Scientific Research Publishing

American Journal of Industrial and Business Management, 2025, 15(8), 1155-1171
https://www.scirp.org/journal/ajibm
ISSN Online: 2164-5175
ISSN Print: 2164-5167

Analyzing Key Factors Influencing Coffee House Revenue: A Predictive Modeling Approach

Saptarshi Chakma

Department of Management, Rangamati Science and Technology University, Rangamati, Bangladesh
Email: saptarshichakma87@gmail.com

How to cite this paper: Chakma, S. (2025). Analyzing Key Factors Influencing Coffee House Revenue: A Predictive Modeling Approach. *American Journal of Industrial and Business Management, 15,* 1155-1171.
https://doi.org/10.4236/ajibm.2025.158057

Received: August 4, 2025
Accepted: August 23, 2025
Published: August 26, 2025

Abstract

Nowadays, understanding and predicting revenue trends is highly competitive, in the food and beverage industry. It can be difficult to determine which aspects of everyday operations have the most impact on income, especially for coffee shops. Transactional and behavioral data are readily available, but numerous small businesses lack the data-driven models necessary to convert these insights into predictions that can be put into action. By using linear regression techniques to forecast daily income based on important business parameters, this study seeks to close the gaps. In order to investigate feature distributions and correlations, exploratory data analysis performed including statistical summaries, box plots, and scatter plots with regression lines. A correlation study determines the most important parameters and are "Number of Customers Per Day", "Average Order Value", and "Marketing Spend Per Day". Secondary elements like location foot traffic, the number of employees and operating hours come after this. A linear regression model is trained using these characteristics, yielding an R^2 score of 0.89, a Mean Absolute Error (MAE) of 244.13. The model's efficacy is validated by comparing actual and projected revenue. This method provides a useful foundation for forecasting revenues and making well-informed decisions in small-scale retail businesses, such as coffee shops.

Keywords

Revenue, Data Analysis, Linear Regression, Business, Coffee House

1. Introduction

In the food and beverage sector, revenue forecasting is an essential aspect of strategic decision-making, especially for small and medium-sized businesses like coffee shops. Numerous variables, such as the number of customers, average ex-

penditure, marketing initiatives, and foot traffic surrounding the company's location, the number of employees, and working hours can affect daily income. For forecasting and planning, many coffee shop managers depend on their gut feelings rather than data-driven insights, even if they have access to key operational information (Ahmad et al., 2020a; Fiori & Foroni, 2020). Coffee shop operators may increase overall profitability, optimize marketing expenditures, and manage resources more effectively with the aid of accurate revenue forecasting. The use of straightforward, interpretable prediction models is lacking in many corporate settings, nevertheless, as resources for intricate systems could be scarce (Munoz & Vassilvitskii, 2017).

To fill this gap, this study models and forecasts daily revenue using linear regression based on important business parameters. I determine the association between each parameter and daily income via an exploratory data analysis that includes statistical summaries, box plots, and scatter plots with regression lines. The most important variables, "Number of Customers Per Day", "Average Order Value", "Marketing Spend Per Day", and "Location Foot Traffic", are chosen to train the linear regression model. The main objectives of the study are as follows:

- Explore descriptive statistics and visualization techniques to understand the impact of each feature.
- Visualize the feature relationship using box plots and scatter plots with a regression line.
- Compute the correlation between input features and the target variable.
- Select the best attributes based on correlated values.
- Develop a linear regression-based model on the top features.
- Evaluate the performance of the model using R^2 value, RMSE and MAE.
- To show and contrast actual and projected income in order to evaluate the model's predictive power and practical correctness.
- To offer data-driven insights that can assist stakeholders and coffee shop managers in making well-informed choices about revenue optimization, customer service, and marketing.

2. Related Works

As predicting revenue is very crucial for all small and large businesses, many researchers are working in this field. Some of them are as follows: Lin and colleagues (Lin et al., 2022) proposed employing Generalized Additive Models (GAMs) and machine learning models based on artificial neural networks (ANNs) to estimate the revenues of hybrid hydropower and energy storage systems in a quick and precise manner. With validation errors often less than 5%, the novel method cuts calculation time from around three hours to just one to four minutes per battery configuration when compared to standard MILP models. The accuracy of ML models was consistently higher than that of GAMs. This approach promotes the wider use of energy storage in renewable systems and provides investors and plant owners

with a useful and effective tool for assessing battery size alternatives.

Researchers (Jian et al., 2020) investigated the prediction of product market revenue using neural networks backed by fuzzy logic and artificial intelligence algorithms. Future sales are predicted using neural networks, which are well-known for their resilience and capacity to represent intricate nonlinear interactions. With prediction errors kept to 4%, the study concludes that the neural network-based prediction model achieves excellent accuracy. Businesses may increase profitability by using this forecasting technique to better understand market trends and make well-informed decisions.

A shortcoming of traditional frequent itemset mining that ignores the differing importance or monetary worth of transactions among various consumers is addressed in this work (Weng, 2017). In order to address this, the authors developed a frequency-monetary (FM) weighting approach that more accurately reflects client value by taking into account both transaction frequency and revenue contribution. For mining high-revenue frequent itemsets from FM-weighted transactions, they put out a unique approach. The efficiency of this strategy in revenue-focused customer research was demonstrated by experimental findings utilizing survey data, which revealed that the top-k itemsets found using this method more correctly forecasted future customer revenues than the usual methodology.

In contract-based service contexts, where it's critical to estimate changes in client or service-level revenues over time, the revenue change prediction problem addressed (Mahajan et al., 2020). Since limited resources, like consultants or salespeople, must be effectively distributed based on anticipated revenue fluctuations, accurate forecasting is essential. By framing revenue change forecast as a classification issue, the authors present a unique paradigm. They present a system that optimizes prediction precision while reducing the loss of overall accuracy, acknowledging the problem of class imbalance in such datasets. The technique outperforms conventional classifiers and provides useful insights for resource prioritization. It is proven using real-world data from a top global cloud services provider.

