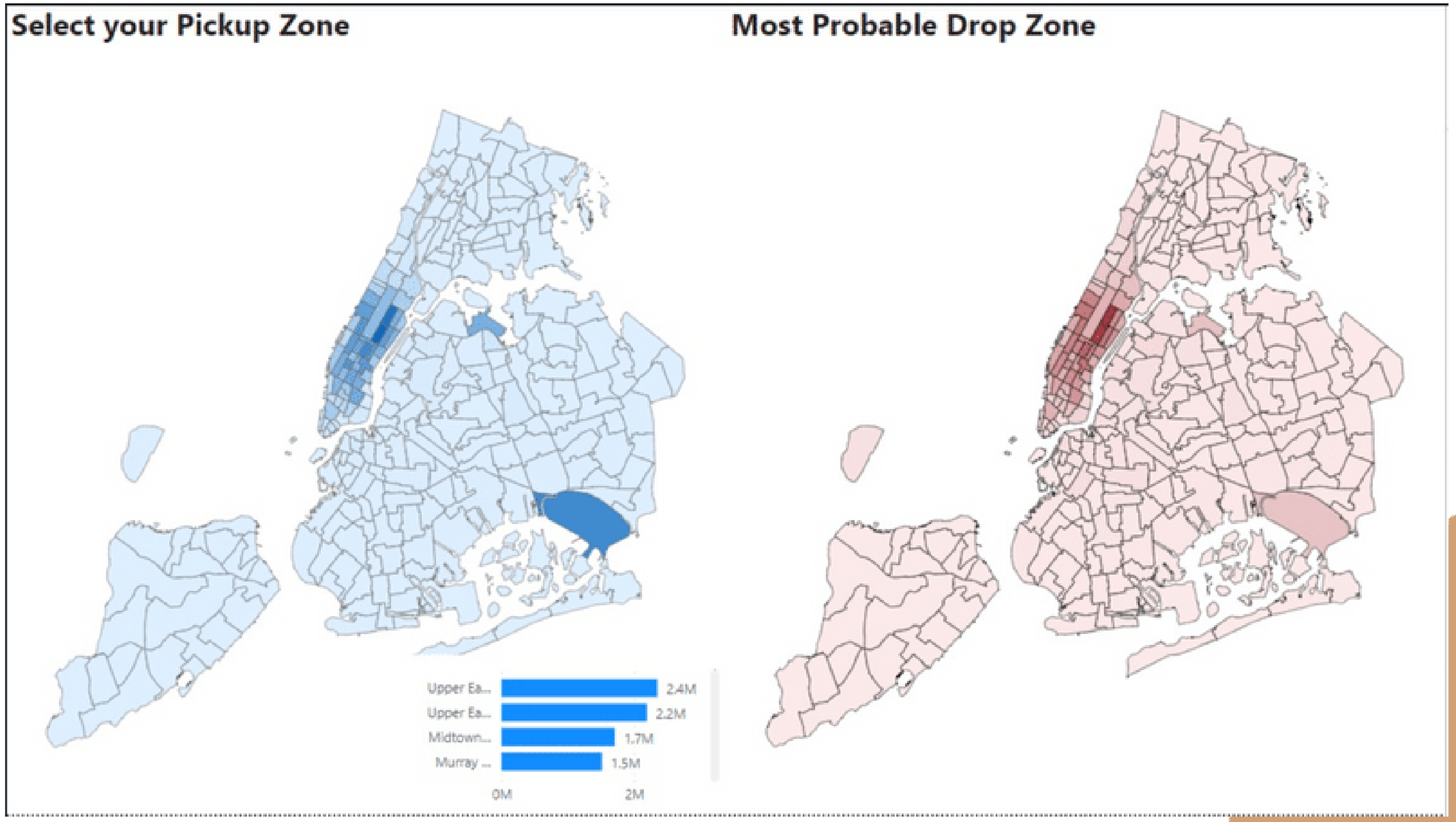
- The aggregate data was also transformed to represent daily views and monthly views for better querying.

