MAHATMA EDUCATION SOCIETY'S

PILLAI COLLEGE OF ARTS, COMMERCE & SCIENCE
(Autonomous)

NEW PANVEL

RESEARCH PAPER ON

"Pizza Sales Analysis: Understanding Customer Choices and Pricing"

IN PARTIAL FULFILLMENT OF

MASTERS OF DATA ANALYTICS

SEMESTER **I** – **2023-24**

PROJECT GUIDE

Name: **Prof. Sabitha Praveen**

SUBMITTED BY: **Athulkrishna Pramod**

ROLL NO: **3105**

Abstract –

This research delves into the dynamic world of pizza ordering and sales, aiming to uncover valuable insights into customer preferences and factors influencing the pizza market. Using a dataset of pizza orders, we applied various statistical tests, including chi-squared tests, t-tests, ANOVA and descriptive analyses, to gain a comprehensive understanding of the pizza industry.

Our findings revealed significant associations between pizza category and size, shedding light on customer choices and preferences. The data-driven analysis also provided insights into pricing strategies and their impact on total sales. Notably, we identified the best-selling and least-selling pizza varieties, allowing for strategic marketing and product development.

Furthermore, the research includes visualizations such as scatterplots, bar graphs, boxplots, and statistical measures such as the standard deviation and coefficient of variation. These visual aids enhance our understanding of the data distribution and variation in pizza prices.

In summary, this research offers a valuable contribution to the pizza industry, guiding businesses in optimizing their product offerings and pricing strategies. The methodologies and insights presented here serve as a foundation for future research in the field of consumer behaviour and market analysis.

Introduction –

The pizza industry, characterized by its ubiquity and diverse product offerings, is a prominent and ever-evolving sector of the food market. With a global appetite for pizza continuing to grow, understanding consumer preferences, sales patterns, and market dynamics is of paramount importance for both pizzerias and foodservice businesses. In this context, our research endeavors to delve into the intricate world of pizza sales and customer choices, leveraging data-driven techniques and statistical analyses.

Background and Rationale –

The popularity of pizza as a go-to meal choice is undeniable, yet the factors influencing the selection of specific pizza varieties, sizes, and pricing strategies are multifaceted. Pizzerias and food establishments must navigate these complexities to cater to customer preferences effectively. By analyzing a comprehensive dataset of pizza orders, we seek to uncover valuable insights that can aid businesses in decision-making processes.

Research Objectives –

The primary objectives of this research are as follows:

1. To investigate the associations between pizza category (e.g., Margherita, Pepperoni) and pizza size (e.g., Small, Large).

2. To assess the impact of pricing strategies on total sales and customer choices.

3. To identify the best-selling and least-selling pizza varieties, informing marketing and product development strategies.

Research Methodology –

Our research methodology involves the application of various statistical tests, including chi-squared tests, t-tests, and descriptive analyses, to the dataset of pizza orders. Visualizations such as scatterplots, bar graphs, and boxplots provide additional context and enhance our understanding of the data.

Pizza Sales Dataset

Pizza Sales.csv is a dataset sourced from Kaggle, offering insights into pizza sales within a restaurant or chain setting. This tabular dataset, stored in CSV format, likely originates from a pizza business and includes key fields such as date, pizza type, quantity sold, price, total sales and potentially order details.

The dataset's primary purpose is to support diverse analyses and research endeavours, from understanding pizza sales trends over time to uncovering the most favoured pizza varieties. It serves as a valuable resource for assessing revenue patterns, profit margins, and the impact of promotional activities. Researchers and data analysts can utilize this dataset for market research purposes, gaining insights into consumer preferences and pizza consumption habits.

Overall, "Pizza Sales.csv" presents an opportunity to delve into the delectable world of pizza economics and consumer behaviour, providing a rich dataset for a variety of data-driven investigations.

Dataset Link:

https://www.kaggle.com/datasets/shilongzhuang/pizza-sales/data

The below table provides a summary of the columns present in the dataset:

Column Names	Data Types	Description
order_details_id	Integer	Unique identifier for each order placed by a table
order_id	Integer	Unique identifier for each pizza placed within each order, pizzas of the same type and size are kept in the same row, and the quantity increases
pizza_id	Object	Unique key identifier that ties the pizza ordered to its details, like size and price
quantity	Integer	Quantity ordered for each pizza of the same type and size
order_date	Object	Date the order was placed
order_time	Object	Time the order was placed
unit_price	Float64	Price of the pizza in USD
total_price	Float64	unit_price * quantity
pizza_size	Object	Size of the pizza (Small, Medium, Large, X Large, or XX Large)
pizza_category	Object	Category of the pizza (Classic, Supreme, Veggie, or Chicken)
pizza_ingredients	Object	Ingredients used in the pizza
pizza_name	Object	Name of the pizza as shown in the menu

