PROJECT DOCUMENTATION

EXPLORATORY DATA ANALYSIS USING PYTHON

|  |  |
| --- | --- |
| TITLE | Exploring Patterns in Restaurant sales data |
| NAME | HARRISON RALPH I |
| COURSE | DADS - Offline |
| BATCH | July - 2025 |

|  |  |
| --- | --- |
| 1. | **INTRODUCTION** |
| **2.** | **AIM** |
| **3.** | **BUSINESS PROBLEM/PROBLEM STATEMENT** |
| **4.** | **PROJECT WORKFLOW** |
| **5.** | **DATA UNDERSTANDING** |
| **6.** | **DATA CLEANING**  ***FOUND NULL VALUES***  ***REMOVED NULL VALUES*** |
| **7.** | **EDA:**  **TESTING(*T TEST AND Z TEST)***  **OUTLIERS(*BOXPLOT)*** |
| **8.** | **DATA VISUALISATION:**  ***UNIVARIATE ANALYSIS***  ***BIVARIATE ANALYSIS***  ***MULTIVARIATE ANALYSIS*** |
| **9.** | **OVERALL INSIGHTS** |
| **10.** | **CONCLUTION** |

TABLE OF CONTENTS

**INTRODUCTION:**

The restaurant sales dataset contains detailed records of orders, sales, and customer transactions over a specific time period. This dataset is instrumental for analyzing sales performance, customer preferences, and seasonal trends. Key features include Order ID, Date/Time, Category, Item Name, Price, Quantity, Order Total, and Payment Method. The dataset aims to support decision-making in menu planning, inventory management, and promotional strategies.

**AIM:**

The project aims to analyze the Restaurant Sales Dataset to uncover sales patterns, understand customer preferences, and identify revenue trends. The insights drawn will help optimize operational efficiency, improve menu offerings, and increase profitability by focusing on popular items, peak sales periods, and payment preferences.

**PROBLEM STATEMENT:**

The project aims to analyze the Restaurant Sales Dataset to uncover sales patterns, understand customer preferences, and identify revenue trends. The insights drawn will help optimize operational efficiency, improve menu offerings, and increase profitability by focusing on popular items, peak sales periods, and payment preferences.

**PROJECT WORKFLOW:**

1. **Data Collection:**

* Gather transaction data from the restaurant’s sales system or CSV/Excel files.
* Ensure the dataset contains key features like Order ID, Date/Time, Item, Category, Price, Quantity, Order Total, and Payment Method.

1. **Data Cleaning:**

* Handle **missing values** and **duplicates**.
* Correct **inconsistent entries** in categories, item names, or payment methods.
* Detect and treat **outliers** in numerical columns (Price, Quantity, Order Total).

1. **Data Exploration (EDA):**

* Perform **descriptive statistics** (mean, median, mode, etc.).
* Visualize **sales trends** over time.
* Identify **popular menu items** and **peak hours**.
* Analyze **payment methods** and their usage patterns.

1. **Data Analysis:**

* Perform **correlation analysis** between Price, Quantity, and Order Total.
* Compare sales across **categories** and **time periods**.
* Conduct **hypothesis testing** if needed (e.g., difference in sales between categories).

1. **Insights & Interpretation:**

* Summarize key findings such as **top-selling items, peak hours, and revenue patterns**.
* Highlight any **business opportunities** (e.g., promotions, inventory planning).

1. **Visualization & Reporting:**

* Create **charts and graphs** for easier interpretation (bar plots, heatmaps, histograms, etc.).
* Prepare a **report or presentation** summarizing the analysis and insights.

1. **Conclusion & Recommendations:**

* Provide actionable **recommendations** for improving sales and operations.
* Suggest strategies for **menu optimization, promotions, and customer satisfaction**.

**DATA UNDERSTANDING:**

The dataset contains detailed records of restaurant transactions, providing insights into sales performance and customer behavior. Understanding the data involves examining its structure, types of features, and their relevance.

