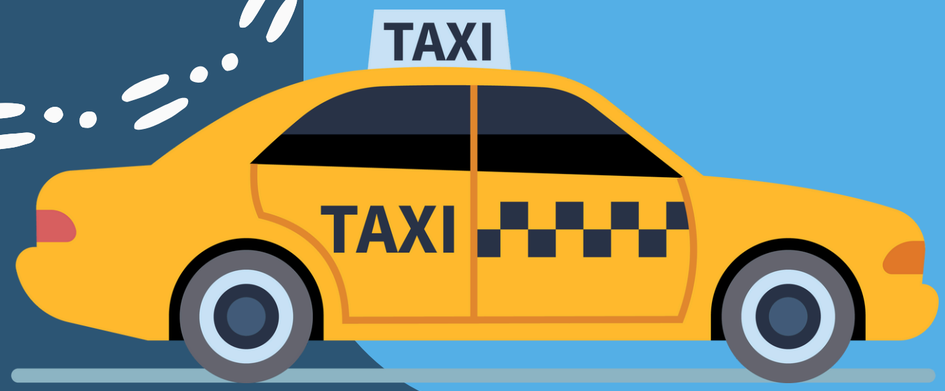
Data Engineer

Analysis of Taxi Trips in Chicago

By Almaidah



Background Project

This project focuses on analyzing taxi trip data in Chicago using a public dataset from the BigQuery Link Data Set. It was created as part of a mini-project assignment in the Dibimbing Data Series 10.0 - Data Engineering class.

The dataset includes various information about taxi trips, such as pickup and drop-off times, trip distance, pickup and drop-off community areas, payment types, and trip costs.

The project also aims to develop technical skills in using data analysis tools like BigQuery and deepen the understanding of data engineering concepts taught in the class. Consequently, this project not only provides practical insights into taxi trip data but also enhances participants' technical competencies in data analysis and engineering.

Questions Project

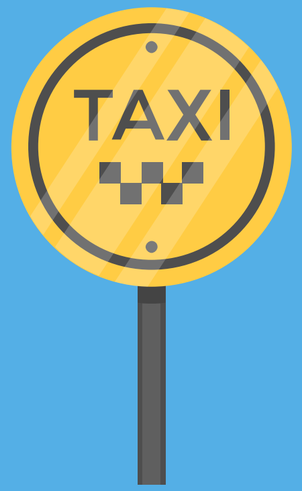
Based on the project, the following questions were formulated:

1. Calculate the average, median, and standard deviation of trip duration (trip_seconds) for trips conducted on Mondays and Saturdays. Compare the results for both days.
2. Identify the five routes (from the starting community_area to the destination community_area) with the highest number of trips in 2023.
3. Compare the average taxi trip costs (fare, tips, and taxes) based on payment methods in 2019.

Tools Project

Google BigQuery

Github



Query

Duration Analysis

```sql
select
  format_date('%A', date(trip_start_timestamp)) as weekday
  avg(trip_seconds) as avg_seconds,
  approx_quantiles(trip_seconds, 2)[offset(1)] as median_s
  stddev(trip_seconds) as stddev_seconds
from
  bigquery-public-data.chicago_taxi_trips.taxi_trips
where
  extract(dayofweek FROM trip_start_timestamp) in (2,7)
group by
  weekday
order by
  weekday
limit 1000
```

Trip Duration Analysis Results

JOB INFORMATION | RESULTS | CHART | JSON | EXECUTION DETAILS | EXECUTION GRAPH

Row	weekday	avg_seconds	median_seconds	stddev_seconds
1	Monday	846.06224089783973	540	1351.4682439442406
2	Saturday	742.91936105793388	555	1154.8569669563567

Insights:

1. The highest average trip duration is recorded on Mondays, indicating that trips on this day tend to take longer compared to Saturdays.
2. The highest median trip duration is recorded on Saturdays, meaning half of the trips on Saturdays are longer than those on Mondays.
3. The highest standard deviation in trip duration is recorded on Mondays, indicating that trip durations on Mondays tend to be more varied.

Query

Most Frequent Routes

```sql
SELECT
  pickup_community_area,
  dropoff_community_area,
  COUNT(*) AS num_trips
FROM
  `bigquery-public-data.chicago_taxi_trips.taxi_trips`
WHERE
  EXTRACT(YEAR FROM trip_start_timestamp) = 2023
  AND dropoff_community_area IS NOT NULL
GROUP BY
  pickup_community_area,
  dropoff_community_area
ORDER BY
  num_trips DESC
LIMIT 5;
```

