DSA 210 Project: Travel Pattern Analysis and Trip Count Prediction

Mehmet Barkın Palabıyık
Student Number: 22424

1 Introduction

This project analyzes Google Maps Timeline data collected between February and April 2025 to understand personal travel behavior. We aim to:

1. Uncover recurring travel patterns across days of the week and hours of the day.
2. Test the hypothesis that mean travel distance on weekdays exceeds that of weekends.
3. Build predictive models for daily trip counts using mobility-derived features.

2 Data Collection and Preprocessing

The raw dataset consists of 312 records labeled as "activity" (movement) or "visit" (stationary). Each record contains start/end timestamps and geocoordinates in the Europe/Istanbul timezone. Key preprocessing steps:

- Parsed ISO-8601 timestamps and computed travel duration in minutes and distance in kilometers for *activity* events.
- Extracted calendar features: date, day-of-week, and departure hour.
- Flattened nested JSON to normalize fields and removed non-*activity* events from the distance analysis.

- Aggregated *activity* events by day to compute total distance, total duration, and trip count.

3 Exploratory Data Analysis

3.1 Average Daily Distance by Day of Week

Figure 1 shows the mean total travel distance for each day of the week. Weekdays (Tuesday through Friday) exhibit higher average distances (around 20–24 km) compared to Monday (3.4 km) and Sunday (19.7 km). The spike on Thursday and Friday suggests regular long commutes or errands.

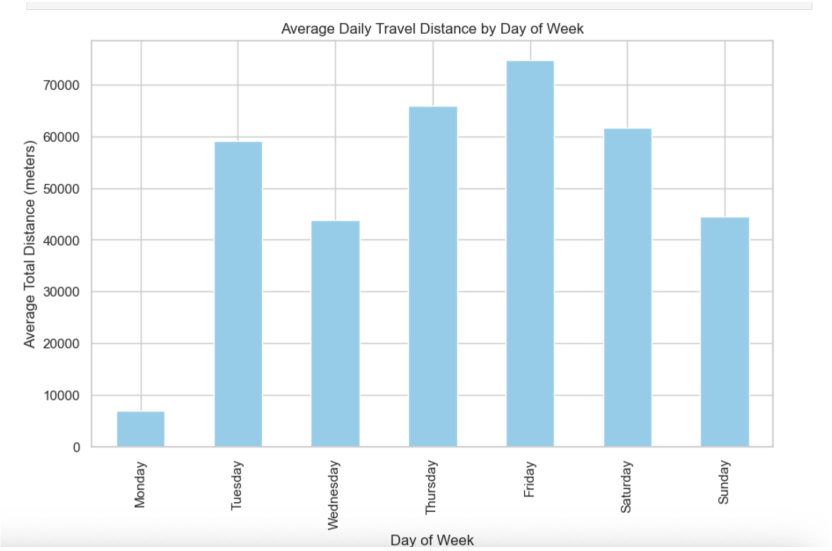

Figure 1: Average Daily Travel Distance by Day of Week

3.2 Distribution of Daily Distance and Trip Count

To understand variability, we plotted boxplots of daily distances and trip counts (Figures 2 and 3). The distance distribution confirms that weekdays have more consistent—but sometimes extreme—travel days, while weekends show wider spread and outliers due to irregular outings. Trip-count variability is higher on weekdays, reflecting routine commute patterns.