The use of social media data, particularly YouTube trailer reviews, to forecast box office receipts before to a film's debut is investigated (Ahmad et al., 2020b). This method concentrates on early-stage prediction by examining public involvement and sentiment, in contrast to earlier approaches that depend on Twitter or IMDb evaluations after publication. The model presents novel indicators such the like-to-dislike ratio, the positive-to-negative sentiment ratio, and purchase intention. According to experimental data, the suggested approach achieves a relative absolute error of 29.65%, outperforming three baseline models. This strategy demonstrates how pre-release social media interactions may be used to anticipate movie income earlier and with more accuracy.

A WiFi-based sensing method presented to better estimate revenue in retail environments and comprehend consumer behavior (Golderzahi & Pao, 2024). The system finds groups of consumers with similar habits by using WiFi access points

placed in cafeterias to track client visiting patterns, including Service Set Identifier (SSID) data. Predictive modeling utilizing machine learning methods such as Random Forest and Support Vector Regression is based on these groupings. Revenue from coffee shops, the quantity of products sold, and the number of consumer devices are the three main outcomes that the models seek to forecast. Predictive accuracy is greatly increased by adding weather and customer group data, with a Mean Absolute Percentage Error (MAPE) improvement of 6% to 10%. This study shows how behavioral data and environmental elements may be effectively integrated to improve retail forecasting accuracy.

Using information from the Ethiopian Commodity Exchange (ECX), it was predicted the future prices of two important Ethiopian export commodities: coffee and sesame (Fofanah, 2021). By using and contrasting three algorithms—Linear Regression (LR), Extreme Gradient Boosting (XGB), and Long Short-Term Memory (LSTM)—it fills a research gap. The study assesses the accuracy of each model using datasets of 7205 for sesame and 1540 for coffee. Ethiopia Coffee Prices Predictor (ECPP), an easy-to-use mobile app, was also created to make price forecasts available, demonstrating the potential of mobile-based forecasting tools in regional commodity markets.

The difficulties in predicting income in the food sector, particularly for restaurants in urban locations, are focused (Sanjana Rao et al., 2021). Three restaurant types—inline, food court, and mobile—are the subject of the study. It suggests a method for forecasting restaurant income that takes into account a number of factors, ranking the input features according to how they affect the desired attribute. To enhance the quality of the dataset, the study uses pre-processing methods such Principal Component Analysis (PCA), feature selection, and label encoding. Following dataset training, the models are assessed, and it is shown that Random Forest (RF) predicts revenue more accurately than Linear Regression. Additionally, the study demonstrates that pre-processing greatly improves model accuracy.

However, there is a need for straightforward, understandable, and useful models made for small businesses like coffee shops because the majority of current research focuses on complicated models or large-scale firms. The purpose of this study is to close the gaps.

3. Methods and Materials

Figure 1 illustrates the main architecture of the proposed revenue prediction system. It consists of seven modules. Module 1 describes the dataset, and the second module shows the statistical summary of the dataset. Feature analysis has been explored in the third module. Module 4 demonstrates the relationship of each feature with the target variable. After that 5th and 6th modules calculated the correlation matrix and assisted in selecting the top input features. Finally, in module 7, I have applied the linear regression-based model.

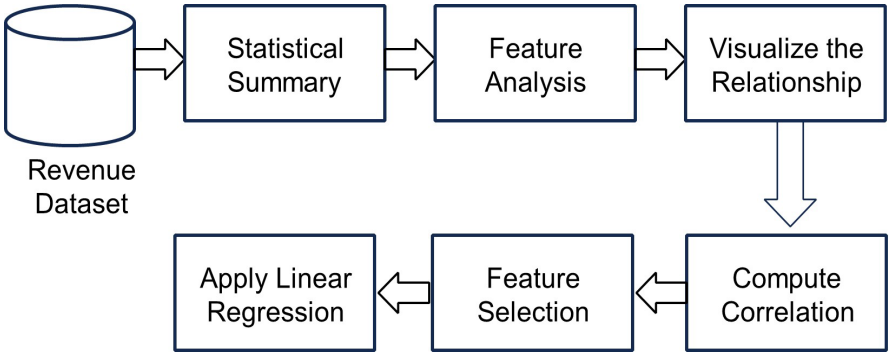


Figure 1. Main architecture of the revenue prediction system.

3.1. Dataset Description

Table 1 shows the samples of the coffee house revenue dataset. The dataset has been collected from a public repository named Kaggle. It replicates a single coffee shop's daily operating data. Although it is appropriate for an exploratory study, it might not accurately reflect the wider range of actual coffee shops. It has six input features—number of customers per day, operating hours, location foot traffic, marketing spend hours, number of employees, average order value, etc. The target variable is Daily Revenue. Location foot traffic is the quantity of people who walk past the coffee shop every day, as determined by street sensors. It doesn't mean you're involved or have entered the establishment. The dataset has a collection of a total of 2000 entities. To know more about the dataset, see https://www.kaggle.com/datasets/himelsarder/coffee-shop-daily-revenue-prediction-dataset.

3.2. Statistical Summary

Table 2 demonstrates the basic statistical summary of the dataset which comprises information from 2000 observations across seven operational and financial variables.

