IDS706 Project #5: Cloud SQL

Nansu Wang

1. Summary

- In this assignment, I use Big Query Planform in GCP to build some useful insights.
- The response variable is trip miles for each taxi order, the predictors I need to explore are the total cost of the trip, taxi company, and payment method.
- An interaction may exist between payment methods and taxi companies. However, there is not enough data in some splitting categories, so this interaction term will not be included in the final model.
- Multicollinearity issue exists between total cost and total tips. Thus, only one of them will be added to my final model.
- Big Query Platform could be fast in SQL queries. Also, it is efficient to make plots using internal UI. Finding insights using the Big Query Platform is efficient and enjoyable.

2. Data

- I select a public dataset in GCP. It is about the Chicago's taxi trip.
- The response variable is trip miles for each taxi order.
- Potential predictors are the total cost of the trip, total tips of the trip, taxi company, and payment method.

3. Insights in EDA

(1) Trips Miles vs. Company

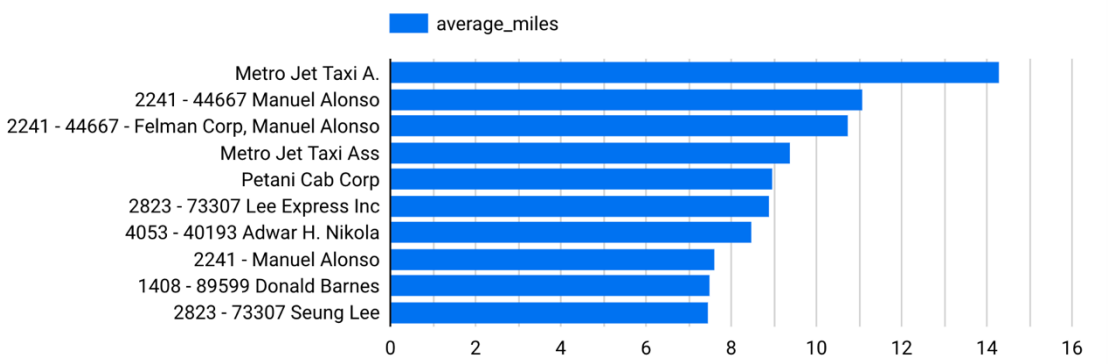
- Insight

Company may be one of the main effects. I should explore this predictor when data modeling. Metro Jet Taxi A. has the biggest average miles.

- SQL

```
SELECT AVG(trip_miles) as average_miles, company
FROM `bigquery-public-data.chicago_taxi_trips.taxi_trips`
GROUP BY company
LIMIT 1000000;
```

- Data Visualization



(2) Trips Miles vs. Payment Type

- Insight

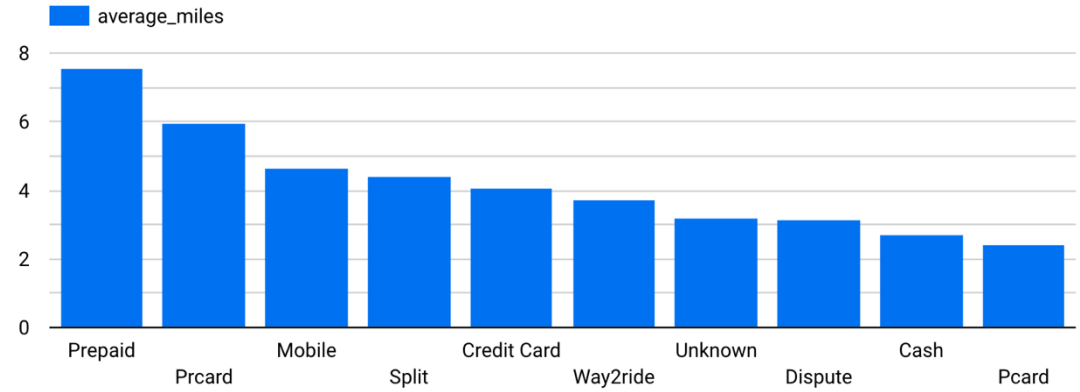
Payment type may be one of the main effects. I should explore this predictor when data modeling.

Prepaid users tend to take longer trips by taxi, while Pcard users tend to take the least.

- SQL

```
SELECT AVG(trip_miles) as average_miles, payment_type
FROM `bigquery-public-data.chicago_taxi_trips.taxi_trips`
GROUP BY payment_type
LIMIT 1000000;
```

- Data Visualization



(3) Trips Miles vs. Total Cost and Total Tips

- Insight

A positive relationship may exist between the trip miles and tips. Also, A positive relationship may exist between the trip miles and the total cost. However, the two trends look similar, which means they might contain similar information.

