DSCI 5260- SECTION 001
BUSINESS PROCESS ANALYTICS

FINAL REPORT

PROJECT TITLE: Trend Analysis in Ride-Sharing Pricing and Commuting Patterns

Professor:
Dr. *Maryam Khatami*
Asst Professor

Submitted by:

- *Gayathri Kankanampati*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The rapid expansion of ride-sharing platforms has significantly transformed urban transportation by introducing dynamic pricing structures and influencing commuter behavior. This study conducts an in-depth analysis of ride-sharing transactions from the District of Columbia in December 2023, utilizing advanced data exploration and machine learning models to uncover key fare determinants. Through Random Forest and Gradient Boosting Regression, the study identified FAREAMOUNT, GRATUITYAMOUNT, and MILEAGE as the most influential variables affecting TOTALAMOUNT. Random Forest achieved the best predictive performance with an R^2 score of 0.96.

Further analyses revealed substantial variation in fare amounts across different cities, states, and times of the day, highlighting complex relationships between trip attributes and fare structures. Additionally, the dominance of credit card transactions and their association with higher gratuities were observed, reinforcing the influence of payment methods on final ride costs.

Based on the findings, the study provides actionable insights into fare prediction strategies, customer behavior patterns, and operational improvements for ride-sharing platforms. It is suggested that future research integrate external factors such as traffic conditions, weather variations, and special event data to enhance fare modeling further. The results contribute to a broader understanding of urban mobility pricing trends and offer a strong foundation for policymaking and business strategy development in the ride-sharing ecosystem.

Keywords: Ride-sharing analytics, Fare prediction, Random Forest Regression, Gradient Boosting Regression, Urban mobility trends

1 | INTRODUCTION

Urban transportation is the backbone of modern-day city infrastructures, facilitating quicker movement to millions of passengers. The taxi operations fill an important gap to connect private and public transport, delivering a point-to-point service to both locals and visitors. However, the arrival of web-based cab-hailing platforms has influenced the classical taxi service tremendously, changing tariff mechanisms, demand for services, and customer preferences. Isolating individual determinants of taxi fares, travel time, distance, and payment modes is necessary to optimize urban transport with higher passenger satisfaction.

Ride-hailing firms offer alternatives to taxis and hence are a great threat to traditional taxi companies. Research indicates that the introduction of these ride-hailing services has actually decreased the utilization of traditional taxi services. For instance, data of 44 Chinese cities, from 2010 to 2016, indicate that ride-hailing services had a strong substitution effect on traditional taxis, especially in eastern Chinese cities [1].

The introduction of Transportation Network Companies (TNCs) in Sweden has placed traditional taxi companies in thinking mode regarding quantifiable designs for impacts on prospective market share [2]. In Sweden, the introduction of TNCs had led traditional taxi operators to rethink and reframe their business models in order to ensure they remained competitive.

Various reasons cut customer satisfaction among taxi services. Service quality, reliability, price, and comfort were the reasons for customer satisfaction with online taxi services in Dhaka [3]. In Ho Chi Minh City, studies have found that comfort and price reasonably affect customer satisfaction, and explained by these variables a large percentage of variation in customer contentment [4].

In Calabar Metropolis, apart from cost, other parameters taken into consideration are product quality, service quality, price, facilities, and emotional factors for the customer's decision-making in choosing between e-cab service and normal taxi [5].

Besides this, customer satisfaction with ride-hailing is influenced by many features of the service such as usability of the app, driver behaviour, and safety procedures. In Sri Lanka, it is discovered that the choice of user in mobile app-based taxi services relies on easy booking, estimated arrival times, and simplicity of digital payment [6]. Similarly, in a comparative analysis of Uber, Careem, and Bolt operating in Qassim City, Saudi Arabia, the role of dynamic pricing mechanisms and promotions was evident in influencing customer choices [7].

Its only growth in markets like that of London has placed in sharp focus concerns regarding its regulation, with questions bordering around the issue of driver license, fare regulation, as well as security protocols for the passengers [8]. Granted that ride-sourcing service behaviors have upended market dynamics for London's taxi sector, an academically reviewed essay of critically informed literature expounds that economically and operationally, the traditional black cab trade appears to be experiencing issues.

A San Francisco data-based study sheds light on the way ride- hailing services not only replace regular taxis but even public transport, altering, thereby, the mobility patterns in cities [9].The emergence of mobile applications that call taxi services has brought new dimensions into fare systems and tipping culture. Many studies found that cashless payments via mobile apps are more likely to encourage tipping behaviour than cash payments, where tipping is generally lower. The dynamic surge pricing and distance-based tariffs also appear to work towards the generation of inequalities in affordability across socio-economic classes. They themselves in turn lead to the call for balanced regulations aimed at minimizing prejudicial competitive tactics while at the same time encouraging the economic profitability of taxi entrepreneurs.

With increased transport infrastructure, predictive models can be used in a way to forecast fares, routes, and stop times. Taxi operators and policymakers can leverage data-driven decision-making systems by taking into account geospatial real-time demand patterns so that eventually it leads to an improved system that is passenger-convenience-oriented.

Competition has thus raised an interesting challenge; whereby traditional taxi policies must react to the changing reality. Competition for 'ride-sharing technology' has set some tough questions for aged taxi business models, which have seen hardcore competition and the need for excellence in a customer experience.

Old-timers cabdrivers are also experimenting with such aspects as technology use, extensively researched fare schemes, and enhanced service quality to gain and retain passengers. Trips data analysis generates identification of city transportation modes, fare schemes, and rider habits. Distance covered, trip length, surcharges, tips, and types of payment are the most significant determinants of fares. Geospatial analysis finds trends such as hot pick and drop-off points, which will reveal where and when demand is highest.

Using these measures, policymakers and transport operators to city planners can influence the efficiency of the service, cost, and customer experience.

The present study therefore attempts to answer numerous doubt-raising questions. Some of them are:

- What is city cab ride price volatility dependent upon?
- How do fares fit into the end price?
- What are the most common payment methods, and how do they influence tipping culture?
- How are distances fitted into the fare models, and how is changing fare distribution shaping cities?
- How can predictive analytics benefit taxi fare systems and route optimization?

To properly respond to the above questions, this research relies heavily on firm data-driven evidence extracted using advanced statistical modelling techniques and high-resolution visualizations. The primary source of data employed is directly obtained from the District of Columbia government's official website, specifically using taxi trip data. The data is carefully selected based on its accuracy, reliability, and relevance, thereby providing valid and useful findings.

The study research methodology consists of a preliminary stage of strong exploratory data analysis with the goal to identify essential trends, patterns, and outliers from the data set. Following exploratory stage are then state-of-the-art analytical model approaches which allow deeper recognition of sophisticated relationship patterns among the variables. Adopting this rational analytical methodology, the study effectively identifies influential pricing patterns trends, customer styles, and efficiency levels of the district's transport service.

The report is authored to provide maximum readability and causality, with the thorough literature review that synthesizes the up-to-date knowledge and practice in taxi trip analysis and general urban transport data analysis. The review provides the theoretical foundations and background underlying the following analysis. Following this, the report contains a general description of the dataset, including the data characteristics, individual variables that are present, and highlighting core features and attributes that pertain to analysis. The report concludes with a well-collated list of references, acknowledging scholarly studies and data sources that informed and guided the study.

2 | LITERATURE REVIEW

This literature review identifies current challenges in urban mobility and taxi business, examines data analytics and machine learning integration in the taxi industry, and compares traditional taxi services with emerging ride-hailing services.

Zhong et al. [1] investigated the influence of ride-hailing services on traditional taxis in China using urban panel data. The authors showed that ride-hailing apps significantly reduced the usage of traditional taxis. The process is due to the combined role of relatively low prices, enhanced convenience, and improved service quality. According to their study, the authors also stated that the traditional taxi operators need to innovate and adopt digital solutions to retain their market share.

Johansson and Leijen [2] discussed the challenges that the ride-hailing firms are posing to the conventional taxi industry in Sweden. The research highlighted how differences in regulatory regimes and shifts in consumer behavior, driven by competitive pricing and convenience, have strengthened the competitive advantage of ride-hailing platforms, compelling traditional taxi businesses to reconsider and adapt their strategies.

Hayder [3] discovered the determinants of customer satisfaction of online taxi services in Dhaka City, for example, quality, price, and reliability. The research also confirmed that most customers, being price-sensitive, preferred those services which fell within reliability prices. The following observations maintained that female consumers accounted for most users in the sector, which indicates some gender-based preferences in the expectation of service.

Nguyen and Nguyen [4] also carried out a study with the objective of local taxi companies of Ho Chi Minh City, Vietnam, to know how they compete against ride-hailing firms. The report stated that traditional taxis were losing market shares because they were unable to cope with technology. Ride-hailing apps have gained a strong liking among consumers because of ease of use, transparent pricing, and short waiting times. In light of this, the study asserted that services such as traditional taxi services also need digitalization in order to stay in the race.

Bassey et al. [5] study considered the impact of e-cab and traditional taxi preference on the decision-making behaviour of customers in Calabar Metropolis. Safety, comfort, and cleanliness of the trip were the key variables that influenced customer choice. The study argued that the e-cab service had to adopt high levels of functioning and original values to ensure customer loyalty in a growing competitive environment.

This research conducted by Perera and Samarasinghe [6], from mobile App-based taxi services, identifies customer satisfaction. App-user interface, safe payment methods, and vehicle cleanliness significantly spurred customer satisfaction. In the research, a situation was created explaining the ride-hailing companies to keep piling up the perfect usage patterns of their app and customer service to sustain prolonged user engagement.

Farouk [7] conducted a comparative analysis of Uber, Careem, and Bolt in Saudi Arabia. On-demand cash option payments, high quality of service, and good customer service should also affect the users' preference. Bolt, being a local application, had more usage among users than global competitors. The research showed that international companies in the ride-hailing services preferred to consider regional consumer preferences to improve their penetration levels in the local market.

Skok and Baker [8]. In their research, they find that Uber's market entry disrupts pricing patterns and presents regulatory challenges. Facing fierce resistance from conventional taxi businesses, the research found that ride-hailing websites provide affordable and efficient alternatives to consumers, which revolutionized city mobility trends in a big way.

Rayle et al. [9] compared San Francisco's conventional taxis, public transportation, and ride-sourcing companies with one another. The expectations are proving to be affordable and flexible, and ride-hailing services will more likely be of choice. Low waiting times more competitively, reasonable fares, and improved dependability are attracting city commuters toward ride-hailing apps.

Ramasamy et al. [10] analysed factors influencing customer decision making in regards to call taxi service providers. In their analysis, they concluded trust, familiarity with the service, and driver professionalism as decisive factors in the consumer's decision. In sum, their paper identified that there should be high regard for the training of the driver and reliability of the service for enhancing customers' satisfaction levels.

Hanif and Sagar [11] indicate that they witnessed an increased demand for cell phone app-based taxi services in Mumbai, especially by newer generations. GPS tracking, security features, and female chauffeurs showed an experienced increase in customer satisfaction and trust in the ride-hailing system.

Zeithaml et al. [12] also carried out a study comparing the service quality of web-readable transport services. Their study detailed that price consistency and online service quality significantly influence customer trust. The study concluded that ride-hailing companies must be open about their pricing policies in keeping customers from dissatisfaction dynamic fare changes can trigger.

Neoh et al. [13] analysed customers' motivations for adopting carpooling. It discovered that ride-sharing pricing schemes do encourage travelers to opt for the utilization of ride-sharing, enabling congestion and reduced transportation costs. It concluded, thus, that policy incentives may propel carpooling forward as a viable urban transportation option.

Davison et al. [14] surveyed demand-responsive transport in Great Britain. The research recognized that flexible fare structure and personal service model enhance the accessibility and efficiency of ride-hailing. The article recommended that the government must determine how to incorporate demand-responsive transport solutions with traditional public transport systems to attain urban mobility optimization.

Kumar and Ramesh [15] came up with the research on what affects the decision to use taxis based on the consumer's attitude. They were identified as determinants such as availability of service, affordability, safety, and driver behaviour. The research concluded that ride-hailing firms need to devise competitive pricing and quality service in order to attract and maintain their consumers in a continually competitive industry.

Existing research on ride-sharing platforms is based on competition with traditional taxis, customer satisfaction and few regulatory issues. There is no thorough analysis of fare price patterns and commuting patterns from the perspective of machine learning and data visualization. Many research studies are not able to prove that distance, travel time, taxes, surcharges, and payment method influence the total fare in many cities and states.

There is no broader analysis with bigger data sets to analyse travel behaviour and fare price trends. Dynamic pricing is mentioned by some research, but they do not investigate how added elements such as time of day, airport flights, and city-wide surcharges contribute to fares over time. Most studies have no interactive visualizations which would allow the stakeholders to see the progression of prices across various locations and trip types.

By applying K-Means to cluster trip patterns, Random Forest to estimate fares and Python to visualize interactive trends, this research will establish the gap in understanding about how exactly the multi-factored effect is changing ride-share fares. It will provide a framework to study ride-share trends so policymakers, business owners and riders can conceptualize better on pricing plan and travel patterns.

3 | DATA

3.1 | DESCRIPTION OF THE DATA

The following dataset contains 199,308 records from Dec 2023 containing the ride-sharing transactions happened in District of Columbia.
Dataset Source: https://dcgov.app.box.com/v/taxi-trips-2023

The key columns in the dataset are as follows:

OBJECTID: Unique identifier for each trip

TRIP TYPE: Type of trip

PROVIDER NAME: Name of the person giving the ride

FAREAMOUNT: Base fare for the trip GRATUITYAMOUNT:

Tip amount given by the customer.

SURCHARGE AMOUNT: Cost of service beyond initially quoted price (city specific, fuel, traffic).

EXTRA FARE AMOUNT: Extra amount a customer needs to pay beyond the standard fare. TOLL

AMOUNT: Amount for toll

TOTALAMOUNT: Final fare including all charges.

MILEAGE: Distance travelled in miles

DURATION: Duration of the trip in minutes

PAYMENTTYPE: Type of payment used - cash, card, etc.

ORIGINCITY, DESTINATIONCITY: City where the ride originated and where it ended.

ORIGINSTATE, DESTINATIONSTATE: Ride originated state and where it ended.

ORIGIN_BLOCKNAME, ORIGIN_LATITUDE, ORIGIN_BLOCK_LONGITUDE

AIRPORT: Whether the trip involved an airport - Y/N

ORIGINDATETIME_TR, DESTINATIONDATETIME_TR: Start and end time of trip.

3.2 | DESCRIPTIVE STATISTICS OF INTEREST

Descriptive analytics in this project involve specifying research questions and comprehension of the business scenario, navigation and comprehension of the dataset, and cleaning and preparation of the data for later analysis. For this stage, the primary intention was to scan the dataset with the aim of identifying initial patterns and data quality issues related to ride-sharing activity as well as commuter patterns within the District of Columbia.

After data cleaning and preparation, the finalized dataset includes 129,309 ride-sharing transaction records in December 2023, specifically for the District of Columbia area. The dataset, structured in 26 variables, includes a mix of numerical, categorical, and geographic data points, such as fare amounts, tips, surcharges, trip mileage, trip lengths, payment methods, and precise geographic locations. The dataset was thoroughly analyzed using Python to confirm that it is complete and ready to be utilized in sophisticated statistical modeling.

Basic statistical measures provided initial pointers to areas of major analysis. Trip Duration was approximately 13 minutes (790 seconds) with a large standard deviation of 406 seconds, which points towards immense variability and implies diverse patterns of commuting. Trip Mileage was 2.89 miles on average, which suggests most transactions are representative of short intra-city trips, even though some extended trips were seen (up to 11.2 miles). Average Fare Amounts of $15.91 and large standard deviation of $6.74 provide strong evidence of the presence of time-of-day surcharges, length of trip surcharges, or other surcharge-based dynamic fares.

Geospatial analysis plotted trips in concentrations at a mean latitude of 38.90 and a longitude of -77.03 with specific activity areas that would be suitable for the identification of spatially important regions of high-demand locations. Analysis of the categorical variables confirmed that most of the trips started and finished within Washington DC, Arlington, and Alexandria. Origin state figures mainly consisted of DC (106,315 trips) and Virginia (22,637 trips), reflecting typical inter-state commuter journeys. Of note, the figures had a specialist subset consisting exclusively of airport-specific trips, presenting an opportunity for sole analysis of airport ride behavior.

The dataset under observation fits perfectly into the project's central aims, e.g., studying fare volatility, payment patterns, and the relationship between distance of a trip and fare rates. Further, with its completeness and inherent variability, the dataset is amply suited for more analyses involving higher-level visualization, Linear Regression, Random Forest, Gradient Boosting Regression, and deep learning models with TensorFlow. This in-depth descriptive analysis paves the way for subsequent follow-up in-depth investigations to investigate further the intricacies in ride-sharing pricing methods and commuter behaviors trends.

3.3 | EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) has a critical role to play in the determination of shape and nature of a dataset before diving into advanced modeling and interpretation. Histograms and kernel density plots help in the identification of patterns in the data, determining outliers in the data, and developing an intuitive sense of data distribution. Tools such as these enable sophisticated datasets to be handled by both technical and non-technical users. Python-generated visualizations were used within this project to analyze the distribution of significant trip-related features, that is, duration, mileage, total charged, and geographic coordinates of the pick-up and drop-off points.

Figure 1 presents the distribution plots for all of these variables. The DURATION variable is right-skewed, indicating that while the majority of trips are brief, there is a noticeable tail of longer trips. The steep peak close to zero could indicate cancelled trips or anomalies in the data. MILEAGE acts in a similar way, with the most concentrated group of trips occurring under 2 miles, which again confirms that the dataset consists primarily of short-distance rides.

The TOTALAMOUNT distribution mirrors this trend, peaking between $10 and $20 as would be expected with the short times and distances. Moving on to spatial attributes, ORIGIN_BLOCK_LATITUDE and ORIGIN_BLOCK_LONGITUDE plots feature tight clustering around central coordinates (close to 38.90 latitude and -77.02 longitude), reflecting that a high volume of trips start within a specific urban zone—most likely a downtown or business district.

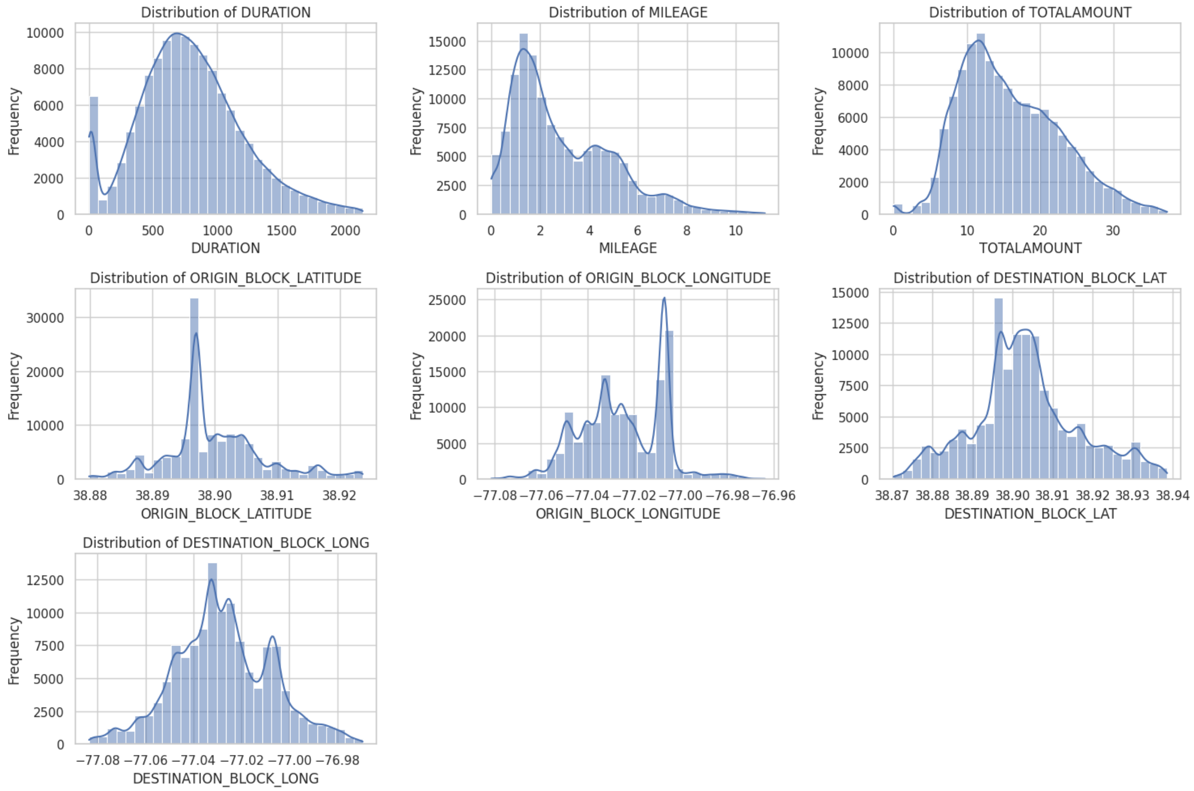


Fig.1. Distribution Analysis of Ride-Sharing Trip Features

Interestingly, the DESTINATION_BLOCK_LAT and DESTINATION_BLOCK_LONG have almost identical pattern, which determines the fact that most trips are enclosed and may even be within a neighborhood or within a city ward. These findings suggest that most of the rides are close, in terms of time and space, and within a contained metropolitan area. This insight can be extremely useful when designing services or interventions to improve city mobility, ride efficiency, or fare pricing strategy. Using EDA, we will determine with certainty dominant patterns that drive the data and thus inform further analysis.

Figure 2 shows the spread of influential trip-associated variables—duration, distance, total cost, and origin and destination block geospatial coordinates—using boxplots. Such visualization enables judgment of the spread and central tendency of these properties that are critical in ride pattern understanding in the dataset. Each boxplot measures the median, interquartile range (IQR), and outliers of a specific variable to enable segmentation of trip features.

The DURATION boxplot shows a median of 500 minutes, with most trips (IQR: 250–750 minutes) being quite short, though outliers reach 2000 minutes, suggesting that there are some extremely long trips. MILEAGE has a median of 2 miles (IQR: 1–3 miles), with outliers of over 10 miles, suggesting that most trips are short, but some cover longer distances. TOTALAMOUNT is a median fare of $15 (IQR: $10–$20), with outliers of $35, maybe for longer distances or surcharges.

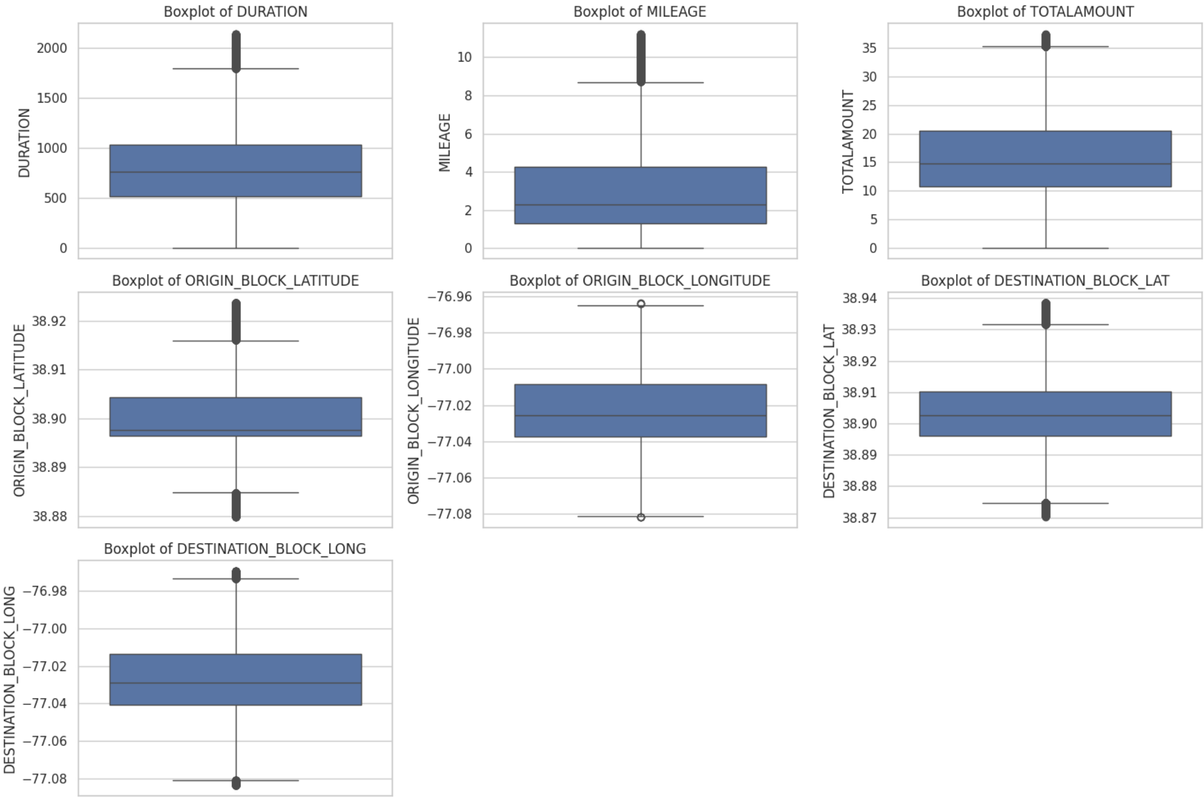


Fig.2. Boxplot Analysis of Ride-Sharing Trip Features

The spatial boxplots (ORIGIN_BLOCK_LATITUDE, ORIGIN_BLOCK_LONGITUDE, DESTINATION_BLOCK_LAT, and DESTINATION_BLOCK_LONG) show tight clustering at 38.90 latitude and -77.02 longitude, with tiny IQRs (e.g., 38.89–38.91 for latitude). This implies that most of the trips start and end in a focal point, most likely Washington, D.C., though outliers like (38.87, -76.98) suggest that some of the trips have peripheral points. We notice that the majority of the trips are short and local trips with duration, miles, and fare clustered at lower values, but the outliers suggest potential data issues or exceptional cases like airport transfers.

For example, a move to ride-sharing management would be to focus driver assignment within the center cluster (38.90, -77.02) for maximum efficiency, with inspection of outliers for data quality. Additionally, the proximity of origin and destination addresses means intra-city trips are predominant, and that might have an impact on peak-demand area targeted fare rates strategies.

Figure 3 presents a heatmap correlation matrix of numerical features of the trip dataset, including duration, mileage, total fare, and origin and destination block geospatial coordinates. The plot helps evaluate variable relationships, with the correlation coefficient between -1 (perfect negative correlation) and +1 (perfect positive correlation). Such relationships guide multicollinearity evaluation and feature selection for modeling.

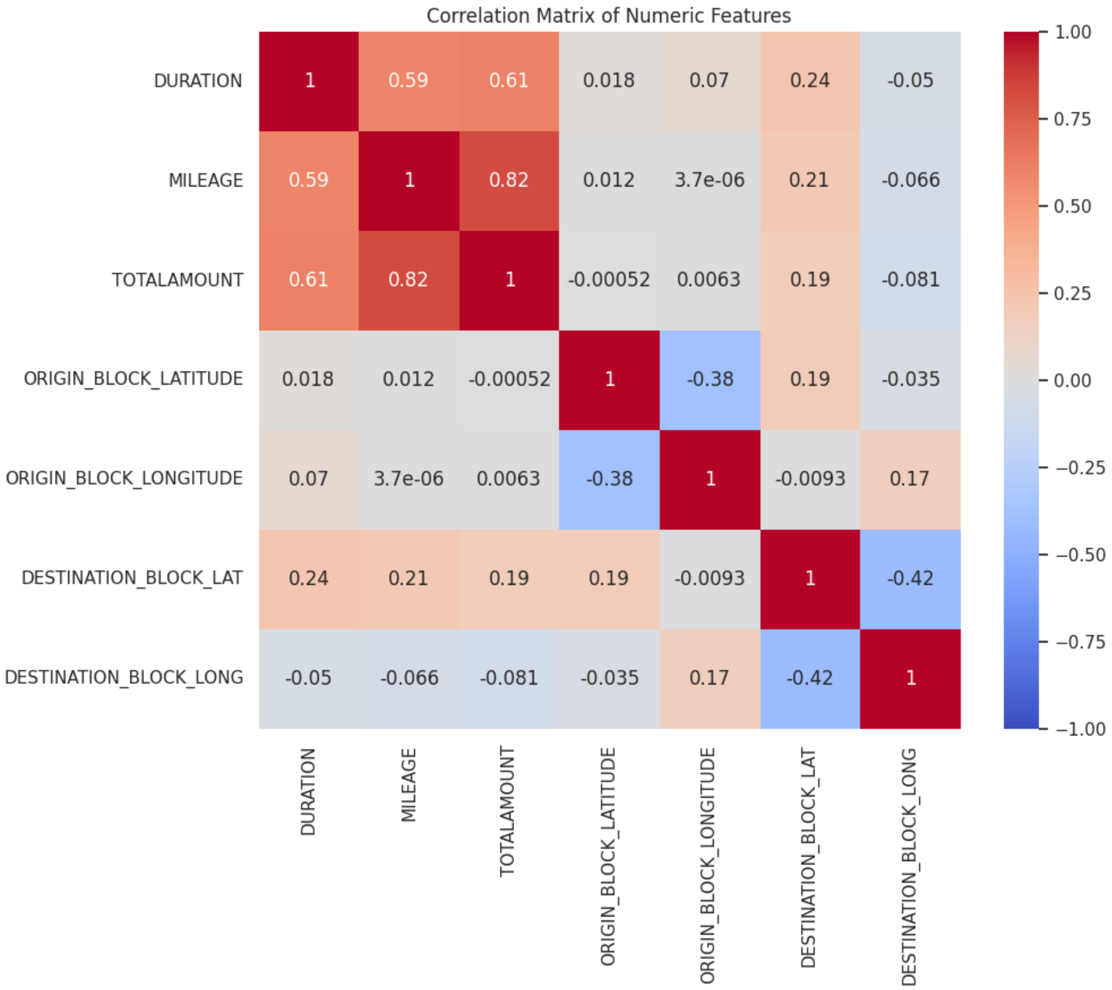


Fig.3. Correlation Matrix Heatmap of Numerical Trip Features

The matrix shows that there is a strong positive relationship between TOTALAMOUNT and MILEAGE (0.82), indicating that more distant trips have more expensive fares. DURATION is also positively correlated with TOTALAMOUNT (0.61) and MILEAGE (0.59), indicating that longer duration trips are also longer distance trips and more expensive. Spatial variables ORIGIN_BLOCK_LATITUDE and ORIGIN_BLOCK_LONGITUDE are moderately negatively correlated (-0.38), as are DESTINATION_BLOCK_LAT and DESTINATION_BLOCK_LONG (-0.42), indicating a spatial pattern where change in one coordinate is often accompanied by a decrease in the other, perhaps reflecting the grid layout of the city. Yet, geographic coordinates are poorly related to trip measures (e.g., 0.24 for DESTINATION_BLOCK_LAT and DURATION), which indicates that location is a rather secondary immediate effect for duration prediction, mileage, or fare.