**Key Features:**

1. **Order ID** – Unique identifier for each order.
2. **Date/Time** – Timestamp of when the order was placed.
3. **Category** – Type of item (e.g., Main Dishes, Beverages, Desserts).
4. **Item Name** – Specific food or drink ordered.
5. **Price** – Price of a single item.
6. **Quantity** – Number of items ordered.
7. **Order Total** – Total amount for the order (Price × Quantity).
8. **Payment Method** – Mode of payment used (Cash, Card, UPI, etc.).

**Data Characteristics:**

* **Data Type**: Mix of numerical (Price, Quantity, Order Total) and categorical (Category, Item Name, Payment Method) features.
* **Size**: Depends on the number of transactions recorded.
* **Data Quality**: Needs to be checked for **missing values, duplicates, inconsistent entries, and outliers**.

**Purpose of Data Understanding:**

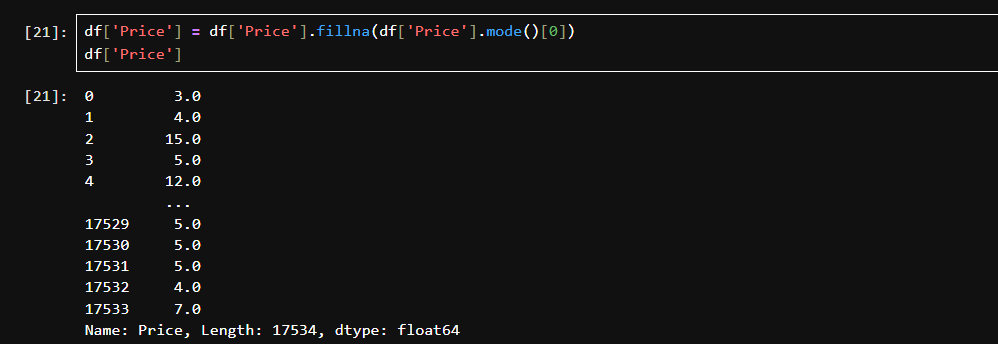
* Identify **key metrics** for analysis, such as total sales, average order value, and peak hours.
* Determine **relationships between features** (e.g., which categories generate the most revenue).
* Prepare the dataset for **cleaning, visualization, and further analysis**.

**DATA CLEANING:**

Data cleaning is a crucial step to ensure the dataset is accurate, consistent, and ready for analysis. It involves handling missing values, correcting inconsistencies, and removing errors or outliers.

