Emir Mirza 30704

Final Report: Uber Trip Data Analysis Project

1. Introduction This project, conducted for the DSA210 Introduction to Data Science course, explores and models ride-sharing behaviors through my personal Uber trip data. The primary goal was to understand trip characteristics, pricing trends, and cost-related factors—particularly the influence of fuel prices—while developing proficiency in data science techniques such as EDA, statistical hypothesis testing, and machine learning.

The dataset consists of Uber trip logs including timestamps, trip distance, fares, and driver response times. Additional LPG fuel price data was manually collected and merged for enriched analysis. Through this project, both user-side insights (how to save time/money) and Uber-side recommendations (pricing optimization, driver deployment strategies) are explored.

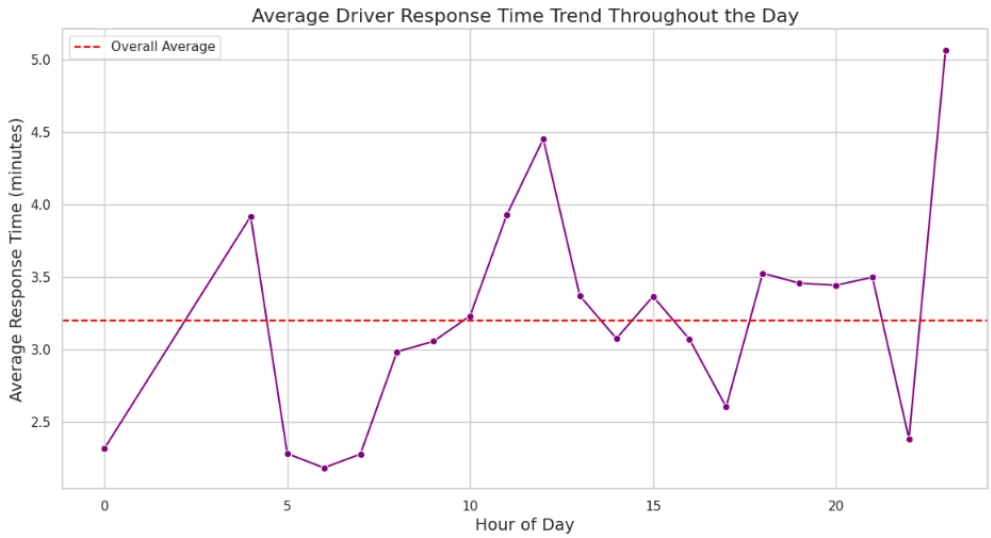
2. Methodology

Data Cleaning & Feature Engineering

- Converted time columns to datetime objects and extracted features: hour, weekday, month.
- Calculated trip durations and driver response times.
- Removed invalid data entries (e.g., fare=0 for completed trips).
- Categorized trip distances into bins for visual comparison.
- Integrated external LPG price data based on request date.

Exploratory Data Analysis (EDA)

- Analyzed trip duration patterns by hour and distance bin.
- Investigated the fare-distance relationship and fare-per-km distribution.
- Explored ride frequency by hour and day.
- Visualized driver response time variations across different hours.



Statistical Hypothesis Testing

1. T-test comparing fare/km between peak and off-peak hours.
2. T-test comparing trip durations on weekdays vs weekends.
3. Pearson correlation between response time and hour of day.
4. Pearson correlation between LPG prices and fare amounts.

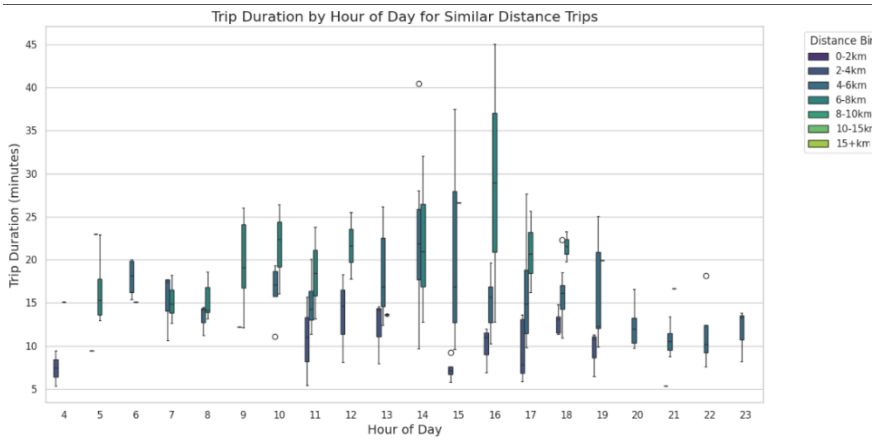
Machine Learning Models

1. **Linear Regression** – Predict fare using distance, time, LPG price.
2. **Random Forest Regressor** – Predict profit margins.
3. **K-Means Clustering** – Segment trips into types based on fare, distance, fuel cost, margin.
4. **Classification Model** – Identify whether a trip happened during peak hours.

