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. I 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 and compare predictive models for daily trip counts using mobility-derived features.

2 Data Collection and Preprocessing

The raw dataset comprised 312 JSON records, each classified as:

- **Activity** – periods of motion (with distance and start/end coordinates).
- **Visit** – stationary stays at points of interest (ignored for distance analysis).

Key preprocessing steps:

1. **Parsing timestamps:** ISO-8601 strings converted to `datetime` (Istanbul time), and durations computed in minutes.
2. **Distance calculation:** `distanceMeters` converted to kilometers for all *activity* events.
3. **Feature extraction:** From each record I extracted:
   - Calendar features: date, day-of-week, hour.
   - Geographic features: start/end latitude and longitude.
4. **Daily aggregation:** Grouped by date to compute:
   - Total distance (km), total duration (min), number of trips.
   - 95th-percentile radius of movement (km).
   - Number of unique POIs visited.

3 Exploratory Data Analysis

3.1 Average Daily Distance by Day of Week

Weekdays (Tuesday–Friday) show higher mean distances (20–24km) than Monday (3.45km) and Sunday (19.74km).

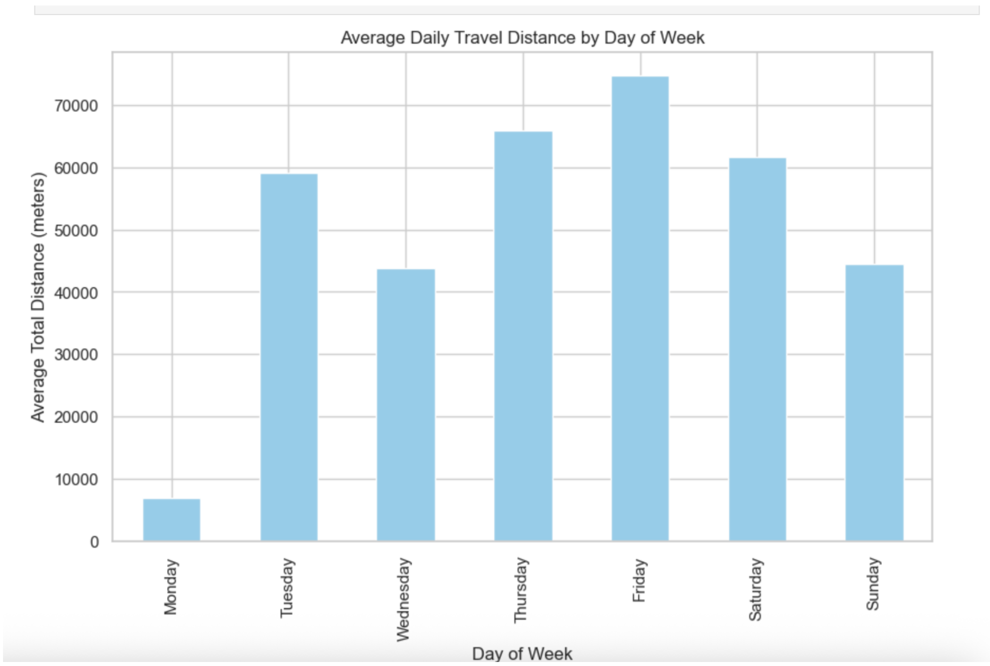

Figure 1: Average Daily Travel Distance by Day of Week

3.2 Distribution of Daily Distance and Trip Count

Boxplots reveal:

- *Distance:* Weekdays are more consistent; weekends have wider spread and outliers.
- *Trip count:* Higher variability on weekdays, reflecting regular commute peaks.

3.3 Temporal Travel Patterns

A heatmap of trip-start counts by weekday and hour shows clear commute windows on weekdays and more irregular, later starts on weekends.

