The Impact of Marketing Spend, Promotions and Other Factors on the Restaurant Revenue

This proposal was submitted to the Department of Mathematics as a partial fulfillment of the Bachelor of Science (Honors) Degree in Financial Mathematics and Industrial Statistics

University of Ruhuna.

By

M.H.M.N.Perera - SC/2021/12511

(Group No.16)

Supervisor :

Ms. R.M.V. Lakmini

Department of Mathematics
University of Ruhuna
Matara

Declaration

I hereby declare that this dissertation titled "The Impact of Marketing Spend, Promotions, and Other Factors on Restaurant Revenue" is my work and, to the best of my knowledge and belief, contains no material previously published or written by another person, nor material that, to a substantial extent, has been accepted for the award of any other degree or diploma at any university or equivalent institution, except where otherwise indicated through proper citation.

Furthermore, this dissertation was conducted under the supervision of
Ms. R.M.V. Lakmini, Lecturer (Probationary), Department of Mathematics, Faculty of Science, University of Ruhuna.

....
M.H.M.N. Perera
(SC_2021_12511)
Faculty of Science
University of Ruhuna
Matara.

....../....../2024

Supervisor:

....
Ms. R.M.V. Lakmini
Lecturer (Probationary)
Department of Mathematics
Faculty of Science
University of Ruhuna
Matara.

....../....../2024

Course Coordinator:

....
Dr. A.W.L. Pubudu Thilan
Senior Lecturer
Department of Mathematics
Faculty of Science
University of Ruhuna
Matara.

....../....../2024

Acknowledgment

At the very beginning of this report, I would like to express my deep and sincere gratitude to our supervisor, Ms. R.M.V. Lakmini, Lecturer (Probationary), Department of Mathematics, Faculty of Science, University of Ruhuna. I am grateful for her consistent direction and invaluable assistance during this study and dissertation, without which I would not have been able to complete my work. Ms. R.M.V. Lakmini has provided me with essential guidance on the approach to conducting this research. Working under her leadership and guidance was a tremendous privilege, an honor, and a joy.

Additionally, I would like to express my deep gratitude to Dr. A.W.L. Pubudu Thilan, Senior Lecturer, Department of Mathematics, University of Ruhuna, for his invaluable guidance on this case study.

I extend my heartfelt gratitude to all my senior members of FMIS. My appreciation goes to all of my friends who directly or indirectly helped me finish this study, all the authors of the references I utilized, and all who have supported the success of my research project, especially to my two team members, G.K.G.C.S. Bandara and A.K.D.A. Dulanjana. Thank you to everyone who encouraged me throughout this journey, even in the smallest ways!

Abstract

In today's competitive restaurant industry, managers are always looking for new ways to grow their customer base and maximize their net profit. Multiple linear regression analysis is a statistical technique for estimating the relationship between variables. Main focus of multiple linear regression in analysis the relationship between a dependent variable and two or more independent variables. It is used in this study to seek at the effects of various important independent factors on restaurant revenue. In this study, the data-set under analysis occurs from an open data hub, and its primary goal is to determine what are the variables effect on revenue. Techniques for quantitative analysis were used to comprehend the connections between variables. The R software has done All calculations and visualizations using tidyverse, ggplot2, lessR, olsrr, and lmtest packages. Based on preliminary findings, just three of the original set's (number of customers, menu price, and marketing spend) characteristics significantly affect restaurant monthly revenue. Furthermore, the investigation comes to the conclusion that promotions don't inspire consumers, and as a result, they don't raise sales. The dataset's limitations were noted, with the limitation that it might not precisely reflect data from the real world.

Keywords: Multiple linear regression, Restaurant revenue, Quantitative analysis techniques

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Overview

First, in the structured introduction chapter, the background of the study is discussed. Second, the research question and research objectives are explained. Finally, the significance of the study under the heading provides the uniqueness and value of this report.

1.1 Background of the Study

The restaurant industry thrives on a constant battle for customer attention and loyalty. In this highly competitive market, attracting new diners and retaining existing ones are fundamental to success. Traditionally, restaurants have relied on various marketing tactics to achieve these goals. However, the effectiveness of these strategies in influencing revenue can be unclear.

This research delves into the combined impact of marketing spend, promotions, and other relevant factors on restaurant revenue. We move beyond a simple focus on total marketing expenditure to explore how strategic investment and targeted promotions interact to influence customer acquisition and retention. We aim to understand if increasing marketing and promotion spending translates into a net gain of new customers while maintaining current customer loyalty. We will analyze a datasets encompassing restaurant performance metrics to answer this question.

By providing data-driven insights into the efficacy of marketing strategies for customer acquisition and retention, this research offers valuable information for restaurant businesses to optimize their marketing efforts and ultimately drive revenue growth.

1.2 Research Question

This research delves into this critical question: **How do marketing spend, promotions, and other relevant factors ultimately impact restaurant revenue?** After factorizing this question, it can be divided into following parts:

- How far does market spending affect restaurant revenue?
- Based on the promotion given to the customer, is there any effect on revenue?
- What could be the most important factor for revenue in a restaurant?

To answer this, we'll embark on an investigative journey, analyzing a comprehensive dataset of restaurant performance metrics. By delving into this data, we aim to illuminate the intricate relationship between marketing expenditure, promotions, and other factors of restaurant revenue.

1.3 Research Objective

By presenting data-driven insights into the effectiveness of marketing techniques for both customer acquisition and retention, this research aspires to empower restaurant businesses. The main objective of this research is **to analyze and identify what makes the most impact on restaurant revenue.**

1.4 Significance of the Study

The restaurant field is never-ending and becomes busier with the growing population. So as an entrepreneur or businessman, it is very advantageous to get an idea of how he/she should spend money on advertisements for their restaurant to make a greater profit. Based on advertisements, they maintain their name and reputation. This research study seeks to bridge this knowledge gap by investigating the interplay between marketing expenditures, promotion, customer acquisition, and retention of revenue.

1.5 Overview of Methodology

The study uses multiple linear regression to analyze the relationship between several factors and restaurant revenue. The dependent variable in this study is the restaurant's monthly revenue; independent variables include menu pricing, number of customers, marketing expenses, promotions, cuisine type, and average customer spending. For this analysis, data is taken from external records. Preliminary steps involve data cleaning to address inconsistencies, exploratory data analysis to visualize relationships, and checking regression assumptions. The regression model is then specified, fitted using statistical software, and validated for accuracy. Next, the regression model is defined, fitted with statistical tools, and its correctness is confirmed. The objective of this study is to get insight into how to optimize restaurant management practices by analyzing how these variables affect monthly revenue.

1.6 Structure of the Report

The format of this report is as follows: The overview of the research's background, research question, objective, and significance is provided in the first chapter, the introduction. A literature review summarizes the variables, prior studies, body of knowledge on this topic, and highlights any gaps in the knowledge that need to be addressed by the current research. The third chapter, "Data Preparation and Analysis," contains information about the metadata and data set. The materials and methodology, which also highlight the use of multiple linear regression for analysis and specify the variables, data sources, and initial steps in data preparation, go into great length about the theories. The data analysis and outcomes section provides the results of the regression analysis.The conclusion summarizes key findings and their practical implications for restaurant management. Finally, the report includes a reference list and, if necessary, appendices with additional supporting materials.

Chapter 2

Literature Review

In this section related works are examined that studies on multiple linear regression, assumptions, research applications, general relationships between factors affect on revenue of a restaurant, challenges, and limitations of multiple linear regression.

Multiple linear regression is a very common statistical technique used in finding the determinants of restaurant revenue, for example (Mohit Tyagi and Nomesh B. Bolia 2020) [11] and (Mun, Sung , Jang, and Soocheong 2018) [5]. The analysis of multiple linear regression often produced low coefficient of multiple determination, or R2 values and the presence of outliers is seen to be a very common problem. Bevans's article "Multiple linear regression (2023)" state that to ensure the validity of the findings, several important assumptions must be satisfied while doing a multiple linear regression analysis. These assumptions include the residuals' normality, homoscedasticity, linearity, and independence [2], [3]. Step wise selection, which makes use of both forward and backward selection methods, is frequently used to determine the most effective predictive model. Furthermore, by helping verify the assumptions, the use of diagnostic plots strengthens the regression model's adaptability [7].

The article Sung Gyun Mun and SooCheong published was "Restaurant Operating Expenses and Their Effects on Profitability Enhancement, (2018)". In their report have been mentioned, that restaurant firms need efficient cost management strategies due to highly competitive market conditions and the weak financial structure of the restaurant industry. Factors that were considered ton chronic industry-wide challenges were restaurant firms' low operating profitability, lack of financial flexibility, and highly competitive market environment. Additionally, they suggested that restaurant managers must identify which operating expenses should pay more attention to improve profitability [5]. It is quite significant and pertinent to restaurant businesses as well other firms.

A company's marketing expenditures are the money set aside for advertising and other marketing communication activities, including press conferences, experiential marketing events, digital and mobile marketing, and sales promotions. Adding value to their goods is the aim of marketing. Creating and sustaining strong brand associations in consumers' hearts and minds in this fiercely competitive business environment calls for a market-oriented management approach that integrates all business units within an organization and paves the way for attracting and retaining engaged loyal customers, which is concluded by the conceptual paper who written, C.M. Sachi, "Customer Engagement, Buyer-Seller Relationships, and Social Media, (2012)" [8].

