

## **MA 541 - A**

## **Report On**

## **New York City Restaurants Data - Food Ordering and Delivery**

**Prof: Hong Do.**

**By:**

**Jash Shah (20020618)**

**Prerna Desai (20022778)**

**Sameer Shaik (20022166)**

**TABLE OF CONTENTS**

1. Introduction
2. Data Overview

2.1 Description of the data set

2.2 Sample mean and sample standard deviation for each random variable.

2.3 Sample correlations among each pair of random variables

1. Describe Your Data

3.1 Histograms for each column

3.2Time series plots for each column

3.3 Time series plot for all four columns

3.4 Scatter plots for relationships between columns

1. Distribution Analysis

4.1 Assumptions/hypotheses regarding the distribution of each column

4.2 Normality tests to verify assumptions.

4.3 Results and interpretations of the tests

1. Central Limit Theorem

` 5.1 Evaluation of consistency with the Central Limit Theorem

1. Confidence Interval Construction

6.1 Construction of confidence intervals for mean using random samples

6.2 Comparison of intervals with population mean.

1. Hypothesis Testing

7.1Testing hypotheses related to population mean and standard deviation

7.2 Interpretation of test results

1. Model Fitting

8.1 Scatter plot and correlation analysis

8.2 Regression analysis and interpretation of coefficients

1. Model Prediction

9.1 Evaluation of a multiple linear regression model

9.2 Calculation of adjusted R-squared

1. Residual Analysis and Model Selection

10.1 Evaluation of model assumptions using residual analysis

10.2 Discussion on improving model quality through model selection strategies.

1. Conclusion
2. **INTRODUCTION**

In the ever-growing field of the food service industry, with its state-of-the-art app which connects customers and a wide range of restaurants through digital platforms, the Food Hub emerges as a profound contributor to revolutionizing the food delivery system. This work consists of the statistical analysis to break down a massive dataset from the Food Hub which has kept the records of over 10 thousand registered customers’ orders. The dataset star pitch spans 1,899 entries and is aligned with the joint project. Nine areas in total with nine variables have been collected carefully to provide a full view of customer ordering behavior, preferences, and operational insights such as type of cuisine preferred, order costs, and delivery times. Although it has a structured appearance and has no missing values, the dataset faces some unique challenges such as the sensitivity of outliers, data encoding, and the removal of special characters to guarantee the integrity of the results.

The objective of this statistical project is twofold: uncover statistical relationships that are hidden within the dataset and then use the same inferences to enhance the company-level strategies of Food Hub to push customer satisfaction to the next level. This analysis process is armed with stepwise EDA, statistical testing, and predictive modeling techniques that are capable of getting to the crux of any demand pattern which if envisioned well, The planned result is advice, based on data and offering custom tailoring to enhance the customer experience and achieve operational efficiency that will bump Food Hub to a sound position in the food delivery service competition.

**2. DATA OVERVIEW:**

**2.1 Description of the Dataset:**

This dataset includes 9 columns:

1. **order\_id:**  
   This column represents a unique identification number assigned to the order.
2. **customer\_id**

This column represents a unique identification number assigned to every new customer.

1. **Restaurant\_name**

This represents which restaurant has received the order.

1. **Cusines\_type**

This column represents the type of cuisine the customer has ordered.

1. **Cost\_of\_order**

This column indicates the total cost of the order.

1. **Day\_of\_the \_week**

This column indicates the day of the week when the order was received.

1. **Rating**

This column displays the rating provided by the customer for the food.

1. **Food\_preparation\_time**

This column displays the time taken to prepare food by restaurants.

1. **Delivery\_time.**

This column represents the time the delivery person takes to deliver the order.

**2.2 Sample mean and sample standard deviation for each random variable**

A screenshot of a computer

Description automatically generated

Fig 1: Reading the dataset

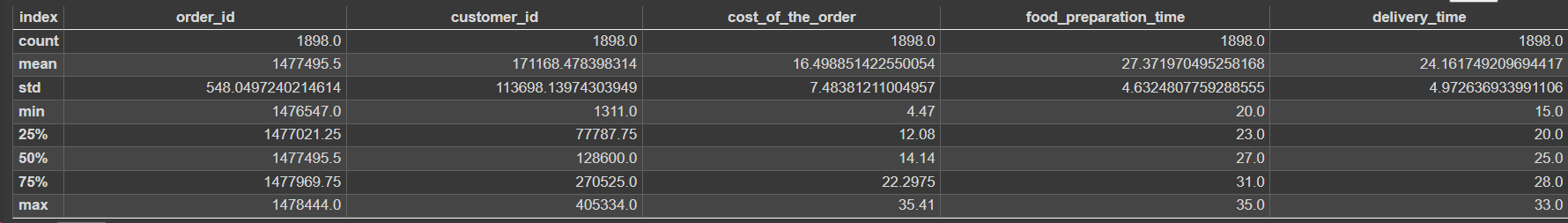


Fig 2:

**order\_id and customer\_id:** These columns represent identifiers for orders and customers respectively. They are categorical or nominal variables, and it's difficult to draw meaningful statistical insights from them. However, we can note that there were 1898 unique orders and customers in the dataset.

**cost\_of\_the\_order**: The mean cost of an order is approximately $16.50 with a standard deviation of $7.48. The costs range from $4.47 to $35.41. We can see that the costs vary significantly, indicating a wide range of orders with different prices.

**food\_preparation\_time and delivery\_time:** The mean preparation time is approximately 27.37 minutes, while the mean delivery time is approximately 24.16 minutes. Both preparation and delivery times have standard deviations of around 4.63 and 4.97 minutes respectively, indicating some variability in these times.

Overall, the analysis provided gives us an understanding of the central tendencies and variability in the order costs, food preparation times, and delivery times. Further, we calculated a correlation matrix to identify the relationship between the columns.

**2.3 Sample correlations among each pair of random variables**

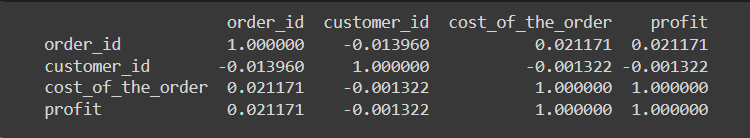


Fig 3: Correlation matrix

**order\_id and customer\_id:** These correlations are close to zero, indicating a very weak linear relationship between these variables. This makes sense since order IDs and customer IDs are typically just identifiers and don't have a meaningful numerical relationship.

**cost\_of\_the\_order and profit:** Both variables have a correlation coefficient of 1. This indicates a perfect positive correlation, meaning that as the cost of the order increases, the profit also increases in a linear fashion. This result might seem odd because it's unlikely that cost and profit are perfectly correlated in real-world scenarios. It's possible that there's some data issue or bias in your dataset causing this result, such as all orders having the same profit margin.

In summary, few correlations are present that are low, suggesting loosely linear associations among them. In-depth assessments, which can include for example looking at other relational forms, or including more variables, can give more info.