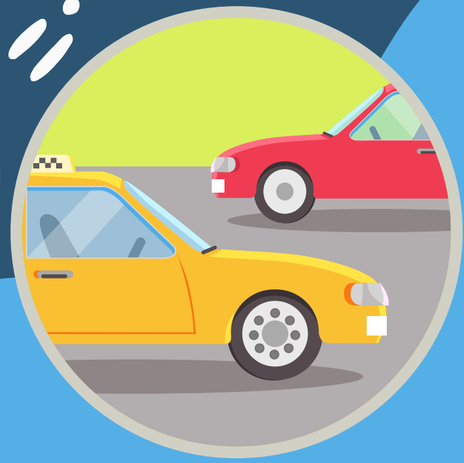
Most Frequent Routes Analysis Results

JOB INFORMATION | RESULTS | CHART | JSON | EXECUTION DETAILS

Row	pickup_community_area	dropoff_community_area	num_trips
1	8	8	464844
2	32	8	291722
3	76	8	274747
4	8	32	267673
5	32	32	241596

Insights:

The route with both pickup and drop-off in area 8 is the most frequent compared to other routes. This suggests that area 8 is a highly active location for taxi activities, both for picking up and dropping off passengers.



Query

Payment Methods

```sql
select
  payment_type,
  AVG(fare) as average_fare,
  AVG(tips) as average_tips,
  AVG(tolls) as average_tolls,
from bigquery-public-data.chicago_taxi_trips.taxi_trips
where extract (year from trip_start_timestamp) = 2019
group by payment_type
order by average_fare desc
limit 1000
```

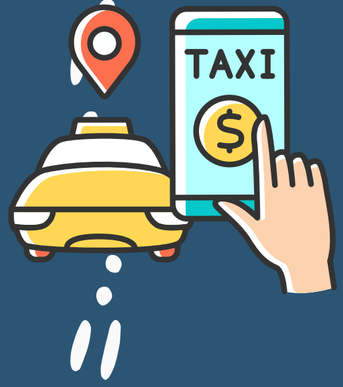
Payment Methods Analysis Results for 2019

JOB INFORMATION | RESULTS | CHART | JSON | EXECUTION DETAILS | EXECUTION GRAPH

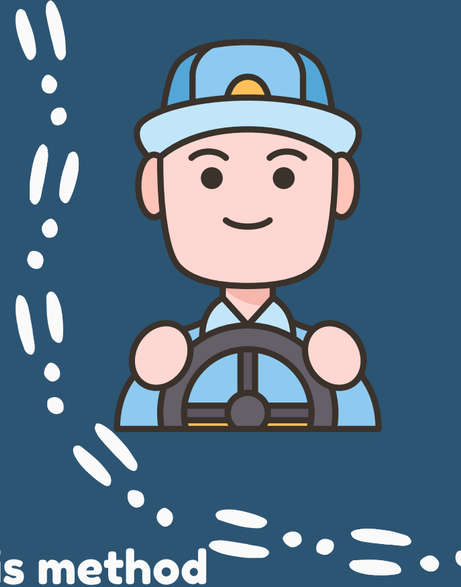
Row	payment_type	average_fare	average_tips	average_tolls
1	Prepaid	19.461414790996788	0.0	0.0
2	Credit Card	16.814137807184956	3.7745446963090816	0.0022669434124616915
3	Prcard	16.130002713879602	0.20352759110881352	0.001529141897131043
4	Mobile	15.973174037338588	3.1126980232537869	4.9680129691285853e-05
5	Unknown	15.872379933364027	0.082129280999720083	0.00014717700065913628
6	No Charge	15.753509638998038	0.24953186394891913	0.019141699410609014
7	Dispute	15.655840224453623	0.0014456585942114572	0.078991435321913786
8	Cash	12.978306351617441	0.002387440611014687	0.0017909260271819813
9	Pcard	11.327471482889733	0.038022813688212927	0.0

Insights:

1. **Average Fare:** The highest fare is recorded for prepaid payments, with an average of $19.40.
2. **Average Tips:** The highest tips are given for payments made using credit cards, with an average of $3.77.
3. **Average Tolls:** The highest tolls are recorded for payments made using mobile, with an average of $4.96.



Recommendations



1. Infrastructure and Services Enhancement in Area 8:
   - Consider increasing the number of taxis operating in area 8.
   - Improve pickup and drop-off facilities.
   - Manage traffic to reduce congestion.
2. Encouraging Higher Tip Payments:
   - Offer discounts or incentives to passengers who choose credit card payments, as this method has shown higher average tips.
3. Driver Education and Training:
   - Enhance driver education and training programs to encourage them to provide excellent service, potentially increasing tip amounts, especially for prepaid payments that currently do not generate tips.
4. Fare and Toll Structure Evaluation:
   - Assess the fare and toll structure to ensure fairness across all payment methods.
   - Consider offering discounts or toll reductions for prepaid and card payments to attract more users.

Thank You!

Click for more

LinkedIn

Github

Email

Whatsapp