Kyongmin Song
Operations Analytics and Data Execution Lead Candidate
November 10, 2022

Summary

The database for this assessment was the Chicago Taxi Trips dataset, available via BigQuery. The database provides information on each taxi trip taken within Chicago. The schema includes fields such as taxi company, timestamps, fare, miles traveled, and various information about payments. Only a portion of the fields were queried as part of the assessment.

For part 1, the general approach for establishing SQL queries for prompts A and B were similar. The steps included leveraging the WITH clause associated with common table expression. This was done so that data could be grouped, and window functions properly leveraged in a systematic manner. Finally, a subquery was used to ensure that the maximum or minimum month over month value could be identified for each company. The SQL code and resulting query are shown below.

In part 2, candidates were given creative freedom to an additional insight or trend that was pertinent to part 1. One important aspect of examining month-to-month data is to see if there are any broader time-based trends in the data, especially with the fields under investigation. For this reason, I wanted to display a time-series visualization to observe whether there was seasonality to fields such as fare and miles per trip. The visualizations revealed clear cyclical trends by month, but also revealed the drastic change in trip data during years 2020-2022, likely due to changes in travel habits from COVID-19.

Of course, with additional time, more insights could be gathered, but the visualizations do provide a starting point for the investigation and demonstrate an aspect of how I assess data, as well as my capabilities in bringing SQL queries into Python for additional analysis. The HTML file of my Python code is included in the email as well. Thank you for your time and for your consideration for me in this role.

Part 1

Which three distinct taxi companies had the largest month-over-month increase in trips, and what were those months and trip amounts?

SQL Code:

```
#StandardSQL
with t1
as
(
  select company, format_datetime("%m-
%Y", trip_start_timestamp) as trip_monthyear, count(*) as trip_count
  from `bigquery-public-data.chicago_taxi_trips.taxi_trips`
  where company is not null
  group by company, trip_monthyear
),

t2 as
(
  select company, trip_monthyear, trip_count, trip_count - lag(trip_count,1) over (order by tr
ip_monthyear) / trip_count as mm_growth
  from t1
)
```

```sql
select company, trip_monthyear, trip_count, mm_growth
from
(
  select company, trip_monthyear, trip_count, mm_growth, max(mm_growth) over (partition by com
pany) as max_mm_growth
  from t2
) t2
where t2.mm_growth = t2.max_mm_growth
order by max_mm_growth desc
limit 3
```

Result:

Row	company	trip_monthyear	trip_count	mm_growth
1	Taxi Affiliation Services	05-2014	924982	924981.841...
2	Flash Cab	10-2014	437031	437030.998...
3	Yellow Cab	10-2015	293286	293285.758...

Which three distinct taxi companies had the largest month-over-month decrease in fare-per-mile, and what were those months and fare-per-mile values?

SQL Code:

```sql
with t1
as
(
  select company, format_datetime("%m-
%Y", trip_start_timestamp) as trip_monthyear, avg(fare / trip_miles) as avg_fare_per_mile
  from `bigquery-public-data.chicago_taxi_trips.taxi_trips`
  where trip_miles > 0 AND company is not null
  group by company, trip_monthyear
),

t2 as
(
  select company, trip_monthyear, avg_fare_per_mile, avg_fare_per_mile - lag(avg_fare_per_mile
,1) over (order by trip_monthyear) / avg_fare_per_mile as change_avg_fair_per_mile
  from t1
  order by change_avg_fair_per_mile asc
)

select company, trip_monthyear, avg_fare_per_mile, change_avg_fair_per_mile
from
(
  select company, trip_monthyear, avg_fare_per_mile, change_avg_fair_per_mile, min(change_avg_
fair_per_mile) over (partition by company) as min_fare_per_mile
  from t2
) t2
where t2.change_avg_fair_per_mile = t2.min_fare_per_mile AND change_avg_fair_per_mile is not n
ull
order by min_fare_per_mile asc
```

```
limit 3
```

Result:

Row	company	trip_monthyear	avg_fare_pe...	change_avg...
1	2733 - 74600 Benny Jona	05-2022	3.54992544...	-471.92110...
2	Top Cab	06-2022	7.87862608...	-436.10502...
3	Sun Taxi	04-2022	11.0368499...	-214.88033...

Part 2

Considering the context of the questions from part I, conduct an additional analysis using the same dataset and design a report that provides at least one additional insight, a trend or any other relevant detail that piques your interest.

For time-series analysis, I queried the average fare, trip miles, and fare per miles by month and year over the entire dataset. I wanted to see whether there were any seasonal trends for these parameters that could provide insight to month-to-month metrics. If there were clear trends, it could provide valuable insights into patterns of growth and decline for key performance indicators. It should be noted that since these were average values across the dataset, more investigation should be done into descriptive statistics of these parameters, such as spread. A boxplot visualization example is also shown below.

Two quick insights can be found in the visualizations provided. There does certainly appear to be seasonality to the data. This is especially consistent when the pandemic years are excluded from the analysis. In general, fares and miles driven are lowest in the winter months, with another drop in July and August. In the pandemic years, the fares and miles per trip are quite higher than previous years, and do not share in a high degree of consistency.

SQL Code:

```
select  format_datetime("%m", trip_start_timestamp) as trip_month,  format_datetime("%Y", trip
_start_timestamp) as trip_year, avg(fare) as avg_fare, avg(trip_miles) as avg_miles, avg(fare
/ trip_miles) as avg_fare_per_mile
from `bigquery-public-data.chicago_taxi_trips.taxi_trips`
where fare >0 AND trip_miles > 0
group by trip_month, trip_year
order by trip_month, trip_year
```

Visualizations: