GROUP 7B

Data and Programming Group Project Report

This report presents an analytical study of Uber fare dynamics, identifying key influencing factors and providing strategies for optimizing booking decisions.

Table of Contents

1. Introduction

Uber, one of the largest taxi services globally, operates in a highly competitive market where pricing is a critical factor for its users. A major challenge faced by Uber is its dynamic pricing model, which often leaves customers confused about fare fluctuations. This unpredictability makes it difficult for users to plan their travels cost-effectively. Understanding and predicting Uber fares based on variables such as time, day, and location can empower users to make more informed and confident booking decisions. Additionally, while Uber offers a fare reservation feature, it is underutilized due to users' lack of confidence in whether the reserved fare aligns with potential future price changes at the end of their ride.

1.1 Why Our Team is Proposing This Topic

The primary motivation for this project is the growing demand for transparency and cost-effective travel solutions in the ride-hailing industry. By analyzing fare patterns and trends, this project aims to enhance user satisfaction by enabling better travel planning and minimizing unexpected charges. Accurate fare prediction can help users book rides at optimal rates, leading to more informed and confident booking experiences. Furthermore, this project will provide valuable insights into Uber's pricing strategy and offer recommendations to improve customer engagement.

1.2 Main Questions

This project seeks to bridge the gap between user expectations and Uber's pricing strategy by helping us find the answers to the following questions:

1. **What are the primary factors influencing Uber fare prices?**
2. **How do fare prices vary based on time of day, day of the week, and geographical location?**
3. **What is the optimal time frame for users to book rides to secure the best rates?**

1.3 Related Work and Inspirations

Research in *Transportation Research Part C: Emerging Technologies* highlights the influence of temporal and spatial factors, such as traffic and demand density, on surge pricing. These findings serve as a basis for our exploration of Uber's fare patterns[1]. Similarly, a Harvard Business Review case study on Uber's surge pricing underscores the importance of balancing revenue optimization and customer loyalty, emphasizing transparency in fare models[2].

Studies from *The Journal of Business Research* demonstrate how machine learning optimizes pricing strategies using historical and behavioral data, which aligns with our objective to predict Uber fares accurately[3]. Additionally, tools like "RideGuru" provide fare estimates but lack the depth needed to account for dynamic price changes, a gap our project seeks to address[4].

Building on these studies and discussions, we aim to analyze Uber fare patterns using a combination of temporal, spatial, and behavioral factors.

2. Data Preparation

2.1 Data Description

The dataset used in this project focuses on transactional data from Uber, the world's largest taxi service. The dataset contains trip-level information, including geographical coordinates, time stamps, and passenger details, enabling comprehensive analysis of fare determinants. The dataset was obtained from Kaggle repository[5] and includes essential trip details. The table below summarizes the features in the dataset:

[1] Transportation Research Part C: Emerging Technologies. 2020. "Dynamic Pricing in Ride-Hailing Services: Factors Affecting Surge Pricing."
[2] Harvard Business Review. 2017. "How Uber's Surge Pricing Really Works." *Harvard Business Review*. https://hbr.org/.
[3] *The Journal of Business Research*. 2021. "Applications of Machine Learning in Dynamic Pricing Strategies."
[4] RideGuru. "Ride Fare Estimator." Accessed November 2024. https://ride.guru/.
[5] https://www.kaggle.com/datasets/yasserh/uber-fares-dataset

Table 2-1 Summarizes the features in the dataset

Feature	Description
key	A unique identifier for each trip.
fare_amount	The cost of each trip in USD.
pickup_datetime	Date and time when the trip began (meter engaged).
passenger_count	The number of passengers in the vehicle (input by the driver).
pickup_longitude	The longitude where the trip started.
pickup_latitude	The latitude where the trip started.
dropoff_longitude	The longitude where the trip ended (meter disengaged).
dropoff_latitude	The latitude where the trip ended.

This dataset contains **200,000 records** spanning trips from **2009 to 2015**, providing a comprehensive view of Uber's fare patterns over time. The coverage allows us to analyze trends and variations in pricing influenced by factors such as distance, passenger count, and time of day.

2.2 Data Cleaning

Cleaning Steps

1. **Handling Missing Values**

   The dataset was initially examined for missing values. Fortunately, one missing value was identified and removed.