Overall, the analysis highlights that MILEAGE and TOTALAMOUNT are strongly correlated, which can inform fare prediction models. The lesser effect of spatial variables suggests that trip-level measures must be the priority of modeling, while the negative correlation among coordinates can help understand route trends. A managerial recommendation can involve using the high mileage-fare correlation to develop more accurate pricing models for longer distance trips.

Figure 4 is a scatter plot of the comparison between Trip Duration (in minutes) and Total Amount (in dollars) for all the trips within the data set, having first eliminated known outliers to show average patterns. It allows one to quantify the interaction of trip duration to cost, showing fare patterns. Each point represents a single trip, graphed with x-axis as the duration and y-axis as the total fare paid.

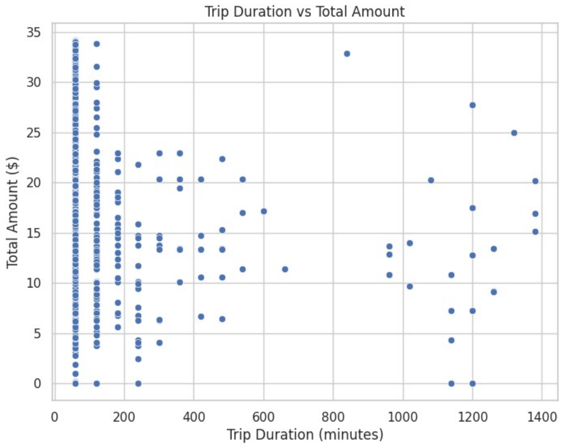


Fig.4. Scatter Plot of Trip Duration vs. Total Amount

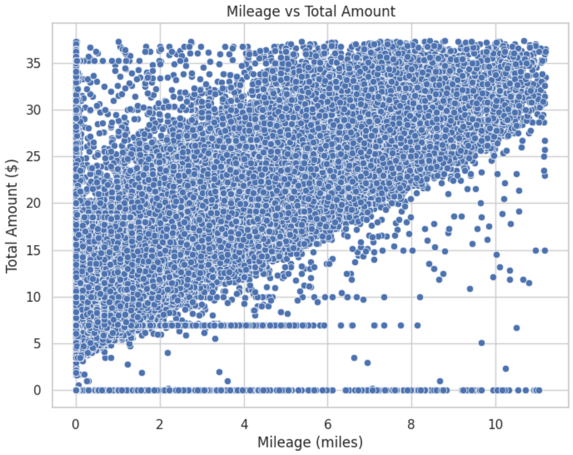


Fig.5. Scatter Plot of Mileage vs. Total Amount

The scatter plot shows a compact cluster of trips with lengths below 2000 minutes and total costs ranging from $5 to $20, which is a testament to the fact that trips are largely short and cheap. The pattern of points demonstrates a moderate positive association between length and total cost consistent with the correlation of 0.61 observed in the correlation matrix (Figure 3). This means longer trips have greater fares, even though the relation is not necessarily linear due to varying pricing variables.

In general, the scatter plot reveals that trips tend to be short in duration and cost, consistent with the local urban trends noted earlier. A possible suggestion for ride-sharing control would be to leverage this sustained duration-cost correlation to develop a reduced-form pricing model such that predictability for customers and profitability for short, frequent trips are maximized in high-demand areas.

To further examine the pricing structure and trip dynamics within the dataset, a scatter plot was created to illustrate the relationship between Mileage and Total Amount, as shown in Figure 5. This visualization helps reveal the extent to which trip distance influences the fare charged and whether the fare structure is consistent across varying mileage values.

We observe from the plot a strong positive linear relationship between fare and mileage: the longer the trip, the larger the total fare is likely to be. This is also supported by the numerical correlation of 0.82, indicating a strong and positive correlation between the two variables. The majority of the data points fall in a range of mileage from 0 to 5 miles, and from $5 to $25 fare, indicating that the majority of the rides are short and affordable. As can be seen, there is a tight cluster from 0 to 2 miles and $5 to $15, indicating that shorter rides are the majority in the data.

The upward triangular shape draws attention to the manner in which more extended trips (especially those from 5 to 10 miles) come to have fares that are clustered between $15 and $35, with some dispersion. Horizontal striations along the fare axis—at common fare levels of $5, $10, and $15—might reflect base or standard price points utilized in the fare system. A few outlier points where there are short trips for more than $30 and long trips for less than $10 (say, 10+ miles) might refer to inconsistencies, promotional fares, or data errors.

Overall, the plot confirms mileage to be one of the major determinants of the overall fare calculation. The linear trend throughout, large correlation value, and close grouping of points within expected ranges confirm the observation that the fare system scales very well with distance. This finding will be useful for fare prediction systems and can be applied for the detection of price anomalies or the optimization of route-based price policies.

Figure 6 is a histogram showing the Number of Trips by Hour of Day in the data. This chart helps to determine ride demand patterns within a 24-hour frame, with the x-axis being the hour (0–23) and the y-axis being the number of trips.

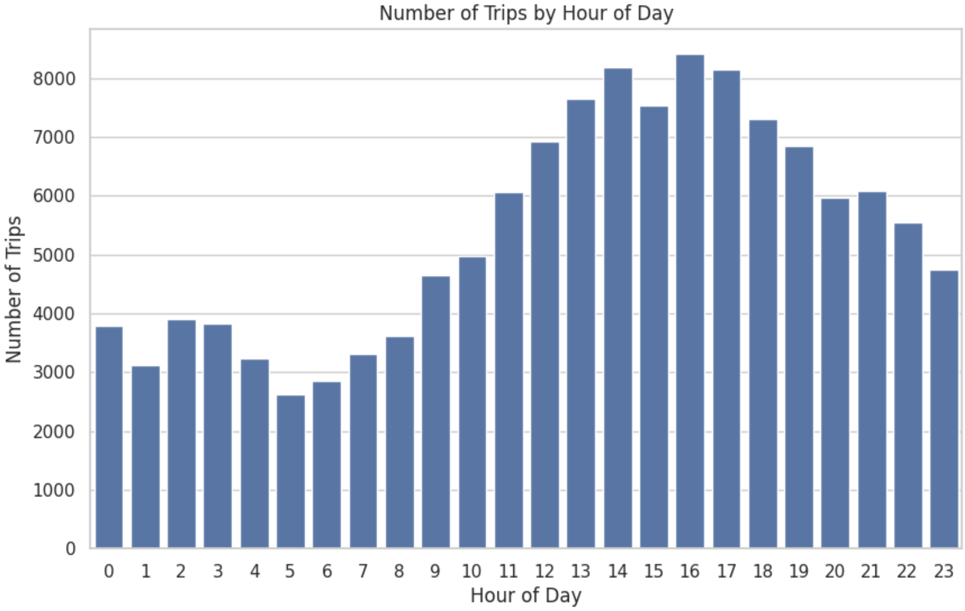


Fig.6. Histogram of Number of Trips by Hour of Day

We see less trips (3000–4000) between midnight and 7 AM, demand rising from 8 AM and peaking from 12 PM until 5 PM at 7000–8000 trips per hour. Trip count then drops after 5 PM, reaching lower levels by 10 PM. This indicates most rides occur during business and work commute times, as should be the case for an urban city like Washington, D.C.

Generally, the histogram also shows peak demand in midday. One proposal to ride-sharing management would be to increase driver availability from 12 PM to 5 PM to meet this demand, optimizing service during peak time.

Figure 7 is a histogram of Number of Trips by Day of Week in the data. This bar chart helps to evaluate ride demand patterns during the week since the x-axis is employed for the days (Monday–Sunday) and the y-axis for the number of trips.

