✓ **Milestone 4** | London Transportation Journey Survey

INTRODUCTION: A great source for datasets to work from can be found from government resources. More and more governing bodies – from the national level all the way down to the city level – are making their data available for the public to download and analyze. The types of data often cover a wide range of topics, such as energy use and conservation, public transportation, and the fine arts and education. Of course, raw data isn't much use without the skills to turn it into useful information — but those skills are what you've been building throughout this entire course!

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.

RESOURCES: If you need hints on the Milestone or are feeling stuck, there are multiple ways of getting help. Attend Drop-In Hours to work on these problems with your peers, or reach out to the HelpHub if you have questions. Good luck!

PROMPT: In this particular Milestone, we will be working with data that has been made available by Transport for London (TfL). More specifically, we'll be looking at data from their Rolling Origin and Destination Survey (RODS). RODS seeks to model typical passenger behaviors on the London Underground system. It goes beyond just knowing how many passengers enter or exit each station, at what time. It tries to model motivations for taking the Tube, such as for work or for leisure. The inclusion of this information can paint a picture of how the system is used that is deeper than standard usage statistics can perform alone.

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 👇)

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

(write your **answer** below 👇)

4878330 is the total amount of journeys

B. Add to your query to return the number of journeys made that originate from each Zone. What percentage of journeys start from a Zone 1 station? (Divide the Zone 1 value by the value you got from part A; you won't calculate this in SQL!)

(paste your query below 👇)

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

(write your **answer** below 👇)

51.72% of journeys start from Zone 1

C. Revise your query to return the number of journeys made in each period of day. Which time period has the highest total volume of passengers?

(paste your query below 👇)

```sql
SELECT
    SUM(daily_journeys) AS "total_journeys",
    time_period
FROM
    tfl.rods
GROUP BY
    time_period
ORDER BY
    total_journeys desc;
```

(write your **answer** below 👇)

PM Peak

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

(paste your query below 👇)

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

(write your **answer** below 👇)

Home has the highest amount of trips, it shows how the subway system is mostly used to get somewhere efficiently from their house, most likely to a place where they have to be day to day like 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. Does this support or change your understanding of what you observed in the previous part?

(paste your query below 👇)

```
SELECT
    SUM(daily_journeys) AS "total_journeys",
    origin_purpose,
    destination_purpose
FROM
```

```
    tfl.rods
GROUP BY
    origin_purpose,
    destination_purpose;
```

(write your **answer** below 👇)

> Supports my theory

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

(paste your query below 👇)

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

(write your **answer** below 👇)

People do travel from home to work at the expected times. Most people leave work during PM Peak or the evening and most people leave home during AM Peak.

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

(paste your query below 👇)

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

(write your **answer** below 👇)

As the zones increase in ranking, Work gets exponentially smaller and Home consistently stays in the hundred thousands

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

(paste your query below 👇)

```
SELECT
    SUM(daily_journeys) AS "total_journeys",
    origin_purpose,
    destination_purpose,
    time_period
FROM
    tfl.rods
WHERE
    origin_purpose LIKE 'Tourist'
    OR
    destination_purpose LIKE 'Tourist'
GROUP BY
    origin_purpose,
    destination_purpose,
    time_period;
```

(write your **answer** below 👇)

> Travel periods for tourism vary widely throughout the day compared to work commute purposes that stick to a certain range of numbers

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.