Data Analysis and Interpretation

Install and Import the Libraries –

```
#install libraries
install.packages("ggplot2")
install.packages("dplyr")
install.packages("tidyr")
install.packages("gridExtra")
```

These lines of code install the necessary R packages ggplot2, dplyr, tidyr, and gridExtra, which are used for data visualization, data manipulation, and plotting.

```
# Load required libraries
library(ggplot2)
library(dplyr)
library(tidyr)
library(gridExtra)
```

After installing the packages, these lines load them into the R environment so that you can use their functions and capabilities.

Read the CSV File –

```
# Read the CSV file (replace 'file.csv' with your file path)
data = read.csv(file.choose())
df = data
```

These lines read a CSV file and store its contents in a data frame called df. You'll need to select the CSV file through a file dialog when prompted.

Data Exploration –

```
# Display the first few rows of the data frame
head(df)
```

```
> head(df)
  order_details_id order_id
1                1        1
2                2        2
3                3        2
4                4        2
5                5        2
6                6        2
       pizza_id quantity
1    hawaiian_m        1
2 classic_dlx_m        1
3 five_cheese_l        1
4   ital_supr_l        1
5    mexicana_m        1
6    thai_ckn_l        1
  order_date order_time
1   01-01-15   11:38:36
2   01-01-15   11:57:40
3   01-01-15   11:57:40
4   01-01-15   11:57:40
5   01-01-15   11:57:40
6   01-01-15   11:57:40
  unit_price total_price
1      13.25       13.25
2      16.00       16.00
3      18.50       18.50
4      20.75       20.75
5      16.00       16.00
6      20.75       20.75
  pizza_size pizza_category
1          M        Classic
2          M        Classic
3          L         Veggie
4          L        Supreme
5          M         Veggie
6          L        Chicken
                                                                  pizza_ingredients
1                                                    Sliced Ham, Pineapple, Mozzarella Cheese
2                                                Pepperoni, Mushrooms, Red Onions, Red Peppers, Bacon
3 Mozzarella Cheese, Provolone Cheese, Smoked Gouda Cheese, Romano Cheese, Blue Cheese, Garlic
4                        Calabrese Salami, Capocollo, Tomatoes, Red Onions, Green Olives, Garlic
5  Tomatoes, Red Peppers, Jalapeno Peppers, Red Onions, Cilantro, Corn, Chipotle Sauce, Garlic
6                              Chicken, Pineapple, Tomatoes, Red Peppers, Thai Sweet Chilli Sauce
                pizza_name
1        The Hawaiian Pizza
2  The Classic Deluxe Pizza
3      The Five Cheese Pizza
4 The Italian Supreme Pizza
5        The Mexicana Pizza
6    The Thai Chicken Pizza
```

```r
# Display the last few rows of the data frame
tail(df)
```

```
> tail(df)
      order_details_id
48615            48615
48616            48616
48617            48617
48618            48618
48619            48619
48620            48620
      order_id      pizza_id
48615    21347  southw_ckn_l
48616    21348 ckn_alfredo_m
48617    21348 four_cheese_l
48618    21348  napolitana_s
48619    21349    mexicana_l
48620    21350     bbq_ckn_s
      quantity order_date
48615        1   31-12-15
48616        1   31-12-15
48617        1   31-12-15
48618        1   31-12-15
48619        1   31-12-15
48620        1   31-12-15
      order_time unit_price
48615   21:14:37      20.75
48616   21:23:10      16.75
48617   21:23:10      17.95
48618   21:23:10      12.00
48619   22:09:54      20.25
48620   23:02:05      12.75
      total_price pizza_size
48615       20.75          L
48616       16.75          M
48617       17.95          L
48618       12.00          S
48619       20.25          L
48620       12.75          S
      pizza_category
48615        Chicken
48616        Chicken
48617         Veggie
48618        Classic
48619         Veggie
48620        Chicken
                                                                      pizza_ingredients
48615      Chicken, Tomatoes, Red Peppers, Red Onions, Jalapeno Peppers, Corn, Cilantro, Chipotle Sauce
48616                         Chicken, Red Onions, Red Peppers, Mushrooms, Asiago Cheese, Alfredo Sauce
48617 Ricotta Cheese, Gorgonzola Piccante Cheese, Mozzarella Cheese, Parmigiano Reggiano Cheese, Garlic
48618                                        Tomatoes, Anchovies, Green Olives, Red Onions, Garlic
48619       Tomatoes, Red Peppers, Jalapeno Peppers, Red Onions, Cilantro, Corn, Chipotle Sauce, Garlic
48620               Barbecued Chicken, Red Peppers, Green Peppers, Tomatoes, Red Onions, Barbecue Sauce
                   pizza_name
```

```r
# Check the dimensions of the data frame
dim(df)
```

```
> dim(df)
[1] 48620    12
```

Standard Deviation & Coefficient of Variation of Total Price –

```
> # Standard Deviation
> # Calculate standard deviation of total price
> sd_total_price <- sd(df$total_price)
> cat("Standard Deviation of Total Price:", sd_total_price, "\n")
Standard Deviation of Total Price: 4.437398

> #Coefficient of Variation (CV)
> # Calculate mean of total price
> mean_total_price <- mean(df$total_price)
> # Calculate coefficient of variation
> cv_total_price <- (sd_total_price / mean_total_price) * 100
> cat("Coefficient of Variation of Total Price:", cv_total_price, "%\n")
Coefficient of Variation of Total Price: 26.37936 %
```

Summary –

```
order_details_id
Min.   :    1
1st Qu.:12156
Median :24311
Mean   :24311
3rd Qu.:36465
Max.   :48620
   order_id
Min.   :    1
1st Qu.: 5337
Median :10682
Mean   :10701
3rd Qu.:16100
Max.   :21350
  pizza_id
Length:48620
Class :character
Mode  :character

    quantity
Min.   :1.00
1st Qu.:1.00
Median :1.00
Mean   :1.02
3rd Qu.:1.00
Max.   :4.00
 order_date
Length:48620
Class :character
Mode  :character

 order_time
Length:48620
Class :character
Mode  :character

  unit_price
Min.   : 9.75
1st Qu.:12.75
Median :16.50
Mean   :16.49
3rd Qu.:20.25
Max.   :35.95
 total_price
Min.   : 9.75
1st Qu.:12.75
Median :16.50
Mean   :16.82
3rd Qu.:20.50
Max.   :83.00
 pizza_size
Length:48620
Class :character
Mode  :character

pizza_category
Length:48620
Class :character
Mode  :character

pizza_ingredients
Length:48620
Class :character
Mode  :character

 pizza_name
Length:48620
Class :character
Mode  :character
```

Data Cleaning –

```
> # Check for missing values
> sum(is.na(df))
[1] 0
> # Check for duplicated rows
> sum(duplicated(df))
[1] 0
```

Exploratory Data Analysis (EDA) –

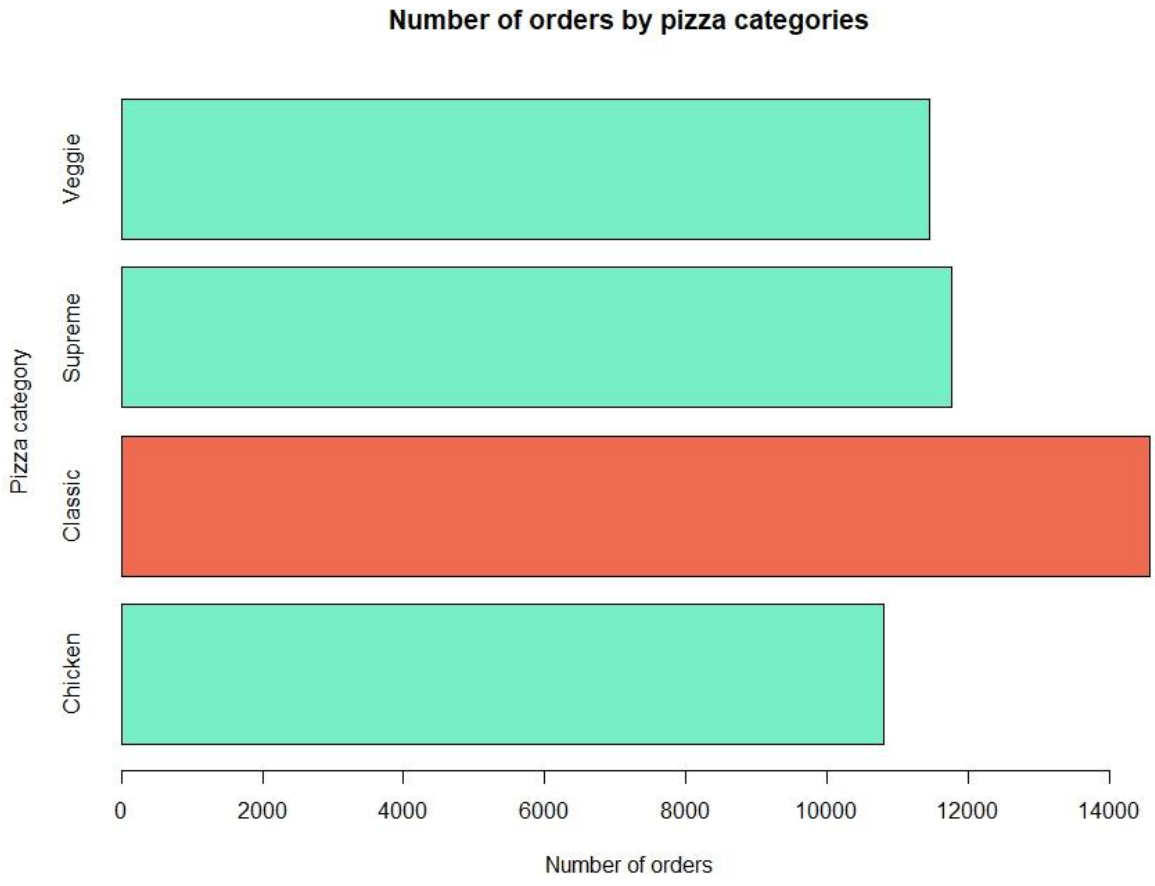
Exploratory Data Analysis (EDA)is a crucial phase in the data analysis process where data professionals delve into a dataset to uncover its underlying patterns, characteristics, and potential issues. Through summary statistics, data visualizations, and data cleaning, EDA offers insights into the data's distribution, relationships, and outliers.