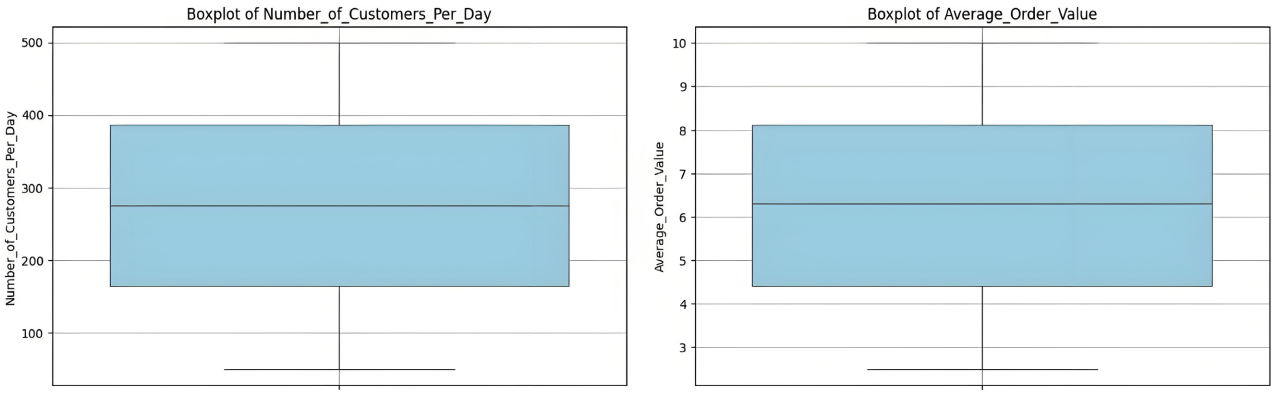
With numbers ranging from 50 to 499, the average daily customer count is around 274, suggesting considerable variation in consumer turnout. The average daily operation hours are around 11.66, with some locations working as little as 4 hours and others up to 17 hours. The average order value is $6.26. There are typically eight employees each site. With a mean of $252.61, the daily marketing cost varies greatly, ranging from $10.12 to $499.74, indicating different advertising tactics in each region. $5114.60, however, possible losses or abnormalities in the data are indicated by a negative minimum value of −$58.95. Significant variation across a number of business aspects is highlighted in this report, and these factors probably affect daily revenue results. Another important consideration is location foot traffic, which varies daily from 110 to 999 people, with an average of around 534.89. Lastly, daily revenue fluctuates a lot, ranging from an average of $1917.32 to a maximum of $5114.60. However, a negative minimum figure of −$58.95 suggests possible losses or irregularities in the data. Significant variation is shown in this overview across a number of business aspects, which probably affect daily revenue results.

Table 1. Sample of the coffee house revenue dataset.

Number of Customers Per Day	Average Order Value	Operating Hours Per Day	Number of Employees	Marketing Spend Per Day	Location Foot Traffic	Daily Revenue
152	6.74	14	4	106.62	97	1547.81
485	4.5	12	8	57.83	744	2084.68
398	9.09	6	6	91.76	636	3118.39
320	8.48	17	4	462.63	770	2912.2
156	7.44	17	2	412.52	232	1663.42
121	8.88	6	9	183.49	484	1155.18

Table 2. Statistical summary of the coffee revenue dataset.

	Number of Customers Per Day	Average Order Value	Operating Hours Per Day	Number of Employees	Marketing Spend Per Day	Location Foot Traffic	Daily Revenue
Count	2000	2000	2000	2000	2000	2000	2000
Mean	274.29	6.26	11.66	7.94	252.61	534.89	1917.32
Std	129.44	2.17	3.43	3.74	141.13	271.66	976.20
Min	50	2.50	6	2	10.12	50	-58.95
25%	164	4.41	9	5	130.125	302	1140.08
50%	275	6.30	12	8	250..99	540	1770.77
75%	386	8.12	15	11	375.35	767	2530.45
Max	499	10	17	14	499.74	999	5114.60



(a) (b)

Figure 2. Boxplots of (a) Number of customers per day, (b) Average order value.

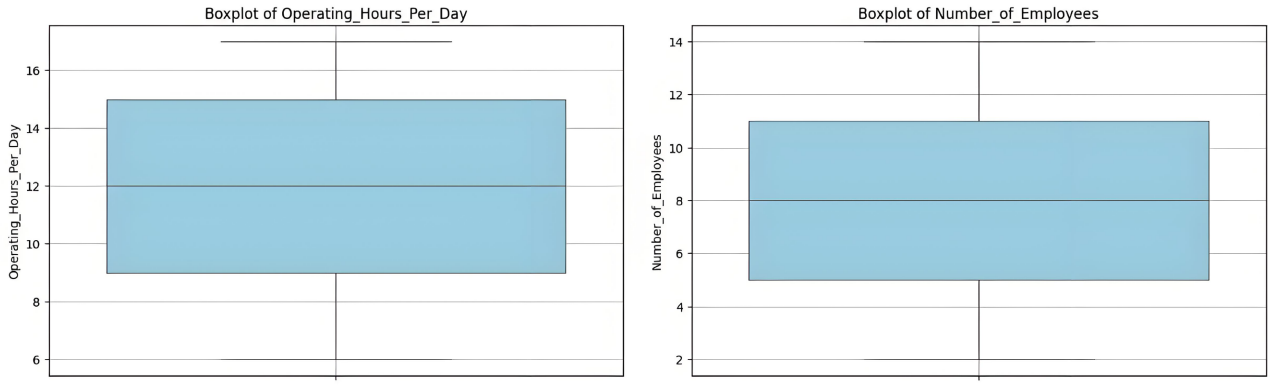
3.3. Preprocessing Stage

Preprocessing revealed that negative revenue figures (such as −$58.95) were data errors and were eliminated from the dataset. Since there were no missing values, imputation was not necessary.

3.4. Feature Analysis

A thorough feature analysis was performed on the dataset in order to estimate restaurant income with accuracy. Every input attribute was evaluated for impact and relevancy (Chang & Yang, 2016).

- Number of Customers Per Day—The dispersion of daily customer counts is seen in **Figure 2(a)**. Since the median is about 300 patrons, most restaurants probably serve this many every day. Variability in the daily client numbers is indicated by the interquartile range (IQR), which ranges from around 175 to 400. Beyond the whiskers, there are a few data points that might indicate outliers or exceptionally high or low consumer days. Daily income may be impacted by this fluctuation, underscoring the significance of steady foot traffic.
- Average Order Value—The distribution of the average order value per customer is seen in **Figure 2(b)**. The majority of values lie between about $4 and $8.5, with the median order value being around $6.5. Order values appear to be quite constant across observations, as indicated by the comparatively tiny IQR. Discounts or days with low expenditure may be indicated by a few low-end outliers. Predicting revenue more precisely can be aided by stable average order values.

