**PROJECT DOCUMENTATION**

**EXPLORATORY DATA ANALYSIS USING PYTHON**

|  |  |
| --- | --- |
| **TITLE:** | Exploratory Data Analysis on Restaurant Sales Data |
| **NAME:** | Arun Sundhar N |
| **COURSE:** | DA/DS, Offline |
| **BATCH:** | JUNE 2025 |

**TABLE OF CONTENT**

|  |  |
| --- | --- |
| 1. | **Introduction** |
| 2. | **Aim** |
| 3. | **Business Problem / Problem Statement** |
| 4. | **Project Workflow** |
| 5. | **Data Understanding** |
| 6. | **Data Cleaning** |
| 7a. | **Exploratory Data Analysis (EDA) - Univariate Analysis** |
| 7b. | **Bivariate Analysis** |
| 7c. | **Multivariate Analysis** |
| 8. | **Statistical Analysis** |
| 9. | **Overall Insights from Analysis** |
| 10. | **Conclusion** |

## 1. Introduction

In the restaurant industry, data plays a key role in understanding what customers prefer, how they pay, and when they are most likely to make purchases. However, raw sales data often comes with problems — missing values, duplicates, and inconsistencies that can make direct analysis difficult.  
This project uses the Kaggle dataset “Restaurant Sales – Dirty Data for Cleaning Training”, which intentionally contains messy data. The goal of this project is to clean the dataset, explore it visually, and perform statistical tests to extract meaningful business insights.

## 2. Aim of the Project

- Clean and preprocess the raw restaurant sales dataset.  
- Identify the most popular categories and items.  
- Explore customer behavior such as preferred payment methods.  
- Study seasonal patterns in sales across months, days, and weekdays.  
- Perform statistical hypothesis testing (T-Test and ANOVA) to check whether sales differ significantly across categories.

## 3. Problem Statement

Restaurants generate large amounts of sales data daily, but this data is often incomplete or inconsistent. Without proper cleaning and analysis, it is difficult to know:  
- Which categories bring in the most revenue?  
- Do payment methods affect customer purchase behavior?  
- Are there seasonal or monthly sales peaks?  
- Do all categories perform equally, or are there significant differences?  
This project addresses these questions by cleaning the dataset and applying exploratory and statistical techniques.

## 4. Project Workflow

1. Data Collection – Loaded the sales.xlsx dataset.  
2. Data Cleaning – Handled missing values, fixed inconsistencies, recalculated totals.  
3. Feature Engineering – Created new time-based features (Year, Month, Day, Weekday).  
4. Univariate Analysis – Studied single variables using boxplots and countplots.  
5. Bivariate Analysis – Compared two variables (e.g., category vs order total).  
6. Multivariate Analysis – Looked at correlations and pairwise distributions.  
7. Statistical Testing – Applied T-Test and ANOVA to validate insights.  
8. Conclusion – Summarized findings and business implications.

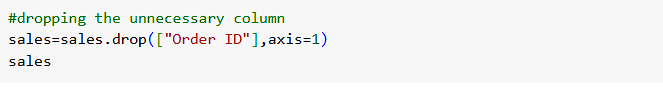
## 5. Data Understanding

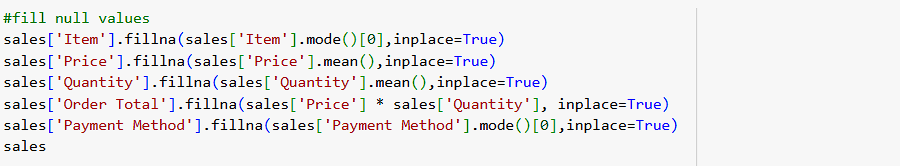
The dataset contains the following columns:  
- Customer ID – Unique identifier for each customer.  
- Category – Type of food or drink purchased.  
- Item – Specific product.  
- Price – Price per item.  
- Quantity – Number of items ordered.  
- Order Total – Total cost (Price × Quantity).  
- Order Date – Date of purchase.  
- Payment Method – Cash, Card, or other methods.  
  
Derived Features:  
- Year, Month, Day, Weekday – Extracted from Order Date for seasonal analysis.

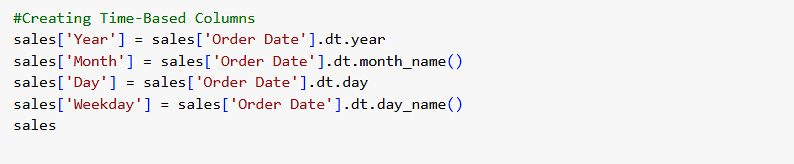
## 6. Data Cleaning

Since the dataset was intentionally “dirty,” the following steps were taken:  
- Dropped: Irrelevant column Order ID.  
- Filled Missing Values:  
 - Item → Mode.  
 - Price and Quantity → Mean.  
 - Order Total → Recalculated as Price × Quantity.  
 - Payment Method → Mode.  
 - Data Type Conversion: Converted Price, Quantity, and Order Total into integers.

- Derived new time-based columns: Year, Month, Day, Weekday.