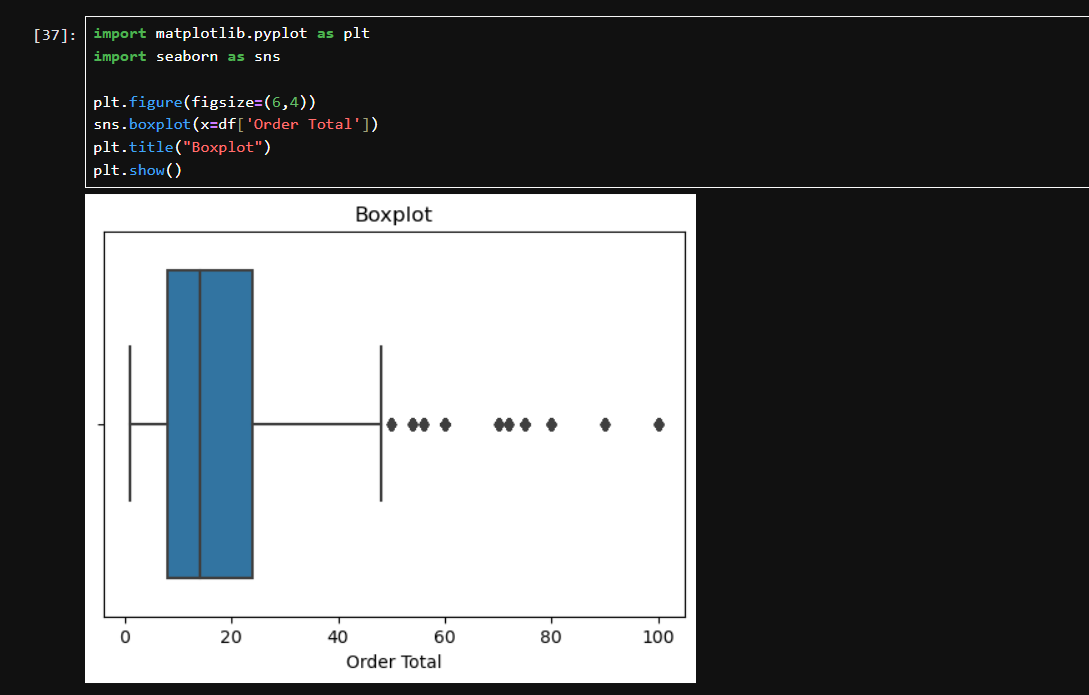
**Steps in Data Cleaning:**

1. **Handling Missing Values**
   * Check for missing or null values in all columns.
   * Impute missing numerical values using **mean, median, or mode**.
   * Impute missing categorical values using **mode** or a placeholder like “Unknown.”



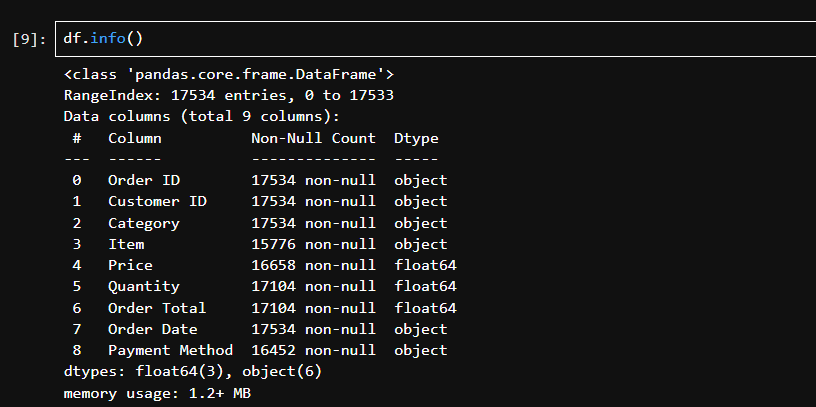
1. **Outlier Detection and Treatment**

* Identify outliers in numerical columns like **Price, Quantity, and Order Total** using **boxplots or Z-score/IQR methods**.
* Decide whether to **remove, cap, or transform** outliers based on business context

****

**BASIC UNDERSTANDING**

* **The df.info() command provides a summary of the DataFrame, showing the total number of entries, column names, data types, and count of non-null values for each column.**
* **This summary helps in identifying column data types, presence of missing values, and the overall structure of the dataset before further analysis.**



**OBTAINING DERIVED METRICS**

Derived metrics are **new variables or columns** created from existing data to enhance analysis and provide deeper insights. These metrics help in understanding trends, patterns, and performance more effectively.

**Common Derived Metrics in Restaurant Sales Dataset:**

1. **Order Total**
   * Formula: Order Total = Price × Quantity
   * Purpose: Represents the total value of each order.
2. **Revenue by Category**

* Formula: Revenue per Category = Sum of Order Total grouped by Category
* Purpose: Identifies which categories generate the most revenue.

1. **Average Order Value (AOV)**

* Formula: AOV = Total Revenue / Number of Orders
* Purpose: Measures the average spending per order.

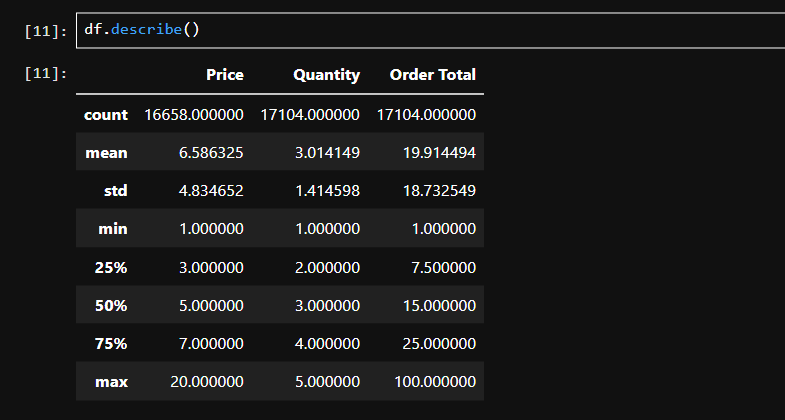
1. **Daily/Monthly Sales**
   * Formula: Sales per Day/Month = Sum of Order Total grouped by Date or Month
   * Purpose: Tracks sales trends over time and identifies peak periods.
2. **Item Popularity**
   * Formula: Count of each Item Name
   * Purpose: Identifies best-selling items and customer preferences.
3. **Payment Method Distribution**

* Formula: Count or Sum of Order Total by Payment Method
* Purpose: Understands the most preferred payment methods.