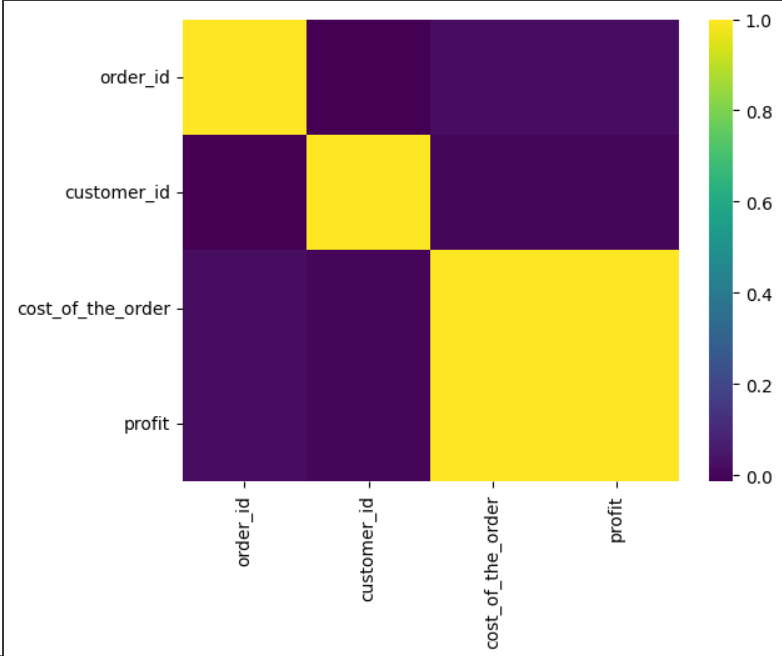
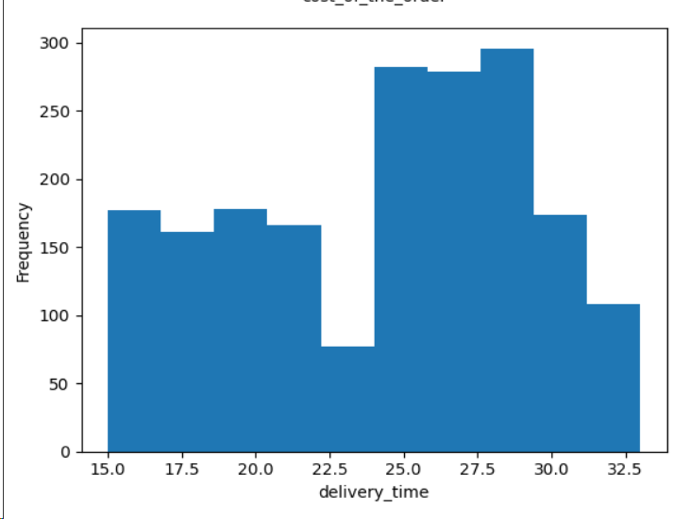
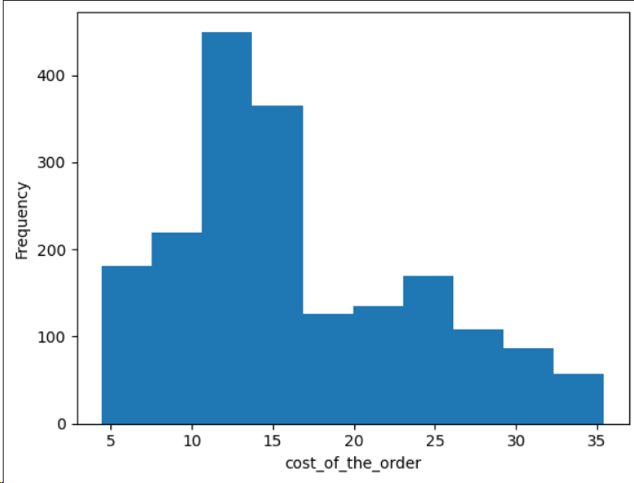


Fig 4: Heatmap correlation matrix

**3. DESCRIBE YOUR DATA**

**3.1 Histogram**



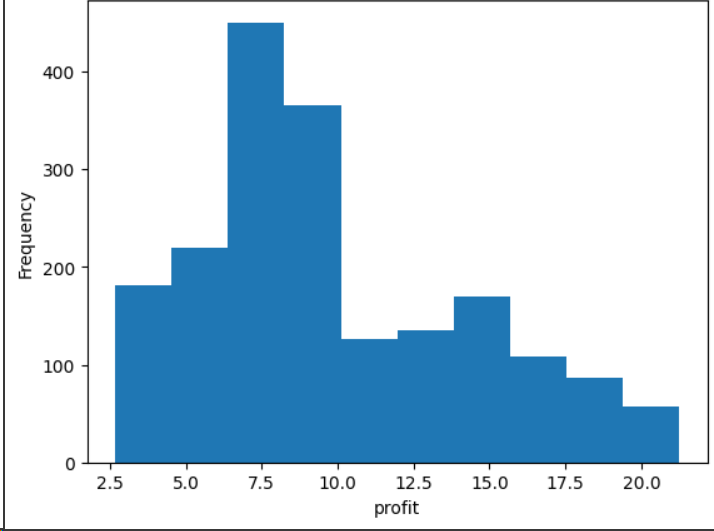
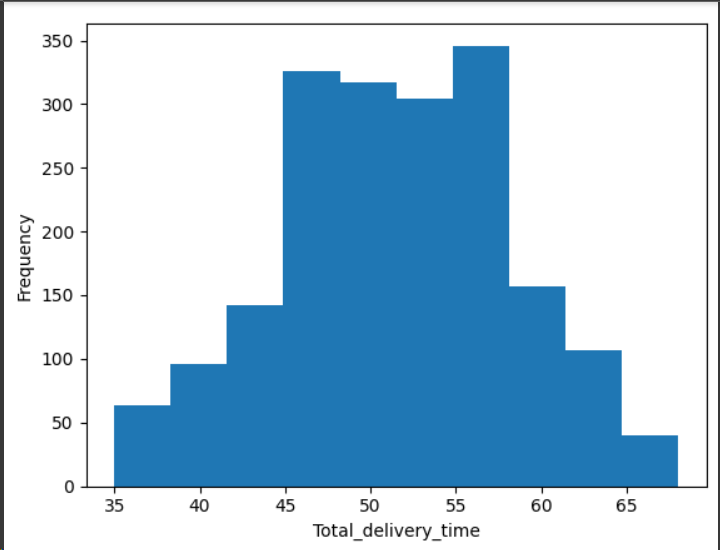


Fig 5: Histograms for all columns

In the 1st histogram, we can see that the range of prices for most goods differs between 10 and 20 units (dollars, for example), with the number of orders dropping as the cost becomes h-factor. This fact proves that there are many moderate prices and few high-priced orders.

The distribution in delivery time is even though the dataset slightly deviates from the median towards the higher values, indicating that delivery times can indeed differ, but they normally fall between 20 and 25 minutes. The total delivery time is representative of the normally distributed data, implying that the delivery times will be around 50 minutes.