Mohit Tyagi and Nomesh B. Bolia's research paper "Approaches for Restaurant Revenue Management, (2020)" have been provided information about the strategic levels of restaurant revenue management (RRM). They have identified three strategic levels for implementing RRM: capacity management, price management, and duration management. Especially they mentioned that Price management involves setting the right prices for all menu items to gain the maximum revenue. Prices may differ by day part, day of the week, or even from day to day if new menus are printed daily. It has also been found that the pricing policy in restaurants can be used to manage customer demand. They concluded that the traditional restaurant revenue management processes have inspired the restaurant business, which has an excellent chance for successful results. More complex approaches are needed, nevertheless, due to special characteristics including limited service capacity and changeable physical limits. Restaurants must combine customer value generation with revenue management strategies [11].

There are many restrictions when using multiple linear regression analysis. The resource book "Introduction to Linear Regression Analysis" by Douglas C. Montgomery has comprehensive information on The assumption of linearity is a major drawback as it could not apply in real-world situations where there may be non-linear correlations between variables. Furthermore, MLR makes the assumptions that the residuals, or errors, have a constant variance and are homoscedastic, or regularly distributed; deviations from these presumptions can result in skewed estimates and inaccurate conclusions. Multicollinearity is another drawback, since strong correlations between independent variables can skew standard errors and make it challenging to evaluate the relative contributions of each predictor. Moreover, MLR is susceptible to outliers and significant data points, which might have a disproportionate impact on the parameters and forecasts of the model [4].

Chapter 3

Data Preparation and Analysis

Overview

In this chapter, expect to provide quantitative and qualitative details of this data set. This chapter is structured as follows: data dictionary, variable type description, metadata, and data preparation and analysis. All necessary R scripts are attached in the appendix.

3.1 Data Source

The Restaurant Revenue Prediction Dataset is a comprehensive collection of simulated data designed to predict monthly revenue for fictitious restaurants.

Metadata

Source: It is an open dataset available in `www.kaggle.com`
Collaborators: MrSimple (Owner)
Authors: MrSimple07
Date: Jan 2024
Provenance: The restaurant revenue prediction dataset is a synthetic dataset created for educational and illustrative purposes. It does not originate from real-world data sources; any resemblance to actual entities or establishments is purely coincidental. The dataset was generated using random data generation techniques to simulate various aspects of restaurant operations.

3.2 Data-set Description

There are 1000 observations and 8 variables in this data set. Refer to the appendix for the CSV data file.

Variables Description

- Number of customers: The count of customers visiting the restaurant.
- Menu Price: Average menu prices at the restaurant. Currency unit: $
- Marketing Spend: Expenditure on marketing activities. Currency unit: $
- Cuisine Type: The type of cuisine offered (Italian, Mexican, Japanese, American).
- Average Customer Spending: Average spending per customer. Currency unit: $
- Promotions: Binary indicator (0 and 1 indicate, no and yes respectively) denoting whether promotions were conducted.
- Reviews: Number of reviews received by the restaurant.
- Monthly Revenue: Simulated monthly revenue, the target variable for prediction. Currency unit: $

Variable name	Type	Description
Number of customers	integer (continuous)	Independent
Marketing Spend	number (continuous)	Independent
Reviews	integer (continuous)	Independent
Promotion	string (categorical) (yes=1,no=0)	Independent
Menu price	number (continuous)	Independent
Cuisine type	string (categorical) (Japanese=1,Italian=2,American=3,Mexican=4)	Independent
Monthly revenue	number (continuous)	Dependent

Table 3.1: List of variables

3.3 Conceptual Model

With this figure, we can understand the mapping between dependent and independent variables.

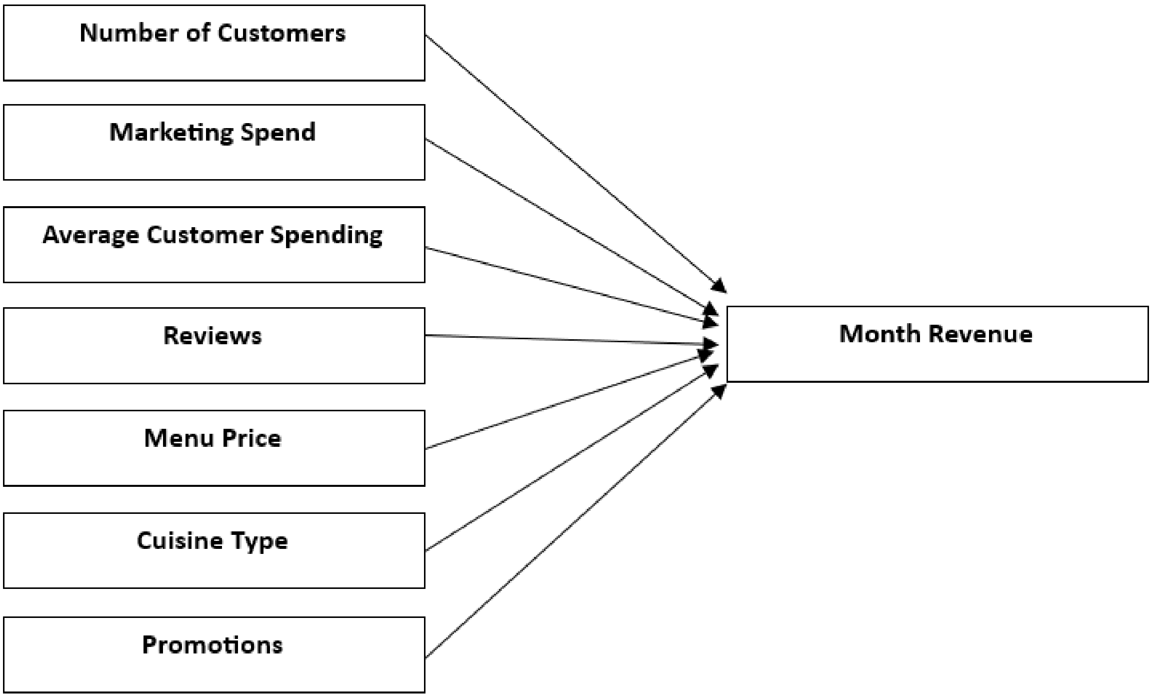


Figure 3.1: Conceptual model

3.4 Data Cleaning and Preprocessing

Check normality of dependent variable

This data set has a maximum of 1000 observations. Before moving forward with calculations, it is necessary to check whether multiple linear regression applies to this data set. Here, it used Shapiro-Wilk to check the normality of the dependent variable.

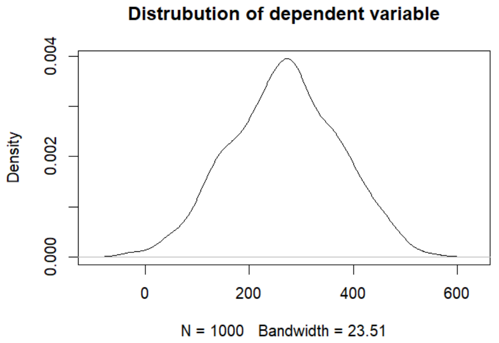


Figure 3.2: Density plot for revenue

According to this plot, we can conclude that the distribution of the dependent variable has an approximately normal distribution. After performing the Shapiro-Wilk test, the result was $W = 0.99756$ and p-value $= 0.1427$. Since the p-value is greater than 0.05, the dependent variable does hold normality. Therefore, the multiple linear regression method applies to this data set.

Missing values

After checking the missing values, I found that the missing values do not hold in this data set. Therefore, does not necessary to deal with missing value handling using techniques like imputation or the random forest approach.

Outliers

Outliers in statistics refer to data points that significantly deviate from the other observations in a data set [1]. These exceptional observations diverge from the typical values or patterns that the majority of the data exhibit.

Impact of outliers: Outliers can distort statistical analysis and data modeling results. Incorrect interpretations may arise if outliers are not properly handled.

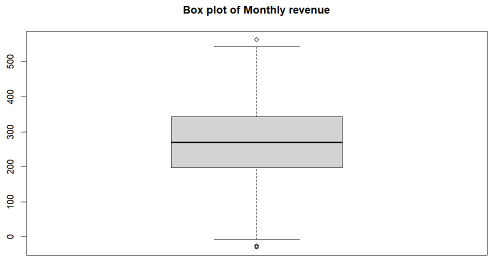


Figure 3.3: Boxplot for revenue with outliers

According to this box plot, there are outliers. Therefore, it should be handled. Here, the technique used is the IQR technique. In this, we remove observed values that are beyond the interquartile range.

Chapter 4

Methods and Methodology

Overview

In this chapter, the applied mathematical theory is discussed. The theoretical part of the upcoming analysis can be easily understood by studying this chapter. The systematic ways related to this study have been discussed in the research approaches section, and the mathematical module used here is described in detail in the research design section.

4.1 Variables

There are eight parameters with one thousand observations. Because of the normality of the parameters, there was no need for any transformation methods like log transformation in this study.