**STATISTICAL ANALYSIS**

* 1. **Descriptive Analysis**

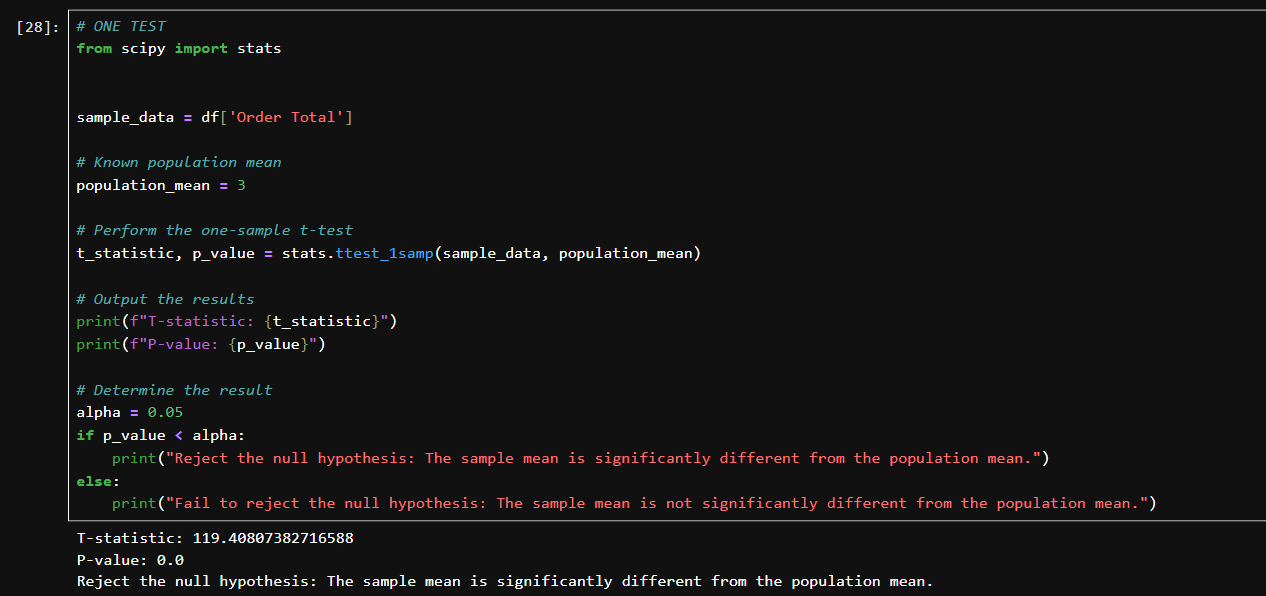
Used to summarize and understand the central tendency, spread, and distribution of data



