Enhancing Taxi Operations : Advanced Demand Forecasting and Dynamic Pricing Strategy Development

A Mid-term report for the BDM capstone Project

Submitted by

Name: Surya Vikram

Roll number: 22f3002751

IITM Online BS Degree Program,

Indian Institute of Technology, Madras, Chennai

Tamil Nadu, India, 600036

Contents :

Declaration Statement

I am working on a Project titled “Enhancing Taxi Operations : Advanced Demand Forecasting and Dynamic Pricing Strategy Development”. I extend my appreciation to Sugam Sawaari, for providing the necessary resources that enabled me to conduct my project.

I hereby assert that the data presented and assessed in this project report is genuine and precise to the utmost extent of my knowledge and capabilities. The data has been gathered from primary sources and carefully analyzed to assure its reliability.

Additionally, I affirm that all procedures employed for the purpose of data collection and analysis have been duly explained in this report. The outcomes and inferences derived from the data are an accurate depiction of the findings acquired through thorough analytical procedures.

I am dedicated to adhering to the principles of academic honesty and integrity, and I am receptive to any additional examination or validation of the data contained in this project report.

I understand that the execution of this project is intended for individual completion and is not to be undertaken collectively. I thus affirm that I am not engaged in any form of collaboration with other individuals, and that all the work undertaken has been solely conducted by me. In the event that plagiarism is detected in the report at any stage of the project's completion, I am fully aware and prepared to accept disciplinary measures imposed by the relevant authority.

I understand that all recommendations made in this project report are within the context of the academic project taken up towards course fulfillment in the BS Degree Program offered by IIT Madras. The institution does not endorse any of the claims or comments.

Signature of Candidate:

Surya Vikram

Name: Surya Vikram

Date: 4th July, 2024

1 Executive Summary and Title

The Sugam Sawaari taxi business data management midterm report offers a deep look into operations from August 2022 to July 2024. Its goal is to boost efficiency by finding hidden patterns in the data.

The main dataset shows trip details like date, demand, locations, distance, profit, and whether it was for a wedding (Lagan). This information is key for making smart business choices. Other datasets add details about fuel prices, weather, and more.

Descriptive statistics revealed flaws in the current pricing system. They also showed popular places, busy times, and how weddings affect profits. Advanced methods like time series forecasting and Random Forest were used to predict monthly demand and suggest profits based on important factors. Winter stood out as the busiest season, proving that weddings do impact business.

Data preparation involved combining datasets and fixing missing information. This ensured the data was clean while keeping its main features. Route analysis found that Patna and Gaya were the most common destinations, with most rides starting from Masaurhi.

Next steps focus on improving demand predictions and adjusting prices, following discussions with the owner. Operational costs will be reduced by optimizing frequently traveled routes. These efforts aim to enhance Sugam Sawaari's efficiency and position the business for future growth.

2 Proof of Originality of the Data

All the evidence substantiating the originality and authenticity of the datasets used for this Business Data Management project can be found at the link below:

https://drive.google.com/drive/folders/1Ee9kwCuekZ7E_XkSMfOX0pNP-eLoi18d?usp=sharing

3 Data

3.1 Data Collection

The data collection process involved gathering information from four distinct sources, covering the period from August 2022 to July 2024. These sources can be categorized as follows:

1. **Primary Data Source:**
    1.1. data.csv: Rides data extracted from physical data records manually.
2. **Secondary Data Sources:**
    2.1. fuel.csv: This dataset is scraped from mypetrolprice.
    2.2. weather.csv: Weather data obtained using the Meteostat Python library.
    2.3. location.csv: Location data retrieved using the Nominatim API.

3.2 Metadata

1. data.csv (722 entries and 7 fields):
    a. Fields:
        i. DATE: Date of the recorded ride in YYYY-MM-DD format
        ii. DEMAND: Boolean value indicating whether the taxi is booked
        iii. FROM: Starting location of the ride
        iv. TO: Destination of the ride
        v. DISTANCE: Distance traveled in kilometers
        vi. PROFIT: Net profit earned on a ride in rupees
        vii. LAGAN: Indicates whether it is a marriage occasion (boolean)
    b. Data type: Numerical and categorical
    c. Limitations:
        i. Data integrity issues with null entries in TO (278 nulls), DISTANCE (196 nulls), and PROFIT (128 nulls) fields.