In this step I would like to explore these questions –

1. Which category of Pizza is ordered the most?
2. Which size of Pizza is ordered the most?
3. What are our best and worst selling Pizza?
4. What’s our average order value?
5. What is the total revenue up to the latest order date?
6. Which month was revenue earned the highest?
7. What is the average unit price and revenue of most sold 5 pizzas?

1. **Which category of Pizza is ordered the most?**

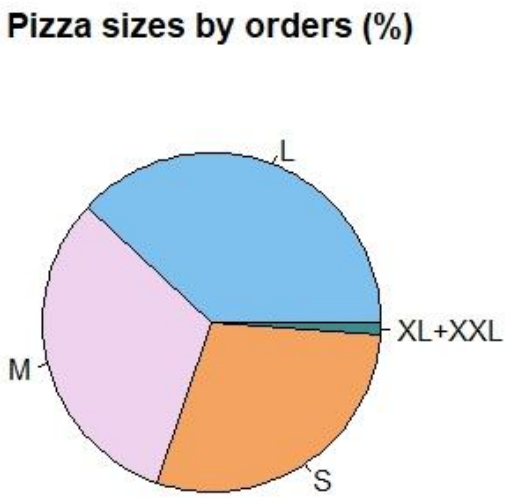
```
> # Show the number of orders for each category of pizza
> categories <- table(df$pizza_category)
> categories

Chicken Classic Supreme  Veggie
  10815   14579   11777   11449
> # Find the category with the highest number of orders
> highest_category <- names(which.max(categories))
> highest_category
[1] "Classic"
> # Set colors for the bar chart
> colors <- ifelse(names(categories) == highest_category, "#EE6A50", "#76EEC6")
> # Plot the bar chart
> bar_plot <- barplot(categories, horiz = TRUE, col = colors,
+                     xlab = "Number of orders", ylab = "Pizza category",
+                     main = "Number of orders by pizza categories")
```



Classic category is mostly preferred by the customer.

2. **Which size of Pizza is ordered the most?**

```
# Total number of orders
total_orders <- nrow(df)
# Calculate the percentages of each pizza size
pizza_sizes <- table(df$pizza_size)
size_percentages <- pizza_sizes / total_orders
# Plot the pie chart
pie_labels <- c("L", "M", "S", "XL+XXL")
pie_colors <- c("#7EC0EE", "#EED2EE", "#F4A460", "#388E8E")
pie(size_percentages, labels = pie_labels, col = pie_colors,
    main = "Pizza sizes by orders (%)")
```



Large sized pizza is sold the most.