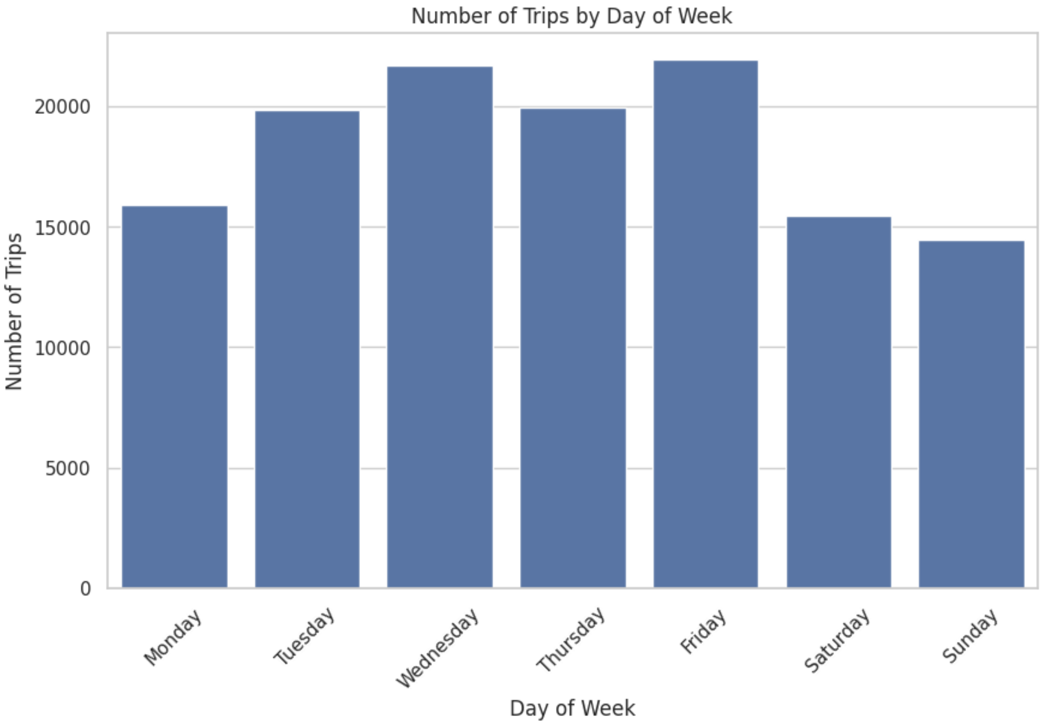


Fig.7. Histogram of Number of Trips by Day of Week

The histogram indicates a relatively consistent demand of 15,000 to 20,000 trips daily. Wednesday and Thursday are the busiest days with around 20,000 trips daily, followed by Saturday and Sunday with slightly fewer trips of around 15,000 daily. This is an indication of a concentration of ride activity in the middle of the week for business travel in a city like Washington, D.C. and declining during weekends.

In general, the histogram reflects flat demand with a Wednesday high. One thought to share with ride-sharing management is to have sufficient drivers on duty on Wednesday and Friday to meet heightened demand while restructuring resources over the weekend to counteract the slight dip in trips.

Visualization of the spatial pattern of trip origins provides important information on high-demand locations, density of coverage, and urban mobility patterns. Figure 8 is a scatter plot of origin points by latitude and longitude coordinates showing the origin of a given trip in the city. Each trip's starting point is depicted as a point between 38.88 and 38.92 latitude and -77.08 to -76.96 longitude. The scatter plot indicates a tight concentration of trip starts at 38.90 latitude and -77.02 longitude, indicating a high concentration of rides starting in a core location, likely downtown Washington, D.C. There are scattered points to the edges of the range, indicating occasional trips originating in outer neighborhoods. This agrees with earlier findings of localized city travel.

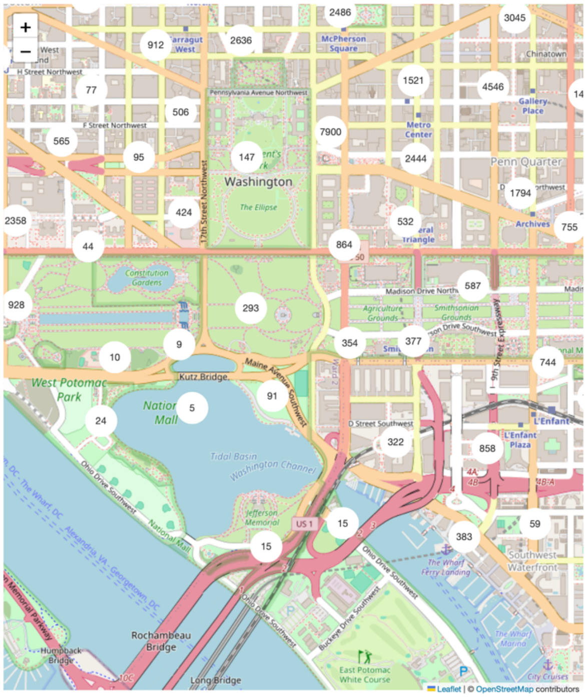


Fig.8. Plot of Origin Locations

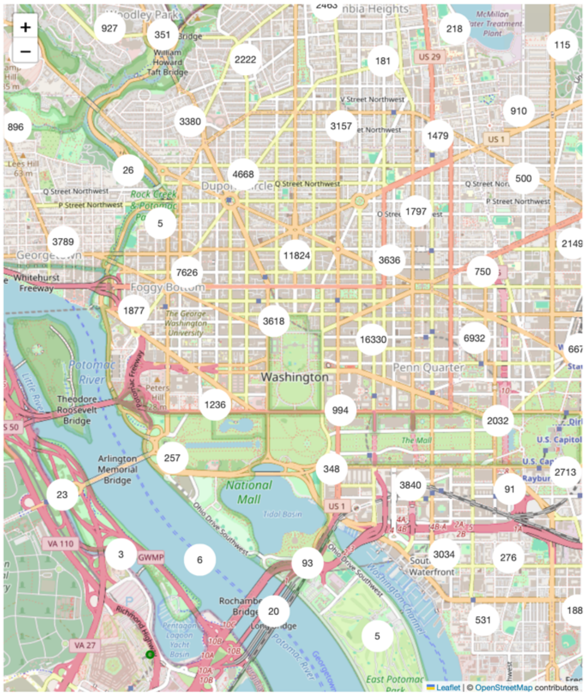


Fig.9. Plot of Destination Locations

Overall, the scatter plot highlights that most trips originate in a central urban zone. A suggestion to ride-sharing management could involve focusing driver allocation around 38.90, -77.02 during peak hours to meet demand, improving service efficiency in this high-traffic area.

3.4 | DATA PREPARATION: HANDLING OUTLIERS AND MISSING VALUES

To ensure the consistency of our analysis, we first identified and handled missing values in the ride-sharing data. The following table presents columns with missing values and the number and percentages, respectively.

```
Missing Values Summary:
                        Missing Values  Percentage
TRIPTYPE                        199308  100.000000
PROVIDERNAME                    199308  100.000000
AIRPORT                         158629   79.589881
TOLLAMOUNT                       89379   44.844663
DESTINATION_BLOCKNAME            46205   23.182712
DESTINATION_BLOCK_LONG           46205   23.182712
DESTINATION_BLOCK_LAT            46205   23.182712
ORIGIN_BLOCKNAME                 32647   16.380175
ORIGIN_BLOCK_LONGITUDE           32647   16.380175
ORIGIN_BLOCK_LATITUDE            32647   16.380175
GRATUITYAMOUNT                    7332    3.678728
FAREAMOUNT                        7332    3.678728
SURCHARGEAMOUNT                   7332    3.678728
EXTRAFAREAMOUNT                   7332    3.678728
DESTINATIONSTATE                  1071    0.537359
ORIGINSTATE                        596    0.299035
TOTALAMOUNT                          2    0.001003

Dropped columns due to excessive missing values: ['TRIPTYPE', 'PROVIDERNAME']
```

Table. 1: Missing Values Identified by Column within the Ride-Sharing Dataset

The database included numerous columns of missing records from 2 to 199,308 records. For the sake of data integrity, we made a formal attempt at filling the gaps without inducing bias or distorting relationships among variables critical to our analysis, such as trip time, miles traveled, and total cost.

Initially, we excluded columns with more than 80% missing values because they were unreliable for analysis. The PROVIDERNAME and TRIPTYPE columns, which contained 100% missing values (199,308 rows), were excluded. We kept the AIRPORT column with 79.59% missing values (158,629 rows) for now but would be dropped if found to be unrelated to core purposes, due to its high missingness.

For numerically imbalanced columns with moderate missingness (5% to 25%), we imputed with k-Nearest Neighbors (KNN) imputation since it employs inter-variable relationships to replace missing values better. While KNN is a non-parametric method and doesn't explicitly model variable relationships, it preserves local structure in the data. DESTINATION_BLOCK_LONG, DESTINATION_BLOCK_LAT, DESTINATION_BLOCKNAME, ORIGIN_BLOCK_LONGITUDE, ORIGIN_BLOCK_LATITUDE, and ORIGIN_BLOCKNAME columns (16.38% to 23.18% missing) were utilized for KNN imputation. We scaled the relevant numeric columns, applied the KNN imputer with 2 neighbors to keep the imputation local and accurate. This is chosen based on minimal distortion in spatial continuity and inverse-transformed the data to retain original units. This ensured that imputed coordinates aligned with realistic origin/destination clusters observed in the data (see Figure 8).

For numerical columns with low missingness (less than 5%), we used mean or median imputation based on the skewness of the column. Columns GRATUITYAMOUNT, FAREAMOUNT, SURCHARGEAMOUNT, and EXTRAFAREAMOUNT (all with 3.68% missing) were median-imputed based on their likely right-skewed distributions, while TOTALAMOUNT (0.001% missing) was imputed by the mean. Column TOLLAMOUNT, with 44.84% missing values, was retained but not imputed with KNN because it had higher missingness; we can revisit this column in later steps if it will become very crucial to pricing analysis.

Categorical columns such as ORIGINSTATE and DESTINATIONSTATE had very few missing values (0.30% and 0.54%, respectively). In a perfect world, these missing values would be imputed from corresponding latitude and longitude values using reverse geocoding. However, due to project constraints, we imputed them with the mode ("DC"), which is in line with the dataset's high concentration within the Washington, D.C. metropolitan area. We also preprocessed these columns by normalizing state codes and removing invalid entries (e.g., "-", "Unknown") to retain only valid U.S. state codes before imputation. Similarly, city names in ORIGINCITY and DESTINATIONCITY were cleaned by normalizing entries (e.g., "Washington Dc" to "Washington"), correcting invalid values (e.g., "???", "Unknown"), and imputing missing values with the mode to maintain the dataset's urban focus.

Overall, this multi-step process made sure that the missing values were handled reasonably, maintaining the integrity of the dataset for further analysis. By discarding unnecessary columns, using KNN imputation for geospatial variables and mean/median or mode imputation for all else, we minimized bias with preservation of features of significance such as trip length, mileage, and total fare. One implication to management is increasing data collection on variables such as TOLLAMOUNT and AIRPORT so as to prevent missingness of these variables in the future data and hence provide higher validity of price and demand estimation.