2. fuel.csv (722 entries and 2 fields):
    a. Fields:
        i. DATE: Date of the fuel price record
        ii. FUEL_COST_PER_LITRE: Cost of diesel per litre on the given date
    b. Data type: Numerical
3. weather.csv (722 entries and 4 fields):
    a. Fields:
        i. DATE: Date of the weather record
        ii. TMIN: Minimum temperature for the day
        iii. TMAX: Maximum temperature for the day
        iv. PRCP: Precipitation amount
    b. Data type: Numerical
    c. Geographic focus: Area around the city of Patna
4. location.csv (29 entries and 3 fields):
    a. Fields:
        i. LOCATION: Name of the location
        ii. LATITUDE: Latitude coordinate of the location
        iii. LONGITUDE: Longitude coordinate of the location
    b. Data type: Categorical and numerical
    c. Geographic focus: Area around Patna

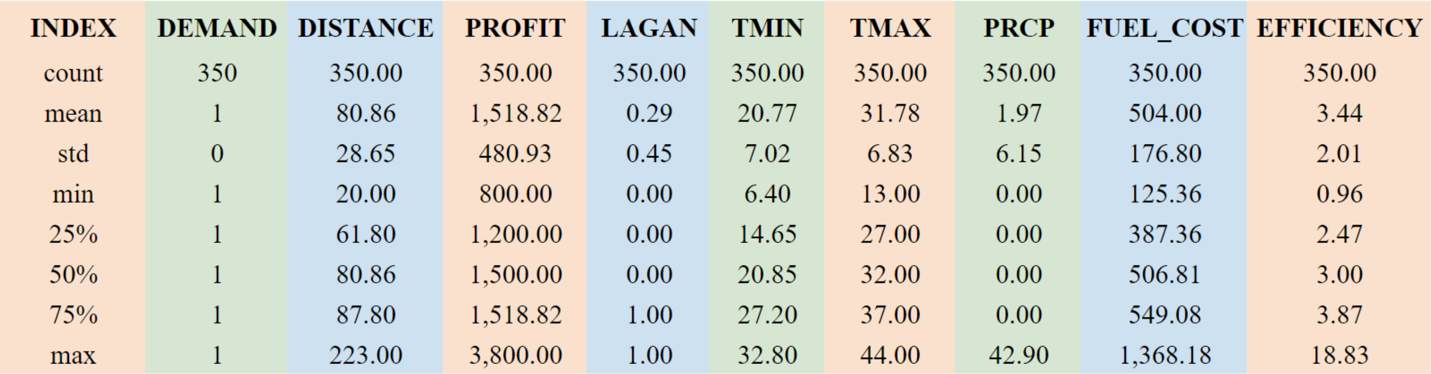
The preprocessed_data.csv file is a refined version of the original datasets, created through data integration and imputation. It merges the original CSV files (data.csv, fuel.csv and weather.csv) based on common field DATE. Missing values were handled using a two-stage process: first, a K-Nearest Neighbors (KNN) imputer was applied to maintain data relationships, then mean imputation was used for any remaining null values, particularly in the DISTANCE and PROFIT fields. This approach addresses data integrity issues while preserving overall data characteristics, though it may affect the variance of imputed fields.

3.3 Descriptive Statistics

As a prelude to the analysis, the fuel cost field has been added to the dataset to provide a clearer understanding of operational expenses. This field estimates the fuel cost for each trip in rupees, calculated using the trip distance (in kilometers), the taxi's mileage (15 kilometers per litre as specified by the owner), and the fuel cost per liter (in rupees). The fuel cost is then used to determine efficiency, defined as the ratio of profit to fuel cost (both in rupees). This unitless efficiency metric serves as a key performance indicator, offering insights into the financial performance of each trip in relation to its fuel consumption and highlighting the most profitable trips.

$$Fuel\ Cost\ = \frac{Distance \times Fuel\ Cost\ per\ Litre}{Mileage}$$

$$Efficiency\ = \frac{Profit}{Fuel\ Cost}$$



INDEX	DEMAND	DISTANCE	PROFIT	LAGAN	TMIN	TMAX	PRCP	FUEL_COST	EFFICIENCY
count	350	350.00	350.00	350.00	350.00	350.00	350.00	350.00	350.00
mean	1	80.86	1,518.82	0.29	20.77	31.78	1.97	504.00	3.44
std	0	28.65	480.93	0.45	7.02	6.83	6.15	176.80	2.01
min	1	20.00	800.00	0.00	6.40	13.00	0.00	125.36	0.96
25%	1	61.80	1,200.00	0.00	14.65	27.00	0.00	387.36	2.47
50%	1	80.86	1,500.00	0.00	20.85	32.00	0.00	506.81	3.00
75%	1	87.80	1,518.82	1.00	27.20	37.00	0.00	549.08	3.87
max	1	223.00	3,800.00	1.00	32.80	44.00	42.90	1,368.18	18.83

Figure 1: Descriptive Statistics Table for Numerical Data with Active Demand

The dataset highlights substantial variability in the taxi service's operational factors. Trip distances range from 20 to 223 km, with an average of 80.86 km, illustrating a broad spectrum of journey lengths that impact service dynamics. Profit margins fluctuate significantly between 800 and 3800, with a mean of 1518.82, indicating variability in profitability. Efficiency also displays considerable variation, from 0.96 to 18.83, reflecting inconsistencies in operational performance. Temperature data remain relatively stable, with average minimum and maximum temperatures of 20.77°C and 31.78°C, respectively, while precipitation levels exhibit considerable variation, potentially affecting service performance. These insights reveal the complex factors influencing the taxi service.

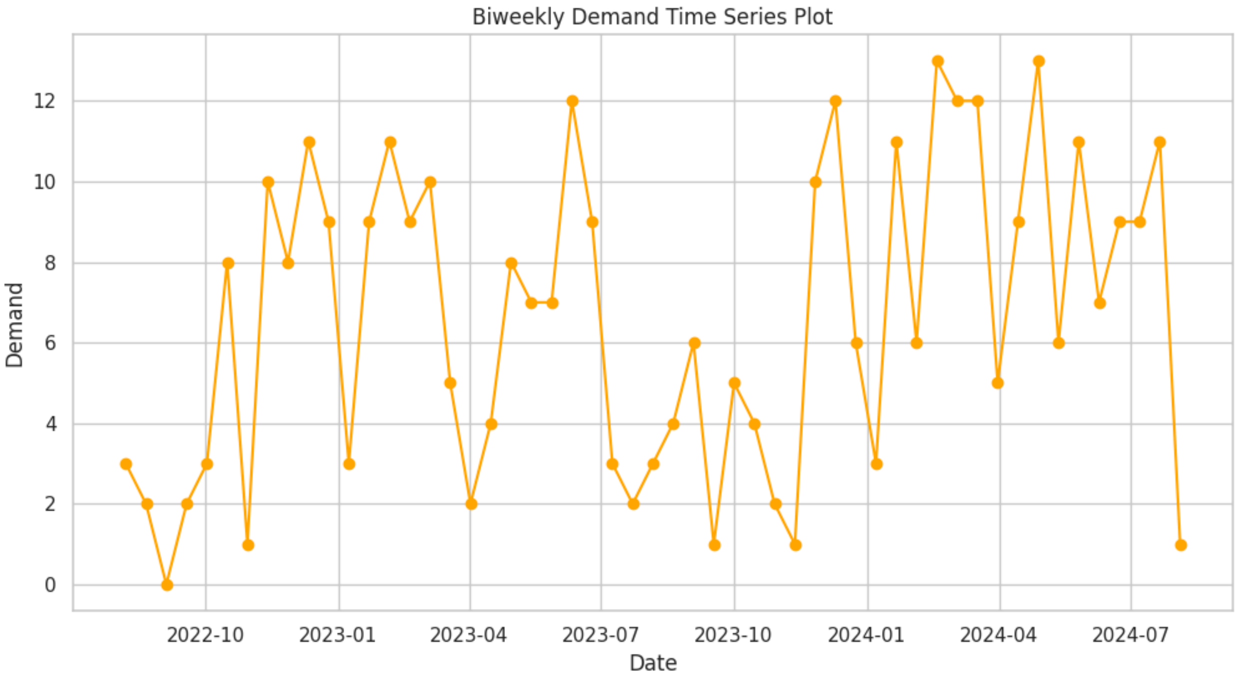


Figure 2: Time Series Plot Depicting the Biweekly Demand Trend Over the Past Two Years

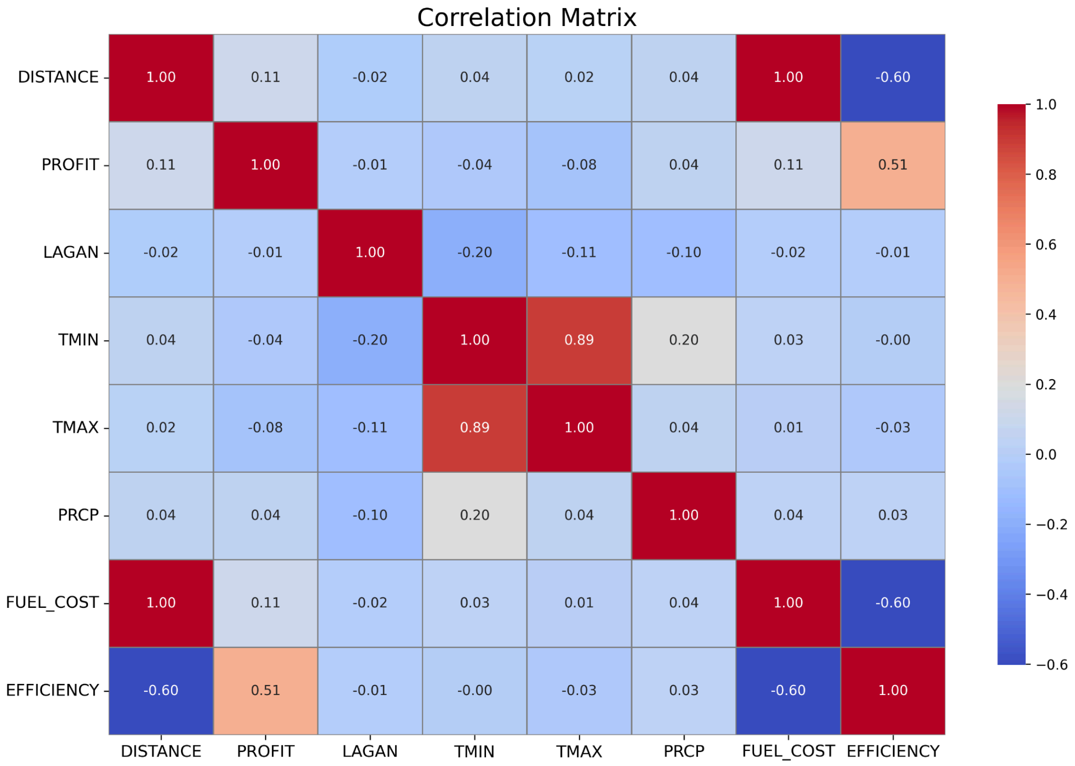


Figure 3: Correlation Matrix Illustrating Relationships Among Key Variables

Due to the high variability observed in the biweekly demand trend and the owner's request for a more stable forecasting approach, transitioning to monthly demand forecasting appears to be a viable option. Additionally, the correlation matrix reveals the relationships between profit and various variables, providing valuable insights that will be utilized to develop a dynamic pricing model.

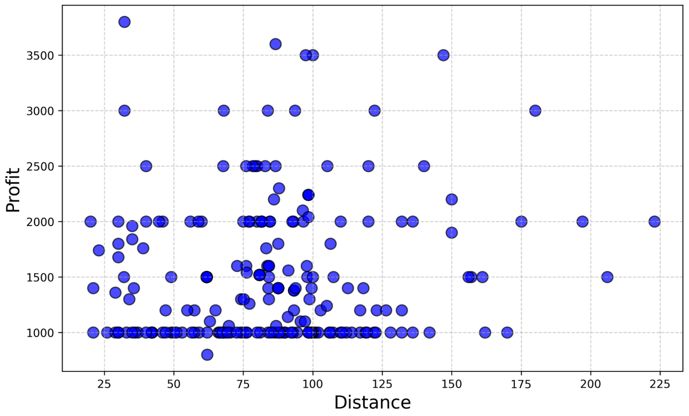


Figure 4

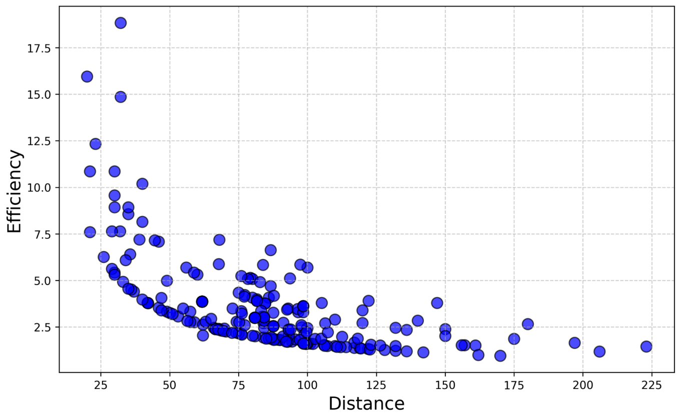


Figure 5

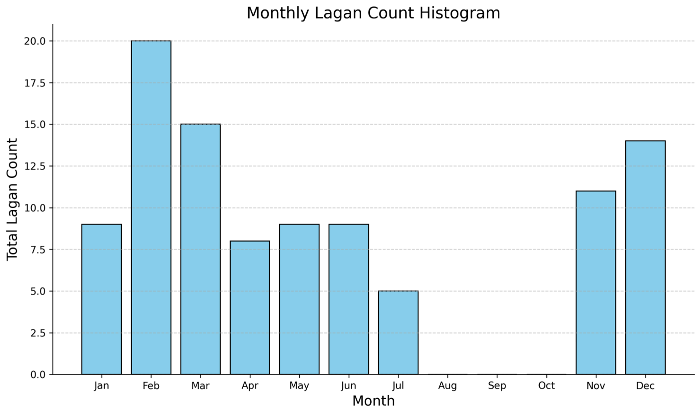


Figure 6

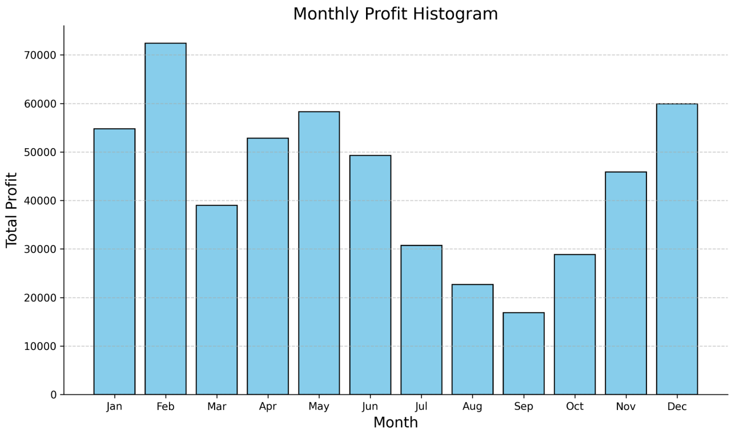


Figure 7

Analysis of these four plots reveal a lack of correlation between distance and profit, indicating a static pricing model that doesn't account for distance-related costs. This suggests potential for optimization in the pricing strategy. Clear seasonality is evident in both Lagan count and profits, with peaks in winter months and troughs in late summer and rainy seasons.