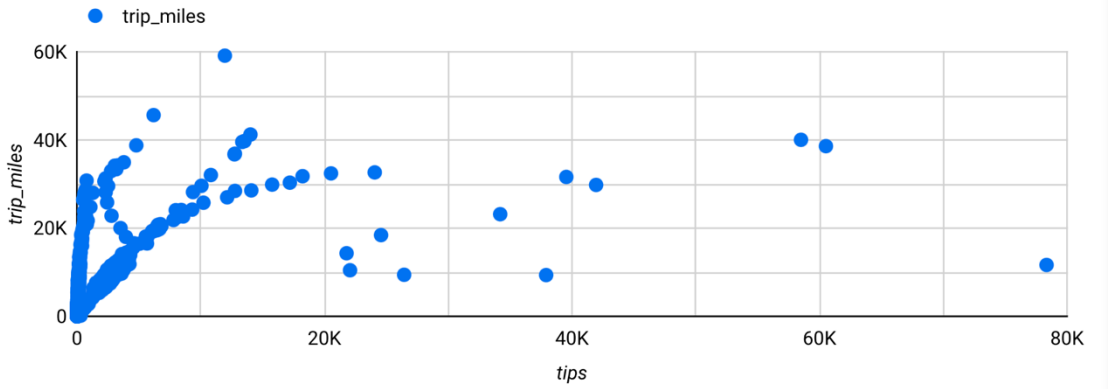
```
SELECT trip_miles, tips
FROM `bigquery-public-data.chicago_taxi_trips.taxi_trips`
WHERE trip_miles != 0
```

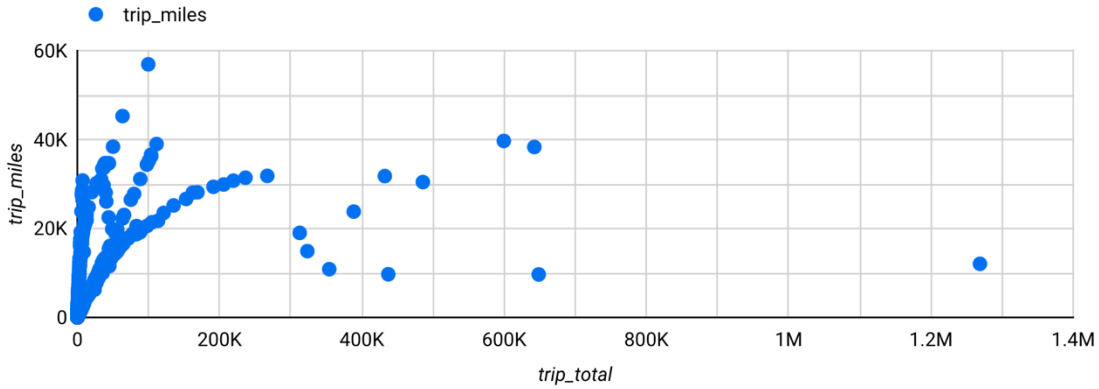
```
LIMIT 1000000;
```

- SQL

```
SELECT trip_miles, trip_total
FROM `bigquery-public-data.chicago_taxi_trips.taxi_trips`
WHERE trip_miles != 0
LIMIT 1000000
```

- Data Visualization





(4) Multicollinearity between Total Cost and Total Tips

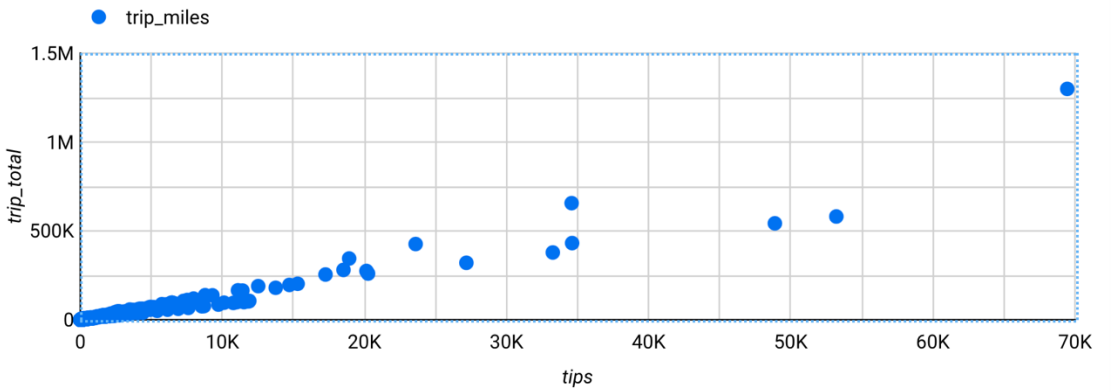
- Insight

As mentioned above, total cost and total tips may contain similar information. As shown below, they have a strong positive relationship. Thus, there may be a multicollinearity issue between these two variables. I would like to just keep one of them in the final model.

- SQL

```
SELECT trip_miles, trip_total, tips
FROM `bigquery-public-data.chicago_taxi_trips.taxi_trips`
WHERE trip_miles != 0
LIMIT 1000000;
```

- Data Visualization



(5) Interaction between Payment Type and Company

- Insight

The trends of trip miles vs. payment types differ by different companies. This means interactions may exist between payment type and company.

However, there is not enough data in some splitting categories, so this interaction term will not be included in the final model.

- SQL

```
SELECT AVG(trip_miles) as average_miles, payment_type, company
FROM `bigquery-public-data.chicago_taxi_trips.taxi_trips`
GROUP BY payment_type, company
```

- Data Visualization