4.2 Model Specification

Multiple linear regression is the statistical technique used. In statistical modeling, multiple linear regression analysis is a set of statistical processes for estimating the relationships between a dependent variable and one or more independent variables [12].It improves an understanding of how adjustments to the independent variables affect the desired result by researchers. The key benefits of using multiple linear regression analysis are that it can :

1. Indicate if independent variables have a significant relationship with a dependent variable.
2. Indicate the relative strength of different independent variables' effects on a dependent variable.
3. Make predictions [4].

Multiple Linear Regression

Multiple linear regression is used to estimate the relationship between two or more independent variables and one dependent variable (especially when the dependent variable is continuous and normally distributed)[2]. The multiple linear regression analysis model is formulated as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon$$

- y represents the dependent variable.
- β_0 is the y-intercept, i.e., the value of y when all other parameters are set to 0.
- $\beta_1, \beta_2, \ldots, \beta_k$ are the coefficients for the independent variables $x_1, x_2, \ldots, x_k$.
- ϵ represents the error term. The errors are assumed to have mean zero and unknown variance σ^2. Additionally, we usually assume that the errors are uncorrelated

It can be written in matrix form within a number of observations as:

$$\text{Where: } y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

In order to accurately estimate the regression coefficients with the use of ordinary least squares (OLS), five different assumptions must be fulfilled. Violation of these assumptions may yield estimators that will differ significantly if applied to different data sets.

1. The relationship between the response variable y and the regressors x are approximately linear.
2. The error term ϵ has a mean of zero: $E(\epsilon) = 0$
3. The error term ϵ has constant variance: $V(\epsilon) = \sigma^2$
4. The errors are uncorrelated.
5. The errors are normally distributed.

Ordinary Least Squares

For estimating the regression coefficients the method of ordinary least squares will be used. The goal is to calculate the vector of least squares estimates $\hat{\beta}$ by minimizing the sum of squares of residuals SS_{Res}:

$$S(\beta) = \sum_{i=1}^{n} \epsilon_i^2 = \epsilon'\epsilon = (y - X\beta)'(y - X\beta)$$

The minimized sum of squares is obtained by deriving and setting equal to zero:

$$\left.\frac{\partial S}{\partial \beta}\right|_{\hat{\beta}} = -2X'y + 2X'X\hat{\beta} = 0$$

Which is simplified to the least-squares normal equations:

$$X'X\hat{\beta} = X'y$$

Multiplying by the inverse of $(X'X)^{-1}$ gives the least-squares estimator of β:

$$\hat{\beta} = (X'X)^{-1}X'y$$

Provided that the inverse exists, i.e. the regressors are linearly independent (assumption 1). The fitted regression model corresponding to the observed values are:

$$\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y$$

The OLS estimator $\hat{\beta}$ is then the best linear unbiased estimator.
When collecting data, there are several important factors to consider.

- Identify variables and type clearly.
- Check the distribution of the dependent variable.
  By using the Shapiro-Wilk test (less than 5000 observations) or the Anderson-Darlin test(for large datasets), check if the p-value of the dependent variable is greater than 0.05. If it is, multiple linear regression applies to that dataset. If the original dependent variable does not hold normality, several approaches can be made such as log transformation, square root transformation, cubic root transformation, and box-cox transformation[11].
- Check missing values and handle
  There are strategies to address missing data and minimize its impact on analysis [10]. Here are two main approaches: Deletion and imputation.
- Handling outliers
  The IQR method for removing outliers is given by:

  $$\text{Lower Bound} = Q_1 - 1.5 \cdot \text{IQR}$$

  $$\text{Upper Bound} = Q_3 + 1.5 \cdot \text{IQR}$$

  Where:

  $$\begin{aligned} &Q_1 \text{ is the first quartile } (25^{th} \text{ percentile}) \\ &Q_3 \text{ is the third quartile } (75^{th} \text{ percentile}) \\ \text{IQR} &= Q_3 - Q_1 \end{aligned}$$

Check Assumptions

The validity of results and assertions in multiple regression studies is questionable due to the uncertainty of whether the statistical tests' assumptions were met, with some assumptions being "robust" to violation and others fulfilling in the study's proper design [6].

Multivariate Normality - Examining Q-Q plots is one way to find out if the residuals' assumption of normality is met. QQ plots, as these plots are called, are useful for determining if the residuals follow a normal distribution. The normalcy assumption holds when the plot points form a straight diagonal line.

Linear Relationship - This method suggests that there is a linear connection between each predictor variable and the response variable. To confirm this, a scatter plot can be generated to show the relationship between each predictor variable and the response variable. If the points on the scatter plot closely follow a straight diagonal line, it indicates a linear relationship between the variables.

No Multicollinearity - Multiple linear regression assumes that there is no significant correlation among the predictor variables. In cases where one or more predictor variables exhibit high correlation, the regression model is affected by multicollinearity, leading to unreliable coefficient estimates. One can calculate the VIF (Variance Inflation Factor) value for each predictor variable to assess whether this assumption is met. Additionally, the correlation matrix can be examined to determine the presence of high correlations. It can be measured using Karl Pearson's or Spearman's rank correlation coefficient. The following table provides a conventional approach to interpreting a correlation coefficient [9].

Absolute Magnitude of the Observed Correlation Coefficient	Interpretation
0.00 - 0.10	Negligible correlation
0.10 - 0.39	Weak correlation
0.40 - 0.69	Moderate correlation
0.70 - 0.89	Strong correlation
0.90 - 1.00	Very strong correlation

Homoscedasticity - The assumption known as homoscedasticity states that there is uniform variance in a regression model's residuals, or errors, at every level of the independent variables. Put more simply, regardless of the projected values, the residuals' "scatter" or spread should be fairly uniform. Plotting the residuals on the y-axis versus the fitted values (or predicted values) on the x-axis allows one to evaluate homoscedasticity. Homoscedasticity is indicated if the residuals show a random distribution with a constant spread around zero.

4.3 Model Fitting

R software was used for data processing, cleaning, model fitting, model analysis and prediction. It is one of the most user-friendly statistical analysis programs, and it will be used to examine the data. These are some of the R's features. Robust data wrangling, broad statistical modeling capabilities, and machine learning algorithm support. For the best model selection, we used three methods: backward elimination, forward elimination, and the best subset of the model criteria.

Forward Elimination

In the forward selection procedure, variables are added to the model step-by-step. It starts with a simple regression model containing only one predictor and evaluates whether additional variables should be included.

- Initial Step: Begin with the simple regression model containing the predictor variable that has the highest correlation with the response variable.
- Subsequent Steps: At each step, compute the F-ratios for each variable not already in the model. Add the predictor variable with the smallest p-value, provided it is smaller than a pre-specified significance level α.
- Iteration: Continue adding variables until no remaining variable produces a significant p-value. A larger α is typically used to allow more variables into the model initially.
- Termination: The process stops when no additional variables meet the chosen significance level.

Traditionally, forward selection was preferred for its simplicity in computation, but it may not always yield the best model.

Backward Elimination

Backward elimination starts with the full model containing all predictor variables and systematically removes the least significant variables.

- Initial Step: Begin with a regression model including all predictor variables.
- Subsequent Steps: Compute the partial F-ratios for each variable and remove the one with the largest p-value, provided it is not significant.
- Iteration: Continue removing variables one by one until all remaining variables have significant F-ratios.
- Recommendation: Use a fairly large α for entry into the model and a more traditional α for a variable to stay in the model.

4.4 Model Evaluation

The process of determining how well regression model generalizes to an independent data-set is known as model validation in multiple linear regression. To determine that the model's assumptions are satisfied and it is accurate and reliable, requires a number of procedures and measurements.

The key aspects of model validation are Coefficient of Determination, F-test, t-tests for Individual Coefficients, and Residual Analysis.

Coefficient of determination (R^2) indicates the proportion of fitted values and residual values and values of (R^2) range from 0 to 1. 0 means there is no relationship among dependent and independent variables. Higher values indicate a better fit of the model. It is defined as:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

The overall F-test and its p-value for the entire model, which tests whether all the predictors together significantly improve the model compared to a model with no predictors. The overall F-statistic in a regression model is

$$F = \frac{SSR/p}{SSE/(n-p-1)}$$

where :
SSR is the regression sum of squares.
SST is the total sum of squares.
SSE is the error(residual) sum of squares.
p is the number of predictors.
n is the number of observations.
and the **p-value** can be obtained using the pf() function in R.
If p-value is less than 0.05 significance level indicates that there is very strong evidence against the null hypothesis.

H$_0$: Non of predictions have a significant linear relationship with the response variable.
H$_1$: At least one of the predictors has a significant linear relationship with the response variable.

4.5 Diagnostics

Evaluating the validity of a regression model involves several diagnostic tests and plots. Here are the methods used: residual plots, Q-Q plots, multicollinearity tests (VIF), and the Shapiro-Wilk test. Using these diagnostic tools helps ensure that the assumptions of the regression model are met, thereby validating the model's reliability and accuracy.

Residual Plot

Residual plots can be used to check the assumptions of linearity, homoscedasticity (constant variance), and independence of residuals.

