⊘ **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.

— 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
   count(daily_journeys) AS n_daily_journeys
FROM
   tfl.rods
```

The total number expected on a typical day of operations for the Underground system is 6295.

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
   entry_zone,
   count(daily_journeys) AS n_daily_journeys
FROM
   tfl.rods
group by
   entry_zone
```

The percentage of journeys starting from Zone 1 is 24.78%.

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(daily_journeys) AS n_daily_journeys
FROM
   tfl.rods
group by
   time_period
```

Midday has the highest total volume of passengers at 1515. A reason for this could be for workers to attend business lunch meetings at a restaurant near the exits of the tunnel. Or maybe it has nothing to do with lunch.

— 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) AS n_daily_journeys
```

```
FROM
   tfl.rods
group by
   origin_purpose
order by
   n_daily_journeys DESC
```

The journey starting from the <u>Home</u> at <u>1040</u> has the highest number of trips. This tells me that there are many people who live near London who also own a car. I assumed only people with cars are allowed in this tunnel to come to that conclusion, but I could and most likely am wrong.

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 n_daily_journeys
FROM
   tfl.rods
group by
   origin_purpose,
   destination_purpose
order by
   n_daily_journeys DESC,
   origin_purpose
```

The data shown tells me that the majority of people are leaving their home to go to work, "Other", or "Unknown", and that people are also leaving work to go home, to work, or to "Other".

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 n_daily_journeys
FROM
    tfl.rods
group by
    time_period,
    origin_purpose
order by
    origin_purpose ASC,
    n_daily_journeys Desc
```

People leave their home equally in AM Peak, PM, Peak, and Midday which is surprising to see how equal the numbers are in relation to one another. Secondly, People leaving work is highest at Midday with 200 journeys, and AM/PM Peaks both equal at 185 journeys.

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 n_daily_journeys
FROM
    tfl.rods
group by
    entry_zone,
    origin_purpose
order by
    entry_zone asc,
    n_daily_journeys desc
```

Zone 1 has Home @ 215 and Work @ 205
Zone 2 has Home @ 215 and Work @ 195
Zone 3 has Home @ 210 and Work @ 175
Zone 4 has Home @ 200 and Work @ 175
Zone 5 has Home @ 200 and Work @ 180

The rankings as seen change so that the difference between Zone 5 and Zone 1 Home journeys is 15 and the difference between Zone 5 and Zone 1 Work journeys is 25.

That is a 7% decrease from Zone 1 → Zone 5 for Home
That is a 13% decrease from Zone 1 → Zone 5 for Work.

Thus with the information I have, it's my prediction that there may be ~5–15% less homes in Zone 5 than in Zone 1.

— 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?

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

------------------------

SELECT
   origin_purpose,
   destination_purpose,
   time_period,
```

```
   count(daily_journeys) AS n_daily_journeys
FROM
  tfl.rods
WHERE
  origin_purpose = 'Work'
  OR destination_purpose = 'Work'
GROUP BY
  origin_purpose,
  destination_purpose,
  time_period
ORDER BY
  n_daily_journeys DESC,
  time_period ASC
```

After comparing Tourist related traveling to work related traveling the conclusion is that tourists are leaving mostly in the AM Peak, Midday, and PM peak.

This is much different to Work related traveling because work is most heavily focused in AM Peak, Early, Evening, Late, and PM Peak.

The difference being that Work related traveling covers a 5 wide range of timezones while Tourist related traveling covers three major time zones.

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.