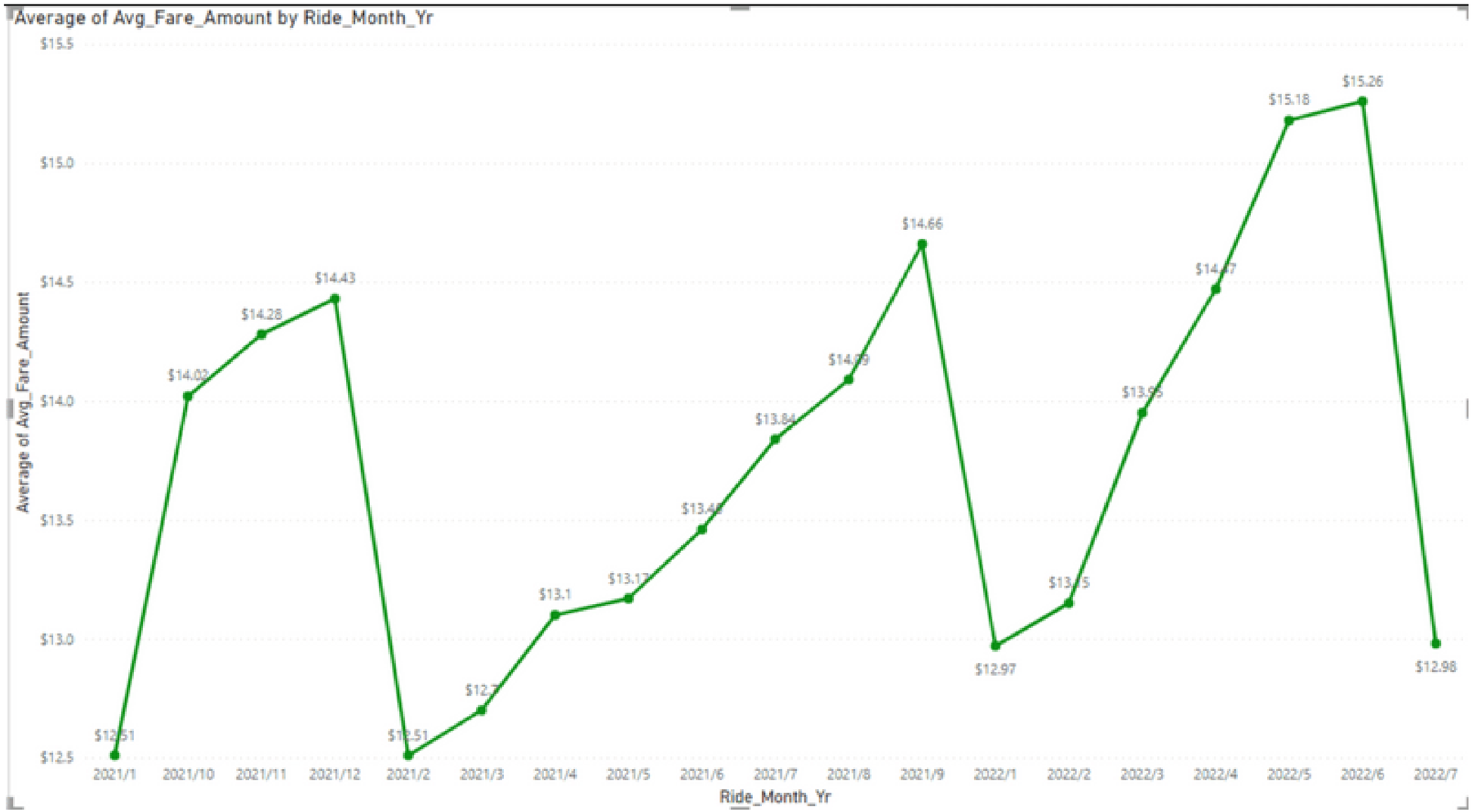
QUERIES

- The database can now handle transaction level queries as well as queries for daily or monthly report