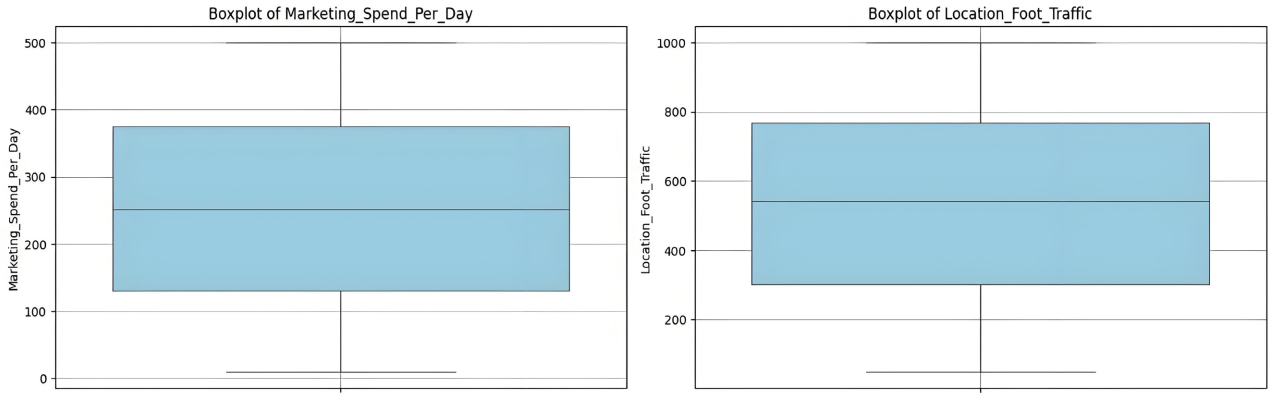


(a) (b)

Figure 3. Boxplots of (a) Operating hours per day, (b) Number of employees.

- Operating Hours Per Day—The number of hours a restaurant is open each day is displayed in **Figure 3(a)**. The majority of establishments operate between 9 and 15 hours, with a median of around 12 hours. There are a few eateries with longer hours, which stand out as minor anomalies. The number of clients serviced and, consequently, income is directly impacted by operating hours. In addition to perhaps increasing revenue, longer hours may also result in increased operating expenses.
- Number of Employees—The median of about ten employees is displayed in the employee boxplot **Figure 3(b)**. There are several outliers on both extremes of the spectrum, which normally ranges from 6 to 13 personnel. For the esti-

mate of labor costs and service quality, this measure is essential. Greater staffing levels might be a sign of busier or larger businesses. Managing operational efficiency may depend on how well employees are allocated.



(a) (b)

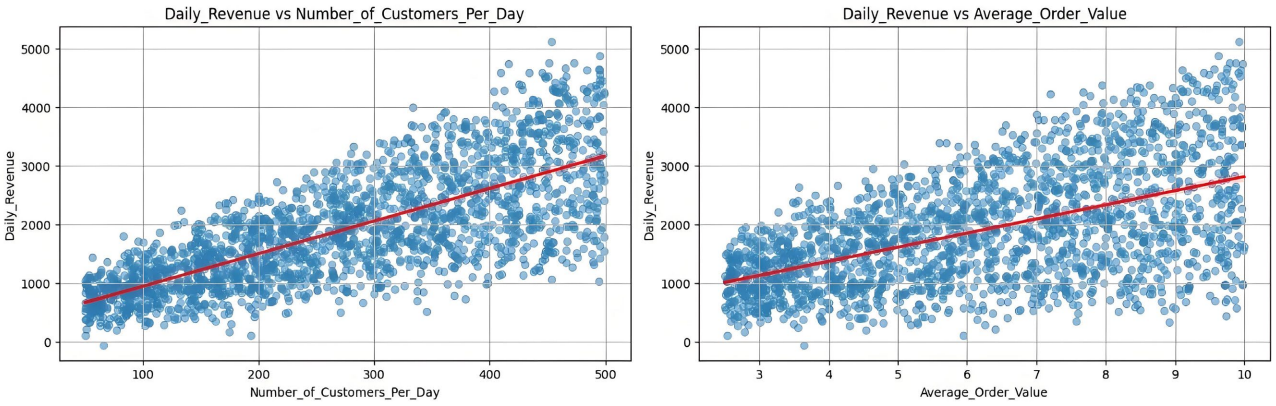
Figure 4. Boxplots of (a) Marketing spend per day, (b) Location foot traffic.

- Marketing Spend Per Day—The distribution of daily marketing costs is seen in **Figure 4(a)**. With a broad IQR ranging from $100 to more than $400, the median is about $250. A few extremely low outliers around 0 can indicate companies that invest little to nothing in marketing. Revenue and consumer engagement are frequently correlated with marketing spending. A wide range of expenditures points to various restaurant tactics.
- Location Foot Traffic—The number of people who pass by the restaurant every day is shown in **Figure 4(b)**. In the IQR, foot traffic ranges significantly from around 300 to 800, with a median of about 550. Outliers are defined as a few extreme values. Although conversion rates rely on other criteria, including location type and marketing efforts, more foot traffic is typically linked to better commercial potential. This factor has a significant impact on revenue forecasting models.

Since each characteristic exhibits significant fluctuation, they may all help formulate predictions. Larger ranges are displayed by some features (such as location foot traffic and marketing budget), suggesting a higher degree of impact or variability. No significant skewness or severe outliers, suggesting that the data behaved well following any necessary cleaning.

3.5. Visualize the Relationship with the Target Variable

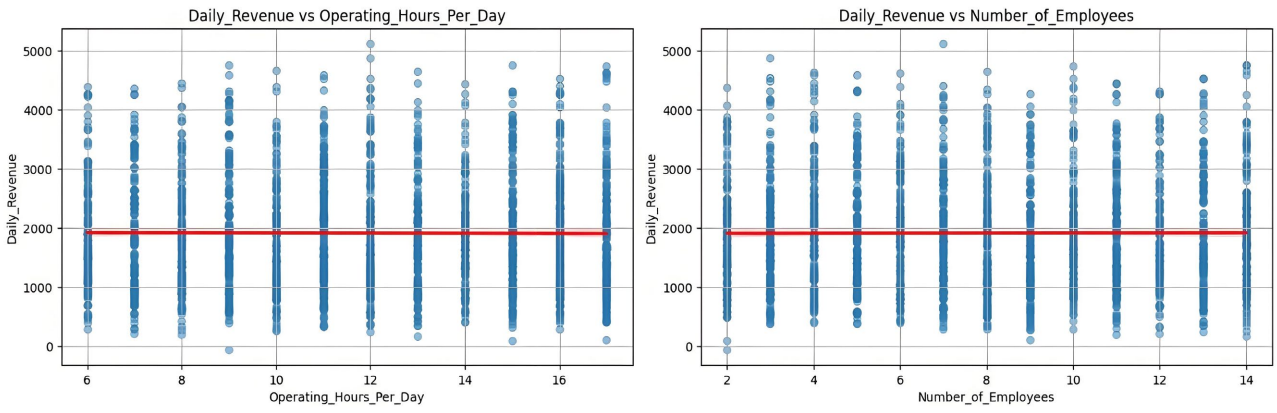
I employ scatter plots with trendlines to assess the impact of each input characteristic (such as the average order value, number of customers, etc.) on Daily Revenue (Waskom, 2021; Sadiku et al., 2016). These images aid in identifying the strength (strong or weak) and direction (positive or negative) of connections. A characteristic is deemed potentially important if it has a distinct upward or downward trend with the target variable.