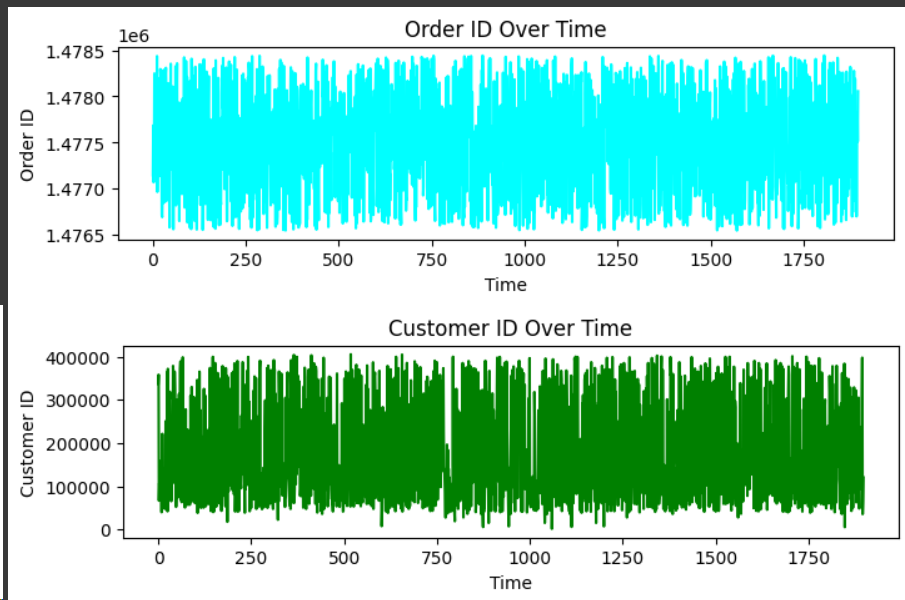
The profit of the restaurants varies mostly from over 200 to 400 units.

**3.2 Time series analysis:**

Order ID Over Time: This graph shows that order IDs increase with the period, which is because it has been regularly set up from the beginning. The high number of places may indicate the order of activity over time and there can be such peaks of the activity shown by the tightly packed bands.

Customer ID Over Time: It portrays customer IDs across the period and enables us to see how the range fluctuated at the time of their ordering. The new customers as well as the regular ones show different dynamics over time. So, the new type of customer and the returning one as well are present throughout time.  
  
Cost Over Time: Another plot involved showing the cost movement on the time scale. The spike energy shows that the fluctuation of the value is so wide that there may be multiple mountains and valleys. On the other hand, there is no certain pattern that could be seen other than a wide fluctuation in the price in the next year, which may mean the total cost per order does not show stable growth or decline over the timeframe portrayed.  
  
Day of the Week Over Time: The last plot shows that Monday and Friday respectively had the maximum number of books borrowed. The checkerboard pattern symbolizes that the column contains multi-week data, where the iron-gray segments represent working days, and the darker grey category probably embraces weekends. This method can be applied to investigate trends or performance metrics by the day of the week.

.



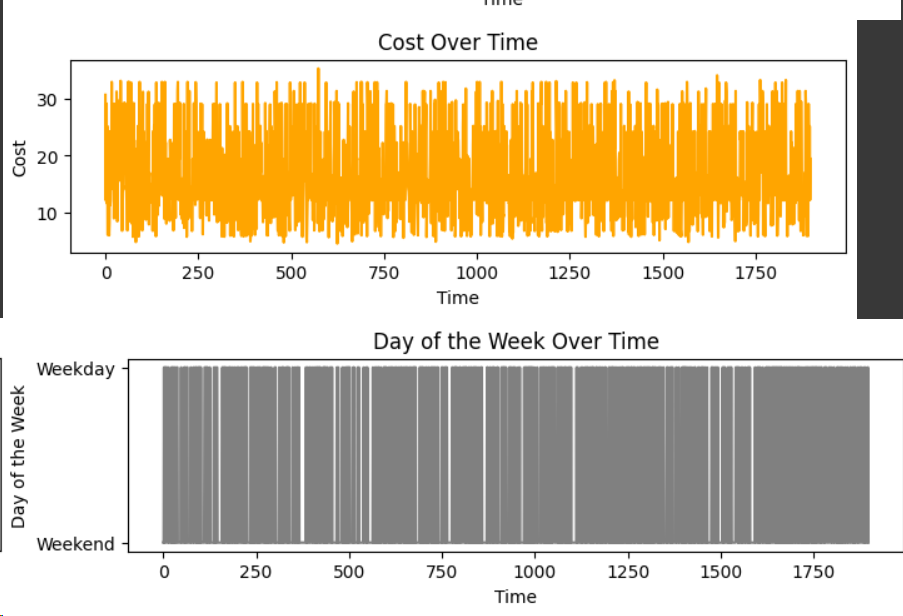


Fig 6: Time series analysis of columns

**3.3 Scatter plots**

This plot shows a random dispersion of points, suggesting that there is no clear pattern or correlation between the Order IDs and the Customer IDs. This is to be expected as Order IDs are usually sequentially assigned regardless of Customer ID.

The distribution of costs across Customer IDs is relatively uniform, indicating that all customers are likely to spend across the range of order costs. There's no visible trend indicating that certain customers consistently spend.

The plot suggests distinct groupings of food preparation times at specific cost intervals, indicating a possible relationship between the type of food ordered (which determines preparation time) and the cost. This could imply menu items with similar preparation times are priced in a similar range.

The QQ plot for rating shows data points lying on a straight line at the far right, indicating that higher ratings are common and fit well with a normal distribution. However, there are deviations from the line at the lower end, suggesting that lower ratings are less frequent or do not fit a normal distribution well.

This plot deviates from the theoretical normal distribution line at both tails, more significantly at the higher end, suggesting that higher costs are more extreme than would be expected in a normal distribution. This could indicate a long tail in pricing, where most orders are of moderate cost with some very high-cost outliers.

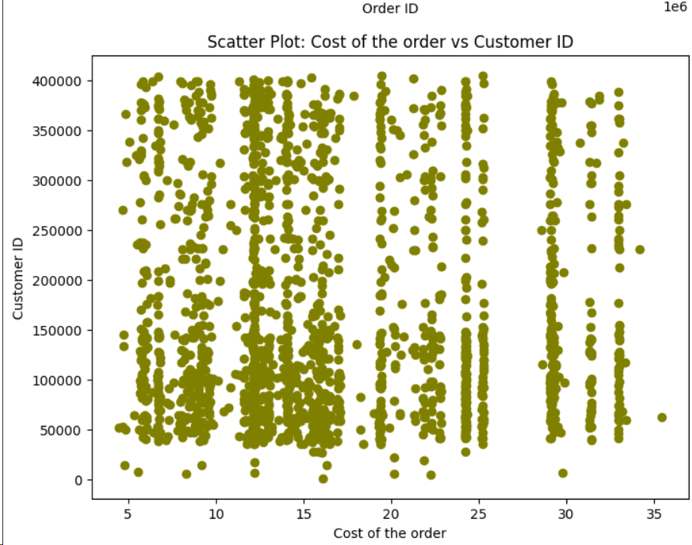
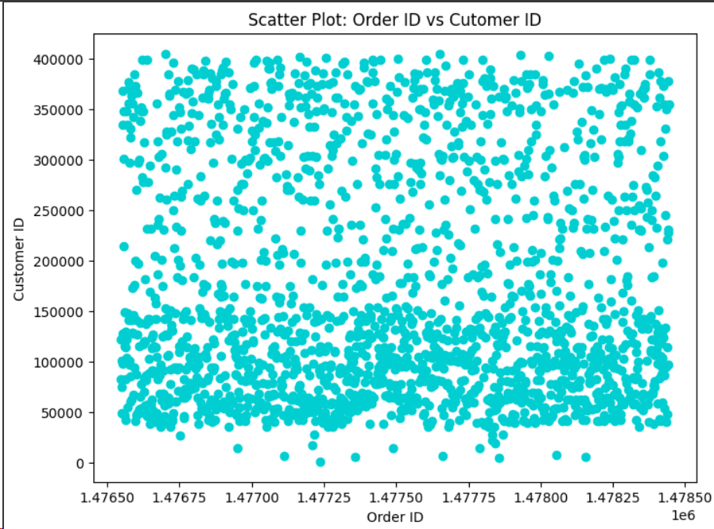
The food preparation time data aligns well with the theoretical line, suggesting that the food preparation times are normally distributed.