3. Results and Discussion

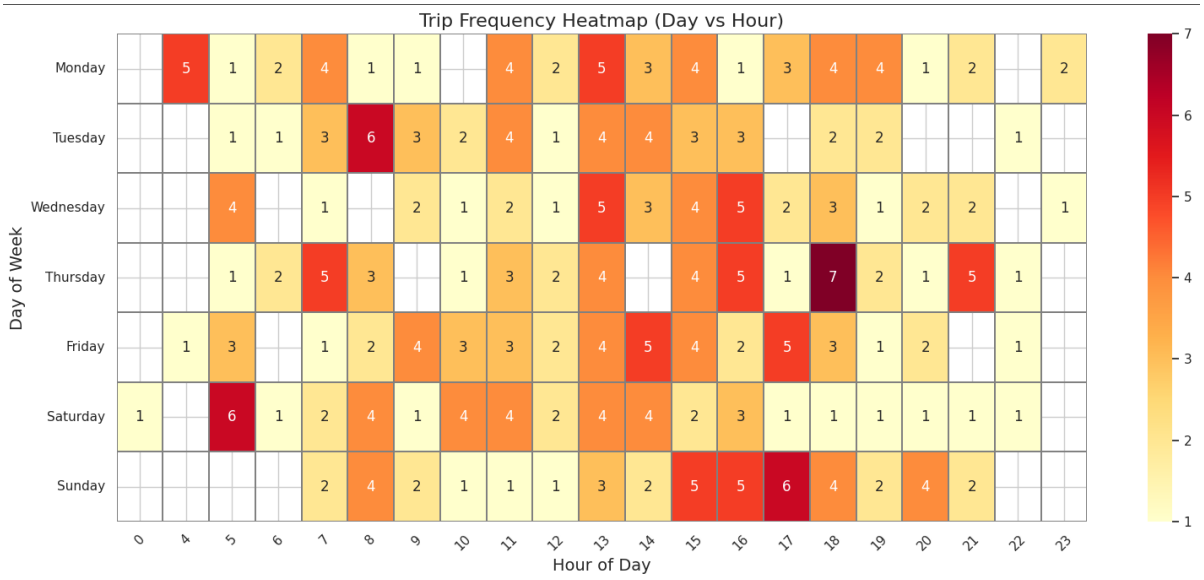
Key Findings

- **Trip Duration:** Peak hours (8-10am, 5-7pm) show higher durations due to traffic. Late-night trips are quicker.



- **Fare Analysis:** Strong positive correlation with distance (r = 0.85). Fare/km rises during peak hours.

- **Trip Frequency:** Most trips occur between 10am–8pm. Weekdays (Wed–Thu) are busier.