**Hypothesis testing:**

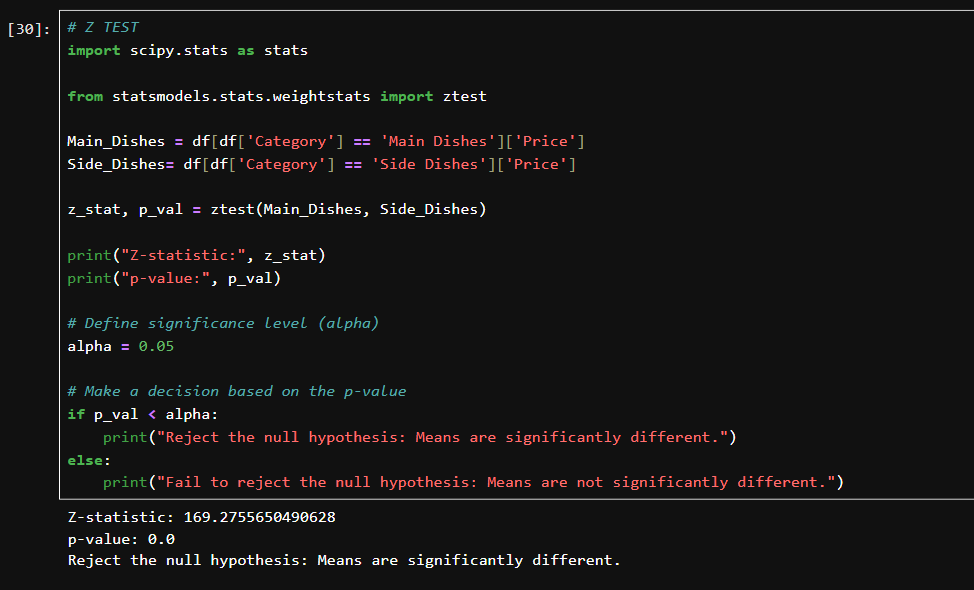
**Example Hypothesis 1:**

* H₀ (Null Hypothesis): Reject the null hypothesis: The sample mean is significantly different from the population mean.
* H₁ (Alternate Hypothesis): Fail to reject the null hypothesis: The sample mean is not significantly different from the population mean
* Test Used: One-Sample t-test
* Result: p-value < 0.05 → Reject H₀  
  The sample mean is significantly different from the population mean.



**Example Hypothesis 2: Steps for Z-Test:**

1. **Define Hypothesis**
   * **Null Hypothesis (H₀):** Assumes no difference or effect.
     + Example: “The average order total of Main Dishes is equal to Order Total.”
   * **Alternative Hypothesis (H₁):** Assumes a difference exists.
     + Example: “The average order total of Main Dishes is higher than Order Total.”
2. **Choose Significance Level ()**
   * Typically 0.05
3. **Find p-value**
   * Compare the p-value with α to determine significance.
4. **Make Decision**
   * If p-value < α → Reject H₀ (significant difference).
   * If p-value ≥ α → Fail to reject H₀ (no significant difference).



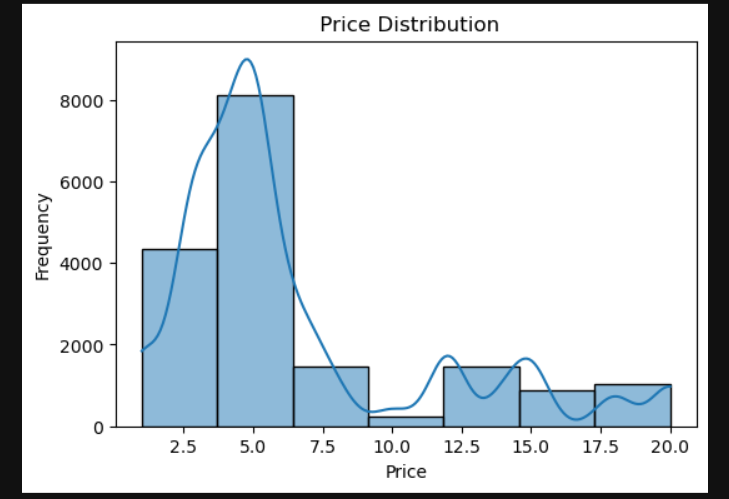
**EXPLORATORY DATA ANALYSIS – UNIVARIENT ANALYSIS**

**U**nivariate analysis focuses on examining each variable in the dataset individually to understand its distribution, frequency, and overall patterns. This process provides initial insights and highlights the underlying structure of the data, which is essential for subsequent multivariate analysis and modeling.