2. **Removal of Partial 2015 Data**

   While the dataset spans from 2009 to 2015, the data for 2015 was incomplete. To ensure consistent monthly analysis, all 2015 records were removed, resulting in a cleaner and more reliable dataset for analysis.

3. **Filtering Invalid Passenger Counts**

   The dataset included a passenger_count variable, which sometimes had invalid values such as 0 or excessively high values of 50 or more. These entries were deemed unrealistic and excluded from the dataset to maintain data integrity.

4. **Filtering Invalid Latitude and Longitude**

   The dataset included pickup and drop-off latitude and longitude coordinates for each trip with a focus on the city of Manhattan. However, some coordinates were incorrectly recorded (e.g., values outside of valid latitude and longitude ranges). These records were identified and removed.

Figure 2.1 Detecting and removing outliers in Uber pickup data

5. **Removing Zero-Distance Records**

   Some entries had identical coordinates for both pickup and drop-off locations, which led to a calculated distance of zero. These entries are removed.

6. **Resulting Dataset**

   After these cleaning steps, the dataset size was reduced from **200,000** to **179,501** records, ensuring a higher level of data quality for the subsequent analysis.

2.3 Data Feature Engineering

To capture the factors influencing Uber fares, the pickup_datetime column was transformed to extract key features such as the hour of the day, day of the week, and month. These features enable a detailed analysis of fare variations during different times of the day, across weekdays and weekends, and between months, helping identify peak and off-peak trends.

To understand the relationship between trip distance and fare amounts, the Haversine formula was used to calculate the shortest distance between pickup and dropoff locations. This formula, which computes distances over the Earth's surface, is essential for deriving an accurate distance feature based on latitude and longitude data while considering the curvature of the earth. The calculated distance serves as a crucial predictor of fare amounts.

The Haversine formula is given by:

$$d = 2 \times r \times \arcsin\left(\sqrt{sin^2\left(\frac{\Delta\emptyset}{2}\right) + \cos(\emptyset_1)\cos(\emptyset_2)\, sin^2\left(\frac{\Delta\lambda}{2}\right)}\right)$$

where:

- ϕ1,ϕ2: Latitude of the pickup and dropoff points in radians.
- Δϕ,Δλ: Differences in latitude and longitude between pickup and dropoff points.
- r: Radius of Earth (~6371 km).

pickup_hour	pickup_month	pickup_weekday	distance_km
19	5	Thursday	1.683323
20	7	Friday	2.457590
21	8	Monday	5.036377
8	6	Friday	1.661683
17	8	Thursday	4.475450
2	2	Saturday	0.000000
7	10	Sunday	11.731015

Figure 2.2 Samples of derived dataset

3. Exploratory Analysis

The exploratory analysis focuses on understanding key patterns in the Uber dataset related to trip distance, fare prices, temporal factors, and passenger behavior. By visualizing data distributions and relationships, we can uncover insights into factors affecting Uber fares. These findings will guide further feature engineering and predictive modeling.