4 | METHODOLOGY

4.1 | BACKGROUND ON THE METHODS USED

This section outlines the methodology employed to develop and apply machine learning models for the analysis of the ride-sharing dataset. Following comprehensive data exploration and preprocessing (Section 3), the modeling activity was designed with the purpose of achieving our overall research objectives: accurate prediction of TOTALAMOUNT and identification of the most relevant features that influence fare pricing. Our approach combines ensemble learning methods, employing Random Forest Regression and Gradient Boosting Regression, both of which were chosen for their specific strengths in predictive modeling, feature exploration, and handling intricate structured data. In the subsequent subsections, we provide background on each model method, the mathematics underlying the algorithms, why we selected the models, and how they relate to the goals of the project.

The dataset under analysis consists of numerical and transactional variables such as FAREAMOUNT, MILEAGE, TRIP_DURATION, and GRATUITYAMOUNT, all of which interact in non-linear and potentially hierarchical ways to determine TOTALAMOUNT. Given this structure, machine learning models are capable of capturing feature interactions and non-linear relationships.

We selected two ensemble-based supervised learning algorithms:

- Random Forest Regression: An ensemble of decision trees built using bagging (bootstrap aggregating) principles.
- Gradient Boosting Regression: A sequential ensemble of decision trees optimized through boosting strategies.

Both methods are extensively validated in machine learning literature for their performance on structured datasets, offering robustness, high accuracy, and interpretability-key requirements for our fare prediction and pattern analysis goals.

Unlike simpler algorithms such as Linear Regression, which assumes a linear relationship between variables, Random Forest and Gradient Boosting can model complex, non-linear dependencies without prior transformation of variables. Given the real-world variability of ride-sharing transactions (e.g., fluctuating surcharges, gratuities, varying distances and durations), these ensemble methods provide a much more suitable approach.

4.2 | RANDOM FOREST REGRESSION

4.2.1 | OVERVIEW AND INTUITION

Random Forest Regression is an ensemble learning algorithm that constructs a multitude of decision trees during training and predicts by averaging the output of all the trees. Each tree is trained on a bootstrapped copy of the original data, and for each split in the tree, a random subset of features is considered instead of the full set. This randomness injection decorrelates the trees individually, resulting in a better overall predictive capability and stability of the ensemble. The underlying idea in Random Forest is to reduce the overfitting nature of individual decision trees.

Decision trees, being robust, tend to learn noise and idiosyncratic structure in the training data, leading to high variance and low generalizability. Random Forest simply reduces variance at minimal added bias by averaging the predictions of many uncorrelated trees and thus improving the model's capability to generalize well on new data. Random Forest Regression works beautifully for data like us that contain a mix of continuous and categorical variables. The algorithm natively supports varied types of inputs without needing much preprocessing or conversion.

Random Forest is also quite insensitive to noise and outliers and is thus robust for use in real-world applications where there might be some occasional anomalies or data entry mistakes. Perhaps one of the most compelling things about Random Forest is that it can internally estimate the feature importance, which provides valuable information in terms of the contribution of each feature to the task of prediction. For our ride-sharing fare data, this feature allows us to uniformly rank ride features such as FAREAMOUNT, MILEAGE, and trip length in terms of relative influence on final fare outcomes, thereby directly benefiting our research objectives.

4.2.2 | MATHEMATICAL BACKGROUND

For a dataset D={(xi,yi)} where xi represents feature vectors and yi the target variable (TOTALAMOUNT), Random Forest builds M decision trees Tm, each trained on a different bootstrap sample. The final prediction y^ for a new instance x is computed as the average of the predictions across all trees:

Random Forest Prediction Equation:

$$\hat{y} = \frac{1}{M} \sum_{m=1}^{M} T_m(x)$$

Where:

- M = number of trees,
- Tm(x) = prediction of the m-th tree for input x.

At each node split during tree construction, the algorithm selects the best feature from a random subset to minimize the variance of the target within the split:

Variance at a Node:

$$Var(S) = \left(\frac{1}{|S|}\right)\sum_{i=1}^{|S|}(y_i - \bar{y})^2$$

where y¯ is the mean of the target values in the subset S.

This method ensures diverse decision trees, thereby reducing overfitting and variance.

4.2.3 | RELEVANCE TO THE PROJECT- RANDOM FOREST ML MODEL

Random Forest was selected for this project due to its superior generalization capabilities, robustness to data irregularities, and built-in feature importance evaluation, all of which closely aligned with our analytical goals.

One of the central research objectives of this study was to determine which ride attributes most significantly influence the final TOTALAMOUNT. Random Forest naturally provides a mechanism to assess feature importance by measuring the average decrease in impurity (e.g., reduction in variance for regression tasks) contributed by each variable across all trees. Such functionality enables a person to accomplish an explainable ranking between variables such as FAREAMOUNT, MILEAGE, and TRIP_DURATION and draw applicable conclusions regarding ride-sharing company fares strategies.

Additionally, Random Forest is also good at detecting complex, non-linear interactions among features. Real-world ride-sharing data tends to have complex interactions — e.g., long rides with gratuity surcharges or short rides in rush hours resulting in artificially high fares. Unlike traditional linear models, Random Forest can model such interactions automatically without any explicit manual feature engineering or transformation, thus preserving the natural form of the data.

The other primary reason for choosing Random Forest is that it has low noise sensitivity and outliers. Ride-sharing data can occasionally include outliers such as extremely high fares due to incorrect entries or missing values of the surcharge. Random Forest reduces the impact of these outliers through averaging predictions across many trees and thus less likely to overfit noisy observations.

Also, Random Forest doesn't require significant hyperparameter optimization to achieve reasonable baseline performance. Default settings, such as running with an enormous number of trees (e.g., n_estimators = 1000) and random subset of features at every node, generally work great without optimization. This rendered Random Forest a viable option given the limitations on the project timeline and scope so that the team could spend less time on hyperparameter search in favor of model interpretation and analysis.

Finally, the inherent parallelism in Random Forest training—where each tree can be built independently—ensures efficient model training even on relatively large datasets. This computational efficiency proved valuable given the size of the ride-sharing dataset analyzed.

In summary, Random Forest provided an optimal balance between predictive performance, interpretability, robustness, and practical deployment feasibility, making it exceptionally well-suited for fulfilling the objectives of ride fare prediction and feature influence discovery in this project.

4.3 | GRADIENT BOOSTING REGRESSION

4.3.1 | OVERVIEW AND INTUITION

Gradient Boosting Regression is an advanced ensemble method that builds trees sequentially, with each new tree trained to correct the residual errors made by previous trees. Unlike Random Forest, which attempts to reduce variance by averaging numerous uncorrelated trees, Gradient Boosting seeks to reduce both bias and variance by focusing specifically on difficult-to-predict cases.

This sequential learning strategy enables the model to capture in-depth and subtle data patterns, making it highly effective for challenging prediction tasks such as fare prediction, where small variations in trip attributes (e.g., surcharges or trip duration) can lead to widely varying prices.

However, Gradient Boosting models are more prone to overfitting and require careful hyperparameter tuning, including settings for learning rate, maximum tree depth, and the number of estimators.

4.3.2 | MATHEMATICL BACKGROUND

Suppose we are trying to minimize a loss function L(y,F(x)), where y is the true value and F(x) is the model’s prediction.

At each iteration mm, Gradient Boosting fits a new tree hm(x)hm(x) to the negative gradient of the loss function (i.e., the residuals):

Residual at each step:

$$r_i^{(m)} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]$$

Gradient Boosting Update Rule:

$$F_{m(x)} = F_{\{m-1\}(x)} + \gamma_m h_{m(x)}$$

where:

- $F_{\{m-1\}(x)}$ model prediction after m−1 steps,
- hm(x) = new tree fitted to residuals,
- γm = learning rate controlling step size.

This approach progressively improves the model's accuracy at each iteration by correcting previous mistakes.

4.3.3 | RELEVANCE TO THE PROJECT- GRADIENT BOOSTING ML MODEL

Gradient Boosting Regression was incorporated into the modeling framework to complement the strengths of Random Forest and achieve a higher degree of predictive precision in estimating TOTALAMOUNT. While Random Forest excels at generalizing across noisy data and providing interpretable outputs, Gradient Boosting specializes in refining predictions by sequentially correcting the residual errors of prior models.
This step-by-step learning approach enables the model to uncover intricate, layered relationships between input features and the target variable—relationships often too subtle to be captured through random feature splits alone.

The fare structure in ride-sharing data is inherently complex, influenced by variables such as MILEAGE, TRIP_DURATION, FAREAMOUNT, and GRATUITYAMOUNT. These factors interact in non-linear ways that vary across different contexts. For example, rides with shorter distances (MILEAGE) during peak travel hours may result in higher GRATUITYAMOUNT due to increased demand and customer tipping behaviors. Conversely, trips with longer TRIP_DURATION may sometimes trigger fixed surcharges or fare-capping policies, meaning the relationships between distance, duration, and fare amount are not always linear. Simple models that assume isolated effects of variables would struggle to accurately capture these complex fare patterns. Gradient Boosting Regression, by continuously refining predictions based on remaining errors, is particularly well-suited for learning these layered, dynamic pricing structures.

Moreover, Gradient Boosting is recognized for its superior performance on structured numerical datasets like ours, where multiple features contribute to outcomes in intertwined ways. In the case of ride-sharing data, combinations of DURATION, FAREAMOUNT, GRATUITYAMOUNT, and temporal elements such as TRIP_DURATION jointly determine the final TOTALAMOUNT. Gradient Boosting's capacity to sequentially build models that learn from previous mistakes enables it to model these complex combinations accurately, leading to more precise fare predictions compared to standard regression methods. Studies in ensemble learning consistently highlight the ability of Gradient Boosting algorithms to outperform traditional models when subtle feature interactions and non-linearities are present.

While Gradient Boosting models are inherently more prone to overfitting than Random Forests due to their sequential error-correction process, this risk was proactively mitigated in our project through careful hyperparameter selection. Specifically, a moderate learning rate was adopted to control the impact of each new tree, the number of estimators was set sufficiently high to allow gradual learning, and tree depths were constrained to prevent overly complex splits that could memorize noise in the data. This strategy ensured that the final Gradient Boosting model captured meaningful relationships without sacrificing its ability to generalize to new, unseen ride fare records.