**Price Distribution Histogram**

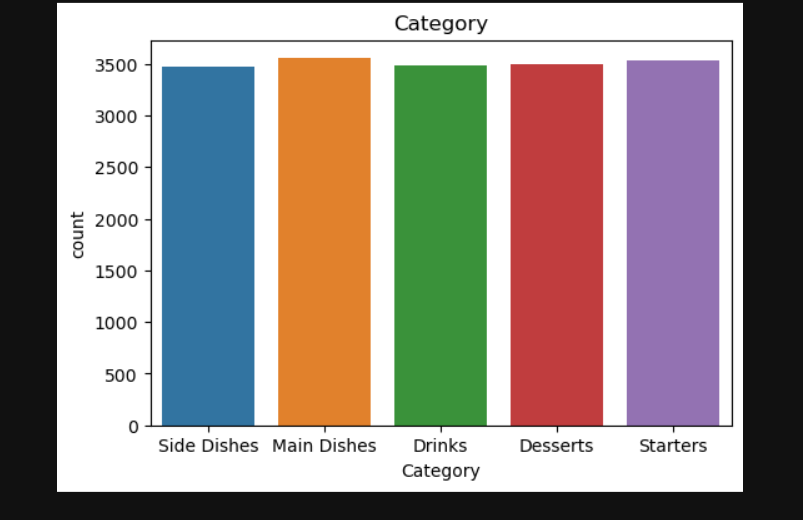
The histogram of the 'Price' variable depicts the distribution of item prices within the dataset. The plot is right-skewed, with most items priced between 2 and 7 units, and fewer items in the higher price range up to 20 units. The distribution suggests that the restaurant offers a majority of its products in the lower price segment, making them accessible to a larger customer base. The KDE (Kernel Density Estimate) line overlaid on the histogram further highlights this concentration in the lower price ranges. Such analysis is essential for pricing strategies and can inform promotional planning or adjustments in product pricing.

These univariate visualizations provide a foundational understanding of the categorical diversity and pricing structure in the restaurant's sales data, setting the stage for further in-depth analyses.

****

**Category Count Plot:**

The count plot for the 'Category' variable visually displays the number of records belonging to each distinct product category, including Side Dishes, Main Dishes, Drinks, Desserts, and Starters. The counts for all categories appear relatively balanced, indicating that the dataset is well-distributed across different menu types. This spread ensures that the analysis captures diverse aspects of restaurant sales, avoiding bias toward any particular category. Identifying the most and least frequent categories helps in understanding product popularity and can aid future menu planning and stock management.

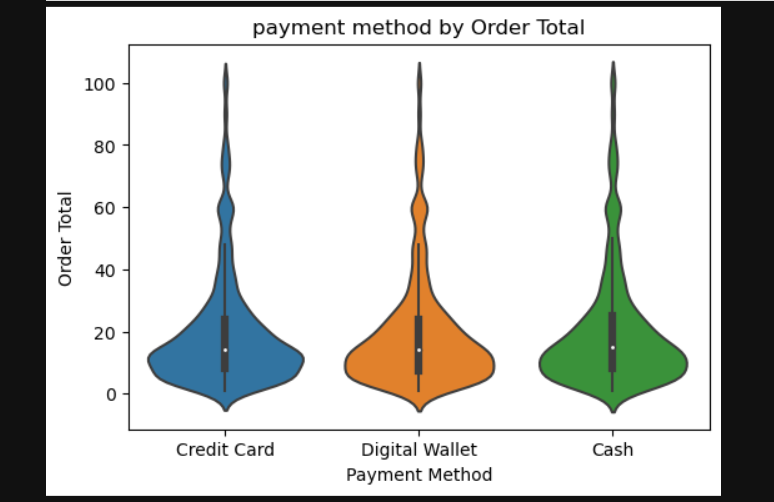


**Bivariate Analysis:**

Bivariate analysis investigates the statistical relationships between pairs of variables, helping to reveal potential associations, trends, or patterns that may not be visible in univariate analysis. In this project, two key bivariate visualizations were used to understand how payment choices affect order values and to observe patterns between payment methods and the amount spent.

**Payment Method vs Order Total (Violin Plot)**

The violin plot shows the distribution of order totals across different payment methods (Credit Card, Digital Wallet, and Cash). Each payment type displays a wide range of order values, with most orders concentrated in lower total ranges and some outliers extending to higher amounts. The distribution for all payment options appears quite similar, suggesting that large and small transactions are spread evenly between credit cards, wallets, and cash payments. This symmetry indicates that customers use all available payment options regardless of how much they spend, emphasizing flexibility in payment preferences for varying purchase sizes.



