



SQL BigQuery Dataset Chicago Taxi Trips

> Data Series 16.0 Data Engineer <



Muhammad Aziz Habiburrahim

CONTENT



OVERVIEW



Q1



Q2



Q3





OVERVIEW



OVERVIEW



- **SQL BigQuery**

Layanan analisis data berbasis cloud yang disediakan oleh Google Cloud Platform (GCP). BigQuery dirancang untuk memungkinkan analisis data dalam skala besar secara cepat dan efisien menggunakan SQL standar. Layanan ini sangat cocok untuk menangani data berukuran terabyte hingga petabyte dengan performa tinggi tanpa perlu mengelola infrastruktur server.

- **Dataset Chicago Taxi Trips**

Bagian dari koleksi data publik di Google BigQuery, yang berasal dari layanan transportasi taksi di Chicago, Illinois, Amerika Serikat. Dataset ini mencatat detail perjalanan taksi dari berbagai perusahaan, menyediakan data yang kaya untuk analisis terkait transportasi, ekonomi, dan perilaku perjalanan.

OVERVIEW



- **Detail Dataset Chicago Taxi Trips**

Nama Dataset: bigquery-public-data.chicago_taxi_trips.taxi_trips

Periode Data: Data mencakup perjalanan taksi yang berlangsung dari tahun 2013 hingga saat ini (periode dapat diperiksa lebih lanjut sesuai kebutuhan).

Sumber: Data ini dirilis oleh Pemerintah Kota Chicago melalui program data terbuka mereka.

- **Kegunaan Dataset Chicago Taxi Trips**

Dataset ini sangat berguna untuk berbagai analisis, antara lain:

Analisis Transportasi: Pola perjalanan, durasi, dan jarak perjalanan.

Analisis Ekonomi: Studi biaya rata-rata perjalanan, tip, dan metode pembayaran.

Perencanaan Perkotaan: Distribusi perjalanan antar area komunitas, identifikasi area sibuk.

Machine Learning: Membuat model prediksi seperti estimasi durasi perjalanan, biaya perjalanan, atau rekomendasi rute.

OVERVIEW



- **Filter Dataset Chicago Taxi Trips**

- i. **trip_id**: ID unik untuk setiap perjalanan taksi.
- ii. **taxi_id**: ID unik untuk setiap taksi.
- iii. **trip_start_timestamp**: Waktu dan tanggal saat perjalanan dimulai.
- iv. **trip_end_timestamp**: Waktu dan tanggal saat perjalanan berakhir.
- v. **trip_seconds**: Durasi perjalanan dalam detik.
- vi. **trip_miles**: Jarak perjalanan dalam mil.
- vii. **pickup_community_area**: Area komunitas tempat perjalanan dimulai (kode numerik).
- viii. **dropoff_community_area**: Area komunitas tempat perjalanan berakhir (kode numerik).
- ix. **fare**: Biaya perjalanan (dalam USD).
- x. **tips**: Jumlah tip yang diberikan (dalam USD).
- xi. **tolls**: Biaya tol (jika ada, dalam USD).
- xii. **extras**: Biaya tambahan, seperti biaya pemesanan.
- xiii. **payment_type**: Metode pembayaran yang digunakan (misalnya, tunai, kartu kredit, dll.).
- xiv. **company**: Nama perusahaan taksi yang mengelola perjalanan.



Q1



Hitung rata-rata, median, dan standar deviasi dari durasi perjalanan (trip_seconds) untuk perjalanan yang dilakukan pada hari Senin dan Sabtu. Bandingkan hasil kedua hari tersebut.