(a) (b)

Figure 5. Scatter plot of daily revenue vs (a) Number of customers per day and (b) Average order value.

- Daily Revenue Vs Number of Customers Per Day—**Figure 5(a)** clearly shows a positive linear association between daily income and the number of clients. Daily income climbs dramatically in tandem with the number of patrons, which is consistent with what is expected in a restaurant or retail establishment. According to the well-fitting regression line, daily income is significantly predicted by the number of customers. This suggests that initiatives to boost foot traffic can improve revenue results directly.
- Daily Revenue Vs Average Order Value—Additionally, there is a significant association between daily income and the average order value in **Figure 5(b)**. The regression line's upward slope suggests that more spending per customer makes a significant contribution to overall revenue, even though the link is not as strong as it is with customer count. Pricing tactics and upselling may still be useful for increasing income, but the data's dispersion indicates that other factors could still be at play.



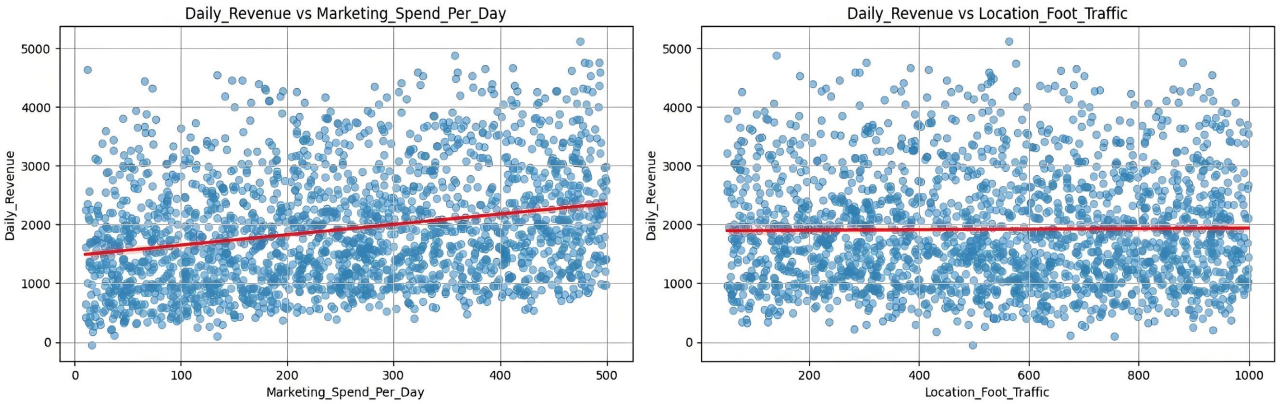
(a) (b)

Figure 6. Scatter plot of daily revenue vs (a) Operating hours per day and (b) Number of employees.

- Daily Revenue Vs Operating Hours Per Day—There is little to no relationship between operating hours and daily income, according to **Figure 6(a)**. The

nearly flat regression line indicates that simply remaining open for a longer period of time does not always result in increased profits. This might indicate that, regardless of the overall number of daily operation hours, customer traffic is concentrated within specific peak periods or that there are diminishing returns to expanding hours.

- Daily Revenue Vs Number of Employees—There doesn't seem to be much of a correlation between daily income and staff count (Figure 6(b)). There is little association, as the regression line is almost flat. This implies that hiring more employees does not always translate into more sales and that labor optimization should be driven by service requirements rather than the hope of immediately increasing revenue.



(a) (b)

Figure 7. Scatter plot of daily revenue vs (a) Marketing spend per day and (b) Location foot traffic.

- Daily Revenue Vs Marketing Spend Per Day—In Figure 7(a), the moderately upward sloping regression line indicates a minor positive trend between daily income and marketing expense. Although not particularly strong, this association implies that marketing initiatives might help boost sales. However, the broad range of data points also suggests that marketing effectiveness varies, maybe based on target audience, time, or technique.
- Daily Revenue Vs Location Foot Traffic—The almost flat regression line indicates a poor relationship between location foot traffic and daily income in Figure 7(b). Although it is anticipated that increased foot traffic will increase revenue, this outcome may point to low conversion rates or inefficient customer service. It emphasizes how crucial in-store appearance and experience are to turning foot traffic into paying clients.

3.6. Compute the Correlation

The correlation analysis between input feature and target variable has been shown in Figure 8. It shows that Daily Revenue has the strongest positive correlation with Number of customers per day (0.74), according to the correlation analysis, suggesting that gaining more customers greatly increases revenue. This is followed

by a moderate correlation with Average Order Value (0.54), indicating that raising per-customer spending also makes a significant contribution. Marketing Spend per day has a lesser positive correlation (0.25) than customer volume or order value, suggesting that it has some influence but not as much.

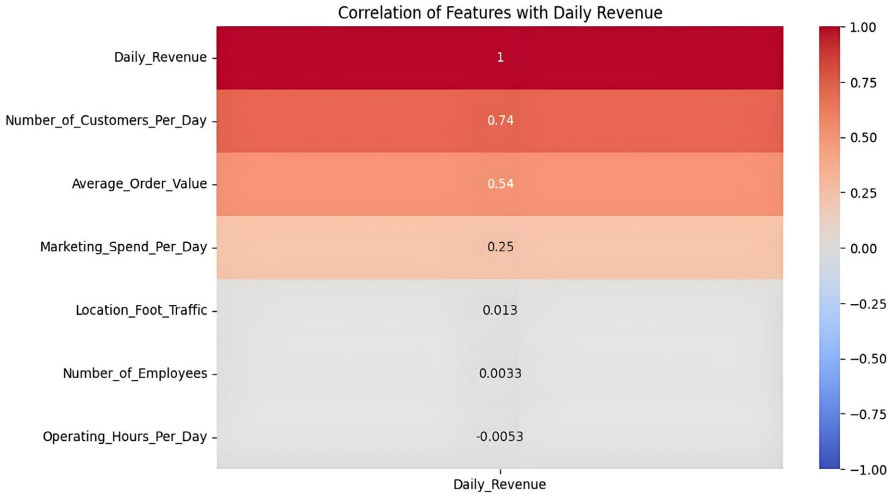


Figure 8. Correlation matrix of the dataset.