3. **What are our best and worst selling Pizza?**

```
> # Identify the best selling pizza
> best_selling_pizza <- names(which.max(table(df$pizza_name)))
> cat("The best selling pizza is:", best_selling_pizza, "\n")
The best selling pizza is: The Classic Deluxe Pizza
> # Identify the worst selling pizza
> worst_selling_pizza <- names(which.min(table(df$pizza_name)))
> cat("The worst selling pizza is:", worst_selling_pizza, "\n")
The worst selling pizza is: The Brie Carre Pizza
```

4. **What's our average order value?**

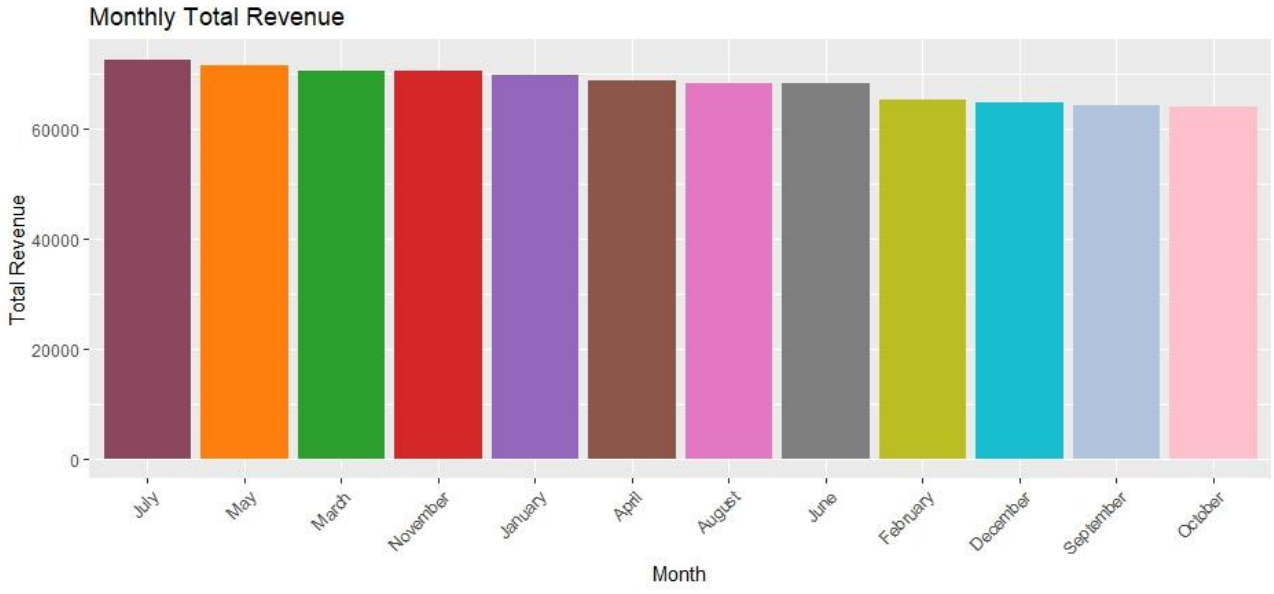
```
> # Calculate mean and median of total prices
> mean_price <- mean(df$total_price)
> median_price <- median(df$total_price)
> cat("Mean: USD", formatC(mean_price, format = "f", digits = 2), "\n")
Mean: USD 16.82
> cat("Median: USD", formatC(median_price, format = "f", digits = 2), "\n")
Median: USD 16.50
```

5. **What is the total revenue up to the latest order date?**

```
> # Calculate total revenue
> total_revenue <- sum(df$total_price)
> cat("Total revenue: USD", formatC(total_revenue,
+                                   format = "f", digits = 2), "\n")
Total revenue: USD 817860.05
```