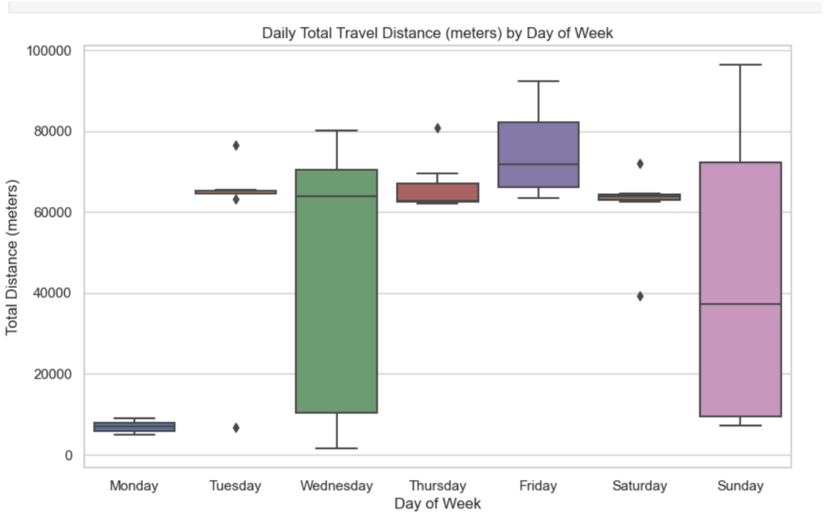

Figure 2: Boxplot of Total Daily Travel Distance by Day of Week

3.3 Temporal Travel Patterns

Figure 4 is a heatmap of trip-start counts by weekday and hour. We observe clear commute peaks between 07:00–09:00 and 15:00–18:00 on weekdays. Weekend activity begins later, after 10:00 AM, and is scattered throughout the day, especially on Saturdays.

4 Hypothesis Testing

We tested:

$$\begin{aligned} H_0 : &\quad \mu_{weekday} = \mu_{weekend} \\ H_1 : &\quad \mu_{weekday} > \mu_{weekend} \end{aligned}$$

Using a one-tailed Welch's t-test at $\alpha = 0.05$, we found:

- Mean weekday distance = 57.90 km
- Mean weekend distance = 55.94 km
- t-statistic = 0.23, one-tailed p-value = 0.41

Since $p > 0.05$, we fail to reject H_0, indicating no statistically significant difference in daily travel distance between weekdays and weekends.

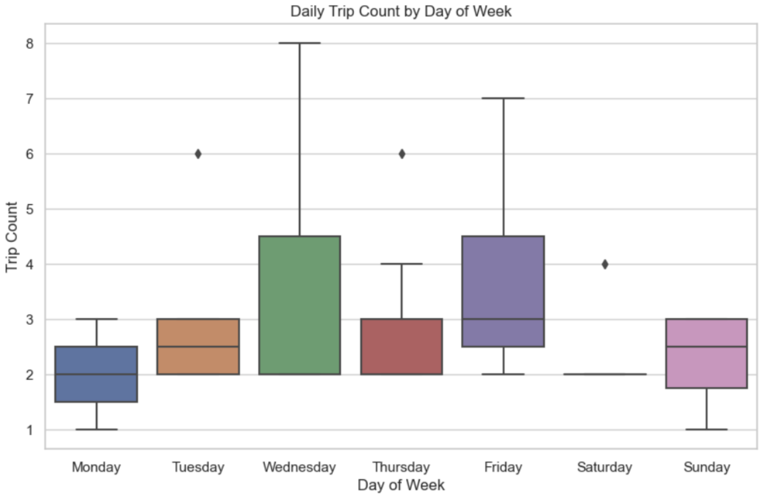

Figure 3: Boxplot of Daily Trip Count by Day of Week

5 Feature Engineering for Prediction

For each day we engineered:

- **Calendar features**: day-of-week (0–6), is_weekend flag.
- **Lag features**: previous-day and previous-week distances (km) to capture short-term trends.
- **Mobility features**: total distance (km), 95$^{\text{th}}$-percentile radius (km), number of unique POIs visited.

Missing values and days without recorded movement were filled with zeros, resulting in 50 days of usable data.