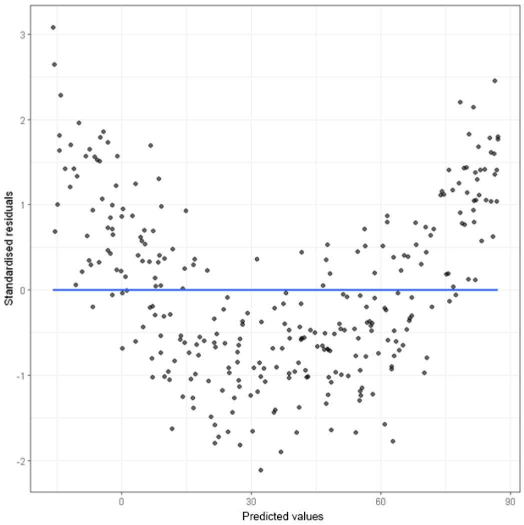


Figure 4.1: Residual plot with non-linearity

When the linearity assumption is violated, the points in the residual plot will not be randomly scattered. Instead, the points will often show some “curvature”. The residuals should have no apparent pattern when plotted against the fitted values or predictor variables.

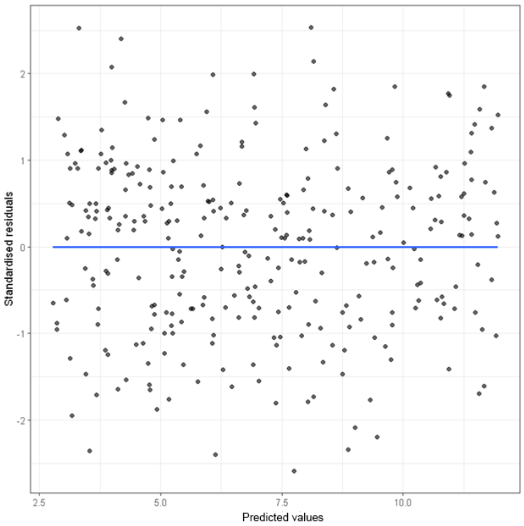


Figure 4.2: Residual plot

When both the assumption of linearity and homoscedasticity are met, the points in the residual plot (plotting standardised residuals against predicted values) will be randomly scattered. If model satisfied homoscedastisity, then he residuals should have a constant spread across the range of fitted values. A funnel shape indicates heteroscedasticity.

Q-Q Plot

By referring to Q-Q plots can be assess whether the residuals are normally distributed.

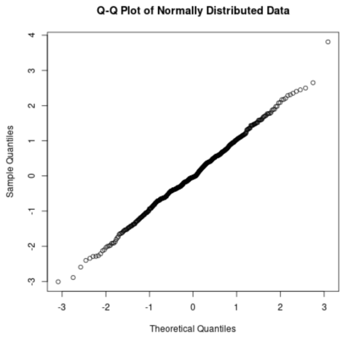


Figure 4.3: Q-Q Plot for Normal Data

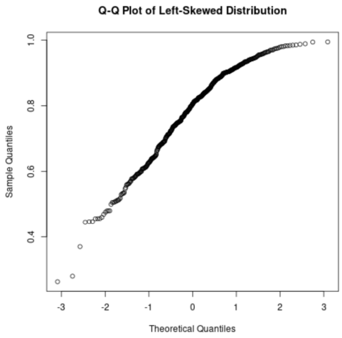


Figure 4.4: Q-Q Plot for Left-Skewed Data

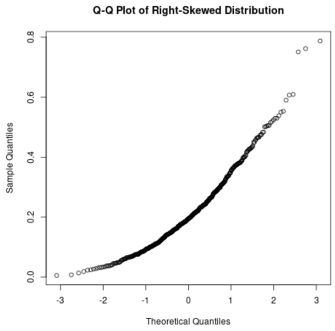


Figure 4.5: Q-Q Plot for Right-Skewed Data

If the residuals are normally distributed, the points should fall approximately along the reference line. Deviations from this line suggest departures from normality.

Variance Inflation Factor (VIF)

The variance inflation factor measures how much the variance of a regression coefficient is inflated due to multicollinearity. Using this, one can identify the presence of multicollinearity among predictor variables, which can affect the stability and interpretation of the coefficients.

- A value of 1 indicates there is no correlation between a given predictor variable and any other predictor variables in the model.
- A value between 1 and 5 indicates moderate correlation between a given predictor variable and other predictor variables in the model, but this is often not severe enough to require attention.
- A value greater than 5 indicates potentially severe correlation between a given predictor variable and other predictor variables in the model. In this case, the coefficient estimates and p-values in the regression output are likely unreliable.

4.6 Software and Tools

In this study, I utilized the following R packages:

- readr: For reading rectangular data, such as csv files.
- ggplot2: For creating advanced and customizable data visualizations.
- caTools: For data splitting and other tools useful in predictive modeling.
- corrplot: For visualizing correlation matrices and exploring relationships between variables.

- dplyr: For data manipulation, including filtering, selecting, and summarizing data.

- tidyverse: A collection of R packages designed for data science, which includes ggplot2, dplyr, and readr, among others.

Chapter 5

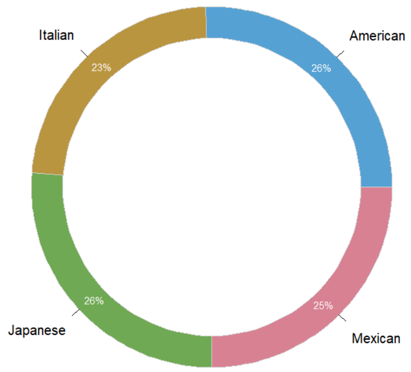
Results

5.1 Exploratory data analysis

	Num of cus.	Menu price	Mark. spend	Avg spending	Review	Revenue
Min	10.00	10.01	0.003768	10.04	0.00	-7.627
1st Qu.	30.50	20.48	4.708948	19.64	24.00	198.352
Median	54.00	30.87	10.148927	29.21	50.00	270.513
Mean	53.39	30.25	9.983004	29.48	49.88	269.620
3rd Qu.	74.00	39.89	14.993962	39.56	76.00	343.429
Max.	99.00	49.97	19.994276	49.90	99.00	542.467

Table 5.1: Five number summary

This table shows the five-number summary for numerical variables. With this identified distribution of reviews, it appears relatively symmetric. After referring to the mean and median of the number of customers, it indicates a consistent customer base. Because, on average customer attendance is stable around its central point. Additionally, the restaurant's menu price is moderately high, with most prices around \$30. (The price range varies from \$10.01 to \$49.97 with a mean of \$30.22 and a median of \$30.86). The mean value and median value of market spending are approximately equal to 10, and it indicates it varies greatly.



Approximately, food types of Italian Italian (23%), American (26%), Japanese (26%), and Mexican (25%) have equal demand in this restaurant.

Figure 5.1: chart of cuisine percentage

5.2 Model Summary

Fitting full model

Implementation of ordinary least squares to fit the full model, i.e. the model where all the regressor variables are used.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \epsilon$$

- y represents the monthly revenue.
- β_0 is the y-intercept, i.e., the value of y when all other parameters are set to 0.
- $\beta_1, \beta_2, \ldots, \beta_7$ are the coefficients for the independent variables $x_1, x_2, \ldots, x_7$.
- x_1 = Number of customers
  x_2 = Marketing spend
  x_3 = Average customer spend
  x_4 = Reviews
  x_5 = Menu price
  x_6 = Cuisine type
  x_7 = Promotions
- ϵ represents the error term.

Full model analysis

Multivariate Normality:
According to this QQ plot, most of the proportion of residuals align on the linear line. This histogram shows an approximate bell shape with moderate kurtosis. Based on that visual analysis, can depicts this QQ-plot and histogram of residuals indicate the normality of residuals of this model. Therefore assumption of normality is satisfied.

Multicollinearity:

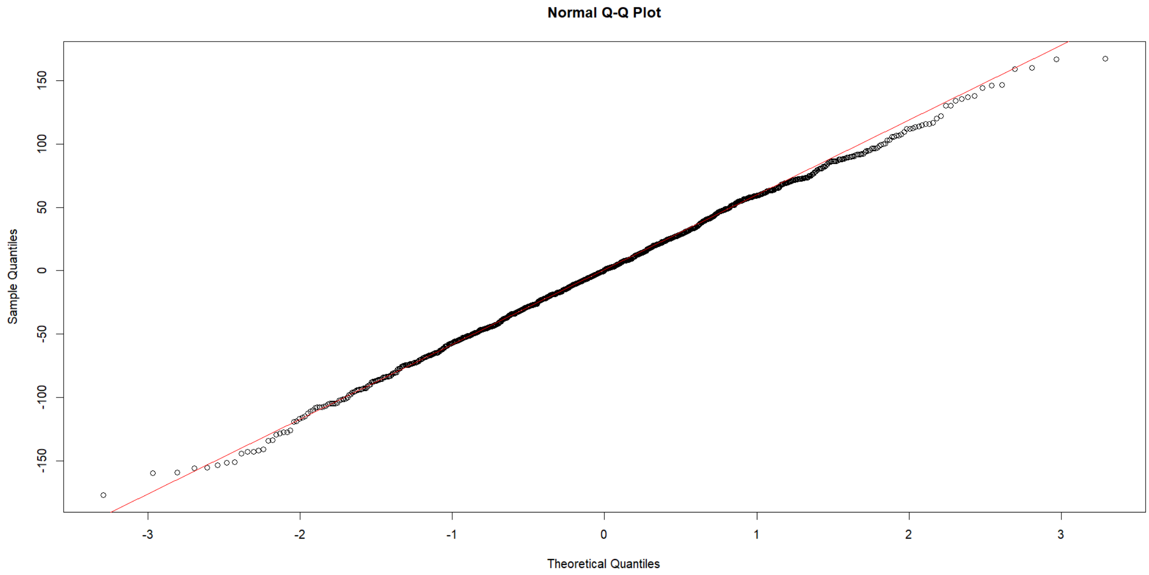


Figure 5.2: QQ plot of the model with all variables

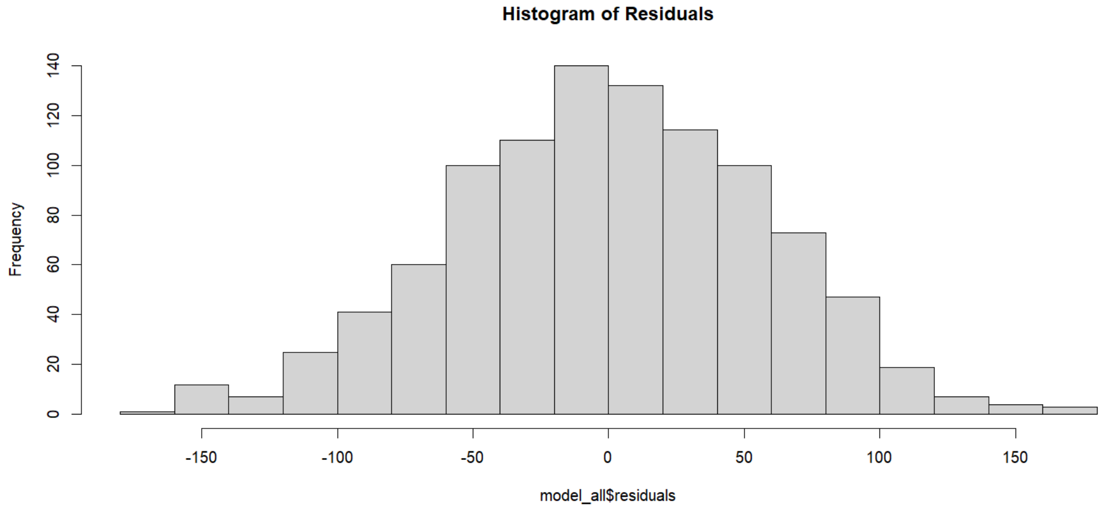


Figure 5.3: Histogram of the model with all variables

	Number of Customers	Menu Price	Marketing Spend	Average Customer Spending	Promotions	Reviews	Monthly Revenue	Cuisine Type num
Number of Customers	1.0000	0.0331	-0.0171	-0.0104	0.0671	-0.0098	0.7427	-0.0075
Menu Price	0.0331	1.0000	0.0155	0.0168	0.0226	0.0018	0.2644	0.0435
Marketing Spend	-0.0171	0.0155	1.0000	-0.0558	-0.0356	-0.0324	0.2580	-0.0512
Average Customer Spending	-0.0104	0.0168	-0.0558	1.0000	0.0042	0.0530	-0.0295	-0.0269
Promotions	0.0671	0.0226	-0.0356	0.0042	1.0000	-0.0220	0.0299	0.0333
Reviews	-0.0098	0.0018	-0.0324	0.0530	-0.0220	1.0000	-0.0248	-0.0546
Monthly Revenue	0.7427	0.2644	0.2580	-0.0295	0.0299	-0.0248	1.0000	0.0071
Cuisine Type num	-0.0075	0.0435	-0.0512	-0.0269	0.0333	-0.0546	0.0071	1.0000

With these results, it can be concluded that there is no or less correlation between each independent variable. Therefore this model satisfied the assumption of no multicollinearity.

Linear relationship:

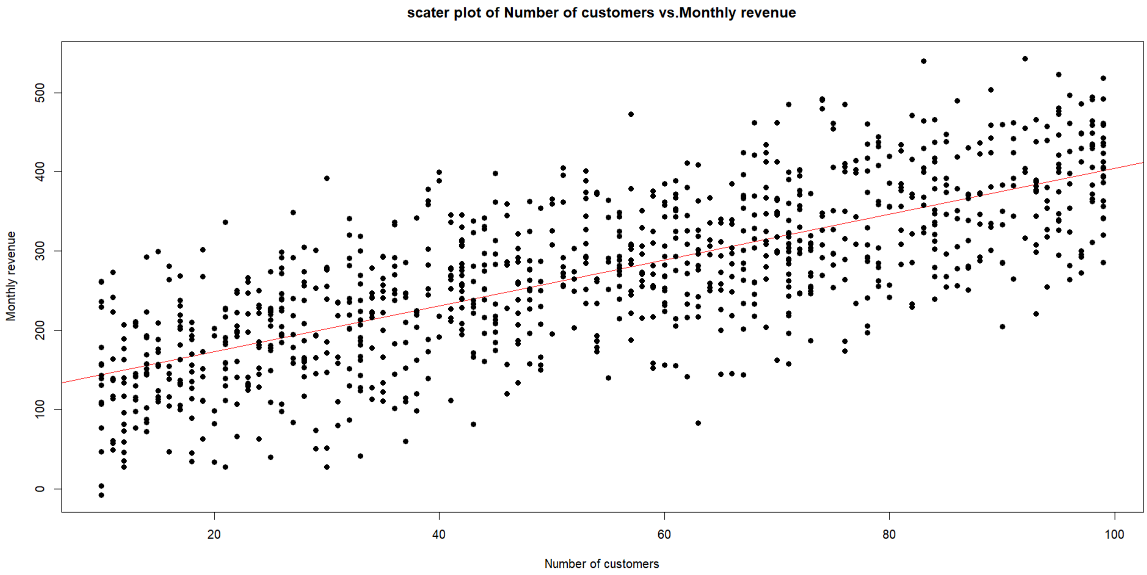


Figure 5.4: Scatter plot of Number of customers vs. Monthly revenue

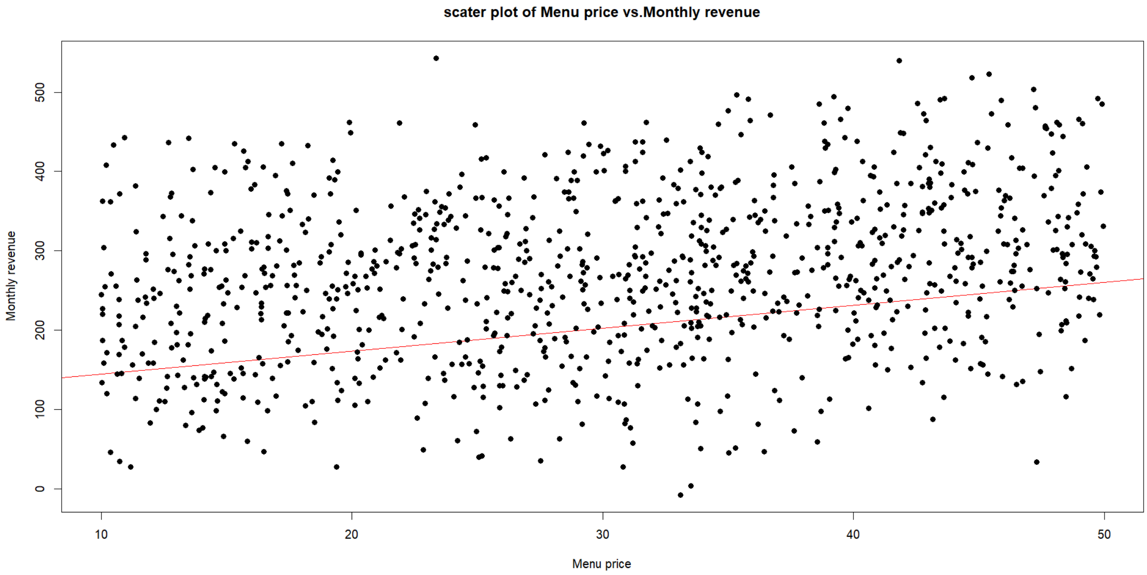


Figure 5.5: Scatter plot of Menu price vs. Monthly revenue

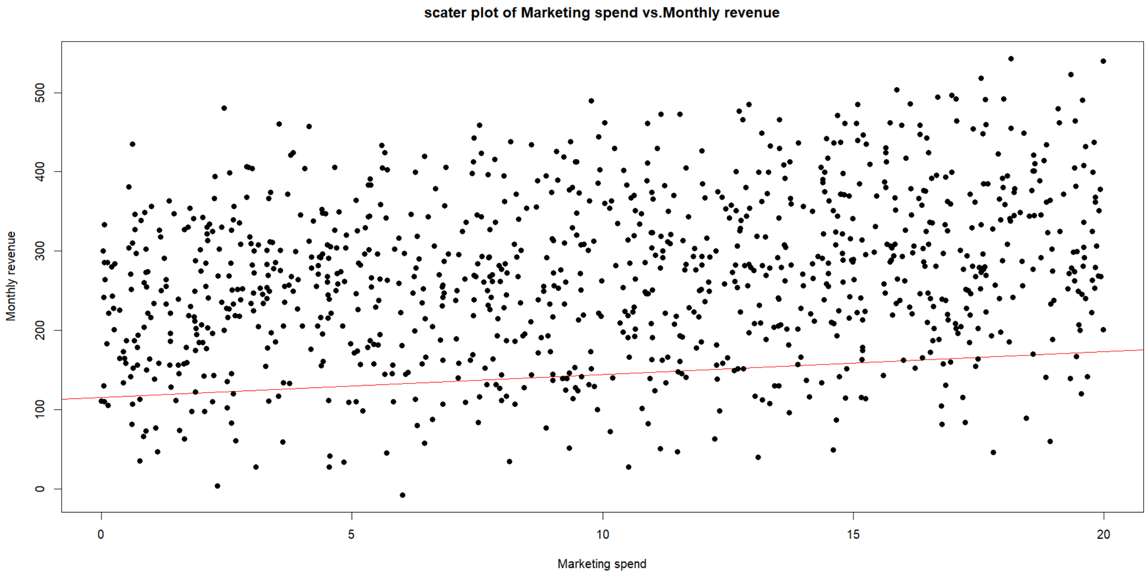


Figure 5.6: Scatter plot of Marketing spend vs. Monthly revenue

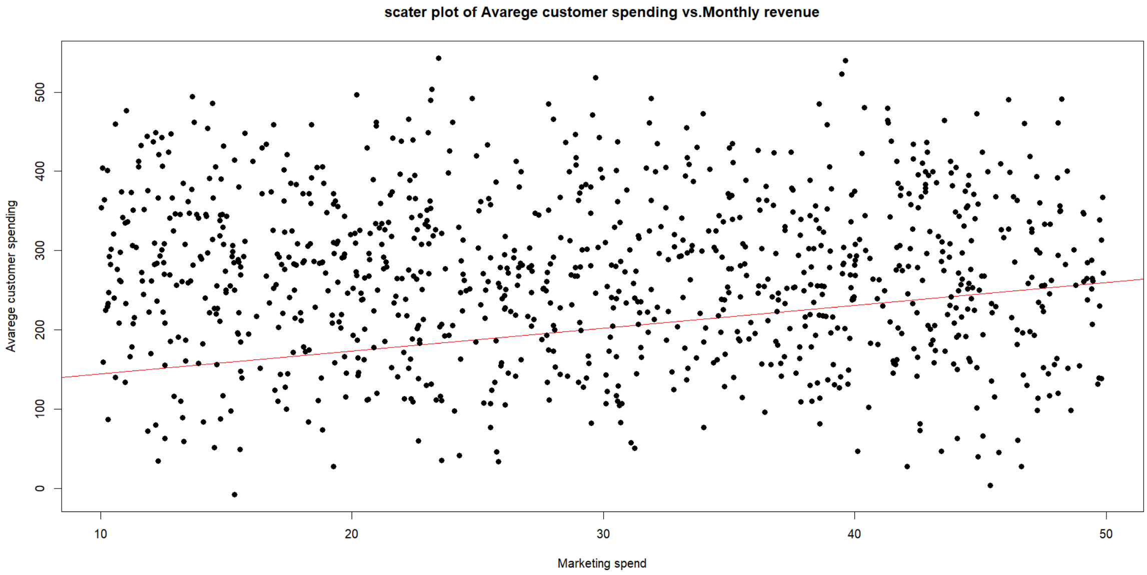


Figure 5.7: Scatter plot of Average customer spending vs. Monthly revenue

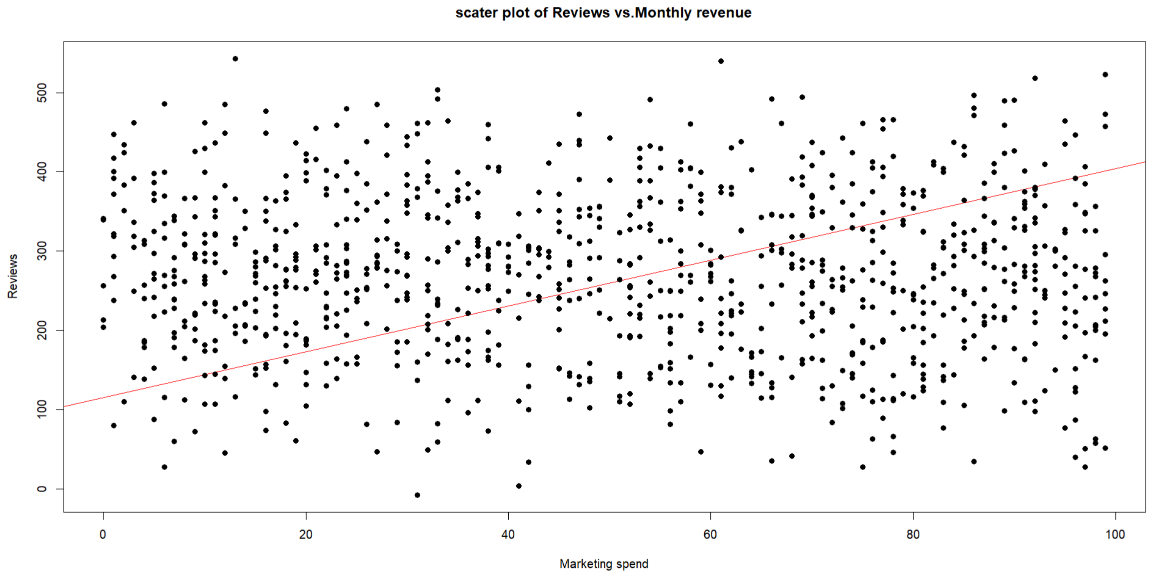


Figure 5.8: Scatter plot of Reviews vs. Monthly revenue

These scatter plots provide information about linear relationship between dependent variable and each is a continuous independent variables. The red lines indicate linear regression lines of relatively. After this visual analysis, it can be concluded that, there is no non-linear relationship among dependent and independent variables.

Homoscedasticity:

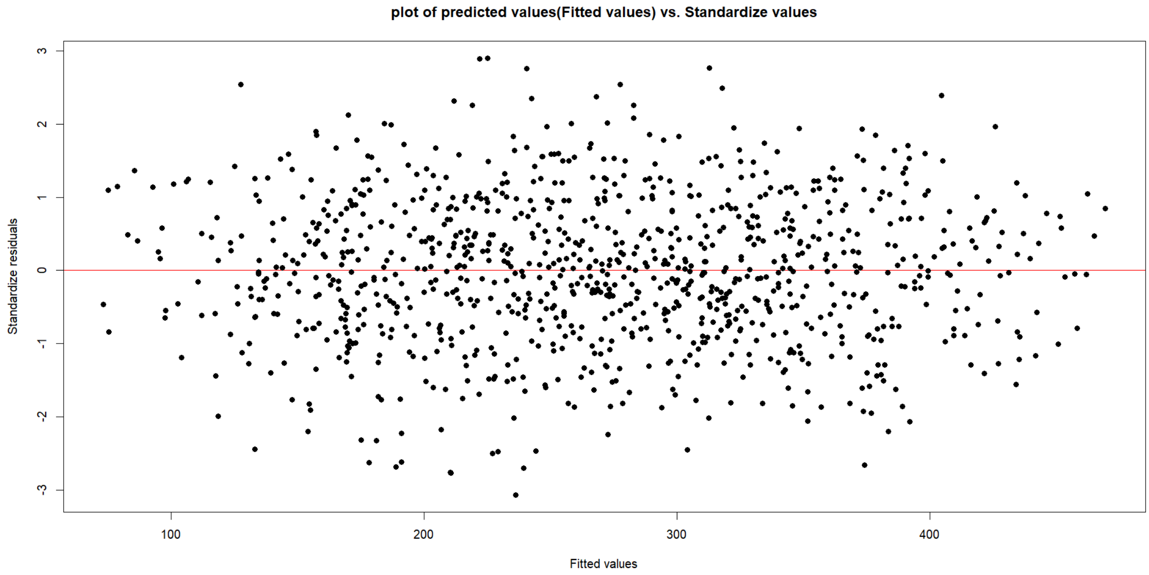


Figure 5.9: Plot of predicted values vs standardized values for the model with all variables

This plot indicates that residuals' "scatter" or spread is fairly uniform. That is the residuals show a random distribution with a constant spread around zero. therefore assumption of homoscedasticity is satisfied.

After considered all results of assumption checking, can depict multiple linear regression is applicable to this data set.

5.3 Regression Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.03106	10.20489	0.689	0.491
Number of Customers	2.87814	0.07007	41.077	<2e-16
Menu Price	2.13319	0.16304	13.084	<2e-16
Marketing Spend	4.66217	0.31588	14.759	<2e-16
Avg Customer Spending	-0.08890	0.16056	-0.554	0.580
Promotions	-3.32778	3.68227	-0.904	0.366
Reviews	-0.02906	0.06301	-0.461	0.645
Cuisine Type num	1.43080	1.63206	0.877	0.381

Table 5.2: Coefficients for the full model

Therefore least square regression line is :

$$y = 7.03 + 2.88x_1 + 2.13x_2 + 4.66x_3 - 0.09x_4 - 3.33x_5 - 0.03x_6 + 1.43x_7 + \epsilon$$

5.4 Best Model Fitiing

There are 1000 observations with 8 variables. If consider all variables for the model it will be less accurate and high cost. In this case, performed variable selection criteria to the approach to finding the active predictors consider all possible choices for variables, and then select the one that optimizes some selection criterion.