6. **Which month was revenue earned the highest?**

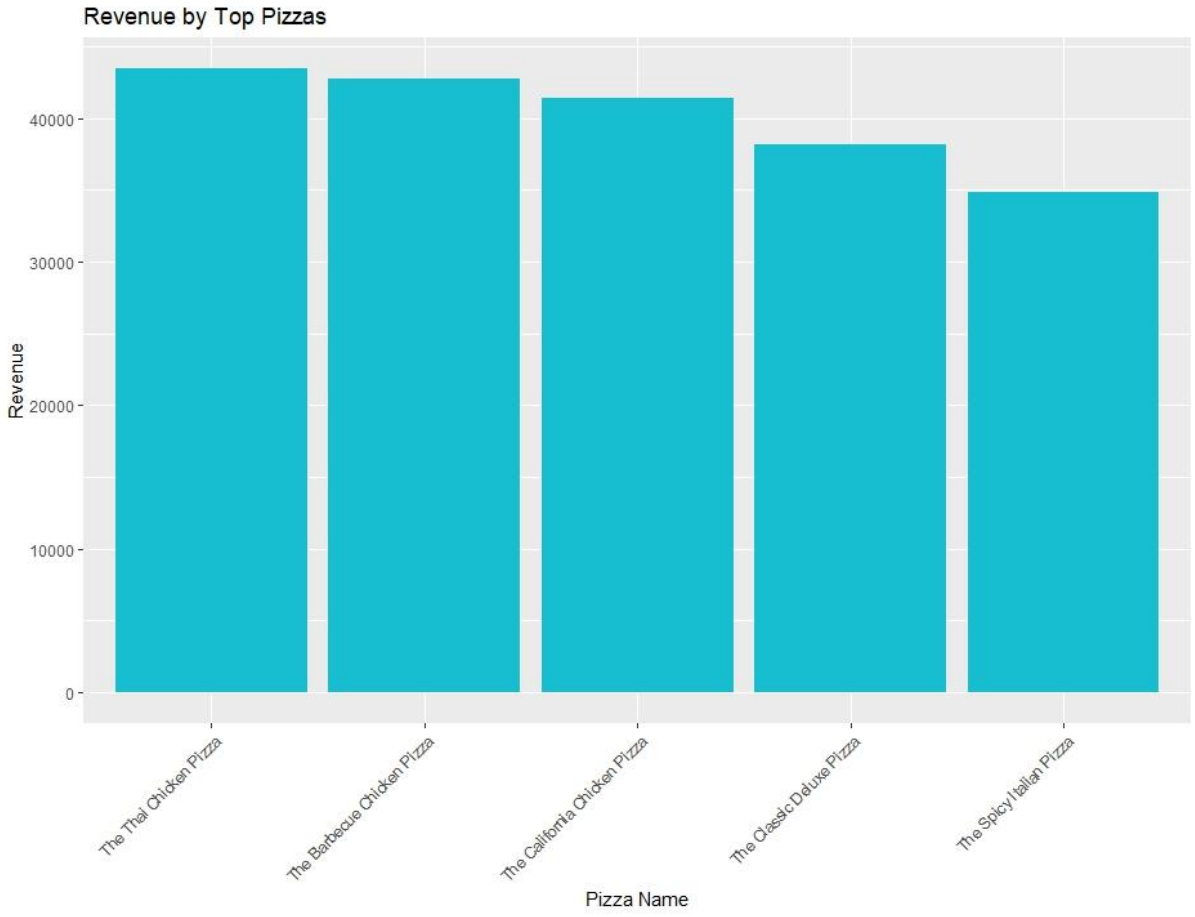
```
> # Extract month and month_name from order_date
> df$order_date <- as.Date(df$order_date)
> df$month <- format(df$order_date, format = "%m")
> df$month_name <- format(df$order_date, format = "%B")
> # Calculate total revenue by month
> revenue_by_month <- df %>%
+   group_by(month_name) %>%
+   summarise(total_revenue = sum(total_price))
> # Rank the months by total revenue
> revenue_by_month$rank <- rank(-revenue_by_month$total_revenue)
> # Sort the data frame by rank
> revenue_by_month <- revenue_by_month %>%
+   arrange(rank)
> # Create a bar plot of monthly total revenue
> custom_colors <- c('#8B475D', '#ff7f0e', '#2ca02c', '#d62728',
+                    '#9467bd', '#8c564b', '#e377c2', '#7f7f7f', '#bcbd22',
+                    '#17becf', '#B0C4DE', '#FFC0CB')
> ggplot(revenue_by_month, aes(x = reorder(month_name, -total_revenue), y = total_revenue)) +
+   geom_bar(stat = "identity", fill = custom_colors) +
+   labs(x = "Month", y = "Total Revenue", title = "Monthly Total Revenue") +
+   theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Most pizzas were sold in the month of July.

7. What is the average unit price and revenue of most sold 5 pizzas?

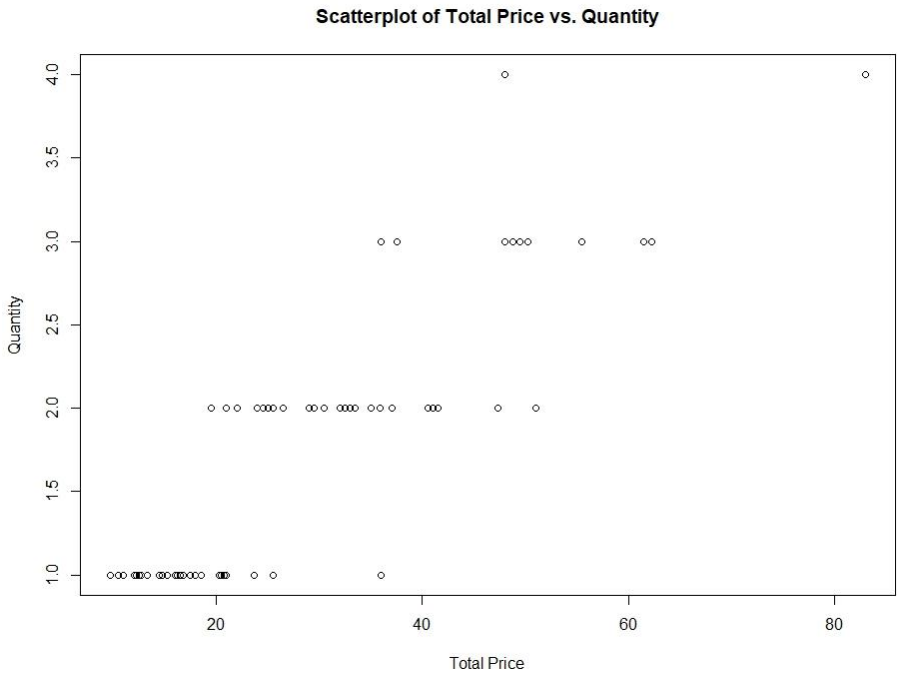
```
> # Top pizza analysis
> top_pizza_analysis <- df %>%
+   group_by(pizza_name) %>%
+   summarise(average_unit_price = mean(unit_price),
+             revenue_per_pizza = sum(unit_price * quantity)) %>%
+   top_n(5, revenue_per_pizza)
> # Create a bar plot for revenue by top pizzas
> ggplot(top_pizza_analysis, aes(x = reorder(pizza_name, -revenue_per_pizza), y = revenue_per_pizza)) +
+   geom_bar(stat = "identity", fill = "#17becf") +
+   labs(x = "Pizza Name", y = "Revenue", title = "Revenue by Top Pizzas") +
+   theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