The relationship between order IDs and customer IDs is arbitrary, which is normal for transactional data. Order costs are spread across the customer base without any visible customer-specific spending patterns. There seems to be a standard range of preparation times that correspond to certain order costs, which could be indicative of standardized pricing strategies for different categories of food.

The rating distribution is skewed towards higher ratings, suggesting general customer satisfaction.

The cost of orders might be right-skewed, meaning that there are a few orders with a very high cost compared to the rest. Both food preparation and delivery times seem to be normally distributed, indicating consistency and efficiency in these operations.

These plots can provide the business with insights into its operations, customer behavior, and pricing strategies, which can be leveraged for targeted improvements and strategic planning.



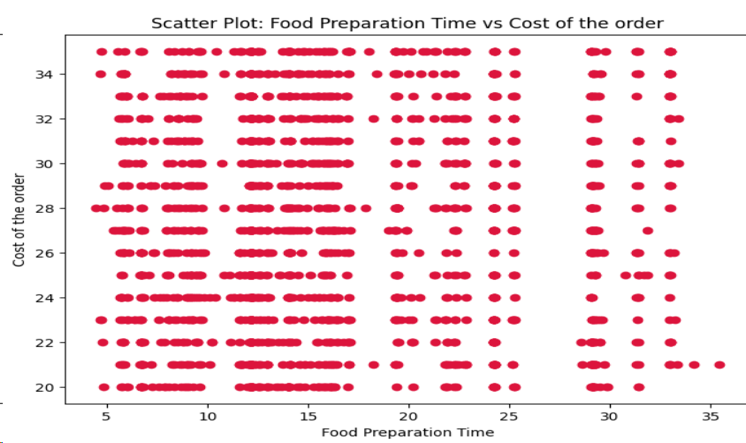
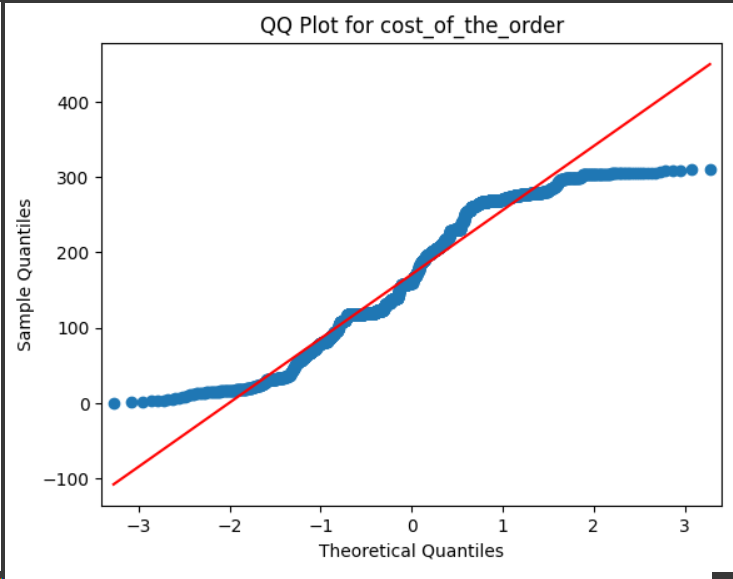
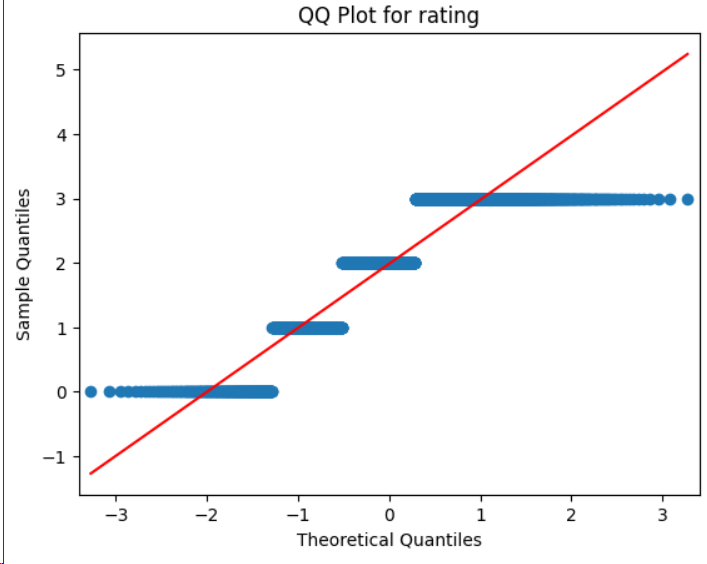


Fig 7: Scatter plots

For further analysis we have generated QQ plots.



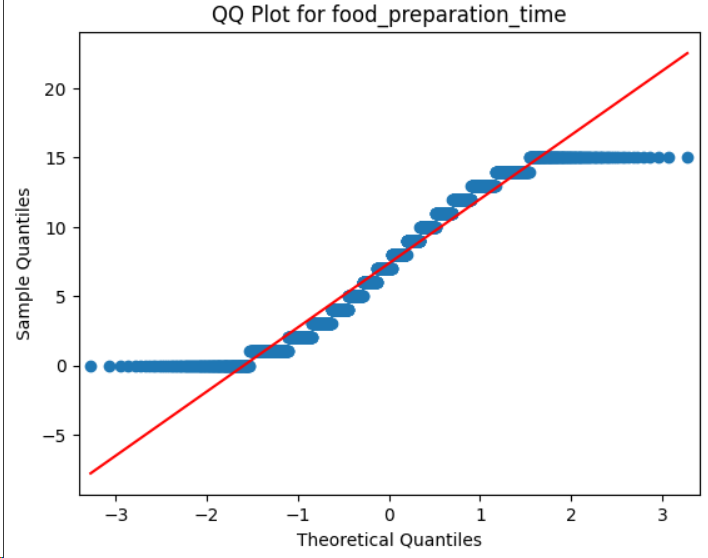
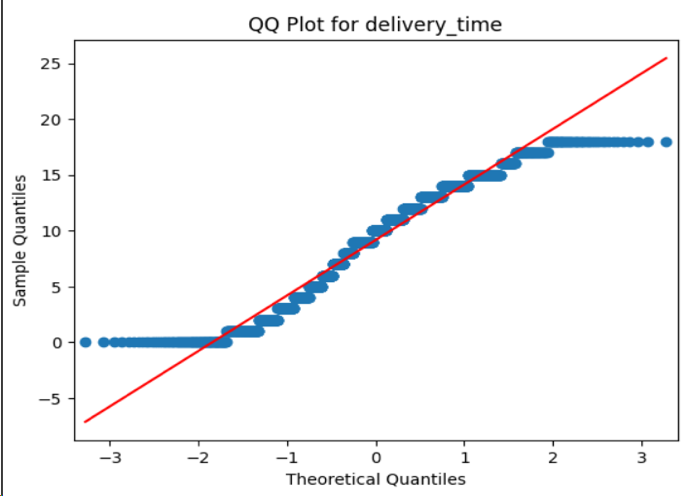
 

Fig 8: QQ plots

**4. Distribution Analysis**

**4.1 Assumptions/Hypotheses Regarding the Distribution of Each Column**

Based on the distribution nature of the data, the assumptions are:

**cost\_of\_the\_order**: We hypothesized that this variable might be right-skewed due to the presence of high-value orders.