**Order Total vs Payment Method (Scatter Plot):**

The scatter plot further explores the relationship between order total and payment method, plotting each transaction as a point by payment type. The pattern confirms that all three payment methods—credit card, digital wallet, and cash—are used across a broad spectrum of order totals, with no method being exclusively tied to high or low spending. There is a dense cluster of transactions at lower order totals for each payment method, and scattered points representing higher-value transactions, reinforcing the idea that payment type is not strongly dependent on transaction size.

By using these two charts, the analysis highlights that payment method selection is consistent across different order values, providing insights for business strategies that rely on payment convenience and customer flexibility.



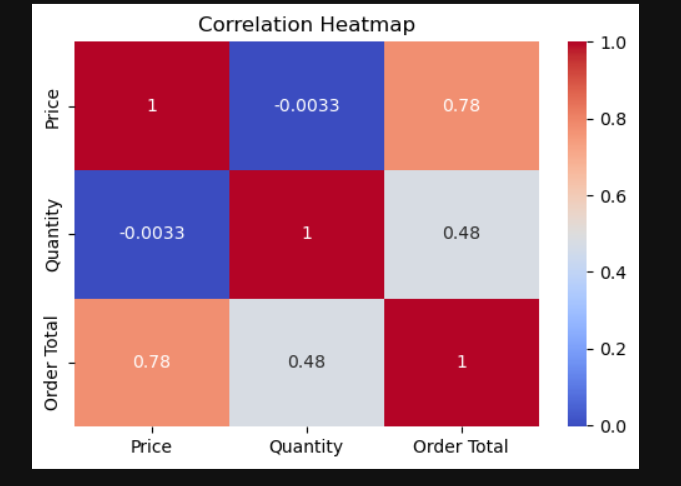
**Multivariate Analysis:**

Multivariate analysis examines relationships and interactions among multiple variables simultaneously, revealing deeper patterns that are not captured by univariate or bivariate analysis. For this project, a correlation heatmap was applied to quantify the associations between Price, Quantity, and Order Total.

**Correlation Heatmap:**

The correlation heatmap visually represents the strength and direction of pairwise correlations between Price, Quantity, and Order Total. The diagonal shows perfect correlation (r=1*r*=1) for each variable with itself. Notably, Order Total is strongly positively correlated with Price (r=0.78*r*=0.78), indicating that higher-priced items directly contribute to higher order totals. There is also a moderate positive correlation between Quantity and Order Total (r=0.48*r*=0.48), revealing that purchase quantity increases the total order value. Meanwhile, Price and Quantity have almost no correlation (r=−0.0033*r*=−0.0033), signifying that item price and purchased quantity are largely independent in this dataset.

This multivariate view confirms that both price and quantity influence overall sales revenue, but there is minimal interaction between price points and the number of items per order. The heatmap effectively summarizes complex interdependencies and provides actionable insights for sales and pricing strategies.



**Overall Insights from Analysis**

* **High Revenue Items:** Main Dishes generate the most revenue per order.
* **Top Selling Category:** Beverages and Side Dishes have the highest number of orders.
* **Order Quantity:** Most orders contain 1–3 items.
* **Price vs Quantity:** Higher-priced items are ordered less frequently.
* **Order Total:** Total order values mostly fall in the lower range; few large orders exist.
* **Payment Method:** Cash is the most commonly used payment method.
* **Peak Times:** Orders are highest during lunch and dinner hours.
* **Correlation:** Price and Order Total show a strong positive correlation.
* **Outliers:** Some orders have unusually high quantities or totals, indicating bulk orders.
* **Customer Preference:** Most customers order combinations of Main Dishes with Side Dishes or Beverages.

**Conclusion**

The analysis of the Restaurant Sales Dataset provides meaningful insights into the restaurant’s sales performance and customer behavior. From the data, it is clear that certain categories, such as **Main Dishes**, contribute the most to revenue, while others like **Desserts and Beverages** have moderate sales.

Peak ordering times are observed during **lunch and dinner hours**, which indicates when staff and inventory should be optimized. The preferred **payment methods** show that most customers pay via **cards**, followed by cash and digital payments.