The Thai chicken pizza has got the highest revenue.

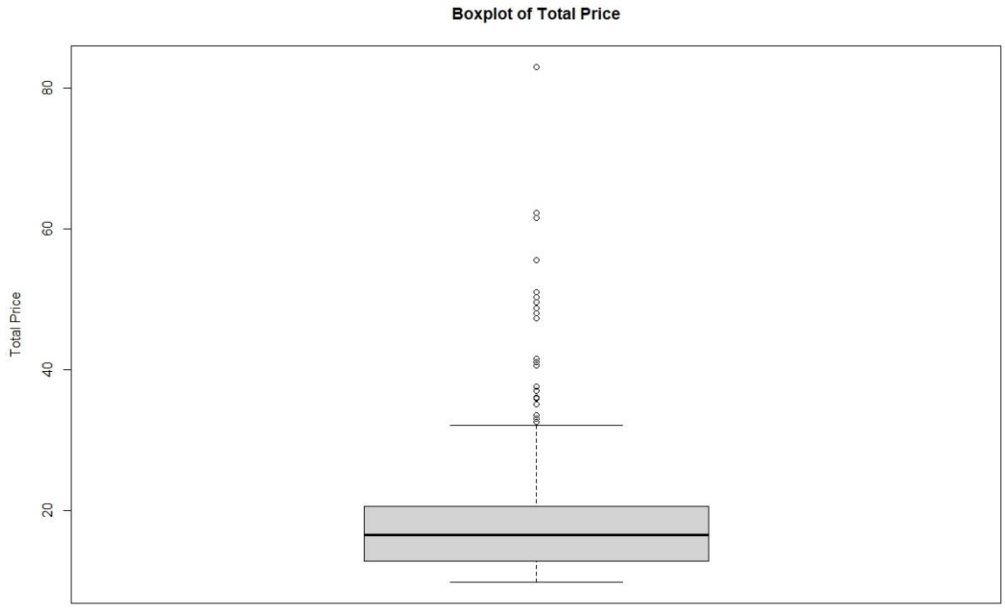
Scatter Plot –

```
> #Scatterplot total price vs quantity
> plot(df$total_price, df$quantity,
+      xlab = "Total Price", ylab = "Quantity",
+      main = "Scatterplot of Total Price vs. Quantity")
```



Box Plot –

```
> # Boxplot for Total Price
> boxplot(df$total_price,
+         main = "Boxplot of Total Price",
+         ylab = "Total Price")
```



Correlation –

```
> correlation_matrix <- cor(df[, c("unit_price", "total_price", "quantity")])
> correlation_matrix
             unit_price total_price    quantity
unit_price  1.000000000   0.8360871 0.007142464
total_price 0.836087087   1.0000000 0.541926225
quantity    0.007142464   0.5419262 1.000000000
```

The correlation matrix indicates that there is a strong positive correlation (0.836) between "unit_price" and "total_price." Additionally, there is a moderate positive correlation (0.542) between "total_price" and "quantity." However, "quantity" shows a very weak correlation (0.007) with "unit_price."