When considering all possible subsets from these variables, there exist 128 models. Because of the difficulty of handling a large number of models, here follows **forward elimination, backward elimination**, and **the best subset of the model criteria**, additionally after that selects the best model by using **cross-validation** for high accuracy.

Intercept	Number of customers	Marketing spend	Menu price
4.946074	2.874164	4.672280	2.134657

Table 5.3: Coefficient values from forward elimination

Table 5.3 indicates that, after performing forward elimination criteria for the selected most suitable regression model, the result is the dependent variable with these 3 predictors. As expected, after performing backward elimination criteria for the same scenario, the result was the same answer.

Therefore, the selected best reduced model was

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

Model Index	Predictors						
1	Number of Customers						
2	Number of Customers	Marketing Spend					
3	Number of Customers	Menu Price	Marketing Spend				
4	Number of Customers	Menu Price	Marketing Spend	Cuisine Type			
5	Number of Customers	Menu Price	Marketing Spend	Promotions	Cuisine Type		
6	Number of Customers	Menu Price	Marketing Spend	Avg Customer Spending	Promotions	Cuisine Type	
7	Number of Customers	Menu Price	Marketing Spend	Avg Customer Spending	Promotions	Reviews	Cuisine Type

Table 5.4: Best Subsets Regression

Model	R-Square	Adj. R-Square	MSEP
1	0.5517	0.5512	4655185.0965
2	0.6250	0.6242	3897877.8503
3	0.6805	0.6795	3324121.3367
4	0.6808	0.6795	3324794.7688
5	0.6810	0.6794	3325454.7133
6	0.6811	0.6792	3327700.7071
7	0.6812	0.6789	3330357.9964

Table 5.5: Subsets Regression Summary

Table 5.4 shows what are the best subsets of regression models after performing the best subset modeling and according to the results where shown in Table 5.5, the selected model is 3. Because it has a relatively high R-square and less MSEP (Estimated error of prediction, assuming multivariate normality). After getting the same model from each criterion, it is not necessary to perform cross-validation.

- $\beta_0 = 4.94607$ is the y-intercept
- $\beta_1 = 2.87416$ is the coefficient of number of customers
- $\beta_2 = 2.13466$ is the coefficient of menu price
- $\beta_3 = 4.67228$ is the coefficient of marketing spend
- y monthly revenue
- x_1 number of customers
- x_2 menu price
- x_3 marketing spend
- ϵ represents the error term.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.94607	7.07149	0.699	0.484
Number of Customers	2.87416	0.06984	41.156	<2e-16
Menu Price	2.13466	0.16266	13.123	<2e-16
Marketing Spend	4.67228	0.31428	14.867	<2e-16

Table 5.6: Coefficients for the reduced model

Furthermore, the partial F test was used to check the hypothesis.
$\mathbf{H}_0$: The reduced model is suitable

H$_1$: Full model is needed.

Test whether the data provides sufficient evidence to support the claim that the monthly revenue depends on the predictors of the reduced model by using a 0.05 significance level.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Number of Customers	1	5716619	5716619	1711.14	< 2.2e-16
Menu Price	1	596691	596691	178.61	< 2.2e-16
Marketing Spend	1	738386	738386	221.02	< 2.2e-16
Residuals	991	3310751	3341		

Table 5.7: ANOVA Table of reduced model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Number_of_Customers	1	5716619	5716619	1707.9457	<2e-16
Menu_Price	1	596691	596691	178.2724	<2e-16
Marketing_Spend	1	738386	738386	220.6064	<2e-16
Average_Customer_Spending	1	1236	1236	0.3694	0.5435
Promotions	1	2512	2512	0.7505	0.3865
Reviews	1	870	870	0.2598	0.6104
Cuisine_Type_num	1	2572	2572	0.7686	0.3809
Residuals	987	3303561	3347		

Table 5.8: ANOVA Table of full model

5.5 Statistical Significance

SSR$_{\text{full}}$ = 7058886, SSR$_{\text{reduced}}$ = 7051696, and MSE$_{\text{full}}$ = 3347. Therefore, the test statistic is 2.148193. The table value $F_{0.05,4,986} = 2.38095761$.
In case of $F_{\text{partial}} < F_{\text{table}}$, there is no sufficient evidence to reject null hypothesis (**H**$_0$) which means the reduced model is suitable at the 0.05 significance level.

According to the model summary, the multiple R^2 is 0.6805 and adjusted R^2 is 0.6795. which means 68.05% of dependent variable can be explained using this three of independent variables.

5.6 Predictor Importance

Let's analyze corresponding variable of x_i is linearly related to the y or not.
H$_0$**:** The β_i does not significant,
H$_0$**:** The β_i does significant.
Reject H_1 if p-value < α; (α=0.05).

With the result of p-value of the number of customer being 2e-16 (<0.05) which is shown in Table 5.7, we have enough evidence to reject null hypothesis. That means for every number of customer increase, monthly revenue increases on average by \$2.874, holding other variables constant.

With the result of p-value of the menu price being 2e-16 (<0.05) which is shown in Table 5.7, we have enough evidence to reject null hypothesis. That means for every unit of menu price, monthly revenue increases on average by \$2.135, holding other variables constant.

With the result of p-value of the marketing spend being 2e-16 (<0.05) which is shown in Table 5.7, we have enough evidence to reject null hypothesis. That means for every unit of marketing spend, monthly revenue increases on average by \$4.672, holding other variables constant.

After referring to the p-values of average customer spending, promotions, reviews, and cuisine type as 0.5435, 0.3865, 0.6104, and 0.3809, respectively, which are shown in Table 5.8, the evidence does not suffice to reject the null hypothesis.

Confidence interval and prediction interval of best model

Assume that the error term ϵ in the MLR model is independent of x_i (i = 1,2,...,n), the interval estimate for the mean of dependent variable is called the confidence interval. When the confidence interval reflects the uncertainty around the mean predictions, the prediction interval gives uncertainty around a single value.

Variable	2.5%	97.5%
(Intercept)	-8.9307	18.8229
Number of Customers	2.7371	3.0112
Menu Price	1.8155	2.4539
Marketing Spend	4.0556	5.2890

Table 5.9: 95% Confidence Intervals for Model Coefficients

Intercept: The interval [-8.93, 18.82] indicates a 95% confidence interval on the location of the true intercept.

Number of customers: The range [2.74, 3.01] shows our level of confidence that, on average, the outcome variable (predicted value) will vary by 2.74 to 3.01 units for each unit of rise in the number of customers.

Menu Price: The range [1.82, 2.45] indicates a 95% confidence interval in the location of the true menu price coefficient.

Marketing Spend: The range [4.06, 5.29] indicates a 95% confidence interval in the location of the actual marketing spend coefficient.

I used a small data set to predict the values according to the selected best model. All values are in the minimum and maximum ranges of each corresponding variable.
Based on the result of Table 5.9, all the predictor variables have positive intervals,

No. Customers	Menu price	Mark. spend	pred. values	lwr conf	upr conf	lwr pred	upr pred
15	11	3	85.55661	75.82502	95.2882	-28.28419	199.3974
25	15	7	141.52600	134.18777	148.8642	27.86478	255.1872
55	20	10	252.44106	247.56921	257.3129	138.91239	365.9697
60	25	12	286.82973	282.55009	291.1094	173.32493	400.3345
75	30	15	354.63232	349.00860	360.2560	241.06890	468.1957
80	40	16	395.02199	388.02000	402.0240	281.38198	508.6620

Table 5.10: Predictions with Confidence and Prediction Intervals

suggesting a positive association with monthly revenue, and based on Table 5.10, The prediction intervals are wider and depict the range within which individual future observations will probably fall, with 95% confidence, than the predicted values, which predict point estimates for the new data points based on the model. The wider intervals show the model's uncertainty.

Chapter 6

Discussion and Conclusion

6.1 Discussion

This study analyses whether or not seven independent variables in the standard model (number of customers, marketing spend, average customer spend, reviews, menu price, cuisine type, and promotions) were significantly predictive of the Monthly revenue, the dependent variable, based on the ANOVA statistics. However result was the number of customers, menu price, and marketing spend, that means reduced model have the most significant effect on monthly revenue. It is the answer for main research question which is mentioned in 1st chapter.

That model successfully passed all the tests in model validation steps, so we can conclude that our model can perform well to predict the monthly revenue of relevant restaurant by using the three independent variables. But still, our model only has R^2 score of 68.05%, which means that there is still about 31.95% unknown factors that are affecting dependent variable.

The average customer spending, promotions, reviews, and cuisine type have been failed to make an effect on monthly revenue of this restaurant. Therefore it would be better, if the owner consider about other factors for increase the revenue.

I depicted that increasing marketing spend is associated with a positive impact on monthly revenue. Also there exists enough evident to argue that based on the promotion given to the customer, there is very low effect on revenue and it is same as cuisine type.

The biggest issue is that this datasets does not represent practical world data. Because of that, we can't make the most accurate decisions based on this. There are more variables highly affecting revenue, such as location, market size, restaurant concept, operating hours, design, data management, staff handling, online sales strategies, etc.