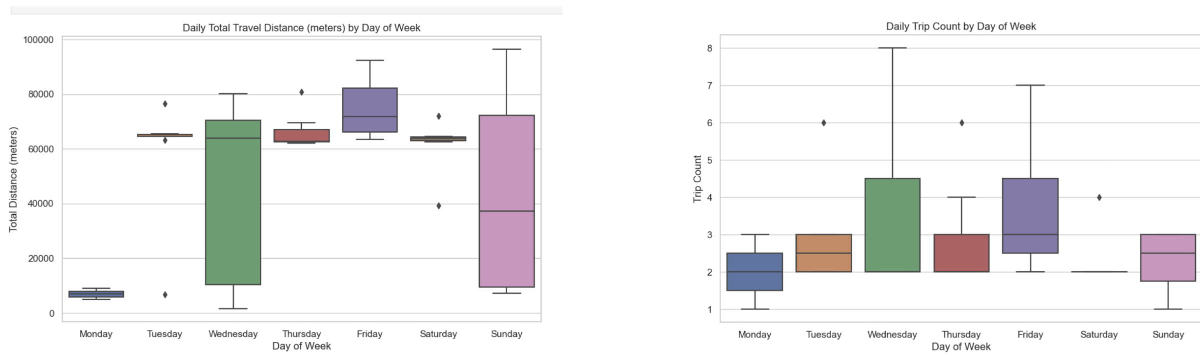

Figure 2: Left: Total Daily Distance; Right: Daily Trip Count by Day of Week

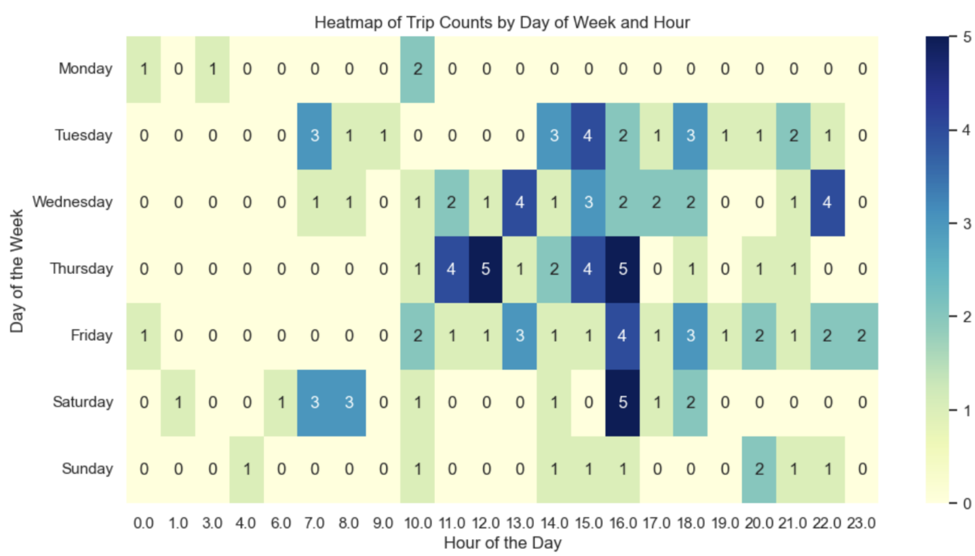

Figure 3: Heatmap of Trip Counts by Day of Week and Hour

4 Hypothesis Testing

I test:

$$H_0:\ \mu_{\text{weekday}} = \mu_{\text{weekend}} \quad \text{vs.} \quad H_1:\ \mu_{\text{weekday}} > \mu_{\text{weekend}}$$

using a one-tailed Welch's t-test at $\alpha = 0.05$.

Results:

- Mean weekday distance: 57.90km
- Mean weekend distance: 55.94km
- t-statistic: 0.23
- Two-tailed p-value: 0.8215
- One-tailed p-value: 0.4108

Since $p_{\text{one-tailed}} = 0.4108 > 0.05$, we **fail to reject** H_0. There is no statistically significant evidence that weekday travel exceeds weekend travel in our sample.

5 Feature Engineering for Prediction

For each day I constructed:

- **Calendar features:** day-of-week (0–6), `is_weekend`.
- **Lag features:** distance (km) from previous day and previous week.
- **Mobility features:** total distance (km), 95th-percentile radius (km), number of unique POIs.

All missing values and non-movement days were filled with zeros, yielding ≈ 50 days of data.

6 Modeling Approach

I used an 80%/20% chronological split and 5-fold `TimeSeriesSplit` within `GridSearchCV` to tune three regressors:

- **Random Forest Regressor**
- **Gradient Boosting Regressor**
- **XGBoost Regressor**

Hyperparameter grids were expanded in three stages to balance depth, regularization, and sampling controls.

7 Results

Table 1 summarizes hold-out performance. Random Forest attains the lowest MAE and highest R^2.

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

Feature importance from the Random Forest (Figure 4) shows that mobility features (distance, radius, POI count) dominate, while calendar and lag features contribute minimally.

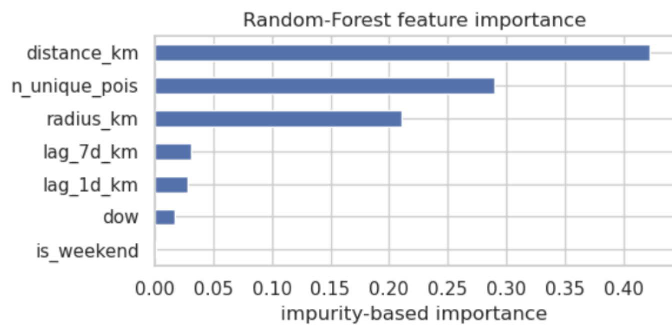

Figure 4: Feature Importance from Random Forest Model

Actual vs. predicted trip counts on the hold-out set (Figure 5) illustrate that most errors lie within ±1 trip.

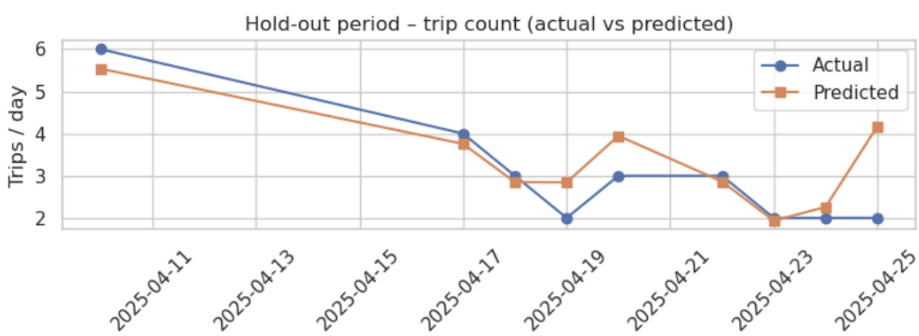

Figure 5: Actual vs. Predicted Trip Counts on Hold-out Set

8 Conclusion

- No significant difference was found between weekday and weekend travel distances (Feb–Apr 2025).
- Mobility-derived features (distance, radius, POIs) are strong predictors of daily trip counts.
- Random Forest achieved the best balance of low MAE (0.61 trips) and high explained variance (60%).
- Future work could use more sample days or additonal variables (e.g. weather, holidays) to capture more variability.