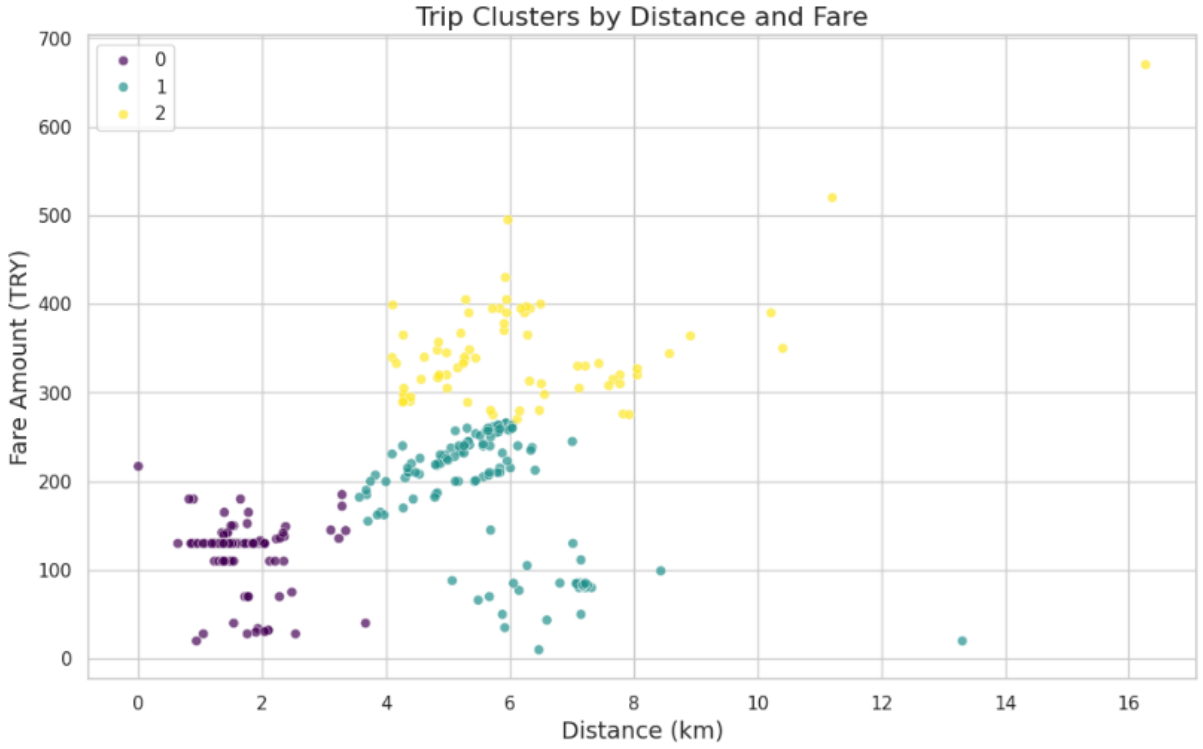
Driver Response: Quickest in early morning (4–6am), slowest in evening peak.

Statistical Tests

- Fare/km is significantly higher during peak hours ($p < 0.01$).
- No significant duration difference between weekdays and weekends.
- Response time increases moderately throughout the day ($r = 0.42$).
- Fare moderately correlates with LPG prices ($r = 0.59$).

ML Model Performance

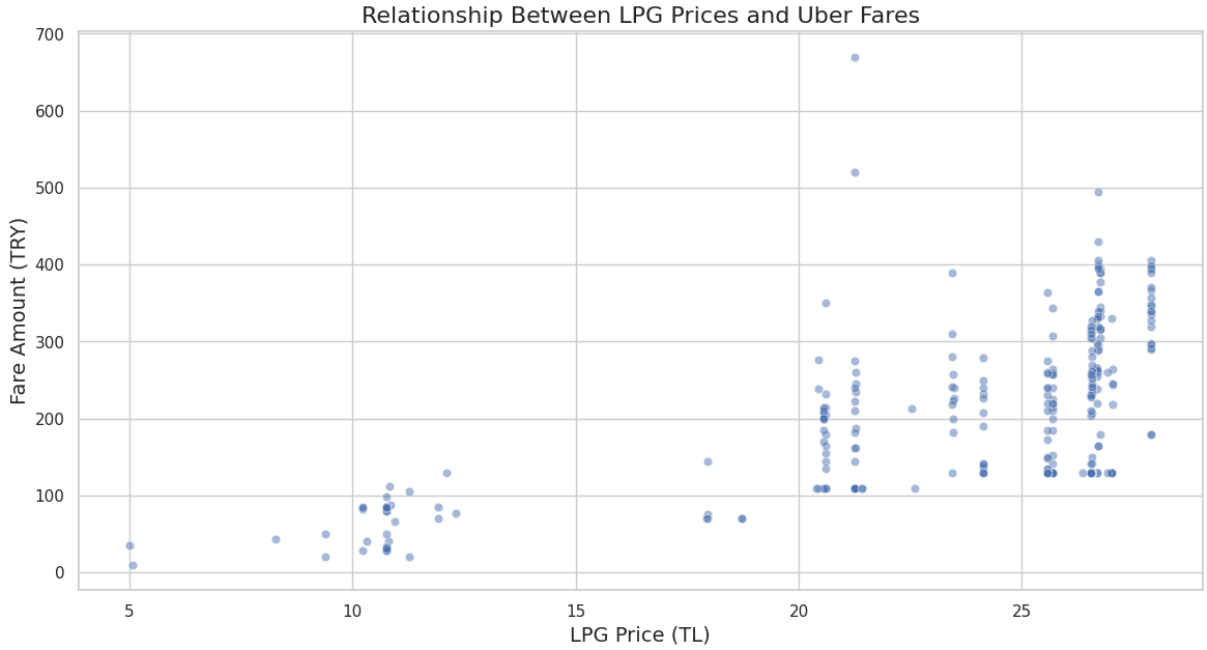
- **Linear Regression (Fare Prediction):** R^2 = 0.84, RMSE = 39.46 TRY.
- **Random Forest (Profit Margin):** R^2 = 0.84, RMSE = 37.23 TRY.
- **Clustering:** Identified 3 segments: short economic, medium standard, long premium trips.

