

## 📍 **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 total number of journeys. This total represents the volume of activity expected on a typical day of operations for the Underground system!

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

The total number of journeys that is expected on a typical day of the underground system is 4878330

- 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!)

```
SELECT
sum(daily_journeys)
FROM tfl.rods
WHERE entry_zone = 'zone 1'
```

The percentage of journeys through the underground system that starts from zone 1 is 51.72% =  $(2522837/4878330)$

- 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?

```
SELECT time_period,  
COUNT(time_period)  
FROM tfl.rods  
GROUP BY time_period
```

The time period that has the highest total volume of passengers is midday (1515 daily journeys were made at midday)

## – **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?

```
SELECT  
origin_purpose,  
COUNT(daily_journeys)  
FROM tfl.rods  
GROUP BY origin_purpose
```

The main journey purpose for using the subway system is to go home. (1040 trips daily) I can infer from the data that the subway is comprised of people who commute to work and back home. The amount of people that use the metro for work is almost equal to the amount that use the metro to go home.

- 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?

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

After running this query I found that the reason people initially use the metro rarely matches where they end up actually going, leading me to believe I wasn't fully correct in my first statement because I was only looking at why they got on the metro, not where they actually ended up going.

- 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?

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

After looking at the data I was able to surmise that people go to home or work at expected times. The data shows that people go to work or home at all times during the day or night, which to me is expected because everyone has a different life and contrasting schedules.

- 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?

```
SELECT
```

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

The amount of people that go to home or work hardly changes across the different zones. Zone 1 has the most people going and coming from work and the other zones aren't far behind. The largest difference between the zones is only separated by 50 trips.

## – Level Up

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?


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

The majority of people go touring around midday which differs from people going to work, which are all times of the day.

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.