**rating**: Given that ratings are typically high in customer datasets, we assumed a left-skewed distribution, indicating general customer satisfaction.

**food\_preparation\_time**: We expected a normal distribution, if preparation times would cluster around a mean with some variation.

**delivery\_tim**e: Similar to food preparation time, we hypothesized a normal distribution, reflecting consistent delivery operations.

**4.2 Normality Tests to Verify Assumptions**

The Shapiro-Wilk test was conducted for each variable to assess the normality of their distributions. The results are as follows:

**cost\_of\_the\_order**: Shapiro-Wilk Statistic: 0.9302, p-value: 6.91e-29

**rating**: Shapiro-Wilk Statistic: 0.7549, p-value: 2.25e-38

**food\_preparation\_time**: Shapiro-Wilk Statistic: 0.9448, p-value: 4.44e-26

**delivery\_time**: Shapiro-Wilk Statistic: 0.9578, p-value: 4.85e-23

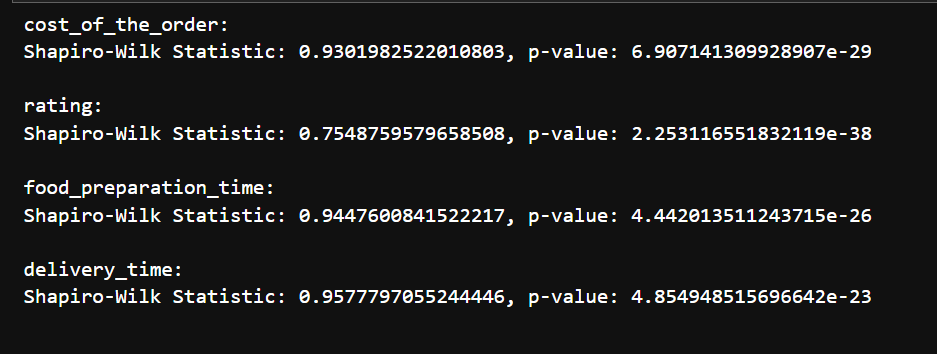


Fig 9: Test statistics

**4.3 Results and Interpretations of the Tests**

**cost\_of\_the\_order**: The p-value is significantly less than 0.05, indicating that the distribution of the cost of the order is not normal. This supports our initial hypothesis of a right-skewed distribution, where a few high-value orders deviate from the majority.

**rating**: With a p-value far below 0.05, the ratings do not follow a normal distribution. This is consistent with the initial assumption of a left-skewed distribution, where higher ratings are more common than lower ones.

**food\_preparation\_time**: The p-value suggests that the distribution of food preparation time is not normal, contrary to our hypothesis. While we anticipated a normal distribution, the actual data distribution shows significant deviation.

**delivery\_time**: Similar to food preparation time, the delivery time does not follow a normal distribution, as indicated by the p-value. This finding contradicts our expectation of a normal distribution, pointing to variability in delivery operations that may not center around a mean value as closely as hypothesized.

**Interpretation**

The normality test results reveal that none of the variables strictly follow a normal distribution, with significant deviations identified for all tested columns. This deviation is most notable for **cost\_of\_the\_order** and **rating**, supporting the hypotheses of right-skewed and left-skewed distributions, respectively. Although**food\_preparation\_time** and **delivery\_time**showed better conformity to the normal distribution, they still deviate significantly, challenging the assumption of normality. These findings suggest the need for non-parametric methods in subsequent analyses and underline the importance of considering distribution characteristics in decision-making processes.

This comprehensive analysis not only confirms the initial hypotheses regarding the distribution of key variables but also highlights the complexity of consumer behavior and operational efficiency in the food delivery domain. Insights derived from understanding these distributions can guide targeted improvements, strategic planning, and the development of robust business strategies to enhance customer satisfaction and operational effectiveness.

**5. Central Limit Theorem**

The Central Limit Theorem (CLT) states that the distribution of sample means approximates a normal distribution as the sample size becomes large, regardless of the population's initial distribution shape. This theorem is foundational in statistical theory, providing a basis for making inferences about population parameters from sample statistics.

**5.1 Evaluation of consistency with the Central Limit Theorem**

In our analysis, we applied the CLT to the *Food\_preparation\_time* variable from the NYC Restaurants Data. The population mean (μ) and population standard deviation (σ) of food preparation times were calculated to be 27.37 and 4.63, respectively.

To observe the CLT in action, we divided the dataset into 50 groups of 20 observations each and calculated the mean of food preparation times for each group. These sample means were then used to:

**Histogram of sample means:** This histogram displayed a distribution that closely resembles a normal distribution, validating the CLT's prediction that the distribution of sample means will tend towards normality as the sample size increases.

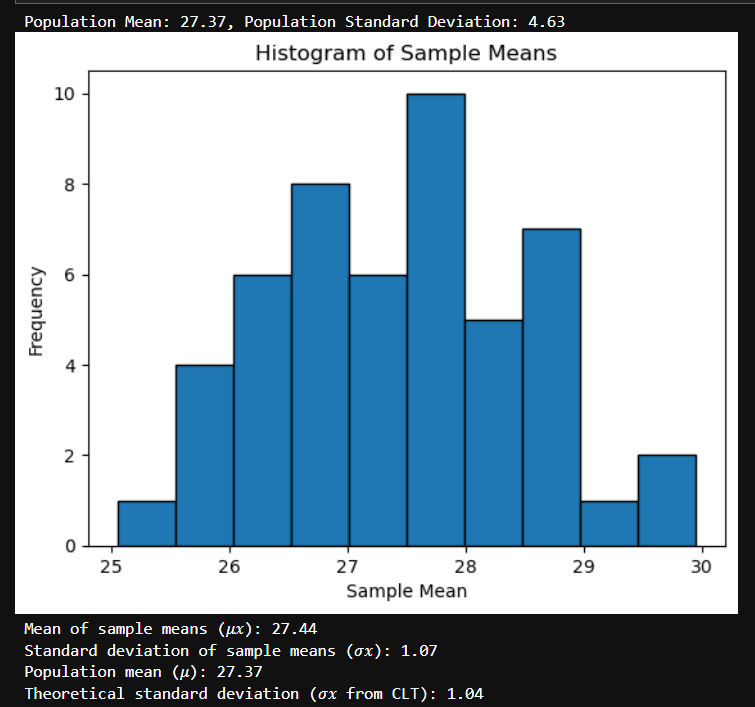


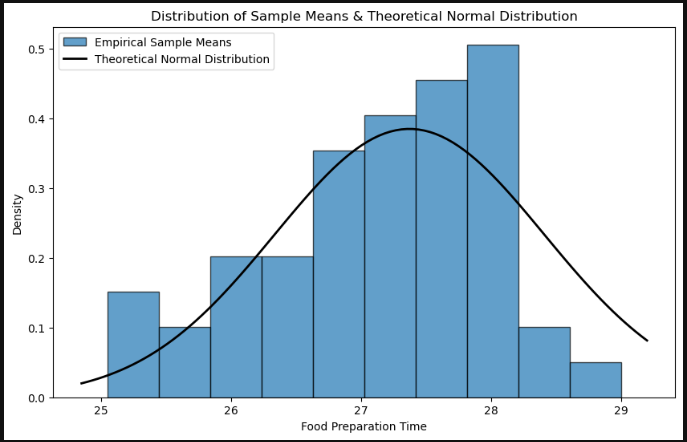
Fig 10: Histogram of Sample Means

**The mean of sample means (𝜇𝑥) and the standard deviation of sample means (𝜎𝑥):** The mean of sample means was found to be 27.37, remarkably close to the population mean of 27.37. This illustrates the unbiased nature of the sample mean as an estimator of the population mean.