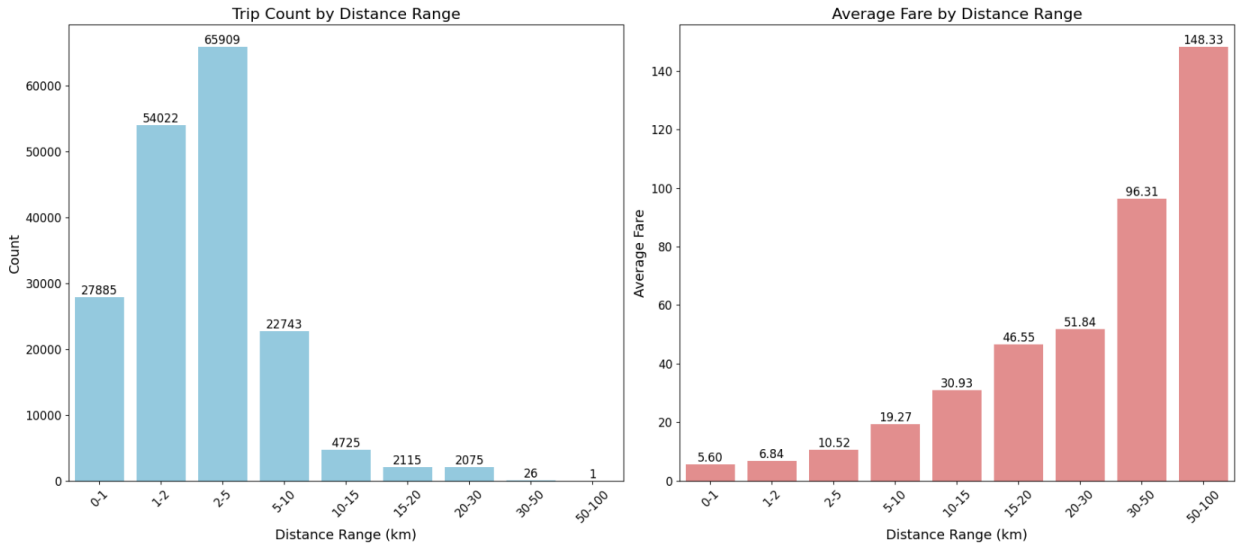

Figure 3.1 Distribution of trip counts and fares by distance range

The figure above highlights the distribution of trips and average fares based on distance ranges. Most trips are short distances (2–5 km), while longer trips above 30 km are rare but correspond to significantly higher fares due to extended distances. This trend reflects a proportional relationship between distance and fare prices.

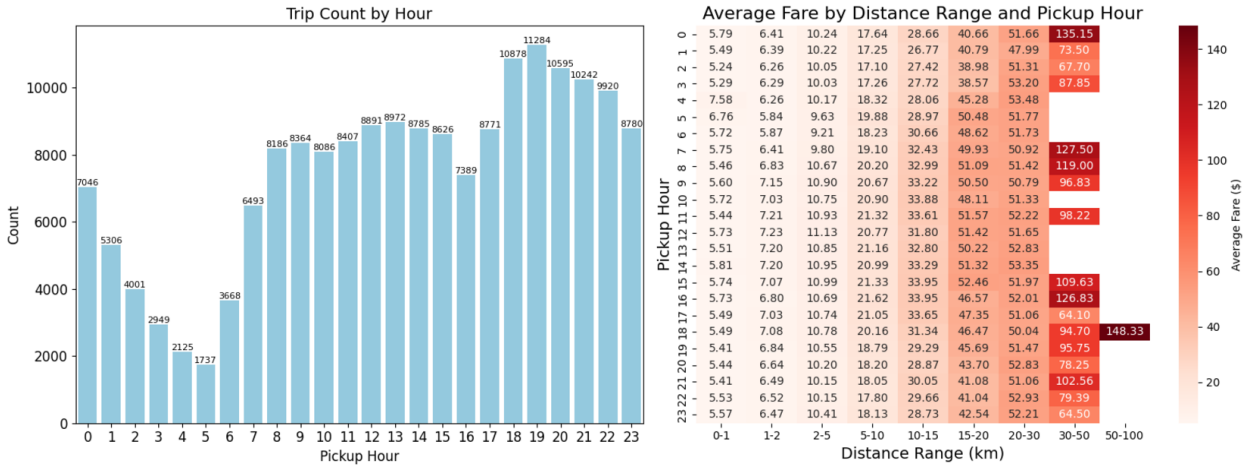

Figure 3.2 Hourly trends in trip counts and fares

Hourly analysis shows that trip counts peak during evening rush hours (6–8 PM), reflecting high demand. Conversely, average fares tend to be higher in the early morning and late-afternoon hours, potentially due to fewer available drivers and longer average trip distances during these times.