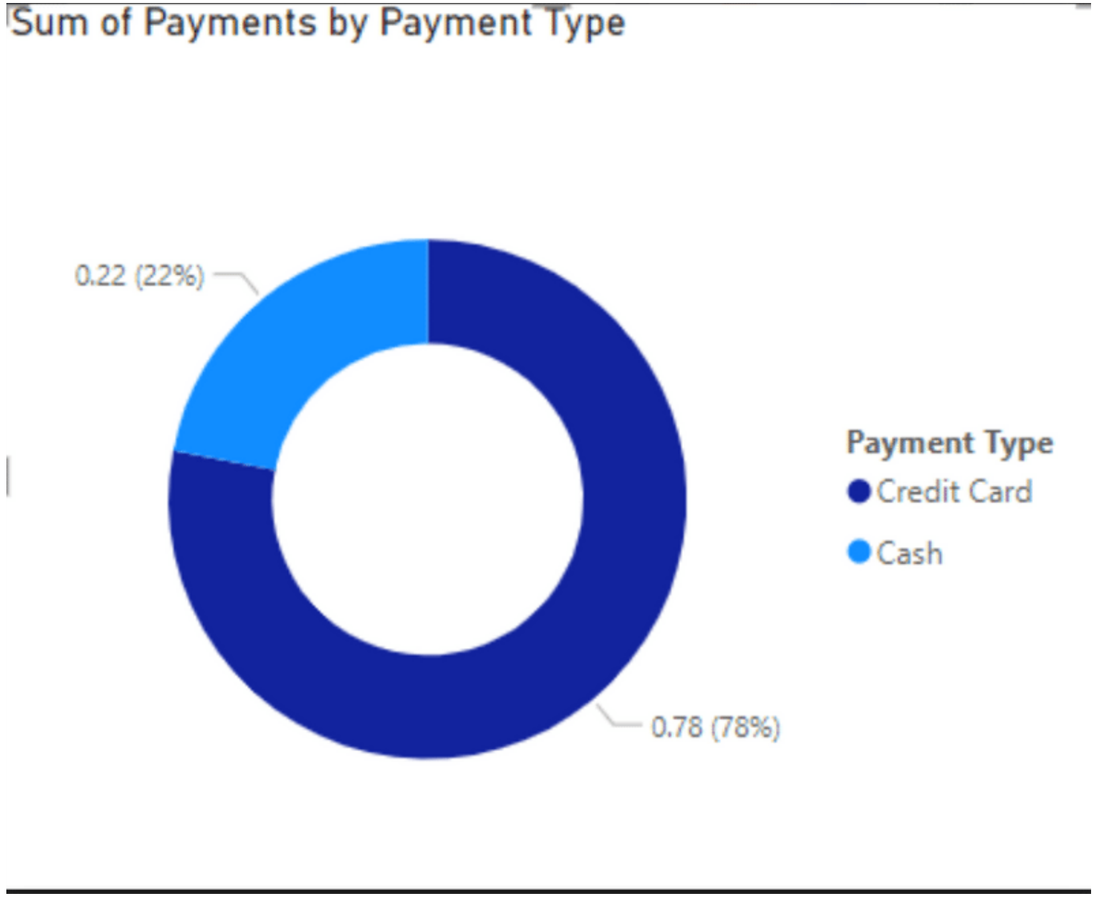
ANALYSIS

This chart captures the popular pick-up and drop zones based on the data





This chart captures the trend of the Average Fare Amount across the months.