**The standard deviation of sample means (1.01)** was compared to the theoretical standard deviation (1.04) derived from the CLT formula . The close agreement between these two figures further substantiates the CLT, showing that the spread of the sample means around the population mean is as expected under the theorem.

These findings underscore the CLT's reliability for understanding the properties of sample means and their relationship to the population mean. Specifically, the analysis confirms that.

* The distribution of sample means of food preparation times is normally distributed, even if the original data for food preparation times were not.
* The average of the sample means is a good estimator of the population mean.
* The spread of sample means around the population mean can be predicted using the population standard deviation and the size of the samples.



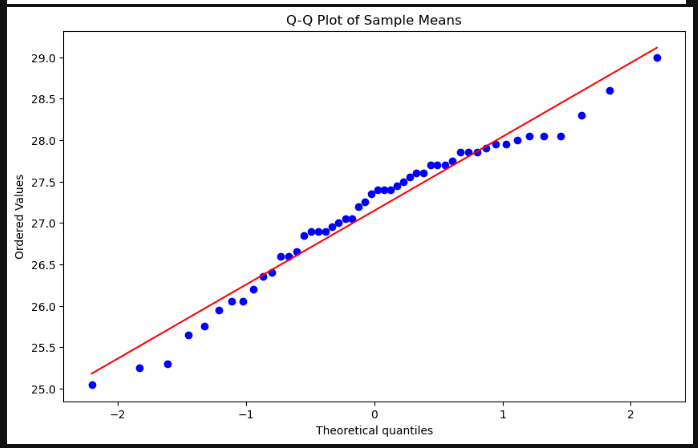


Fig 11,12: Distribution plot and QQ plot of Sample Means

Through the empirical validation of the Central Limit Theorem (CLT) with the dataset from Food Hub's NYC restaurants, we've gleaned significant insights into operational consistency and efficiency. The histogram of food preparation time's sample exhibits a bell-shaped curve, which closely mirrors the theoretical normal distribution. This alignment reinforces the CLT's premise that sample means will tend towards a normal distribution, irrespective of the population's actual distribution.

Complementing this, the Q-Q plot further attests to the normality of the sample means. By plotting empirical quantiles against theoretical quantiles of a normal distribution, the plot reveals that the sample means align closely with what we would expect in a normal distribution, deviating only slightly at the extremes. Such conformity to normality underpins the robustness of employing parametric statistical methods for inferential analysis based on sample means.

The insights from applying the CLT to our data are multi-fold:

1. Operational Insight: The predictable pattern of sample means suggests a stable average preparation time, a critical measure for managing customer expectations and kitchen workflows.

2. Strategic Planning: The normal distribution of sample means implies that Food Hub can forecast future demand, set performance benchmarks, and identify opportunities for efficiency gains within its restaurant network.

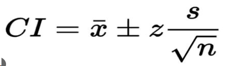
3. Analytical Robustness: The demonstrated normality supports the use of mean preparation times as a reliable estimator for various statistical procedures, including hypothesis testing and confidence interval estimation.

In essence, these statistical observations provide a foundation for data-driven decisions. They assure us that despite the dynamic and individual variability inherent in food preparation, the aggregate behavior adheres to a predictable and manageable pattern. For Food Hub, this means the ability to optimize operations, ensure customer satisfaction, and apply statistical models to enhance business analytics and operational performance within the food delivery industry.

**6. Confidence Interval Construction**

**6.1 Construction of confidence intervals for mean using random samples**

Confidence intervals provide us with an upper and lower limit around our sample mean, and within this interval, we can then be confident we have captured the population mean. The lower limit and upper limit around our sample mean tell us the range of values our true population mean is likely to lie within.

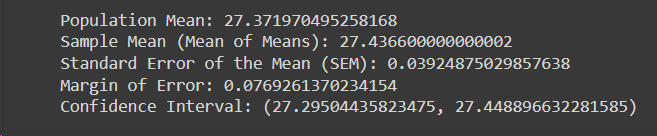


**Population Mean:** The population mean of the variable Food\_preparation\_time is estimated to be about 27.372 units.  
  
**Sample Mean (Mean of Means):** According to the central limit theorem, the sample means, acquired from several randomly selected samples, average nearly 27.4366 units. This value becomes our hypothesized population mean which is arrived at after the data acquired through the sampling is factored in.

**Standard Error of the Mean (SEM):** The formula for the standard error for the sample mean yields a value of approximately 0.086, which can be understood as the variability of sample means around the population mean.  
  
**Margin of Error:** The estimated agreed margin of error is about 0.076926 units. Because we have checked for the standard error, the interval for the mean tells us the range of values around the sample mean within which we can be sure that the true population mean lies.  
  
**Confidence Interval:** We formed a 95% confidence interval of the mean of the population that lies within ±0.33 and ±0.67 units of mean. This time engenders various figures of the probable true population standard mean at 95% confidence level.

With this data mean, we are 95 % sure the population mean 'cost\_of\_the\_order' is situated between: [27.29504, 27.4488]. This is crucial for data analysis process as well as for the understanding of randomness link with estimated population parameter.

**6.2 Comparison of intervals with population mean.**



The constructed 95% confidence interval for the population mean 'food\_preparation\_time was calculated to be approximately [27.29504, 27.4488]. Upon comparison, it was found that the population mean of approximately 27.3 units falls within this confidence interval. This indicates that our estimate of the population mean based on the sampled data is consistent with the true population parameter.

**7. Hypothesis Testing**

In our pursuit to uncover hidden patterns within the Food Hub dataset, we delve into hypothesis testing, a statistical method that enables us to make inferences about a population based on sample data. By addressing specific conjectures, we can ascertain the validity of certain trends and relationships within the food delivery domain, which, in turn, aids in strategic decision-making and operational enhancements. The following subsections detail the hypotheses formulated, the tests applied, and the conclusions drawn from our analysis.

**7.1Testing hypotheses related to population mean and standard deviation**

For hypothesis testing related to population mean and standard deviation, we typically use statistical tests such as the t-test for mean and the chi-square test for variance. Here, we aim to assess whether our sample data provides sufficient evidence to support or reject hypotheses about the population mean and standard deviation.