While Operating Hours Per Day exhibits a minimal negative correlation (−0.0053), suggesting that merely extending business hours does not consistently increase revenue, Location Foot Traffic (0.013) and Number of Employees (0.0033) show negligible correlations, meaning they have virtually no effect on revenue. These results imply that instead than concentrating on elements like foot traffic, personnel numbers, or operation hours, firms should give priority to methods that increase average order values and boost consumer traffic.

3.7. Select the Top Features

I have selected the top features based on the correlation analysis between input features and the target variable. The top features of the dataset are the number of customers per day, average order value, marketing spend, and location foot traffic. The number of employees has less impact on the daily revenue, whereas operating hours have no positive impact on the daily revenue of the coffee house.

3.8. Apply Linear Regression Model

A basic statistical technique for simulating the connection between a dependent variable and one or more independent variables is linear regression. Predicting and explaining how the dependent variable will behave depending on the values of the independent variables is its main goal (James et al., 2023; Hope, 2020).

The goal of linear regression is to fit a hyperplane (in multiple linear regression) or a straight line (in basic linear regression) that best captures the relationship between the independent variable or variables and the dependent variable. The

fundamental linear regression equation is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \quad (1)$$

The dependent variable, represented by y in a linear regression model, is the goal or result I hope to predict. The independent variables or predictors that are utilized to estimate the value of y are $x_1, x_2, \cdots, x_n$. The value of y when all independent variables are zero is shown by the word β_0, which is the intercept. The regression coefficients for each related independent variable are represented by the coefficients $\beta_1, \beta_2, \cdots, \beta_n$, which indicate the strength and direction of the association between each predictor and the target variable. Last but not least, ε is the error term or residual that explains the variation in y that the linear connection with the independent variables is unable to account for.

Reducing the variation between the expected and actual values is the aim of linear regression. Ordinary Least Squares (OLS) is a technique used to minimize the sum of squared residuals (Maulud & Abdulazeez, 2020; Kumari & Yadav, 2018).

Many different areas use linear regression extensively to evaluate correlations and make predictions. It is employed in economics to predict demand patterns and pricing. It aids in sales forecasting in marketing by taking into account variables like advertising budget. In the medical field, linear regression calculates how various risk variables affect a person's chance of contracting an illness.

4. Result

4.1. Evaluation of the Performance Metrics

Table 3, Performance Metrics of the Model, which is shown in this table, gives a numerical overview of a predictive model's performance (Liu et al., 2014). Regression models that might forecast daily income are frequently assessed using these measures.

Table 3. Performance metric model.

R^2 Score:	0.895398
MAE:	244.126
Top Features: Number of Customer, Average Order, Marketing Spend Per Day, Location Foot Traffic	

The percentage of the variation in the dependent variable that can be predicted from the independent variables is shown by the R^2 score. It has a range of 0 to 1. With a R^2 score of 0.895398, the model can account for around 89.54% of the volatility in the result (Padilla et al., 2020). In general, a larger R^2 means that the model fits the data better. Given that 89.54% is a rather large number in this instance, the model appears to have significant explanatory power.

The mean absolute difference (MAE) between the actual observed values and the expected values is computed. It provides a sense of the average deviation of the forecasts (Chicco et al., 2021). MAE makes it clear how inaccurate forecasts

are. An MAE of 244 indicates, for instance, that estimates are, on average, 244 units off from the actual values. There is no direction bias in it. MAE evaluates over- and under-predictions equally since it considers the absolute difference.

In summary, this table shows that the model fits the data well generally (R^2 of 89.54%), with average prediction errors ranging from $244 to $313 (in revenue currency units). The number of customers, average order value, daily marketing expenditure, and foot traffic at the location are the most important aspects affecting the model's projections. These measurements offer a thorough grasp of the model's performance when paired with visual aids such as the scatter plot you previously showed.

4.2. Contrast between Actual and Predicted Revenue

Figure 9 explores the contrast between actual and predicted revenue. A high proportion of points above the red line indicates that the model frequently overestimates daily income. In the event that most of the points are below the red line, the model is likely to underestimate the daily revenue (Wang et al., 2017; Shao et al., 2017). The points appear to be spread rather equally around the "Perfect Prediction" line in this specific figure, indicating neither a significant systematic overestimation nor an underestimation.

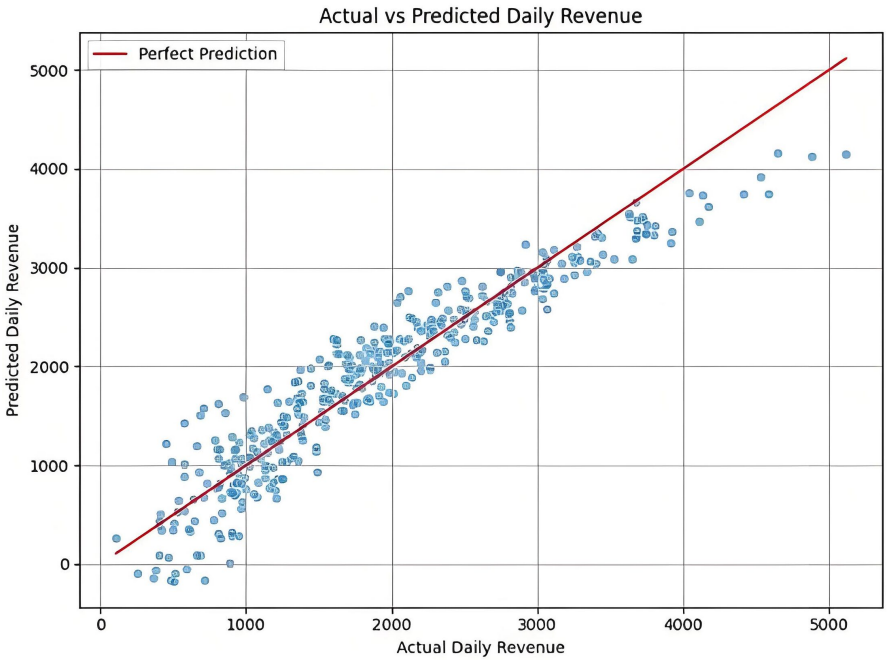


Figure 9. Contrast between actual and predicted revenue.