4 Detailed Explanation of Analysis Methods and Justification

4.1 Demand Forecasting Using SARIMA Time Series Analysis

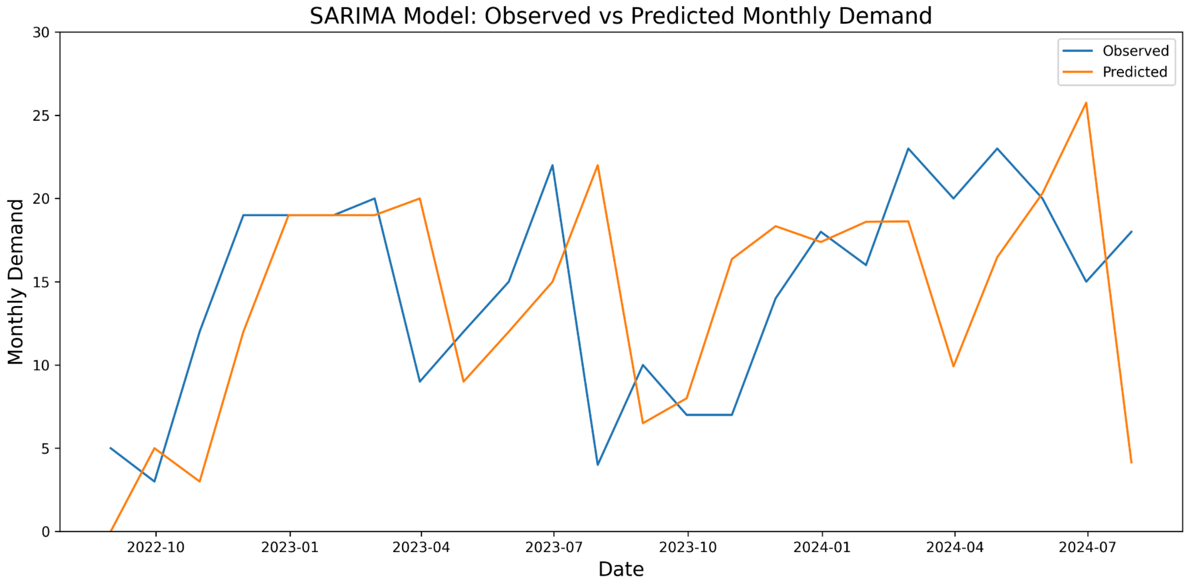


Figure 8: SARIMA model forecast for monthly demand over 2 years period

The SARIMA (Seasonal Autoregressive Integrated Moving Average) model is used for forecasting time series data with trend and seasonal patterns. It combines non-seasonal parameters (1, 1, 1) for autoregression, differencing, and moving averages, with seasonal parameters (1, 1, 1, 12) to capture yearly cycles in monthly data. The model is fitted to predict monthly demand, and the results are visualized by plotting both observed data and predictions to evaluate performance.

SARIMA is ideal for datasets with both trends and seasonal patterns, which can be observed here due to the impact of LAGAN (marriage occasions) especially in the months of July to September. Unlike simpler models, it explicitly handles seasonality, making it more effective for data with clear seasonal effects. Here, it captures both trends and seasonal peaks better than basic ARIMA models. Its flexibility allows adaptation to various seasonal periods, providing more accurate and insightful forecasts. This justifies the usage of SARIMA as a powerful tool for understanding and predicting monthly taxi demands.

4.2 Dynamic Pricing Model Using Random Forest Analysis

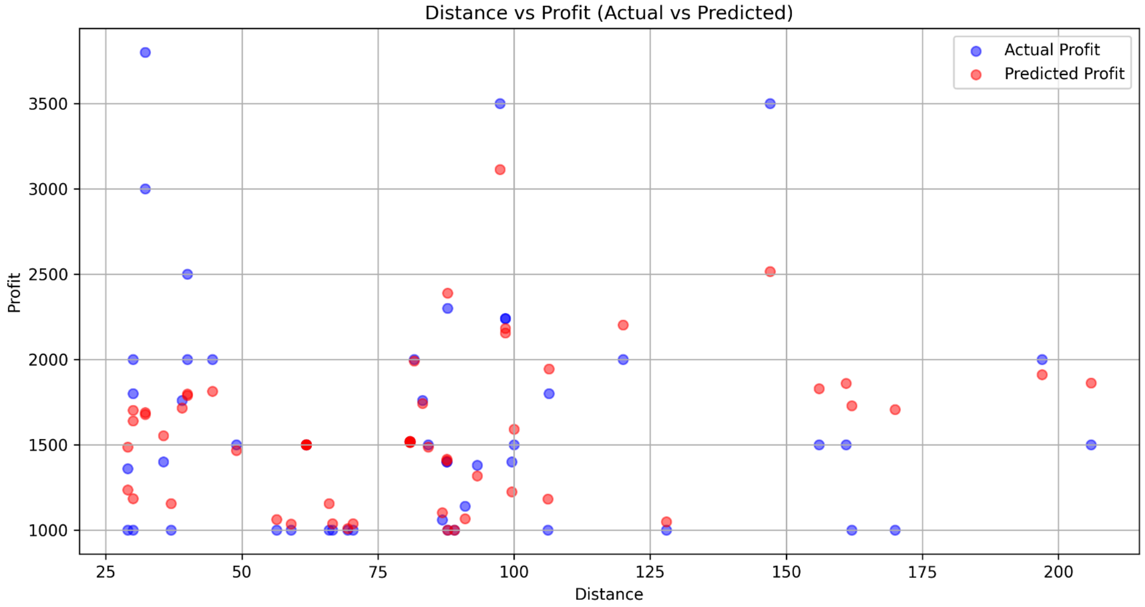
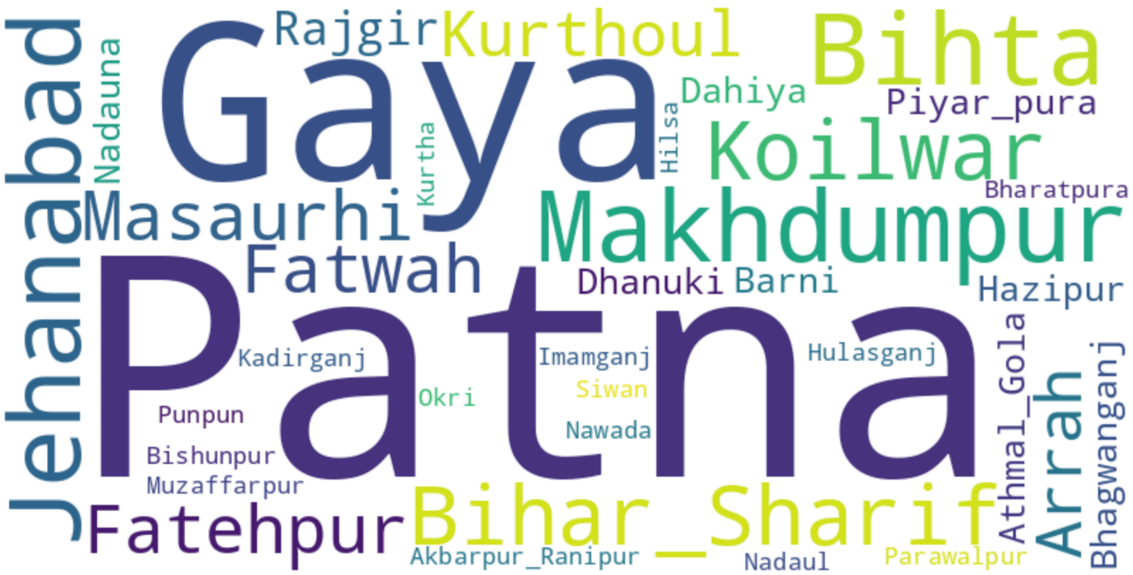