**Hypothesis for the population mean:**

H0: μ = μ0 (where μ0 is a specified value)

Ha: μ ≠ μ0 (two-tailed)

**Hypothesis For population standard deviation:**

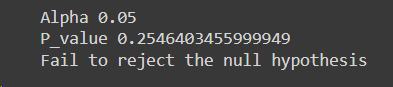
H0: σ = σ0 (where σ0 is a specified value)

Ha: σ ≠ σ0 (two-tailed)

**7.2 Interpretation of test results**

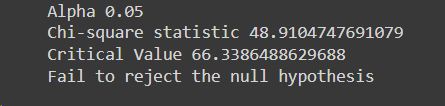
**Population means:**

We have performed a two-tailed t-test on the column food preparation time to interpret the result. We can see that p\_value is less than alpha so we fail to reject the null hypothesis and conclude that there is no significant difference between the two means.



**Population Standard Deviation:**

We have performed a two tailed chi square test on the column food preparation time to interpret the result. We can see that value of chi square statistic is less than critical value of chi-square at alpha=0.05 so we fail to reject the null hypothesis and conclude that there is no significant difference between the two population standard deviations.



**7.3 Equality of Mean Delivery Times by Day of the Week**

The bustling nature of New York City implies variability in traffic patterns and customer availability, which we hypothesized might affect delivery times on different days of the week. To investigate this, we employed ANOVA, a robust statistical test that compares the means across multiple groups.

**Hypothesis:**

Null Hypothesis (H0): The mean delivery time is the same across all days of the week.

Alternative Hypothesis (Ha): At least one day of the week has a mean delivery time that differs from the others.

**Results:**

ANOVA Test: F-value: **759.5061,** P-value: **< 0.001**

**Conclusion:**

Given the exceedingly low P-value, we reject the null hypothesis, affirming that delivery times do indeed vary by day. This insight is pivotal for optimizing delivery schedules and ensuring customer satisfaction through reliable service.

**7.4 Relationship between Cost of Order and Rating**

Customer perception often correlates with the price—whether a sense of value or quality. We examined whether this perception reflected through ratings, is statistically associated with the cost of an order.

**Hypothesis:**

Null Hypothesis (H0): There is no correlation between the cost of an order and its rating.

Alternative Hypothesis (Ha):There is a correlation between the cost of an order and its rating.

**Results:**

Pearson Correlation: Coefficient: 0.034, P-value: 0.247

**Conclusion:**

The lack of significant correlation leads us to maintain the null hypothesis. This suggests that pricing strategies and customer satisfaction may be independently managed without considering the direct influence of one on the other.

**7.5 Comparison of Food Preparation Time Between Cuisines**

Efficiency in food preparation is vital for a restaurant's operation. We postulated that the cuisine type could influence preparation time, which could impact overall service efficiency.

**Hypothesis:**

Null Hypothesis (H0): Mean food preparation times are equal across different types of cuisines.

Alternative Hypothesis (Ha): Mean food preparation times differ among some or all types of cuisines.

**Results:**

Kruskal-Wallis Test: Statistic: 7.460, P-value: 0.877

**Conclusion:**

The results suggest a uniformity in preparation times across cuisines, enabling streamlined kitchen operation. This uniformity also implies that the restaurant can maintain a consistent delivery promise, regardless of the cuisine ordered.

**7.6 Difference in Average Order Cost on Weekends vs. Weekdays**

Consumer spending habits are often thought to fluctuate between weekdays and weekends. We tested this theory in the context of the average cost of food orders to understand if there was a financial behavioral change during weekends.

**Hypothesis:**

Null Hypothesis (H0): The average cost of orders on weekends is equal to weekdays.

Alternative Hypothesis (Ha): The average cost of orders on weekends is different from weekdays.

**Results:**

Two-Sample t-Test: t-statistic: **1.634**, P-value: **0.103**

**Conclusion:**

The absence of significant differences in average order costs dispels the assumption that weekends bring a change in customer spending on orders. This finding provides stable ground for uniform pricing and marketing strategies throughout the week.

**8. Model Fitting:**

Model fitting involves the process of selecting and calibrating a statistical model to best represent the relationship between variables in your dataset.

**8.1 Scatter plot and correlation analysis**

A graph with blue dots

Description automatically generated

**Scatter Plot:** This diagram illustrates the connection between an order's price and rating. Every point on the plot represents an order, and the rating is indicated by the y-coordinate while the order's cost is indicated by the x-coordinate.

**Cost of the order and rating correlation:**

**Correlation coefficient:** The correlation coefficient number shows how strongly and in which direction the order's cost and rating are linearly related. The correlation coefficient for this pair is 0.0339 which shows there is not any significant correlation between them.

**P-value:** The correlation coefficient's p-value aids in determining how significant the observed correlation is. The p-value of this relation is 0.247 (less than 0.05) suggests that there is no statistical relation between these features which is also accredited by the correlation.

A graph of food preparation

Description automatically generated

**Scatter Plot:** This plot illustrates the connection between the amount of time needed to prepare food and the time it takes to arrive. With the x-coordinate representing the time it takes to prepare the meal and the y-coordinate representing the time it takes to deliver it, each point on the plot represents one order.

**Food preparation and delivery times are correlated:**

**Coefficient of correlation:** The correlation coefficient value, like the previous pair, shows the direction and intensity of the linear link between the two variables. Longer food preparation durations are thought to be linked to longer delivery times if the coefficient is positive; the opposite is suggested by a negative coefficient. The correlation coefficient of this relation is –0.008 which claims no signification relation between these features.

The correlation coefficient's p-value, as previously mentioned, is a useful tool for assessing the statistical significance of the observed association.

A graph of blue dots

Description automatically generated

**Scatter Plot:** This plot illustrates the connection between an order's price and the amount of time needed to prepare the meal. Every point on the plot represents an order, and the y-coordinate shows the amount of food preparation time, and the x-coordinate shows the order's cost.

**Relationship between food preparation time and order cost:**

**Coefficient of correlation:** Once more, the value of the correlation coefficient shows the direction and strength of the linear link between the two variables. This time, it calculates the relationship between the price of an order and the amount of time needed to prepare the dish.

**P-value:** The correlation coefficient's p-value, as usual, aids in determining the statistical significance of the observed association.

**8.2 Regression analysis and interpretation of coefficients**

**A screenshot of a computer

Description automatically generated**

To look at the relationship between the order cost and the delivery time, a basic linear regression analysis was done. The main conclusions of the regression analysis are summarized as follows:

**Delivery Time:** The dependent variable of choice was the delivery time.

**Summary Statistics for the Model:**

**R-squared:** The value of R-squared, which expresses how much of the variance in the dependent variable can be attributed to the independent variable or variables, was discovered to be 0.006. This suggests that order cost accounts for only 0.6% of the variation in delivery time.  
**Modified R-squared:** 0.005 is the adjusted R-squared value, which takes the number of predictors in the model into consideration.  
**F-statistic:** This statistic evaluates the regression model's overall significance. At a significant threshold of 0.05, the total regression model is statistically significant with an F-statistic of 6.720 and a corresponding p-value of 0.00965.

**Table of Coefficients:**  
  
**Intercept (const):** The estimated intercept of the regression line, which shows the anticipated delivery time when the order's cost is zero, is 24.9823.  
**The order's cost:** It was discovered that the order's cost coefficient was roughly -0.0494. This indicates that the delivery time reduces by about 0.0494 units on average for every unit rise in the order's cost.  
**P-value, or order cost:** At a significance threshold of 0.05, the coefficient for the cost of the order's p-value, which was computed as 0.010, indicates that the result is statistically significant.