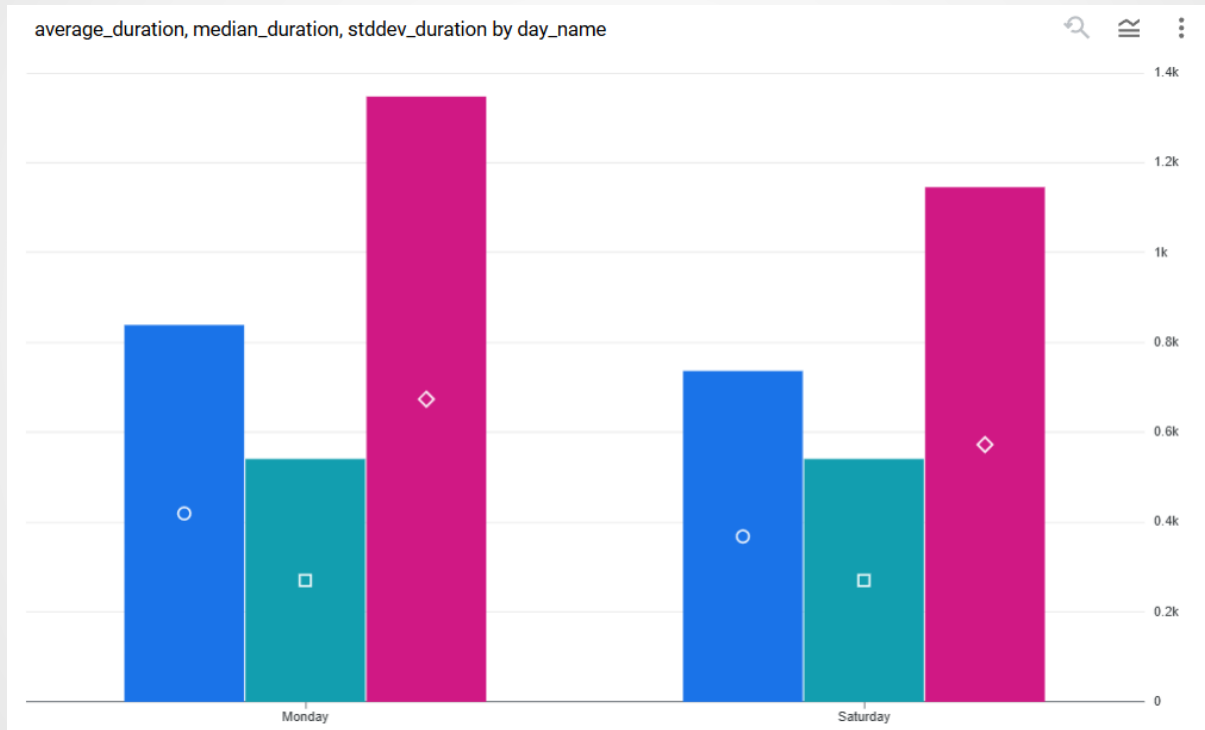
- Code SQL

```
1  -- query untuk menghitung statistik trip_seconds berdasarkan hari senin dan sabtu
2  WITH filtered_trips AS (
3      SELECT
4          EXTRACT(DAYOFWEEK FROM trip_start_timestamp) AS day_of_week,
5          trip_seconds
6      FROM
7          `bigquery-public-data.chicago_taxi_trips.taxi_trips`
8      WHERE
9          EXTRACT(DAYOFWEEK FROM trip_start_timestamp) IN (2, 7) -- 2=senin, 7=sabtu
10         AND trip_seconds IS NOT NULL
11  ), statistics AS (
12      SELECT
13          day_of_week,
14          AVG(trip_seconds) AS average_duration,
15          APPROX_QUANTILES(trip_seconds, 2)[OFFSET(1)] AS median_duration,
16          STDDEV(trip_seconds) AS stddev_duration
17      FROM
18          filtered_trips
19      GROUP BY
20          day_of_week
21  )
22  SELECT
23      CASE day_of_week
24          WHEN 2 THEN "Monday"
25          WHEN 7 THEN "Saturday"
26      END AS day_name,
27      average_duration,
28      median_duration,
29      stddev_duration
```

- Result SQL

Row	day_name	average_duration	median_duration	stddev_duration
1	Monday	839.4797835193...	540	1346.060208432...
2	Saturday	737.4105515098...	540	1146.335888011...