6.2 Conclusion

Restaurants have the potential to incorporate revenue management practices into their operations but cannot simply apply the same revenue management strategies as those used by airlines and hotels. The unique business characteristics of restaurants, rather than these seven independent variables, such as a relatively fixed service capacity due to variable meal durations and elastic physical constraints, require restaurants to develop more sophisticated revenue management. Restaurants also need to educate their customers by providing information about the uniqueness of the restaurant.

The article "New performance indicators for restaurant revenue management (2022)" highlighted the importance for restaurants to sell the right menu in order to maximize profitability. Restaurant operators will increase their total gross profit and bottom line by selling more profitable menu items during high demand periods. The success of the RRM approach depends on the availability of historical data on demand patterns (customer arrival), sales of specific menu items and price. Therefore, it is important for restaurant operators to have reliable data available to them when they need it so they can analyze these factors correctly. The goal of RRM should be about selling the right menu item to the right customer at the right time (and meal duration, as well) for the right price by using the right table mix in order to maximize profit.

Although our approach offers insightful information, it might fail to consider interactions or non-linear effects among variables since it assumes linear relationships. Beyond the scope of our approach, unaccounted-for variables like macroeconomic conditions or consumer demographic information may have an impact on revenue uncertainty. To more precisely capture complicated interactions, future research options might involve experimenting with other modeling methodologies or adding new factors, such as consumer behavior data. By taking these factors into account, the model's resilience and forecast precision may be improved, leading to a better understanding of the factors that influence monthly revenue in our particular setting.

Bibliography

[1] H. Aguinis, R. K. Gottfredson, and H. Joo. Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*, 16(2):270–301, 2013.

[2] R. Bevans. Multiple linear regression | a quick guide (examples). *Scribbr*, June 2023.

[3] Z. Bobbitt. The five assumptions of multiple linear regression. *Statology*, Nov. 2021.

[4] D. C. Montgomery.

[5] S. Mun and S. Jang. Restaurant operating expenses and their effects on profitability enhancement. *International Journal of Hospitality Management*, 71:68–76, 04 2018.

[6] J. Osborne and E. Waters. Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research Evaluation*, 8, 01 2002.

[7] M. Sarstedt and E. Mooi. Regression analysis. pages 193–233, 03 2014.

[8] C. Sashi. Customer engagement, buyer-seller relationships, and social media. *Management Decision*, 50:253–272, 03 2012.

[9] P. Schober, C. Boer, and L. Schwarte. Correlation coefficients: Appropriate use and interpretation. *Anesthesia Analgesia*, 126:1, 02 2018.

[10] N. Tamboli. Effective strategies for handling missing values in data analysis (updated 2024). *Analytics Vidhya*, May 2024.

[11] M. Tyagi and N. Bolia. Approaches for restaurant revenue management. *Journal of Revenue and Pricing Management*, 21, 02 2022.

[12] Wikipedia contributors. Regression analysis — Wikipedia, the free encyclopedia. 2024. [Online; accessed 14-May-2024].

Appendix

To obtain access to CSV related to this analysis, **click here**.

R codes for this analysis.

```
library(tidyverse)
library(lessR)
library(olsrr)
library(lmtest)
library(ggplot2)

df <- read.csv("test.csv")
head(df$Cuisine_Type,50)
summary(df)

# assigning nnumerical values for categorical variable

Cuisine_Type_map <- c("Japanese"=1,"Italian"=2,"American"=3,"Mexican"=4)
df$Cuisine_Type_num <- Cuisine_Type_map[df$Cuisine_Type]

# removed charactor type observations after assign numerical values

df <- df %>% select(-Cuisine_Type)

# check for missing values

any(is.na(df))

# peform shapiro test

shapiro.test(df$Monthly_Revenue)
boxplot(df$Monthly_Revenue,main="Boxplot of dependent variable")
plot(density(df$Monthly_Revenue),main="Distrubution of dependent variable")

# outliers handling

Q1 <- quantile(df$Monthly_Revenue, 0.25)
Q3 <- quantile(df$Monthly_Revenue, 0.75)
IQR_val <- Q3 - Q1
lower_bound <- Q1 - 1.5 * IQR_val
```

```
upper_bound <- Q3 + 1.5 * IQR_val
df <- df[df$Monthly_Revenue >= lower_bound &
  df$Monthly_Revenue <= upper_bound, ]
boxplot(df$Monthly_Revenue,
  main="Box plot of Monthly revenue after handling outliers")

# draw a ring chart

cui <- data.frame(cu = df$Cuisine_Type)
PieChart(cu,hole=0.825,stat="%",data=cui,main = "")

# boxplots

par(mfrow=c(1,2))
boxplot(df$Number_of_Customers, main="Number of Customers")
boxplot(df$Menu_Price, main="Menu Price")

par(mfrow=c(1,2))
boxplot(df$Marketing_Spend, main="Marketing Spend")
boxplot(df$Average_Customer_Spending, main="AVG customer spending")

par(mfrow=c(1,1))
boxplot(df$Reviews,main="Reviews")


### model building

# full model
model_all <- lm(Monthly_Revenue ~., data = df)

# initial model
model_initial <- lm(Monthly_Revenue ~ 1, data = df)

# Forward selection

forward <- step(model_initial, direction = "forward",
   scope = formula(model_all), trace = 0)
forward$coefficients

# Backward selection

backward <- step(model_all, direction = "backward",
    scope = formula(model_all), trace = 0)
backward$coefficients

# best subset of model

ols_step_all_possible(model_all)
```

```
ols_step_best_subset(model_all)


# selected best model: Number_of_Customers Menu_Price Marketing_Spend

model <- lm(Monthly_Revenue ~ Number_of_Customers
   + Menu_Price+ Marketing_Spend,data=df)
summary(model)

### assumption checking

# Normality - model_all

qqnorm(model_all$residuals)
qqline(model_all$residuals,col="red")
hist(model_all$residuals, breaks = 20, main = "Histogram of Residuals")

# linearity - model all

mod1 <- lm(df$Monthly_Revenue ~ df$Number_of_Customers, data = df)
plot(df$Number_of_Customers,df$Monthly_Revenue,
ylab = "Monthly revenue",xlab="Number of customers",
main="scater plot of Number of customers vs.Monthly revenue",
pch=16,
abline(coef(mod1)[1],coef(mod1)[2],col="red"))

mod2 <- lm(df$Monthly_Revenue ~ df$Menu_Price, data = df)
plot(df$Menu_Price,df$Monthly_Revenue,
ylab = "Monthly revenue",xlab="Menu price",
main="scater plot of Menu price vs.Monthly revenue",
pch=16,
abline(coef(mod1)[1],coef(mod1)[2],col="red"))

mod3 <- lm(df$Monthly_Revenue ~ df$Marketing_Spend, data = df)
plot(df$Marketing_Spend,df$Monthly_Revenue,
ylab = "Monthly revenue",xlab="Marketing spend",
main="scater plot of Marketing spend vs.Monthly revenue",
pch=16,
abline(coef(mod1)[1],coef(mod1)[2],col="red"))

mod4 <- lm(df$Monthly_Revenue ~ df$Average_Customer_Spending, data = df)
plot(df$Average_Customer_Spending,df$Monthly_Revenue,
ylab = "Avarege customer spending",xlab="Marketing spend",
main="scater plot of Avarege customer spending vs.Monthly revenue",
pch=16,
abline(coef(mod1)[1],coef(mod1)[2],col="red"))

mod5 <- lm(df$Monthly_Revenue ~ df$Reviews, data = df)
```

```
plot(df$Reviews,df$Monthly_Revenue,
ylab = "Reviews",xlab="Marketing spend",
main="scater plot of Reviews vs.Monthly revenue",
pch=16,
abline(coef(mod1)[1],coef(mod1)[2],col="red"))

# Homoscedasticity

plot(model_all$fitted.values,model_all$residuals,
ylab = "residuals",xlab="fitted values",
main="scater plot of residuals vs. fitted values",
pch=16,
abline(h=0,col="red"))

# multicollinearity - model_all

correlation_matrix <- cor(df[,],method = "pearson")
round(correlation_matrix,4)

# Independence

dwtest(model_all)


# perform partial F test

Anova(model_all)
Anova(model)
confint(model)

### prediction

# New data frame for predictions

new_df <- data.frame(
Number_of_Customers = c(15, 25, 55, 60, 75, 80),
Menu_Price = c(11, 15, 20, 25, 30, 40),
Marketing_Spend = c(3, 7, 10, 12, 15, 16)
)

# Get predictions, confidence intervals, and prediction intervals

confidence_intervals <- predict(model, newdata = new_df,
      interval = "confidence")
prediction_intervals <- predict(model, newdata = new_df,
      interval = "prediction")

# Combine predictions into a data frame
```

```
pred_df <- data.frame(
Number_of_Customers = new_df$Number_of_Customers,
Menu_price = new_df$Menu_Price,
Marketing_spend = new_df$Marketing_Spend,
predicted_value = confidence_intervals[, "fit"],
lwr_conf = confidence_intervals[, "lwr"],
upr_conf = confidence_intervals[, "upr"],
lwr_pred = prediction_intervals[, "lwr"],
upr_pred = prediction_intervals[, "upr"]
)
```