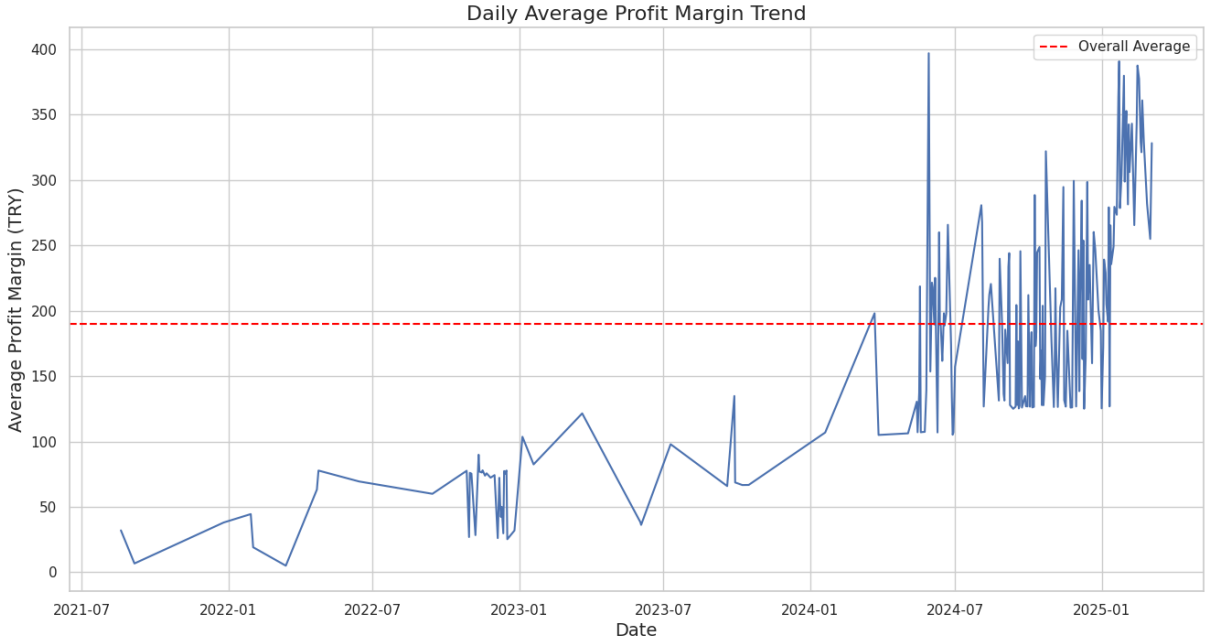


- **Classification (Peak Hour):** Accuracy = 100%. Most predictive feature: request_hour.

Fuel Sensitivity Insight

- LPG price is a major determinant of fare and profit.
- When LPG > 25.68 TRY, profit margin drops by 28.5%.
- Linear regression shows LPG price coefficient = +77.15 TRY.





4. Future Work To deepen analysis and improve predictive accuracy, future iterations of this project could:

- Integrate weather and traffic congestion data to explain more variance in trip duration.
- Analyze longer-term trends by acquiring additional Uber data from other months/years.
- Employ deep learning models for more complex fare prediction.
- Explore city-wide demand maps using geolocation data.

Conclusion This project successfully demonstrated the application of data science tools on real-world Uber data. Key insights included traffic-driven delays, cost implications of fuel price volatility, and high-accuracy fare and peak classification models. These findings are valuable for both riders and platform operators. The integration of external fuel cost data added a critical layer of business relevance. Going forward, the inclusion of more contextual data and temporal expansion will offer even more powerful insights.

Data Sources

1-Uber Trip Data:

Obtained directly from my personal Uber account via Uber's Data Download Tool.

2-: https://www.epdk.gov.tr/Detay/Icerik/3-0-158/akaryak%C4%B1tfiyat

Manuelly collected from EPDK (Turkish Energy Market Regulatory fuel price archive.