- Chart SQL





Q2



**Temukan lima rute (dari
community_area awal ke
community_area tujuan) dengan jumlah
perjalanan terbanyak pada tahun 2023.**

-
-
-

- Code SQL

```
1  -- query untuk menemukan lima rute dengan perjalanan terbanyak
2  WITH trips_2023 AS (
3      SELECT
4          pickup_community_area,
5          dropoff_community_area,
6          COUNT(*) AS trip_count
7      FROM
8          `bigquery-public-data.chicago-taxi-trips.taxi-trips`
9      WHERE
10         EXTRACT(YEAR FROM trip_start_timestamp) = 2023
11         AND pickup_community_area IS NOT NULL
12         AND dropoff_community_area IS NOT NULL
13     GROUP BY
14         pickup_community_area, dropoff_community_area
15 )
16 SELECT
17     pickup_community_area,
18     dropoff_community_area,
19     trip_count
20 FROM
21     trips_2023
22 ORDER BY
23     trip_count DESC
24 LIMIT 5;
25
```

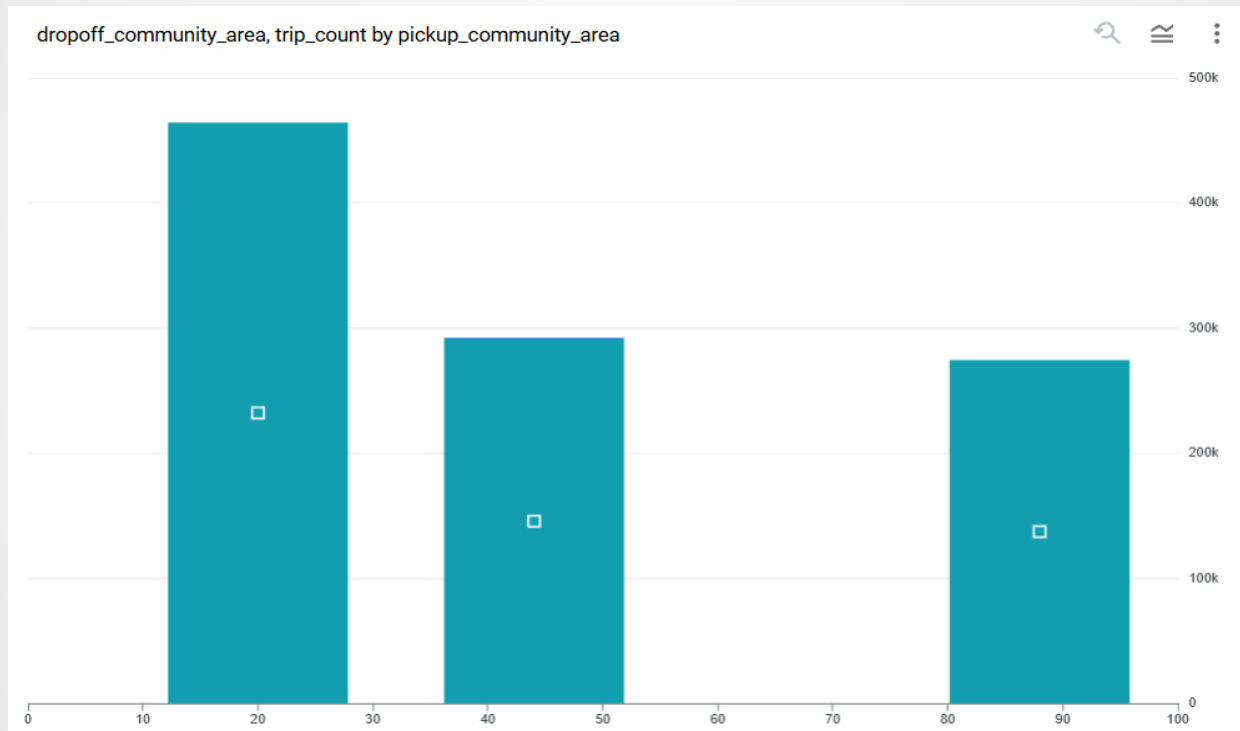
Q2



- Result SQL

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

- Chart SQL





Q3



**Bandingkan biaya rata-rata perjalanan
taksi (fare, tips, tolls dan extras)
berdasarkan metode pembayaran pada
tahun 2019.**