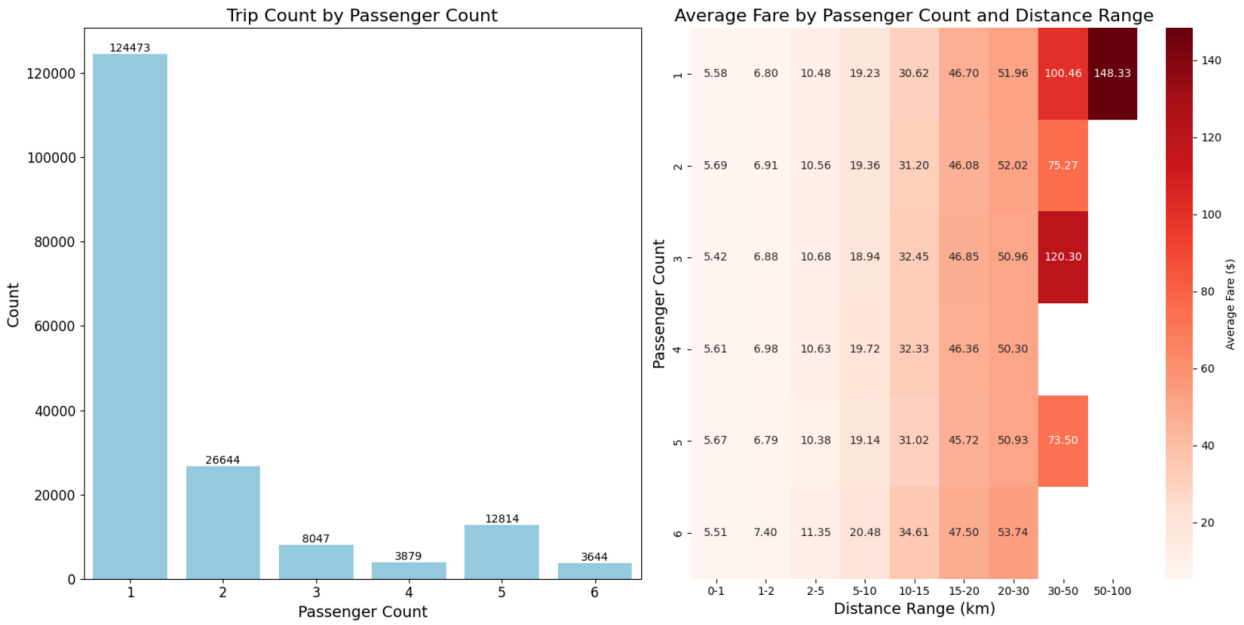

Figure 3.3 Passenger count distribution and fare patterns

Most trips involve a single passenger, indicating that Uber's services cater primarily to individual users. Average fares show a moderate increase with more passengers, likely reflecting the longer distances or larger vehicle requirements for group travel.

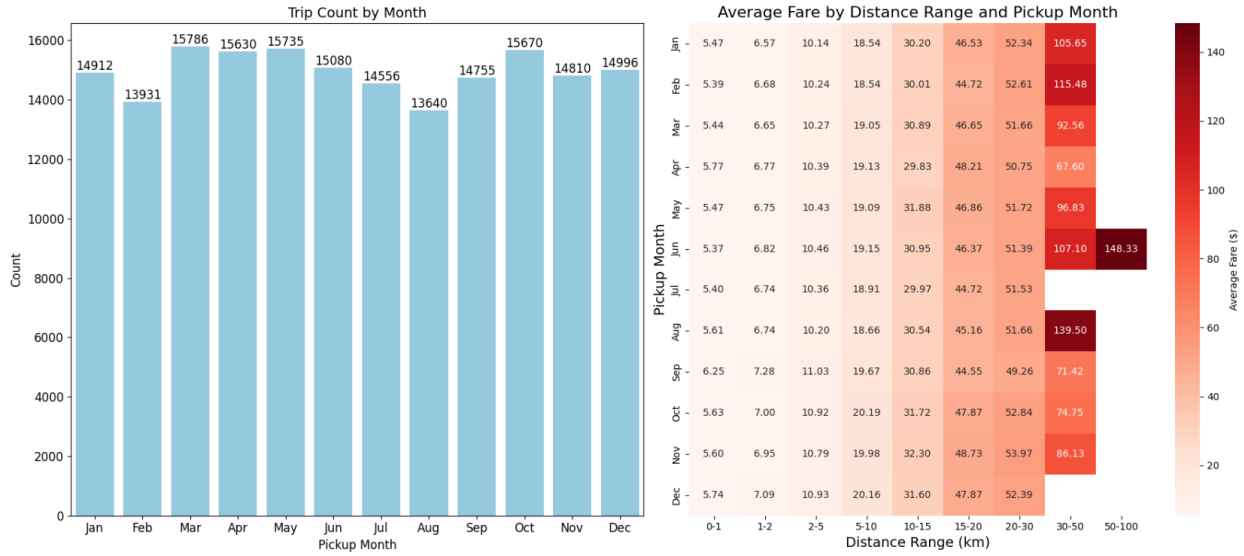

Figure 3.4 Monthly trends in trip counts and fares

Trip counts remain consistent throughout the year, with February showing a slight dip due to fewer days. Seasonal trends are observed in fares, with higher fares during the spring, late fall and winter months, potentially due to increased tourism or travel demand.