The accuracy variance of the model is shown by the degree of dispersion of the dots around the red line. Prediction mistakes may be higher when the spread is broader. In summary, a regression model—a model that forecasts continuous quantities like revenue—is frequently evaluated using this figure. It offers a rapid visual evaluation of how well the model's predictions match reality.

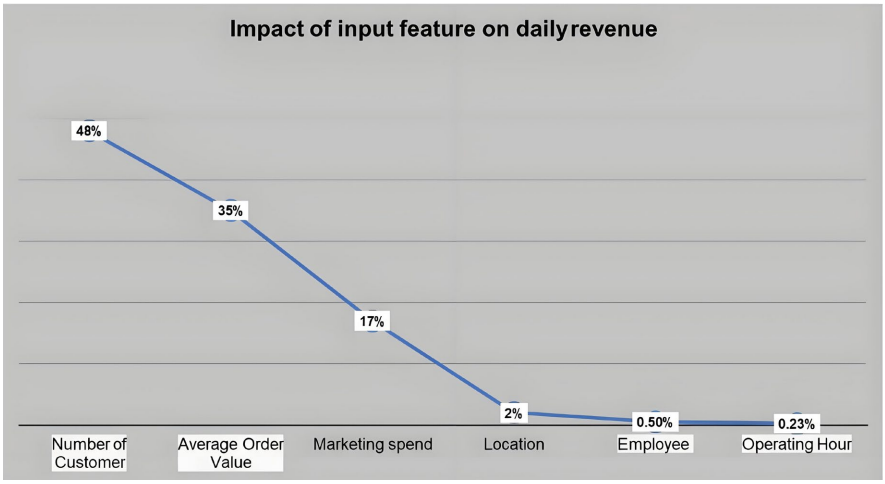


Figure 10. Contrast between actual and predicted revenue.

5. Discussion

The objective of this study was to develop a prediction model that could anticipate desired outcomes like demand or revenue using structured data and machine learning techniques, especially linear regression. It gave interpretability, practical application, and model correctness top priority throughout the process. Box plots were utilized to examine feature distributions and find outliers, skewed variables, and underlying trends throughout the data exploration stage. Important details on the behavior of particular traits and their possible impact on income were uncovered by this visual analysis. By ranking the most important predictors using feature relevance analysis, we were able to simplify the model and enhance performance. It is clear from Figure 10 that the most important indicators for predicting revenue are the number of customers per day, average order value, and daily marketing spend. Secondary elements like location foot traffic, the number of employees and operating hours come after this.

With a high R^2 value of 0.895, the chosen linear regression model explained over 89.5% of the variation in daily income. A strong linear association between the chosen characteristics and the target variable is shown by this high explanatory power. The accuracy of the model's predictions is further supported by the Mean Absolute Error (MAE) of 244.12. From a commercial perspective, these results provide small coffee shop owners with useful, data-driven insights that they can implement. Marketing efforts should prioritize customer acquisition strategies, such as local advertising, social media campaigns, or partnerships with foot-traffic-heavy locations, over initiatives that aim to increase average order value alone, for example, since the number of daily customers had the highest impact on revenue.

Additionally, even while passive location measures show a smaller association, investing in foot traffic visibility enhancements (such as outside signs or marketing) may still pay off. These tactics provide business stakeholders with precise, doable advice and immediately mirror the model's conclusions.

In a broader sense, this approach facilitates revenue forecasting, demand analysis, and strategic resource allocation. It encourages data-driven decision-making and reduces risks, which is useful not just in the retail industry but also in the restaurant business and even in commodity-based industries like agriculture. Stakeholders may now more accurately forecast changes in income and make proactive adjustments to operations or budgetary allocations, for instance.

Overall, the study shows that it is both possible and efficient to model and forecast important business indicators using supervised learning approaches, particularly linear regression. The robustness and predictive power of the model might be further enhanced for further research by including time-series components, experimenting with ensemble learning techniques, advanced machine learning, or utilizing deep learning models.

6. Conclusion and Future Plan

This study uses correlation-based feature selection and linear regression to explore how well a data-driven method can predict coffee shop income. This analysis identified the key predictors of daily revenue by examining operational factors such as average order value, marketing expenses, number of employees, and daily customer volume. A substantial amount of the revenue variance can be explained by the linear regression model that was created using these highly correlated characteristics, as seen by its high R^2 score of 0.89. Further confirming the correctness and dependability of the model were evaluation measures such as MAE. A substantial amount of the income variance can be explained by the linear regression model that was created using these highly correlated characteristics, as seen by its high R^2 score of 0.89. Assessment indicators such as RMSE and MAE provided additional confirmation of the model's accuracy and dependability. The results emphasize how crucial order value optimization, customer flow management, and focused marketing campaigns are to increasing revenue. By providing business leaders with practical information, this method helps them make better strategic decisions.

The model was trained and evaluated on a single public dataset, which is a limitation of the study, even if it performs well on the dataset that is currently accessible. As a result, it might not be as applicable in other retail settings. To guarantee wider application, future studies should test the model utilizing private data from various retail establishments and coffee shops. The model's generalizability is still constrained because the dataset does not reflect multi-location data from coffee shops. Future research should test this model using real or private data from various business kinds and geographical areas.