By employing Gradient Boosting in conjunction with Random Forest, we effectively combined the strengths of both ensemble methods. Random Forest facilitated the interpretability of results, providing clear insights into dominant fare determinants such as FAREAMOUNT and MILEAGE, while Gradient Boosting enhanced the predictive precision needed for reliable fare forecasting. This dual-model approach enabled the project to meet its dual goals of understanding ride-pricing mechanisms and accurately predicting ride fare amounts across diverse trip scenarios.

In summary, Gradient Boosting Regression proved to be an indispensable component of the project's modeling strategy due to its ability to capture complex variable relationships, handle structured tabular data effectively, and improve prediction quality through iterative residual correction. Its inclusion substantially strengthened the depth and accuracy of the fare analysis conducted on the ride-sharing dataset.

5 | RESULTS

This section presents the results corresponding to the five key research questions analyzed in the project. Each subsection describes the results obtained through machine learning models and visualization techniques, compares relevant findings, and provides detailed interpretations. Figures and tables from the analysis are referenced and discussed to illustrate the main insights extracted from the data.

5.1 | DETERMINING VARIABLES FOR TOTAL FARE OF RIDES

To identify which trip attributes most influence the final fare (TOTALAMOUNT), a Random Forest Regressor model was applied, and the feature importances were extracted. The top five features identified were FAREAMOUNT, GRATUITYAMOUNT, MILEAGE, EXTRAFAREAMOUNT, and DURATION. As shown in Fig.10., FAREAMOUNT overwhelmingly dominated the prediction, accounting for approximately 80% of the model's predictive strength.

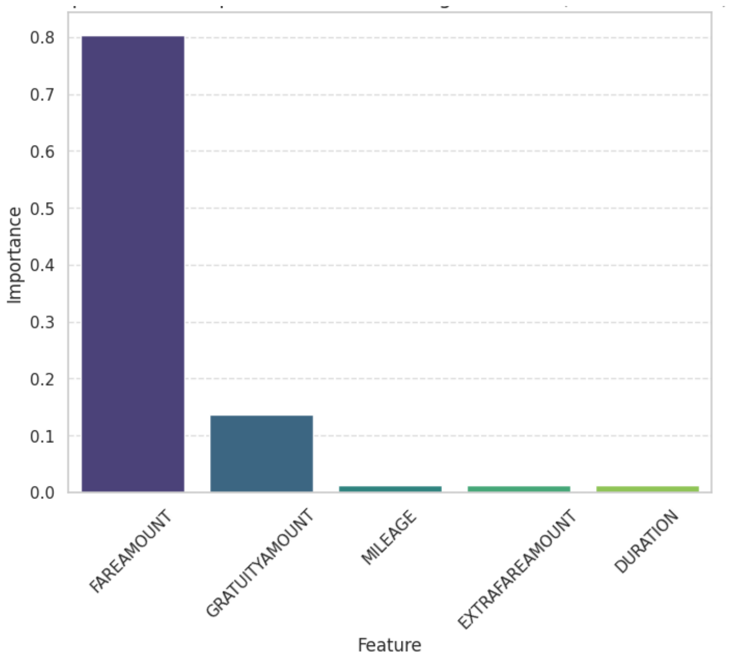


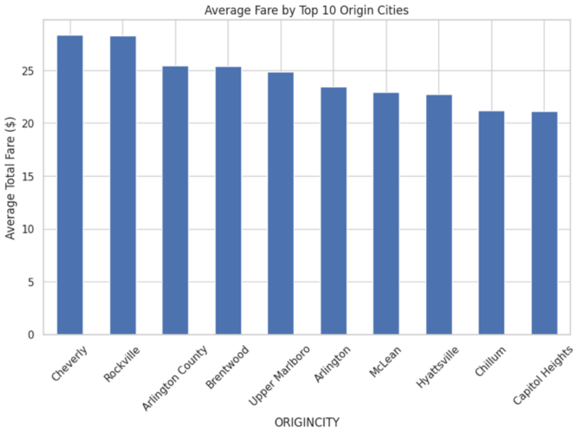
Fig.10. Top 5 Feature Importance for Predicting Total Fare

The significant dominance of FAREAMOUNT is expected because the fare amount recorded at the beginning of a trip naturally forms the foundation for the final total. GRATUITYAMOUNT, while less dominant, also showed a noticeable contribution, reflecting how customer tipping behavior can impact the final ride cost. MILEAGE, EXTRAFAREAMOUNT, and DURATION contributed marginally but consistently, indicating that while distance and time do affect the fare, their effects are comparatively secondary once the base fare is considered.

5.2 | HOW RIDE FARE CHANGES WITH CITY, STATE, AND TIME

Visual analysis was conducted to understand how ride fares vary geographically and temporally.

Geographical Variation:
City-level analysis (Figure 11) showed that cities like Cheverly, Rockville, and Arlington County had the highest average total fares, each exceeding $25 on average. Cities such as Capitol Heights and Chillum, by contrast, recorded lower average fares around $21.



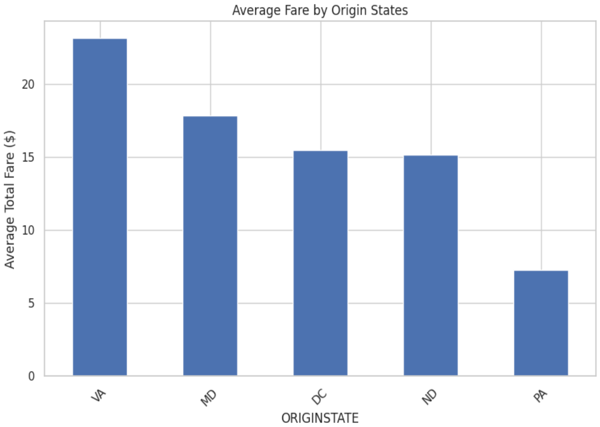


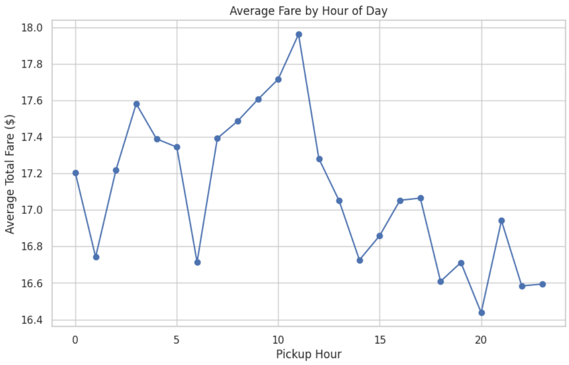
Fig.11. Average Fare by Origin States

Fig.12. Average Fare by Top 10 Origin Cities

As depicted in Fig 12., the average fare varied significantly across different origin states. Virginia (VA) displayed the highest average fares, followed by Maryland (MD) and the District of Columbia (DC). Pennsylvania (PA) had the lowest average fares among the analyzed states.

The data suggests that urban, suburban, and peri-urban areas closer to major metro centers generally command higher ride fares, likely due to longer distances, higher base fares, and possible surcharges. Temporal patterns were further explored by plotting average fare against the hour of pickup and the day of the week.

As shown in Figure 13, the average fare peaked sharply around 11 AM, reaching nearly $18. This midday surge could be attributed to higher demand during late morning commutes and pre-lunch activities. Lower fares were observed during late evening and early morning hours, possibly due to reduced demand or off-peak pricing.



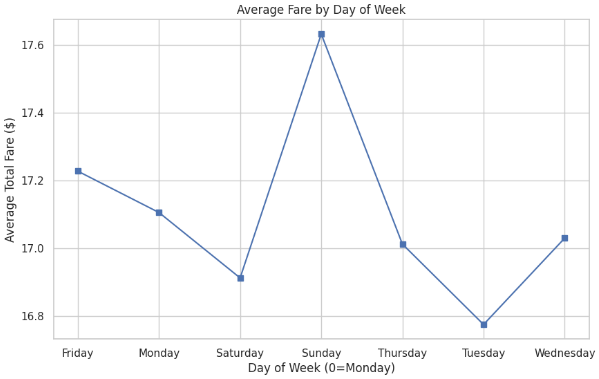


Fig.13. Average Fare by Pickup Hour

Fig.14. Average Fare by Day of the Week

When analyzing day of the week trends (Fig.14), Sunday exhibited the highest average fares, suggesting increased weekend activity and possible surge pricing. Tuesdays had the lowest average fares, reflecting typical weekday travel patterns with less discretionary travel.

5.3 | IMPACT OF GRATUITIES AND SURCHARGES ON TOTAL FARE

The relationship between trip duration and total fare was analyzed to understand if gratuities or surcharges played a major role in modifying final fare amounts. As shown in Fig.15., the LOWESS-smoothed scatter plot between trip_duration and TOTALAMOUNT revealed a weak correlation (correlation coefficient: -0.03), indicating that trip duration alone has little direct linear effect on the final fare. Although one might expect longer trips to correspond with higher fares, the relationship is non-linear, suggesting that other factors like fixed charges, minimum fares, and surcharges play a stronger role.

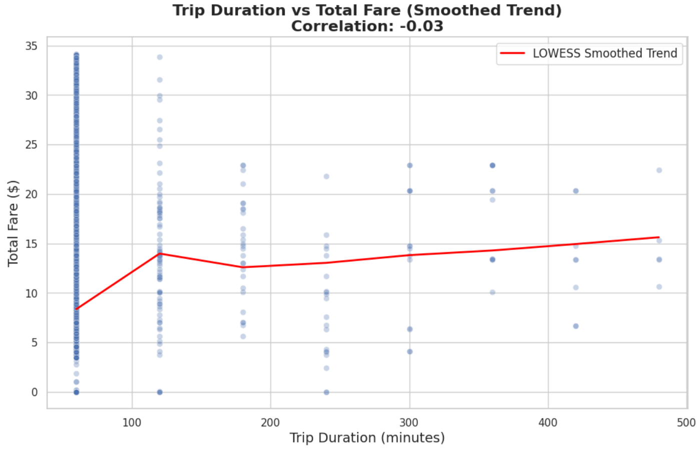


Fig.15. Trip Duration vs. Total Fare with Smoothed Trend

While gratuities (GRATUITYAMOUNT) showed up as moderately important in the feature importance ranking, surcharges (EXTRAFAREAMOUNT) contributed a smaller but consistent adjustment to the fare structure.