■

■

■

Q3



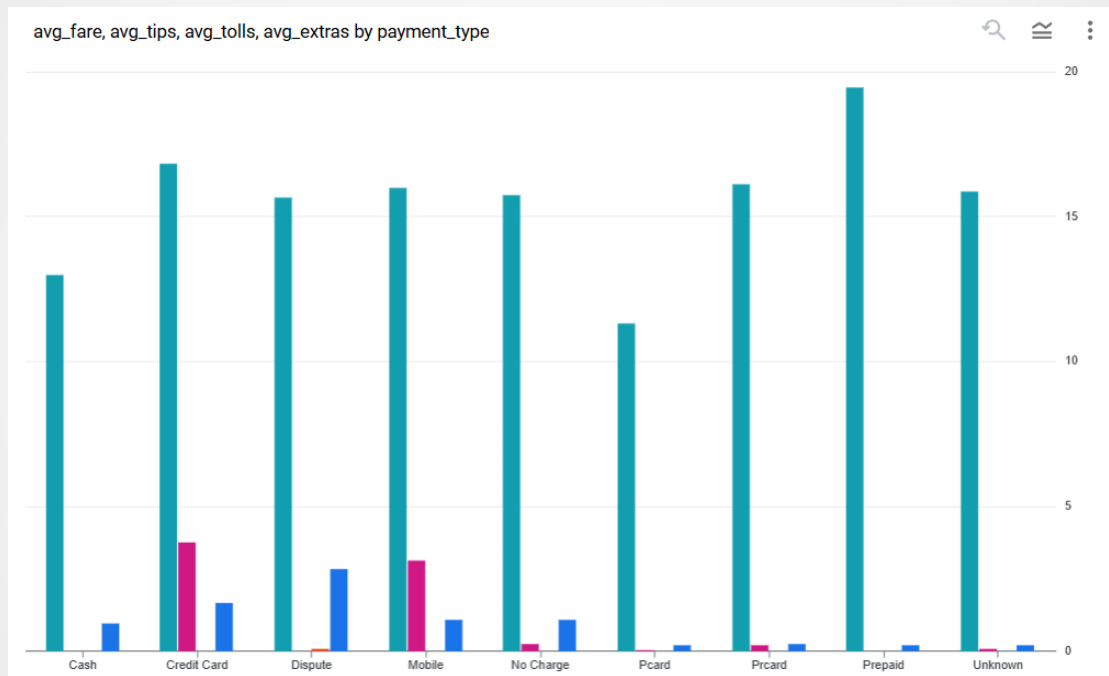
- Code SQL

```
1  -- query untuk membandingkan biaya rata-rata berdasarkan metode pembayaran
2  SELECT
3  |   payment_type,
4  |   AVG(fare) AS avg_fare,
5  |   AVG(tips) AS avg_tips,
6  |   AVG(tolls) AS avg_tolls,
7  |   AVG(extras) AS avg_extras
8  FROM
9  |   `bigquery-public-data.chicago_taxi_trips.taxi_trips`
10 WHERE
11 |   EXTRACT(YEAR FROM trip_start_timestamp) = 2019
12 GROUP BY
13 |   payment_type
14 ORDER BY
15 |   payment_type;
```

- Result SQL

Row	payment_type ▼	avg_fare ▼	avg_tips ▼	avg_tolls ▼	avg_extras ▼
1	Cash	12.97830635161...	0.002387440611...	0.001790926027...	0.967582228072...
2	Credit Card	16.81413780718...	3.774544696309...	0.002266943412...	1.661699447860...
3	Dispute	15.65584022445...	0.001445658594...	0.078991435321...	2.843499704666...
4	Mobile	15.97317403733...	3.112698023253...	4.968012969128...	1.071116146913...
5	No Charge	15.75350963899...	0.249531863948...	0.019141699410...	1.077899680746...
6	Pcard	11.32747148288...	0.038022813688...	0.0	0.215779467680...
7	Pcard	16.13000271387...	0.203527591108...	0.001529141897...	0.234292194365...
8	Prepaid	19.46141479099...	0.0	0.0	0.220257234726...
9	Unknown	15.87237993336...	0.082129280999...	0.000147177000...	0.196470189885...

- Chart SQL



The background is a dark blue gradient with a complex geometric pattern of overlapping triangles in various shades of blue and grey. Scattered throughout are small, out-of-focus white and light blue circular bokeh lights, giving it a sense of depth and a modern, digital feel.

THANK YOU