Figure 9: Random Forest predictions for optimal profit on distance

The Random Forest model was implemented to support dynamic pricing by predicting the profit based on various influential factors. Key features included FROM, TO, DISTANCE, LAGAN, TMIN, TMAX, PRCP, FUEL_COST, and EFFICIENCY, with PROFIT as the target variable. The dataset was enriched with additional date-related features such as MONTH, enhancing the model's ability to capture temporal patterns. The preprocessing steps involved scaling numeric features and one-hot encoding categorical features to ensure uniformity and model readiness. The model was trained and tested using an 80-20 train-test split, followed by fitting the Random Forest algorithm and making predictions on the test set.

The Random Forest model was chosen for its robustness and ability to handle complex, non-linear relationships. The model's flexibility allows for more accurate and dynamic pricing strategies, maximizing profitability by considering a comprehensive set of factors influencing demand and cost. The scatter plot comparing actual and predicted profits against distance demonstrates the model's effectiveness, providing insightful forecasts that may help optimize pricing decisions. This justifies the use of Random Forest as a powerful tool for enhancing dynamic pricing in the taxi service.

4.3 Word Cloud Analysis of Destinations



To enhance route efficiency for Sugam Sawaari, a word cloud was generated to visualize the most frequently traveled destinations from Masaurhi. Analysis of the journey data revealed that Patna and Gaya are the most common destinations. By focusing on optimizing routes to these key locations, Sugam Sawaari can significantly improve operational efficiency and increase profitability.

5 Results and Findings

The dataset analysis revealed significant variability in operational factors, including trip distances, profit margins, and efficiency. The correlation between distance and profit increased from -0.03 to 0.3, a relationship that may be further refined based on the owner's feedback. For monthly demand forecasting, the SARIMA model has been selected but still requires fine-tuning. Additional feature engineering, such as creating lag and rolling average features, can enhance the model's performance. Route analysis shows that Patna accounts for 27% of the destinations, while Gaya accounts for 11%, with Masaurhi being the starting point for rides 95% of the time. Clear seasonal patterns are evident, with peaks in profit and Lagan counts during winter months and troughs in late summer and rainy seasons, justifying the use of SARIMA modeling to capture these trends. These findings provide valuable insights for improving the operational efficiency and profitability of the taxi service.