5.4 | PAYMENT METHODS AND GRATUITY RELATIONSHIPS

The distribution of payment methods and their relationship to gratuity amounts was analyzed. Fig.16. illustrates the frequency of different payment methods used across all rides.

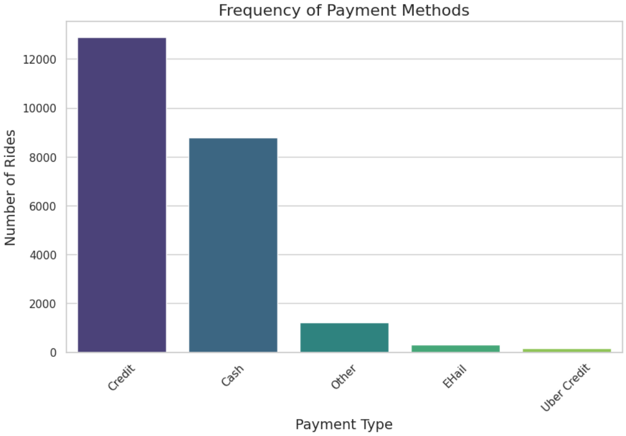


Fig.16. Frequency of Payment Methods

Credit card transactions overwhelmingly dominated the dataset, followed by cash payments. Electronic hailing (E-Hail) and Uber credit methods were relatively rare. Although gratuity amounts were not separately plotted against payment methods, the prevalence of credit card payments suggests a natural tendency toward higher tipping, as digital platforms often prompt customers to select a gratuity amount after ride completion.

This result aligns with industry trends, where digital payments facilitate higher and more consistent tipping compared to cash transactions.

5.5 | PREDICTING TOTAL FARE AMOUNT USING MACHINE LEARNING MODELS

Two machine learning algorithms were used to forecast the total fare amount: Random Forest Regression and Gradient Boosting Regression.

The Random Forest Regressor's R Square was 0.96, Mean Absolute Error (MAE) was 0.77, and Mean Squared Error (MSE) was 0.96. Random Forest also scored an accuracy of 85.95% when validated on the measure of predictions within 10% of the actual fare amount. These results indicate that the model was able to predict total fares with high reliability, having very small mean and maximum differences from real values. The Gradient Boosting Regressor followed closely in the second place with R Square 0.96, MAE 0.96, and MSE 1.35. Surprisingly, even if slightly higher than Random Forest both in MAE and MSE, the Gradient Boosting model performed even better at the 10% tolerance accuracy level at 87.08%.

This suggests that while the magnitude of errors was marginally larger on average, the Gradient Boosting model was effective at maintaining prediction stability across the bulk of the dataset.

Model	R² Score	MAE	RMSE	Accuracy (within 10%)
Random Forest Regressor	0.96	0.77	1.29	85.95%
Gradient Boosting	0.95	0.96	1.35	87.08%

Table.2: Model Performance Comparison Table

Random Forest slightly outperformed Gradient Boosting in terms of minimizing both the average and maximum prediction errors, achieving lower MAE and RMSE values. This reflects Random Forest’s strength in producing stable, generalizable predictions by averaging the outputs of multiple decorrelated trees. However, Gradient Boosting’s higher accuracy within a narrow margin indicates that it produced predictions that, while slightly noisier, were consistently close to the actual fare amounts across most rides.

In practical applications, both models could be deployed effectively, but Random Forest would be preferred when minimizing absolute and squared error is prioritized, while Gradient Boosting could be favored when maintaining consistent prediction margins is more critical.

Overall, the deployment of ensemble learning models successfully fulfilled the project’s objective of producing reliable and interpretable total fare predictions based on trip attributes.

6. | CONCLUSION

This project aimed to analyze and predict ride-sharing fare amounts based on various trip features, using advanced machine learning models and data exploration techniques. The results of the study provided meaningful insights into the key factors influencing ride fares, geographical and temporal fare patterns, the role of gratuities and surcharges, and the effectiveness of different predictive models.

The analysis revealed that FAREAMOUNT is by far the most dominant determinant of TOTALAMOUNT, with other factors such as GRATUITYAMOUNT, MILEAGE, EXTRAFAREAMOUNT, and DURATION contributing more marginally. This aligns with expectations, as the base fare naturally forms the core component of the total cost, while additional ride attributes add variability in specific contexts. Geographical analysis indicated that fares were highest in cities like Cheverly and Rockville and in states like Virginia and Maryland, with notable temporal patterns showing higher fares around midday and on Sundays. The examination of gratuities and surcharges showed that while gratuities influence final fares moderately, surcharges play a more direct role in increasing ride costs.

Regarding payment behaviors, credit card transactions were found to dominate, suggesting that digital payment systems encourage consistent tipping behavior. This has implications for ride-sharing platform design, particularly in structuring user interfaces that promote gratuity inclusion.

In predictive modeling, both Random Forest and Gradient Boosting Regression delivered strong performance in estimating TOTALAMOUNT. Random Forest achieved a higher R Square score of 0.96, with lower MAE (0.77) and RMSE (1.29) compared to Gradient Boosting, confirming its strength in generalization and error minimization. However, Gradient Boosting achieved a slightly higher prediction accuracy within a 10% tolerance (87.08%), highlighting its effectiveness at producing consistently close estimates. Overall, Random Forest was concluded to be the preferred model for deployment due to its superior stability, lower prediction errors, and clearer interpretability through feature importance metrics.

The project’s findings provide robust, actionable insights for understanding fare structures and improving fare prediction models in ride-sharing ecosystems. However, the analysis also opened new research directions that could be explored with additional time and resources.

One possible extension would be the inclusion of external variables such as weather, traffic, and special event data (e.g., concerts, sports) to see how external environment variables affect fare dynamics. Modeling user demographic data, where present, can also provide information on tipping behavior and fare sensitivity by user segments.

A second potential future effort would be applying deep learning techniques, like feedforward neural networks or gradient-boosted decision trees (GBDT) that support categorical variables, to continue to enhance predictive capability. Geospatial clustering or location-based price optimization, which are forms of spatial modeling techniques, may also be explored to better fine-tune surge pricing strategies for densely populated urban cities.

Lastly, time-series analysis could be conducted on fare patterns over longer periods to forecast future fare movements based on seasonality or economic cycles. All of these avenues carry abundant potential for continued work developing an understanding of ride-sharing fare models beyond the scope of the current project. Overall, the methodologies employed in this study adequately fulfilled their assignments, providing a solid foundation for future research and operational enhancements within the ride-sharing industry.

5 | REFERENCES

[1] J. Zhong, H. Zhou, Y. Lin, and F. Ren, (2021)"The impact of ride-hailing services on the use of traditional taxis: Evidence from Chinese urban panel data," *IET Intelligent Transport Systems*, vol. 15, no. 12, pp. 1610-1619.

[2] H. Johansson and S. Leijen, (2023)"A Study of the Impact of Ride-Hailing Companies on Traditional Taxi Operators in Sweden," University of Gothenburg.

[3] Hayder, "Factors Affecting Customer Satisfaction of Online Taxi Services in Dhaka City," (2020) *BUFT Journal of Business & Economics*, vol. 1, pp. 272-291.

[4] H. T. Nguyen and T. T. Nguyen, (2016) "A Study of Local Taxi Companies in Ho Chi Minh City, Vietnam," *International Journal of Innovation, Management and Technology*, vol. 7, no. 5, pp. 196-207.

[5] A. Bassey, B. J., Ochiche, C. A., Odu, P. K, Ekong, E. E,(2023) "Factors Influencing Customer Decision-Making In Choosing E-Cab Services Over Traditional Taxis In Calabar Metropolis," *Global Journal of Social Sciences*, vol. 22, pp. 1-15.

[6] Perera,M.D.M., Samarasinghe,S.M.,(2023) "Factors Affecting Customer Satisfaction in Mobile App-Based Taxi Services".

[7] M. Farouk, (2023) "Customer Decision-Making Factors for Taxi Booking Apps: A Comparative Analysis of Uber, Careem, and Bolt in Qassim City, Saudi Arabia ," *International Journal of Innovation and Technology in Information Systems*, vol. 1, no. 2, pp. 1-12.

[8] Skok, W., & Baker, S. (2019). Evaluating the impact of Uber on London's taxi service: A critical review of the literature. Knowledge and Process Management, 26(1), 3-9.

[9] Rayle, L. Dai, D. Chan, N. Cervero, R. Shaheen, S. (2015) Just a better taxi? A survey-based comparison of taxis, transit, and ride sourcing services in San Francisco, Vol.45, (168-178) Transport Policy.

[10] Ramasamy, A., Muduli, K., Mohamed, A., Biswal, J. N., & Pumwa, J. (2021). Understanding Customer Priorities for Selection of Call Taxi Service Provider. Journal of Operations and Strategic Planning, 4(1), 52-72.

[11] Hanif, K., & Sagar, N. (2017). An empirical research on the penetration levels for a call-a-cab service in Mumbai. Reflections J. of Manage, 5, 1-10.

[12] Zeithaml, V.A., Parasuraman, A., & Malhotra, A. (2002). Service quality delivery through web sites: A critical review of extant knowledge. Journal of the Academy of Marketing Science, 30(4), 362–375.

[13] J. Neoh, M. Chipulu, Alasdair Marshall (2017). What encourages people to carpool? An evaluation of factors with meta-analysis. Volume 44, pages 423–447

[14] Davison, L., Enoch, M., Ryley, T., Quddus, M., Wang, C., (2014). A survey of demand responsive transport in Great Britain. Transp. Policy 31, 40–55.

[15] Kumar, P Kishore. Kumar, N Ramesh. (2016) A Study on Factors Influencing the Consumers in Selection of Cab Services, International Journal of Social Science and Humanities Research, Vol. 4, Issue 3: 556-560.

[16] Jason-M-Richards, "Imputation-Techniques," GitHub repository, 2023. [Online]. Available: https://github.com/Jason-M-Richards/Imputation-Techniques