Uber Trip Analysis — Project Report

Author: Khushi Agrawal

---

1. Executive Summary

This project presents an end-to-end analysis of Uber trip data using SQL, Python, and Power BI. The goal is to clean and enrich raw trip logs, explore spatial and temporal patterns, detect anomalies, cluster trip types, and demonstrate a short-term forecasting approach. The interactive Power BI dashboard visualizes hotspots, peak hours, trip purpose breakdowns, and model outputs to support operational decision-making.

Key finding: 503 trips (\~43%) lacked a recorded trip purpose — a significant data-quality gap that affects segment-level insights and downstream analyses.

---

2. Project Purpose and Objectives

Purpose: Convert messy Uber trip logs into actionable business insights via a reproducible pipeline and interactive dashboard.

Objectives:

- Ingest and store trip data using SQL.
- Clean and feature-engineer the dataset using Python.
- Explore patterns in time, space, purpose, distance, and duration.
- Detect anomalies and flag data-quality issues.
- Cluster trips to identify ride-types and behaviour segments.
- Produce an interactive Power BI dashboard for stakeholders.
- Provide recommendations for operations and data collection improvements.

---

3. Data Description

Primary fields (examples):

- trip_id — Unique identifier
- pickup_*datetime, dropoff_*datetime — Timestamps
- duration_min — Trip duration in minutes
- purpose — Trip purpose (business, personal, etc.)

---

4. Methodology

4.1 SQL

- Created normalized tables and indexes for efficient querying "create_table.sql"
- Performed initial transformations and filtering (e.g., remove duplicates, standardize columns) using load_data.sql and queries/analysis_queries.sql.

4.2 Python (ETL & EDA)

- Used pandas and numpy for cleaning & feature engineering.
- Key steps:
- Parse and standardize timestamps, derive date, day_of_week, hour, time_of_day.
- Compute duration_min and validate positive durations.
- Handle missing purpose values: flag for imputation or exclusion.
- Flag anomalies: extremely long distances, negative durations, duplicated trips, or implausible coordinates.
- Export a cleaned CSV (UberDataset_Cleaned.csv) for Power BI consumption.

4.3 Power BI

- Imported cleaned CSV / connected to the SQL view.
- Built visuals: map with clustering/heatmap, hourly heatmap, trip purpose breakdown (bar/treemap), distance/duration histograms, clustering visuals, and forecasting card.
- Created slicers for date range, purpose, and distance bucket for interactive exploration.

---

5. Data Cleaning & Quality Issues

Major issues identified:

- **Missing trip purpose:** 503 records (\~43%) without purpose
- **Timestamp inconsistencies:** Mixed formats and timezone ambiguity.
- **Outliers:** Trips with very long distances or zero/negative durations.
- **Duplicate records:** Some trips repeated due to ingestion errors.

Actions taken:

- Standardized timestamps and truncated timezone ambiguity by documenting the assumed timezone.
- Filtered or flagged anomalies; added an anomaly_flag column to the cleaned dataset.
- Kept missing-purpose records but flagged them — recommendations include collecting purpose at booking or adding a mandatory field.

---

6. Exploratory Data Analysis (high-level findings)

- **Temporal patterns:** Clear commute peaks during morning (7–10 AM) and evening (5–8 PM). Weekday/weekend patterns differ, with more leisure trips on weekends.
- **Spatial patterns:** Map visualizations show concentrated hotspots in business districts and transit hubs. Some unexpected origin/destination clusters were investigated as anomalies.
- **Trip purpose distribution:** Business and Personal categories dominate among labeled trips; however, missing-purpose entries skew distribution.
- **Clustering:** Using distance, duration, and purpose (when present) produced distinct clusters resembling "short city trips", "medium commute trips" and "long intercity trips".
- **Forecasting:** A short demo forecast (simple ARIMA / Exponential Smoothing) shows predictable cyclical demand for short horizons — appropriate for staffing/fleet decisions.

---

7. Visuals & Dashboard Guide

Placeholders for visuals are included in screenshots. Recommended dashboard pages:

1. **Overview:** Key KPIs (total trips, average distance, avg duration, percent missing purpose)
2. **Map & Hotspots:** Interactive map with pickup density and top zones
3. **Time Analysis:** Hourly heatmap and weekday comparison
4. **Trip Segments:** Purpose breakdown, distance buckets, clustering results
5. **Anomalies & Data Quality:** Table of flagged anomalies and steps taken
6. **Forecasting:** Short-term demand forecast with confidence intervals

Each visual should have clear tooltips and short notes on interpretation.

---

8. Recommendations

Operational:

- Improve booking flows to capture purpose reliably (mandatory field or post-ride quick survey).
- Use forecasted demand to allocate drivers to identified hotspots during peak times.
- Investigate top anomaly clusters — they may indicate data collection problems or fraud.

Data collection & governance:

- Standardize timestamp formats and store timezone metadata.
- Implement validation rules for duration and distance at ingestion.
- Keep a data-quality dashboard to monitor missing fields and anomaly rates.

---

9. Limitations

- The dataset may not include fare/driver supply and city-level demand context.

- Missing purpose in a substantial portion of records limits segmentation accuracy.
- Forecasting demo is short-horizon and illustrative; production forecasts require robust modeling, cross-validation, and external features (weather, events).

---

10. Contact & Credits

Author: Khushi Agrawal

- Email: khushiagrawal0893@gmail.com
- LinkedIn: www.linkedin.com/in/khushiagrawal1