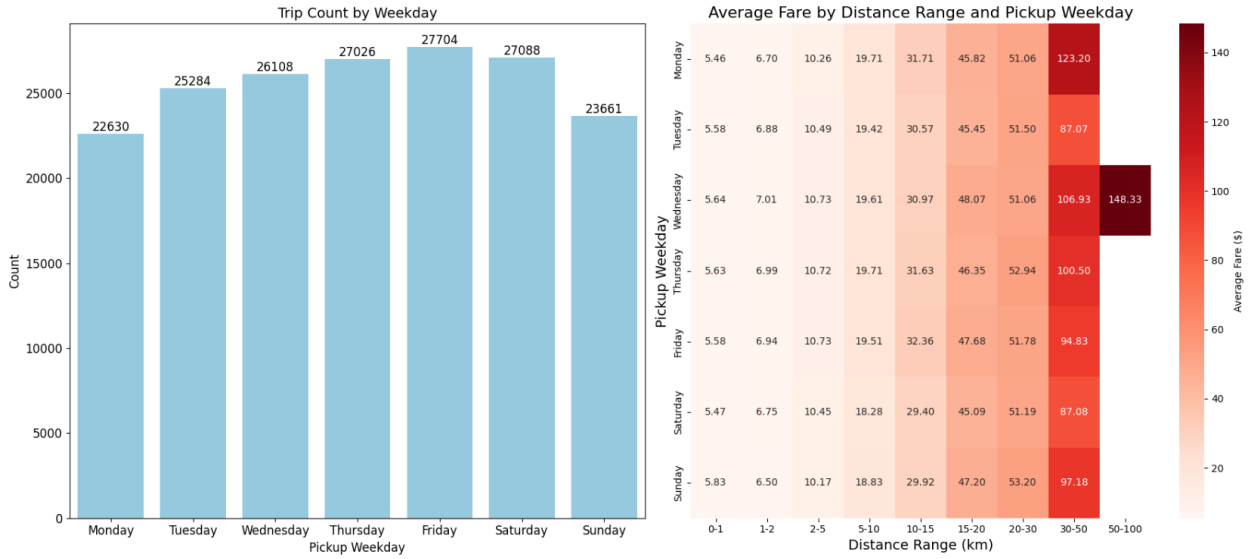

Figure 3.5 Weekly trip and fare distribution trends

The analysis reveals higher trip volumes on weekdays, particularly Fridays and Saturdays, indicating regular commute patterns and work-related travel. Average fares remain steady throughout the week but show a slight increase on Wednesdays and Fridays, possibly reflecting higher leisure travel demand.

4. Modeling

4.1 Regression Analysis on Fare Amount

The dataset was divided into training and testing sets with an 80:20 split to ensure robust model evaluation. To include categorical variables in the regression model, features such as passenger_count, pickup_hour, pickup_month, and pickup_weekday were converted into dummy variables. This transformation allowed the model to analyze categorical features as independent predictors. A linear regression model was then applied to predict the fare_amount using both the dummy variables and continuous variables such as distance.

The regression model achieved an R^2 of 0.73, explaining 73% of the variance in fare amounts. These results demonstrate a reasonably good fit, with room for improvement in capturing fare variability.

Feature *Importance*

Feature	Coefficient
distance_km	2.322500
pickup_hour_14	1.388262
pickup_hour_12	1.315604
pickup_hour_15	1.275267
pickup_hour_11	1.216853
pickup_hour_9	1.205832
pickup_hour_13	1.203347
pickup_hour_16	1.118514
pickup_hour_17	1.100708
pickup_hour_10	1.069960
pickup_hour_8	0.899490
pickup_hour_18	0.877144
passenger_count_6	0.712035
pickup_month_9	0.675899
pickup_month_10	0.663666
pickup_month_12	0.650400
pickup_month_11	0.647841
pickup_hour_19	0.575226
pickup_weekday_Thursday	0.412422
pickup_weekday_Friday	0.411675
pickup_weekday_Wednesday	0.399370
pickup_hour_5	0.353682
passenger_count_4	0.279412
pickup_hour_7	0.251765