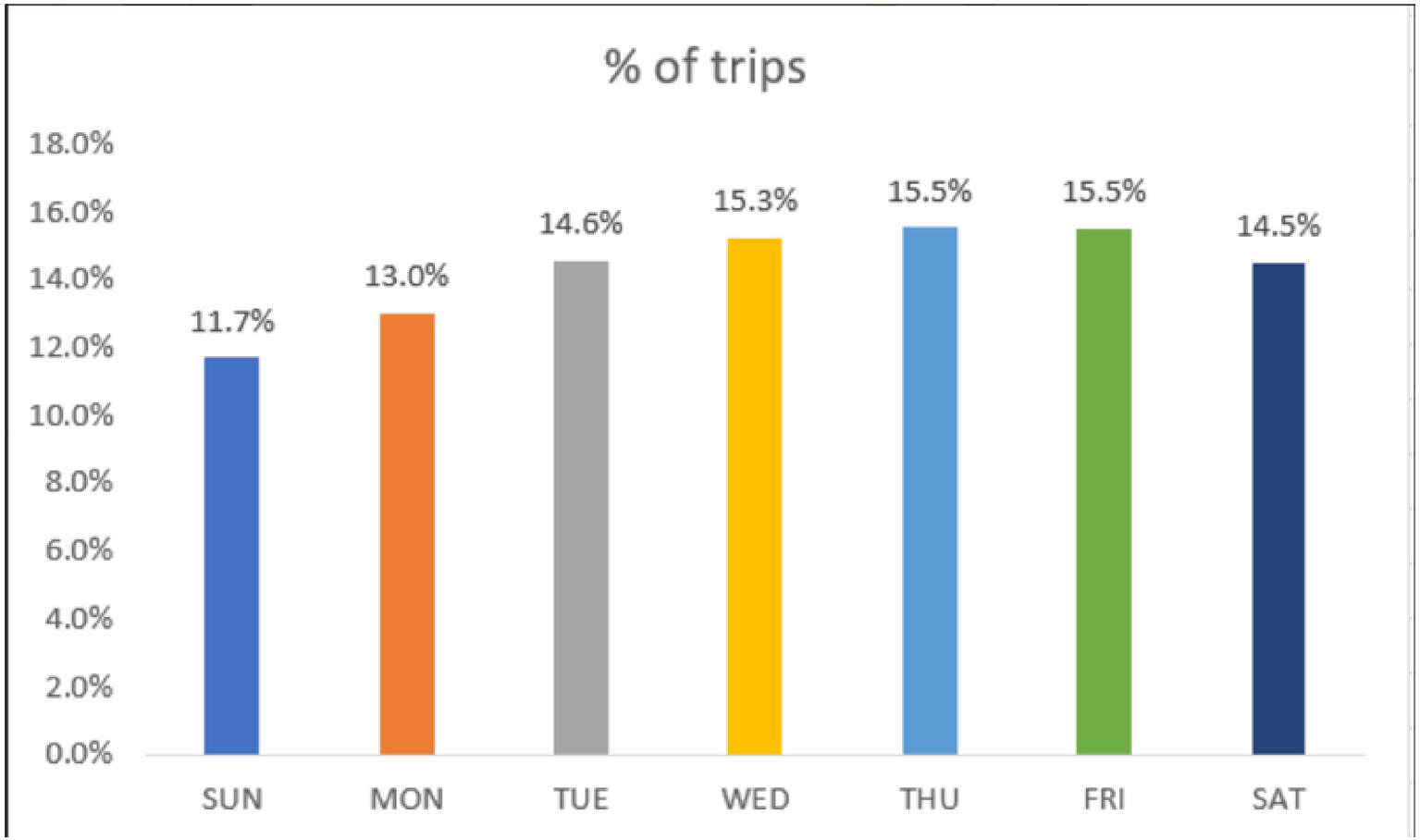


This chart captures the Sum of Payments by Payment Type at the Transaction Level:
Credit card v. cash

pickup_hour	1	2	3	4	5	6	7	Total
0	12.26%	10.85%	10.89%	11.36%	11.68%	11.98%	12.32%	**11.62%**
1	12.32%	10.32%	10.33%	10.80%	11.31%	11.79%	12.45%	**11.33%**
2	12.21%	9.83%	9.52%	10.09%	10.67%	11.31%	12.34%	**10.85%**
3	11.89%	9.38%	8.64%	9.17%	9.75%	10.43%	11.96%	**10.17%**
4	10.97%	9.09%	8.56%	8.65%	9.34%	9.81%	10.67%	**9.58%**
5	9.76%	9.40%	9.47%	9.44%	9.78%	9.93%	9.30%	**9.58%**
6	9.96%	10.53%	10.81%	10.92%	10.95%	10.83%	9.97%	**10.57%**
7	10.03%	11.49%	11.81%	11.89%	11.87%	11.57%	10.30%	**11.28%**
8	10.70%	11.77%	12.03%	12.09%	12.10%	11.78%	10.85%	**11.62%**
9	11.15%	11.38%	11.60%	11.67%	11.70%	11.43%	11.29%	**11.46%**
10	11.42%	10.97%	11.17%	11.22%	11.27%	11.11%	11.45%	**11.23%**
11	11.68%	10.88%	11.01%	11.14%	11.16%	11.07%	11.61%	**11.22%**
12	11.77%	10.91%	10.99%	11.07%	11.14%	11.16%	11.66%	**11.24%**
13	11.75%	10.93%	11.01%	11.06%	11.11%	11.13%	11.63%	**11.23%**
14	11.67%	11.07%	11.12%	11.17%	11.22%	11.23%	11.52%	**11.28%**
15	11.65%	11.17%	11.23%	11.30%	11.31%	11.28%	11.54%	**11.35%**
16	11.74%	11.26%	11.31%	11.38%	11.40%	11.34%	11.60%	**11.43%**
17	11.82%	11.76%	11.88%	11.94%	11.91%	11.76%	11.77%	**11.83%**
18	11.87%	12.11%	12.29%	12.30%	12.24%	11.99%	11.88%	**12.10%**
19	11.84%	12.11%	12.33%	12.35%	12.28%	12.06%	11.96%	**12.13%**
20	11.82%	12.22%	12.47%	12.47%	12.40%	12.00%	11.83%	**12.17%**
21	11.90%	12.36%	12.69%	12.72%	12.62%	12.14%	11.99%	**12.35%**
22	11.73%	12.16%	12.49%	12.64%	12.62%	12.32%	12.21%	**12.31%**
23	11.19%	11.43%	11.89%	12.10%	12.28%	12.31%	12.26%	**11.92%**
Total	**11.46%**	**11.06%**	**11.15%**	**11.29%**	**11.42%**	**11.41%**	**11.51%**	**11.33%**

This heatmap captures the cross-section of the average percentage of tips by the hour of the day v. the day of the week

The percentage of tipping is the lowest in the early hours of the week, especially during the work week



This graph captures the proportion of the % of the trips across each day of the week.

There is a higher percentage of trips at the latter half of the work week - Wednesday through Friday.

REFLECTIONS

WHAT WE LEARNED

- How to deal with an immense amount of data
- Deal with data inconsistency issues in real-life data

MOST VALUABLE

- Different levels of data (transactional, daily, monthly) give different insights

FUTURE SCOPE

- Connect Data warehouse to PowerBi using the BQ connector for Business 360 view
- More In-depth analysis with data from additional raw sources

THANK YOU!

Questions?

APPENDIX

TEAM WORK DISTRIBUTION

ROCHAN
DATA MODELING + ANALYSIS

MUSKAAN
DATA MODELING + ANALYSIS

TANVI
PRESENTATION + REPORT

PRANAV
PRESENTATION + REPORT

SNEHAL
ERD + REPORT