Performing Hypothesis –

Chi-Squared Test for independence –

Hypothesis:

H0: There is no association between pizza size and pizza category

H1: There is an association between pizza size and pizza category

```
> # Create a contingency table of pizza size vs. pizza category
> contingency_table <- table(df$pizza_size, df$pizza_category)
> # Perform chi-squared test for independence
> chi_squared_result <- chisq.test(contingency_table)
> # Print chi-squared test result
> print(chi_squared_result)

        Pearson's Chi-squared test

data:  contingency_table
X-squared = 3347.8, df = 12, p-value < 2.2e-16
```

The chi-squared test statistic (X-squared) is calculated to be 3347.8 with 12 degrees of freedom.

The p-value associated with this test statistic is less than 2.2e-16, which is extremely small.

Based on the results of the chi-squared test, we reject the null hypothesis (H0) that there is no association between pizza size and pizza category.

Therefore, we conclude that there is a significant association or dependency between the size of pizzas ordered and their respective categories.

T-Test for Pizza Category (Two-Sample Independent T-Test) –

Hypothesis:

Null Hypothesis (H0): There is no significant difference in the mean total prices of pizzas between category A (Chicken) and category B (Classic)

Alternative Hypothesis (H1): There is a significant difference in the mean total prices of pizzas between category A (Chicken) and category B (Classic)

```
> # T-Test for Pizza Category (Two-Sample Independent T-Test):
> # Extract data for two pizza categories (e.g., Category A and Category B)
> category_a <- df$total_price[df$pizza_category == "Chicken"]
> category_b <- df$total_price[df$pizza_category == "Classic"]
> # Perform two-sample independent t-test
> t_test_result <- t.test(category_a, category_b)
> # Print t-test result
> print(t_test_result)


        Welch Two Sample t-test

data:  category_a and category_b
t = 54.052, df = 23847, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2.912118 3.131269
sample estimates:
mean of x mean of y
 18.11553  15.09384
```

The t-test statistic is calculated to be approximately 54.052, with 23847 degrees of freedom.

The p-value associated with this test statistic is less than 2.2e-16, which is extremely small.

Based on the results of the t-test, we reject the null hypothesis (H0) that there is no significant difference in the mean total prices of pizzas between the "Chicken" and "Classic" categories.

Therefore, we conclude that there is a statistically significant difference in the mean total prices of pizzas between these two categories.

ANOVA Test for Pizza Category –

Hypothesis:

Null Hypothesis (H0): There is no significant difference in the quantity of orders among different pizza categories.

Alternative Hypothesis (H1): There is a significant difference in the quantity of orders among different pizza categories.

```
> # Perform ANOVA test for pizza category
> anova_result <- aov(quantity ~ pizza_category, data = df)
> # Print ANOVA summary
> summary(anova_result)
                  Df Sum Sq Mean Sq F value Pr(>F)
pizza_category     3    0.2 0.05831   2.849  0.036 *
Residuals      48616  995.1 0.02047
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The ANOVA test results show that the p-value associated with the F-statistic is 0.036.

Since the p-value (0.036) is less than the commonly used significance level of 0.05 ($\alpha = 0.05$), we reject the null hypothesis (H0).

Therefore, we conclude that there is evidence to support the alternative hypothesis (H1), indicating that there is a statistically significant difference in the quantity of orders among different pizza categories.

Conclusion –

In this context, if we look on the EDA part, the customer mostly orders Large sized Classic Thai Chicken Pizzas because these pizzas have bought most revenue to the restaurant so the restaurant must focus on improving the quality of the mentioned and the most sales of the pizzas are on Friday in every month. With these insights the restaurant can improve its sales.

Hypothesis Tests –

- **T-Test for Pizza Category:**
  There is a statistically significant difference in mean total prices between the "Chicken" and "Classic" pizza categories.

- **ANOVA Test for Pizza Category:**
  There is a statistically significant difference in the quantity of orders among different pizza categories.

- **Chi-Squared Test for Independence (Pizza Size vs. Pizza Category):**
  There is a significant association between pizza size and pizza category, indicating that customers' choices of pizza size are not independent of pizza category.

References –

Dataset Reference: https://www.kaggle.com/code/kohjerry/analyzing-pizza-sales

Visualization: https://www.geeksforgeeks.org/boxplots-in-r-language/

Hypothesis Testing: https://www.scaler.com/topics/hypothesis-test-in-r/

ChatGPT: https://chat.openai.com/