Future research can improve prediction power by incorporating seasonal trends, consumer segmentation, and external contextual factors like weather, holidays, and seasonal patterns are not taken into consideration in the present dataset. Such characteristics should be taken into account in future research as they have the potential to greatly increase model accuracy, even if the model works well with

the information provided. Overall, this work offers a solid basis for using straightforward but powerful machine learning models to anticipate retail sales.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

Ahmad, I. S., Bakar, A. A., Yaakub, M. R., & Muhammad, S. H. (2020a). A Survey on Machine Learning Techniques in Movie Revenue Prediction. *SN Computer Science, 1,* Article No. 235. https://doi.org/10.1007/s42979-020-00249-1

Ahmad, I. S., Bakar, A. A., & Yaakub, M. R. (2020b). Movie Revenue Prediction Based on Purchase Intention Mining Using Youtube Trailer Reviews. *Information Processing & Management, 57,* Article ID: 102278. https://doi.org/10.1016/j.ipm.2020.102278

Chang, X., & Yang, Y. (2016). Semisupervised Feature Analysis by Mining Correlations among Multiple Tasks. *IEEE Transactions on Neural Networks and Learning Systems, 28,* 2294-2305. https://doi.org/10.1109/tnnls.2016.2582746

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The Coefficient of Determination R-Squared Is More Informative than SMAPE, MAE, MAPE, MSE and RMSE in Regression Analysis Evaluation. *PeerJ Computer Science, 7,* e623. https://doi.org/10.7717/peerj-cs.623

Fiori, A. M., & Foroni, I. (2020). Prediction Accuracy for Reservation-Based Forecasting Methods Applied in Revenue Management. *International Journal of Hospitality Management, 84,* Article ID: 102332. https://doi.org/10.1016/j.ijhm.2019.102332

Fofanah, A. J. (2021). Machine Learning Model Approaches for Price Prediction in Coffee Market Using Linear Regression, XGB, and LSTM Techniques. *International Journal of Scientific Research in Science and Technology, 8,* 10-48.

Golderzahi, V., & Pao, H. K. (2024). Revenue Forecasting in Smart Retail Based on Customer Clustering Analysis. *Internet of Things, 27,* Article ID: 101286. https://doi.org/10.1016/j.iot.2024.101286

Hope, T. M. H. (2020). Linear Regression. In A. Mechelli, & S. Vieira (Eds.), *Machine Learning* (pp. 67-81). Elsevier. https://doi.org/10.1016/b978-0-12-815739-8.00004-3

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). Linear Regression. In G. James, D. Witten, T. Hastie, R. Tibshirani, & J. Taylor (Eds.), *An Introduction to Statistical Learning* (pp. 69-134). Springer International Publishing. https://doi.org/10.1007/978-3-031-38747-0_3

Jian, Z., Qingyuan, Z., & Liying, T. (2020). Market Revenue Prediction and Error Analysis of Products Based on Fuzzy Logic and Artificial Intelligence Algorithms. *Journal of Ambient Intelligence and Humanized Computing, 11,* 4011-4018. https://doi.org/10.1007/s12652-019-01650-2

Kumari, K., & Yadav, S. (2018). Linear Regression Analysis Study. *Journal of the Practice of Cardiovascular Sciences, 4,* 33-36. https://doi.org/10.4103/jpcs.jpcs_8_18

Lin, Y., Li, B., Moiser, T. M., Griffel, L. M., Mahalik, M. R., Kwon, J. et al. (2022). Revenue Prediction for Integrated Renewable Energy and Energy Storage System Using Machine Learning Techniques. *Journal of Energy Storage, 50,* Article ID: 104123. https://doi.org/10.1016/j.est.2022.104123

Liu, Y., Zhou, Y., Wen, S., & Tang, C. (2014). A Strategy on Selecting Performance Metrics for Classifier Evaluation. *International Journal of Mobile Computing and Multimedia*

Communications, 6, 20-35. https://doi.org/10.4018/ijmcmc.2014100102

Mahajan, P. D., Maurya, A., Megahed, A., Elwany, A., Strong, R., & Blomberg, J. (2020). Optimizing Predictive Precision in Imbalanced Datasets for Actionable Revenue Change Prediction. *European Journal of Operational Research, 285,* 1095-1113. https://doi.org/10.1016/j.ejor.2020.02.036

Maulud, D., & Abdulazeez, A. M. (2020). A Review on Linear Regression Comprehensive in Machine Learning. *Journal of Applied Science and Technology Trends, 1,* 140-147. https://doi.org/10.38094/jastt1457

Munoz, A., & Vassilvitskii, S. (2017). *Revenue Optimization with Approximate Bid Predictions.* arXiv: 1706.04732.

Padilla, R., Netto, S. L., & da Silva, E. A. B. (2020). A Survey on Performance Metrics for Object-Detection Algorithms. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)* (pp. 237-242). IEEE. https://doi.org/10.1109/iwssip48289.2020.9145130

Sadiku, M., Shadare, A. E., Musa, S. M., Akujuobi, C. M., & Perry, R. (2016). Data Visualization. *International Journal of Engineering Research and Advanced Technology (IJERAT), 2,* 11-16.

Sanjana Rao, G. P., Aditya Shastry, K., Sathyashree, S. R., & Sahu, S. (2021). Machine Learning Based Restaurant Revenue Prediction. In V. Suma, N. Bouhmala, & H. Wang (Eds.), *Evolutionary Computing and Mobile Sustainable Networks* (pp. 363-371). Springer Singapore. https://doi.org/10.1007/978-981-15-5258-8_35

Shao, L., Mahajan, A., Schreck, T., & Lehmann, D. J. (2017). Interactive Regression Lens for Exploring Scatter Plots. *Computer Graphics Forum, 36,* 157-166. https://doi.org/10.1111/cgf.13176

Wang, Y., Han, F., Zhu, L., Deussen, O., & Chen, B. (2017). Line Graph or Scatter Plot? Automatic Selection of Methods for Visualizing Trends in Time Series. *IEEE Transactions on Visualization and Computer Graphics, 24,* 1141-1154. https://doi.org/10.1109/tvcg.2017.2653106

Waskom, M. (2021). Seaborn: Statistical Data Visualization. *Journal of Open Source Software, 6,* Article 3021. https://doi.org/10.21105/joss.03021

Weng, C. (2017). Revenue Prediction by Mining Frequent Itemsets with Customer Analysis. *Engineering Applications of Artificial Intelligence, 63,* 85-97. https://doi.org/10.1016/j.engappai.2017.04.020