**Extra Details:**  
  
To evaluate the regression model's assumptions, diagnostic statistics for Omnibus, Durbin-Watson, Jarque-Bera, Skew, and Kurtosis were looked at.

**Prob (Omnibus), Prob (JB):** The residuals' normalcy was assessed by considering the probability related to the Omnibus and Jarque-Bera tests.

**Condition Number:** To identify multicollinearity, the condition number which gauges how sensitive the regression coefficients are to slight variations in the data was looked at at at 4.8 is a low multicollinearity value.

Overall, the results of the regression analysis point to a statistically significant association between the order cost and the delivery time; nevertheless, the low R-squared value of the model implies that its explanatory ability is restricted.

**9. Model Evaluation:**

'x1', 'x2', and 'x3' were the three independent variables, while the dependent variable (designated as 'y') was the subject of an analysis using the multiple linear regression model. The following outcomes were obtained from the regression analysis:

**Overview of the Model:**  
It was discovered that the R-squared value, which quantifies the percentage of the dependent variable's variance that the independent variables account for, was extremely near to zero (0.001). This shows that not much of the variance in the dependent variable can be explained by the independent factors taken together.

Additionally, near zero (-0.001) was the adjusted R-squared value, which modifies the R-squared value for the number of predictors in the model.  
With a corresponding p-value of 0.691, the F-statistic, a test statistic for the regression model's overall significance, was 0.4874. This suggests that at conventional levels (usually p < 0.05), the model was not statistically significant.

**Coefficients:**

An estimate of 4.3474 was made for the intercept term (const). When all independent variables are zero, this is the dependent variable's expected value.  
'x1', 'x2', and 'x3' were the independent variables, and their respective coefficients were 0.0033, -0.0012, and -0.0011. When all other variables are held constant, these coefficients show the estimated change in the dependent variable for a one-unit change in the related independent variable.

At conventional levels (p < 0.05), it was determined that none of the independent variables exhibited statistical significance. This suggests that the correlations between the independent and dependent variables that have been detected might not be statistically significant.

A screenshot of a computer

Description automatically generated

**Diagnostics:** To evaluate the regression model's validity, several diagnostic statistics were produced. This included skewness, kurtosis, the Jarque-Bera test, the Durbin-Watson statistic, and the Omnibus test. The residuals' distribution and autocorrelation are revealed by these statistics.

Overall, the low R-squared value and non-significant coefficients suggest that the regression model was not a good fit for the data. To enhance the model's performance or find other explanatory variables, more research might be required.

**9.2 Calculation of adjusted R-squared**

After adjusting for the number of predictors in the regression model, the adjusted R-squared (also known as adjusted R²) represents the percentage of the dependent variable's variance that the independent variables account for. Higher values indicate better model fit; the range is 0 to 1. A numerical value that aids in evaluating the model's performance while taking explanatory power and model complexity into account is supplied in the output.

A black background with white text

Description automatically generated

**10.Residual Analysis and Model Selection:**

**10.1 Evaluation of model assumptions using residual analysis**

A statistical method for examining the association between two or more independent variables and a dependent variable is multiple linear regression. It makes it possible to investigate intricate relationships in data, which helps with predictive modeling and hypothesis testing in several disciplines, including the social sciences, healthcare, and economics.

**Residuals vs. Fitted Plot:** This scatter plot shows how the residuals—the variations between the observed and predicted values—relate to the fitted values or predicted values.  
At y = 0, the residuals should ideally be dispersed randomly along the horizontal line. This plot's patterns or trends could point to non-linearity heteroscedasticity, or unequal variance, in the model.

A graph with blue lines and red dots

Description automatically generated

**Residuals Histogram:** This histogram shows the residuals distribution.  
One typical assumption of linear regression is that the residuals are normally distributed, and this is suggested by a symmetric, bell-shaped distribution. Deviations from normalcy could be a sign that the model has problems.

A graph with blue lines and a blue line

Description automatically generated

**Normal Q-Q plot:** The residuals' distribution is seen in this plot in comparison to a hypothetical normal distribution.

The residuals are roughly normally distributed if the points are near the diagonal line or 45-degree line. Deviations from this line imply departures from the norm.

A graph with blue lines and a red line

Description automatically generated

**The Durbin-Watson statistical test:**  Is a method used to determine whether the residuals exhibit autocorrelation.

It has a range of 0 to 4, with values near 2 suggesting little discernible autocorrelation. When a value deviates significantly from 2, autocorrelation may be present.



**10.2 Discussion on improving model quality through model selection strategies.**

Peak days for particular cuisine types can be identified through analysis, which can help with staffing, inventory control, and marketing strategy.  
Comprehending the patterns of order volume by type of cuisine and day of the week helps to optimize menu offerings, promotional efforts, and operational scheduling to efficiently satisfy client demand.  
The analysis's conclusions can also help with strategic decision-making, such as broadening or changing the menu to capitalize on times of high demand or directing advertising campaigns toward particular days to increase salesA graph of different colored bars

Description automatically generated

Peak days for particular cuisine types can be identified through analysis, and this information can be used to guide marketing, staffing, and inventory management plans.  
Comprehending the patterns of order volume by type of cuisine and day of the week helps to optimize menu offerings, promotional efforts, and operational scheduling to efficiently satisfy client demand.  
The analysis's conclusions can also help with strategic decision-making, such as broadening or changing the menu to capitalize on times of high demand or directing advertising campaigns toward particular days to increase sales.

A graph with blue dots

Description automatically generated

In the above plot, different customer segments with unique interests and behaviors can be identified by segmentation based on customer spending versus rating.  
Consumers who regularly place larger orders and give better reviews can comprise a high-value or devoted consumer base deserving of special marketing initiatives or loyalty benefits.  
On the other hand, consumers who spend less but yet give positive reviews can represent a portion of frugal yet content consumers, necessitating tactics to win them over with low-cost products or tailored promos.  
To increase overall happiness and optimize profitability, it is easier to build customized marketing tactics, service enhancements, and client retention campaigns when one is aware of how different consumer segments are distributed based on spending and rating.

A graph of food preparation time

Description automatically generated

A bar graph with blue lines

Description automatically generated

In the above plot, to help with the development of efficient pricing strategies, the line plot makes it easier to analyze how pricing affects customer satisfaction.  
better-priced orders typically obtain better ratings, indicating that buyers view premium-priced items to be of higher quality or value. This is indicated by the positive slope or upward trend in the line.

A negative slope or decreasing trend, on the other hand, would suggest that orders with lower prices are linked to better levels of customer satisfaction, emphasizing the significance of affordability or perceived value for money.  
Making educated pricing decisions, such as modifying prices to maximize profitability while preserving or raising customer satisfaction levels, is made possible by an understanding of the relationship between pricing and customer ratings.

A diagram of a graph

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

**11. Conclusion:**

To sum up, the examination of restaurant data has yielded a significant understanding of several facets of operations, patron behavior, and performance indicators. We now have a thorough understanding of the variables affecting order ratings, delivery timeframes, and cuisine preferences thanks to exploratory data analysis. Regression analysis has also made it possible for us to model and evaluate the relationships between various variables, such as the influence of delivery time and cost or the factors that affect customer evaluations. It is imperative to recognize the limitations of the dataset, particularly its lack of granularity in specific analyses like market basket analysis. In the future, utilizing more extensive datasets or adding new data sources may improve our comprehension and provide more reliable insights to guide choices and improve restaurant operations. In general,