6 Modeling Approach

We used an 80%/20% chronological split to train and hold out data. Three models were tuned via 5-fold TimeSeriesSplit and GridSearchCV:

- Random Forest Regressor
- Gradient Boosting Regressor

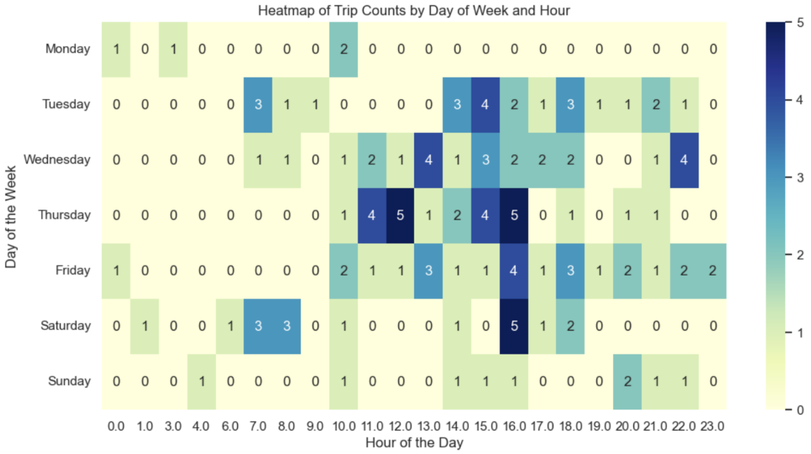

Figure 4: Heatmap of Trip Counts by Day of Week and Hour

- XGBoost Regressor

Three iterative hyperparameter grids progressively expanded model complexity and regularization options.

7 Results

Table 1 summarizes hold-out performance. Random Forest achieved the lowest MAE (0.61 trips) and highest R^2 (0.60), explaining 60% of daily variability. XGBoost was a close second (MAE 0.61, R^2 0.44), while Gradient Boosting underperformed (R^2 0.23).

Model	MAE (trips)	R^2
Random Forest	0.61	0.60
XGBoost	0.61	0.44
Gradient Boosting	0.62	0.23

Table 1: Hold-out MAE and R^2 for predictive models

Figure 5 shows the Random Forest feature importances: distance, radius, and POI count drive most of the predictive power, while calendar flags and lags contribute minimally.

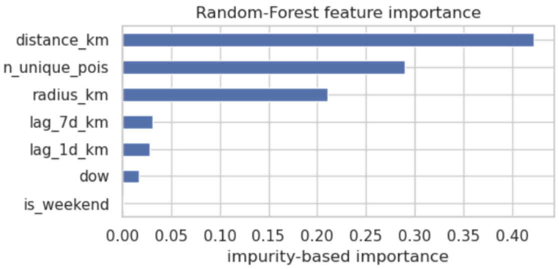

Figure 5: Feature Importance from Random Forest Model

Figure 6 compares actual vs. predicted trip counts on the hold-out set, illustrating that most predictions fall within ±1 trip of the true value.

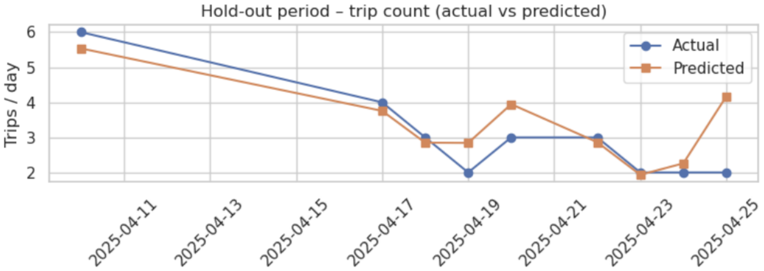

Figure 6: Actual vs Predicted Trip Counts on Hold-out Set

8 Conclusion

Our analysis reveals no significant difference between weekday and weekend travel distances for Feb–Apr 2025. Mobility-derived features (distance, radius, POIs) form strong predictors of daily trip counts, with Random Forest providing the best performance in terms of error and explained variance. Future improvements could incorporate exogenous data such as weather or holiday indicators to capture additional variability.