Feature	Coefficient
pickup_hour_20	0.238512
pickup_month_5	0.223154
pickup_month_6	0.191641
pickup_weekday_Tuesday	0.185068
pickup_month_4	0.182997
passenger_count_3	0.150697
pickup_hour_23	0.149647
pickup_month_3	0.122947
pickup_hour_21	0.106449
passenger_count_2	0.095818
pickup_month_8	0.071224
pickup_month_7	0.046760
pickup_hour_22	0.031887
pickup_weekday_Saturday	0.006383
pickup_month_2	-0.011474
pickup_weekday_Sunday	-0.053707
passenger_count_5	-0.061613
pickup_hour_4	-0.219378
pickup_hour_1	-0.271396
pickup_hour_3	-0.292117
pickup_hour_6	-0.404686
pickup_hour_2	-0.407524

Figure 4.1 Feature importance from regression analysis

The most significant factor affecting fare prediction is the trip distance. For each additional kilometer traveled, the fare increases by 2.32 units. This highlights the direct and proportional relationship between distance and fare, making it the primary determinant in the pricing model. Temporal factors like the time of day strongly influence fares. Peak hours such as 2 PM, 12 PM, and 3 PM contribute to fare increases of 1.39, 1.32, and 1.28 units, respectively. These time slots reflect higher demand, leading to increased pricing. Seasonal and weekly patterns play a notable role in fare variability. For instance, trips in September add 0.68 units to the fare, likely due to seasonal demand. Similarly, Thursdays added 0.41 units, indicating increased travel activity on this day of the week. Trips involving larger groups, such as those with six passengers, raise fares by 0.71 units. This increase likely reflects the need for larger vehicles or additional services, which are more costly.

4.2 Model Performance Comparison

To identify the best-performing model for fare prediction, three models were trained using Decision Tree, Random Forest, and XGBoost. The dataset was split into an 80:20 ratio for training and testing. Using 10-fold cross-validation ensured robust performance evaluations, while Grid Search optimization was applied to tune hyperparameters for each model. The objective was to minimize Mean Absolute Error (MAE) and maximize R^2, ensuring accurate and reliable prediction.

		MAE	R2
Decision Tree	Before params tuning	3.237140	0.537366
	After params tuning	2.299720	0.739349
	best params	'max_depth': 5, 'min_samples_split': 10	
Random Forest	Before params tuning	2.375256	0.727114
	After params tuning	2.204537	0.747190
	best params	'max_depth': 10, 'n_estimators': 200	
XGBoost	Before params tuning	2.236225	0.730663
	After params tuning	2.210073	0.742470
	best params	'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 50	
Regression (baseline model)		2.261416	0.728671

Figure 4.2 Model performance comparison result

The Random Forest model emerged as the best-performing model for fare prediction, achieving the lowest MAE of $2.20 and the highest R^2 of 0.747, effectively explaining fare variability. While the Decision Tree showed significant improvement after tuning (MAE: 2.30, R^2: 0.739), and XGBoost performed comparably (MAE: 2.21, R^2: 0.742), Random Forest outperformed both by capturing complex relationships between features. The baseline regression model (MAE: 2.26, R^2: 0.728) provided a solid starting point but lacked the flexibility of tree-based models, reinforcing Random Forest's robustness as the optimal choice.

5. Conclusion

This analysis provides critical insights into factors influencing Uber fare prices and strategies to optimize booking. The most significant predictor of fare amounts is distance traveled, followed by factors such as pickup hour, day of the week, and month. Peak hours (e.g., early mornings and late afternoons) and weekdays (especially Wednesdays and Fridays) show higher fares due to increased demand, while non-peak hours offer the lowest fares. The passenger count has a moderate impact, with larger groups slightly increasing fares due to vehicle size requirements.

Urban and high-demand areas exhibit higher fare variability, reflecting the influence of location. To secure the best rates, take morning trips that aren't too early but early enough and avoid rush hours or weekends. These findings not only enhance fare predictability for users but also provide valuable insights for Uber's pricing strategies and demand management.