Fakhri Nugraha Pratama

PROJECT DATA ENGINEERING

ibimbing

Data Series 16.0

2025 Presentation

fakhrinugrahap25@gmail.com

Fakhri Nugraha Pratama

DATA SERIES

This training discusses the basics of Data Engineering, from concept introduction to data manipulation using SQL. The material covers the use of BigQuery, WHERE clause, aggregation statement, CASE WHEN, JOIN, UNION, to advanced concepts such as subquery, rank, and other SQL operations. This program is designed to build a strong foundation in data management and analysis.

DATASET

`bigquery-public-data.chicago_taxi_trips.taxi_trips`

2025 Presentation

Fakhri Nugraha Pratama

QUESTION 1

Calculate the mean, median, and standard deviation of trip duration (trip_seconds) for trips made on Monday and Saturday. Compare the results of the two days.

QUESTION 2

Find the five routes (from the starting community_area to the destination community_area) with the highest number of trips in 2023.

QUESTION 3

Compare the average cost of a taxi ride (fare, tips, and taxes) by payment method in 2019.

Fakhri Nugraha Pratama

SOLUTION QUESTION 1

```sql
-- Calculate the mean, median and standard deviation of trips_seconds on Monday and Saturday.
WITH trip_data AS (
    SELECT
        EXTRACT(DAYOFWEEK FROM trip_start_timestamp) AS day_of_week,
        trip_seconds
    FROM
        `bigquery-public-data.chicago_taxi_trips.taxi_trips`
    WHERE
        trip_seconds IS NOT NULL
        AND EXTRACT(DAYOFWEEK FROM trip_start_timestamp) IN (2, 7)  -- 2: Monday, 7: Saturday
)

SELECT
    CASE
        WHEN day_of_week = 7 THEN 'Saturday'
        WHEN day_of_week = 2 THEN 'Monday'
    END AS weekday,
    AVG(trip_seconds) AS avg_seconds,
    APPROX_QUANTILES(trip_seconds, 100)[OFFSET(50)] AS median_seconds,
    STDDEV(trip_seconds) AS stddev_seconds
FROM
    trip_data
GROUP BY
    day_of_week

```

Fakhri Nugraha Pratama

RESULT SOLUTION 1:

Row	weekday	avg_seconds	median_seconds	stddev_seconds
1	Saturday	737.4105515098...	551	1146.335888011...
2	Monday	839.4797835193...	540	1346.060208432...

This SQL analyzes taxi trip durations in Chicago on Mondays and Saturdays by creating a temporary table (`trip_data`) that filters out trip durations (`trip_seconds`) that are not NULL for both days. From this table, three important statistics were calculated: average trip time (avg_seconds), median trip time (median_seconds), and standard deviation (stddev_seconds). The results provide a snapshot of taxi travel patterns on Mondays and Saturdays, and provide insight into the habits of taxi users in Chicago.

Fakhri Nugraha Pratama

SOLUTION QUESTION 2

```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  -- Filtering the year 2023
    AND pickup_community_area IS NOT NULL
    AND dropoff_community_area IS NOT NULL
GROUP BY
    pickup_community_area,
    dropoff_community_area
ORDER BY
    num_trips DESC  -- Sort by most number of trips
LIMIT 5;  -- Taking the top five results

```

Fakhri Nugraha Pratama

RESULT SOLUTION 2:

Row	pickup_community_a	dropoff_community_	num_trips
1	8	8	464844
2	32	8	291722
3	76	8	274747
4	8	32	267673

This SQL was used to analyze taxi trip data in Chicago in 2023 by calculating the number of trips between pick-up and drop-off community areas. This query filters the data to ensure that only trips with valid timestamps and non-NULL pick-up and drop-off community areas are included. The results are grouped by pick-up and drop-off community areas, then sorted by the highest number of trips in descending order. Finally, this query retrieves the top five community area combinations with the highest number of trips.

Fakhri Nugraha Pratama

SOLUTION QUESTION 3

```sql
SELECT
  payment_type,
  AVG(fare) AS average_fare,
  AVG(tips) AS average_tips,
  AVG(tolls) AS average_tosls,
  AVG(extras) AS average_extras
FROM `bigquery-public-data.chicago_taxi_trips.taxi_trips`
WHERE
    EXTRACT(YEAR FROM trip_start_timestamp) = 2019
group by
  payment_type
ORDER BY
  payment_type;
```

Fakhri Nugraha Pratama

RESULT SOLUTION 3:

Row	payment_type	average_fare	average_tips	average_tosls	average_texas
1	Cash	12.97830635161...	0.002387440611...	0.001790926027...	0.967582228072...
2	Credit Card	16.81413780718...	3.7745446963091	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	Prcard	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...

This SQL is used to analyze data on taxi trips in Chicago in 2019 by calculating the average cost of rides, tips, tolls, and extras by payment type. This query filters the data to only include trips from 2019, then calculates averages for each cost category (fare, tips, tolls, and extras) grouped by payment type. The results are sorted by payment type, providing insight into the spending patterns of taxi users based on the payment method used.