Data Analysis Report : London Transportation Journey Survey

Introduction

Project Background: This report analyzes data provided by Transport for London (TFL) from their Rolling Origin and Destination Survey (RODS). RODS models typical passenger behaviors on the London Underground system, providing insights beyond standard usage statistics by including motivations for travel, such as commuting for work or leisure activities.

Purpose: The objective of this analysis is to understand the volume and patterns of journeys made on the London Underground, identify the primary reasons for travel, and explore travel behaviors based on time periods and zones.

Data source: The data used in this analysis comes from the RODS dataset provided by TFL, focusing on daily journey counts, entry zones, time periods, and travel purposes.

Data Cleaning and Preparation

- **Data Cleaning Steps:**
  - List any steps you did to ensure the data was clean and ready for analysis
- **Data Transformation:**
  - List any transformations you did

Exploratory Data Analysis (EDA)

Through this initial exploration, we aim to lay the groundwork for a more detailed analysis and ensure that we are well-prepared to address the project's key questions. The insights gained during EDA will guide our approach to more complex statistical analyses and predictive modeling in the subsequent sections.

First we find the sum total of journeys. This total represents the volume of activity expected on a typical day of operations for the underground system.

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

This showed that a typical day on the London Underground sees about 4,878,330 journeys. This total volume indicates a high level of daily activity.

Then we were curious to what percentage of journeys start from a Zone 1 station.

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

This revealed that 51.7% of trips originate from Zone 1, amounting to 2,522,837 journeys. The number of trips decreases as we move outward from the city center.

Analysis

While the Exploratory Data Analysis (EDA) provided an initial understanding of the data's structure and key trends, the focus of this report will shift to a more detailed analysis.

The questions we aimed to answer were :

- What are the peak travel periods on the London Underground?
- What are the primary purpose for travel?
- How do travel patterns vary by zone?
- How do travel patterns on the London Underground vary by time of day and the purpose of the journey?
- How do travel purposes vary across different entry zones of the London Underground?

What are the peak travel periods on the London Underground?

```
SELECT
  time_period,
  SUM(daily_journeys) AS total_journeys_by_time_period
```

```
FROM
  tfl.rods
GROUP BY
  time_period
ORDER BY
  total_journeys_by_time_period DESC
```

Insight : The PM Peak period has the highest number of journeys (1,367,309), representing 28.0% of all trips. The AM Peak and Midday periods also show significant activity, with the combined total of other periods (Evening, Late, Early) accounting for 21% of all trips.

What are the primary purposes for travel?

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

Insight : The majority of journeys originate from Home (1.8 million) or Work (1.4 million), indicating that most people use the Underground for commuting. Hotel and Tourist purposes have the lowest counts, suggesting potential variations on weekends or warmer months.

How do travel patterns vary by zone?

```
SELECT
  origin_purpose,
```

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

Insight : The most common journey pairs are Home-Work (1.2 million) and Work-Home (1 million), supporting the primary use for commuting. A significant number of journeys have unknown purposes.

How do travel patterns on the London Underground vary by time of day and the purpose of the journey?

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

Insight : The most common journey times from home are during the AM peak, as expected, followed by an unexpected rise during the Midday period. From work, the PM peak sees the highest number of journeys, followed by Midday, which also shows a significant number of trips.

How do travel purposes vary across different entry zones of the London Underground?

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

Insight : From zone 1, work is the most common destination, followed by home. In zone 2, home is the highest, with work coming next. For zones 3, 4, and 5, home is the most common, while the next most frequent category is unknown or not given.

Results

Our objective was to uncover significant patterns and insights regarding passenger behavior on the London Underground system. By leveraging statistical techniques and predictive modeling, we aimed to answer key questions derived from our initial exploratory data analysis (EDA).

We found that a typical day on the London Underground sees approximately 4.9 million journeys, with 51.7% originating from Zone 1, the PM Peak being the busiest time with 28.0% of trips, the majority of journeys starting from home (1.8 million) or work (1.4 million), and the most common journey pair being Home-Work (1.2 million), with travel patterns varying by time of day and zone, showing the AM peak busiest for trips from home and the PM peak from work, and home being the most common destination in Zones 2 to 5, with unknown purposes following in Zones 3, 4, and 5.

We would recommend further analysis focusing on specific time periods and travel zones to optimize transport planning, as well as implementing targeted measures to address the high volume of commuters during peak hours and exploring the underlying reasons for the significant number of unknown journey purposes in outer zones.

Future analysis could include weekend and seasonal data to understand variations in travel behavior, and more granular studies on specific user demographics.