✅ **Milestone 3** | London Transit Analysis

INTRODUCTION: In this Milestone, you will work with data provided by Transport for London (TfL), specifically from the Rolling Origin and Destination Survey (RODS). As a data analyst for Transport for London (TfL), your role is to support efforts to improve public transit operations through data-driven insights. TfL officials rely on a clear understanding of ridership patterns, peak travel times, and line usage to keep the system running smoothly and efficiently.

You're tasked with analyzing ridership data to uncover these trends. Your findings will help guide decisions on service scheduling, resource allocation, and infrastructure planning, ensuring that London's transport network continues to meet the needs of its millions of daily passengers.

HOW IT WORKS: Follow the prompts in the questions below to investigate your data. Post your answers in the provided boxes: the **yellow boxes** for the queries you write, **purple boxes** for visualizations and **blue boxes** for text-based answers. When you're done, export your document as a pdf file and submit it on the Milestone page – see instructions for creating a PDF at the end of the Milestone. Please don't ever remove (paste your query below 👇) or (write your **answer** below 👇). These help your Evaluator!

SQL App: Here's that link to our specialized SQL app, where you'll write your SQL queries and interact with the data.

— Data Set **Description**

The TfL RODS data (`tfl.rods`) models activity on the London Underground that would take place on a typical November weekday. The slice of the data that has been pulled out from the survey consists of 6295 rows across six columns:

- **entry_zone**: Zone of the station in which a passenger starts their journey. Zone 1 encompasses the central part of London, and each higher-numbered

Zone is a ring around the previous. In other words, Zone 5 represents stations that are furthest out from the central part of London. [See here for a visualization of Zones in London.](#)

- **time_period**: Time period in which the passenger started their trip. There are six periods of day: Early (5am-7am), AM Peak (7am-10am), Midday (10am-4pm), PM Peak (4pm-7pm), Evening (7pm-10pm), and Late (10pm-5am).
- **origin_purpose**: The reason for the passenger to have chosen the station from which they begin their journey. There are eight categories: Home, Work, Shop, Education, Tourist, Hotel, Other, and Unknown/Not Given.
- **destination_purpose**: The reason for the passenger to have chosen the station from which they end their journey. The possible values for this feature are the same eight categories as for the origin_purpose feature.
- **distance**: Approximate distance between the passenger's origin and destination stations. Distances are grouped into five levels: <3 km, 3–8 km, 8–16 km, 16–24 km, and over 24 km.
- **daily_journeys**: Number of daily journeys matching the entry, time period, purpose, and distance profile indicated by the data row. This number is derived from the RODS model, rather than a specific day of data collection.

— Task 1: General Usage Statistics

Although we'd like to eventually understand why passengers use the rail system, we should start by making some summaries of the rail system in general.

A. Write a query that returns the sum total of journeys. This total represents the volume of activity expected on a typical day of operations for the Underground system!

(paste your query below 👇)

```
SELECT
    SUM(daily_journeys) AS total_journeys
FROM
    tfl.rods
```

What is the total number of journeys expected on a typical day?

(write your **answer** below 👇)

```
4878330
```

B. Add to your query to return the number of journeys made that originate from each Zone.

(paste your query below 👇)

```
SELECT
    entry_zone,
    SUM(daily_journeys) AS total_journeys
FROM
    tfl.rods
GROUP BY entry_zone
```

What percentage of journeys start from a Zone 1 station? **HINT:** (Divide the Zone 1 value by the value you got from part A; you won't calculate this in SQL!)

(write your **answer** below 👇)

51.7 %

C. Revise your query to return the number of journeys made in each period of day.

(paste your query below 👇)

```
SELECT
```

```
    time_period,
    SUM(daily_journeys) AS total_journeys
FROM
    tfl.rods
GROUP BY time_period
ORDER BY total_journeys DESC
```

Which time period has the highest total volume of passengers?

(write your **answer** below 👇)

PM Peak – `1367309`

— Task 2: For what reasons do people use the London Underground?

Let's start adding in the survey information about the reasons why passengers take trips on the subway system.

A. Write a query that returns the number of journeys made grouped by their reasons for the origin station.

(paste your query below 👇)

```
SELECT
    origin_purpose,
    SUM(daily_journeys) AS total_journeys
FROM
    tfl.rods
GROUP BY origin_purpose
ORDER BY total_journeys DESC
```

Which journey purposes have the highest number of trips, and what does this tell you about how the subway system is used?

(write your **answer** below 👇)

Home and Work have the highest origin-purpose trips and it tells me that the London Underground is primarily used for commuting between home and work.

B. Change the grouping on your query to be on both the origin purpose and the destination purpose, so that you get the number of journeys by each origin-destination purpose pair.

(paste your query below 👇)

```
SELECT
    origin_purpose,
    destination_purpose,
    SUM(daily_journeys) AS total_journeys
FROM
    tfl.rods
GROUP BY
    origin_purpose,
    destination_purpose
ORDER BY
    total_journeys DESC
```

Does this support or change your understanding of what you observed in the previous part?

(write your **answer** below 👇)

> This query supports my understanding of what the London Underground is primarily used for as it shows the top 2 common pairs are from home to work and then work to home.

C. Is there a bias in when people make their trips, depending on why they make a trip?

Modify your query to get the number of trips grouped by origin purpose and time of day. Sort by origin purpose so that all of the trips for a specific reason are returned together.

(paste your query below 👇)

```
SELECT
    origin_purpose,
    time_period,
    SUM(daily_journeys) AS total_journeys
FROM
    tfl.rods
GROUP BY
    origin_purpose,
    time_period
ORDER BY
    origin_purpose,
    total_journeys DESC
```

Interpret the output: Do people travel from Home or Work at the expected time periods?

(write your **answer** below 👇)

Yes, because it shows Home origin at the top with most journeys in the AM peak period and right under that is the Work origin in the PM Peak period.

D. Is there a difference in travel purposes based on which zone is the trip origin?

Modify your query to get the number of trips grouped by origin purpose and entry zone. Sort by entry zone so that all of the frequency counts for a single zone are in consecutive rows.

(paste your query below 👇)

```
SELECT
    origin_purpose,
    entry_zone,
    SUM(daily_journeys) AS total_journeys
FROM
    tfl.rods
GROUP BY
    origin_purpose,
    entry_zone
ORDER BY
    entry_zone,total_journeys DESC
```

Interpret the output: how does the ranking of Home and Work purposes change as we change Zone?

(write your **answer** below 👇)

Zone 1 has trips from work at the top, while zone 2 has more from home than work. This shows Central London is primarily a work hub and outer zones more residential.

E. Now that you've explored patterns in the RODS dataset, it's time to think like a data analyst working with a real client. Transport for London (TfL) uses this type of data to improve how the transit system functions day-to-day.

> **Try this prompt:** What does it mean if Zone 1 sees more people starting trips from work, while Zones 2–5 see more people starting from home? How might this pattern help city planners or transit authorities better understand urban flow?

Based on ChatGPT's response, what specific recommendation would you make to Transport for London (TfL) to enhance the efficiency, accessibility, or user experience of the transit system?

(write your **answer** below 👇)

> To enhance the efficiency of the user experience of the transit system we will have to optimize service by increasing capacity during peak commute times and having efficient connections focusing between residential outer zones and central work zones.

— LevelUp

There's a lot of finer investigations that you can do with the RODS data, but it is most useful when you can focus your attention on just part of the data. We learned that the majority of rides for home/work happened during the peak times. Let's investigate how that changes for tourism related travel.

A. Write a query that returns the total number of journeys grouped by origin purpose, destination purpose, and time period. Filter to trips where either origin or destination is done for tourism purposes.

(paste your query below 👇)

```
SELECT
    origin_purpose,
    destination_purpose,
    time_period,
    SUM(daily_journeys) AS total_journeys
FROM
    tfl.rods
WHERE
    origin_purpose = 'Tourist'
    OR destination_purpose = 'Tourist'
GROUP BY
    origin_purpose,
    destination_purpose,
    time_period
ORDER BY
    total_journeys DESC
```

How do travel periods for tourism related travel differ from those for work commute purposes?

(write your **answer** below 👇)

Tourism trips peak during midday instead of am or pm for work commute purposes showing more flexible travel compared to work commuting.

Next, you will learn about how to apply two different kinds of clauses to filter aggregated data in two different ways. But if you're excited about this dataset or want to think ahead, you can try your hand at applying the WHERE keyword you learned about previously. The WHERE clause comes after FROM and before GROUP BY. Try to see how adding a WHERE clause on one or two different journey purposes cleans up the output, and see if it makes it easier to see trends on some of the less-common trip reasons.

— Submission

Great work completing this Milestone! To submit your completed Milestone, you will need to download / export this document as a PDF and then upload it to the Milestone submission page. You can find the option to download as a PDF from the File menu in the upper-left corner of the Google Doc interface.