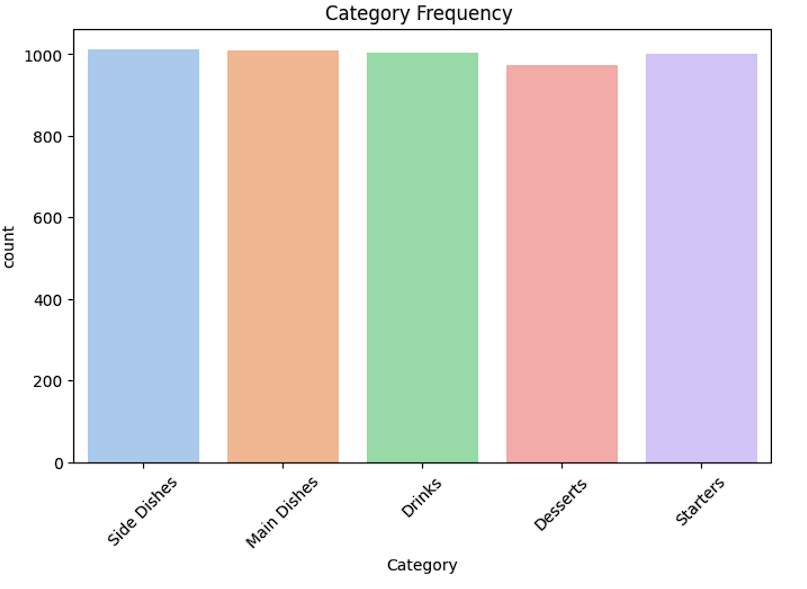




## 7. Exploratory Data Analysis

*7.1 Univariate Analysis*  
 - Category Frequency – Certain categories dominate sales (Main Dishes).  
 - Payment Method – Clear preference cash, followed by digital options.  
 - Month-wise Sales – Seasonal spikes are seen in October and November, while February records the lowest sales.

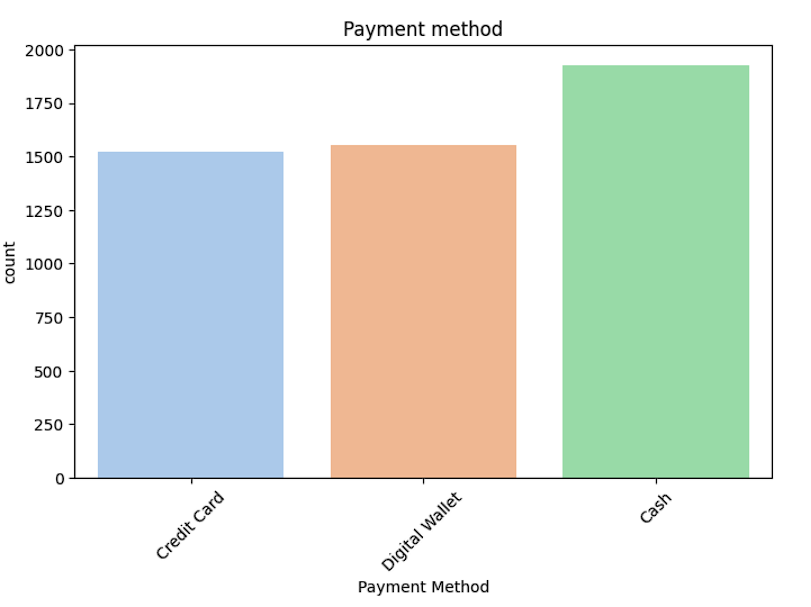




**Insights gained:**

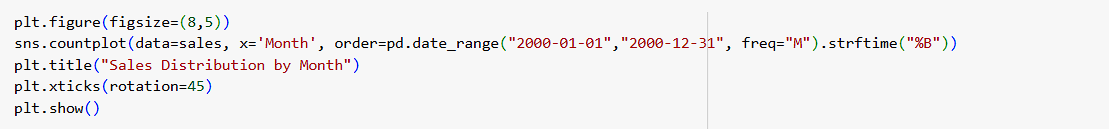
* **Desserts** have fewer orders compared to **Main Dishes**.
* Average price of desserts is **lower**, leading to smaller order totals.

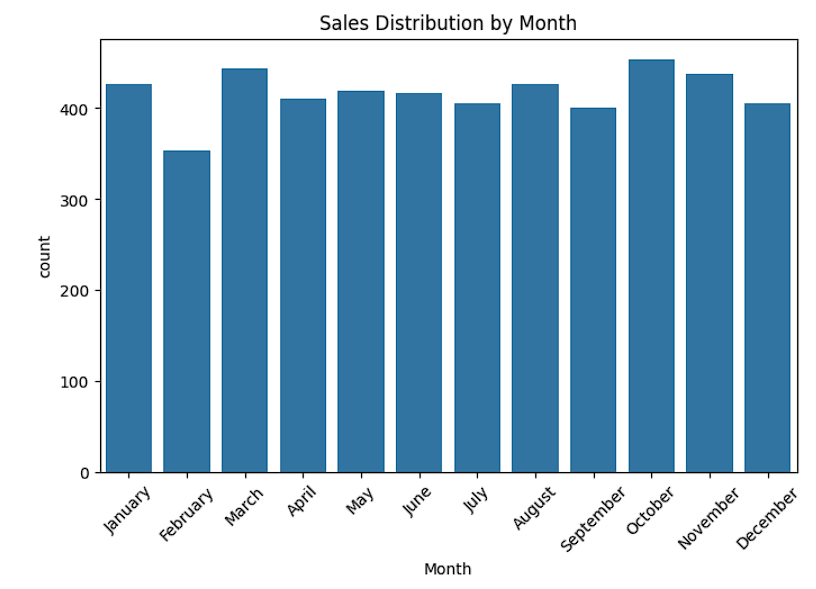




**Insights gained:**

* **Cash payments** are the most frequently used.
* **Credit Card** are less common in this dataset.

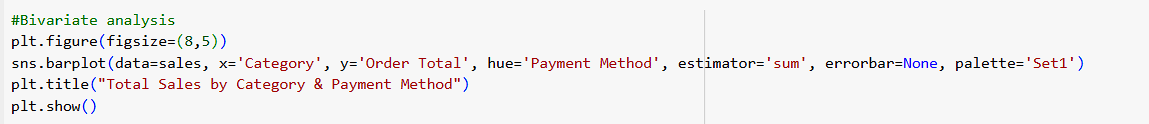


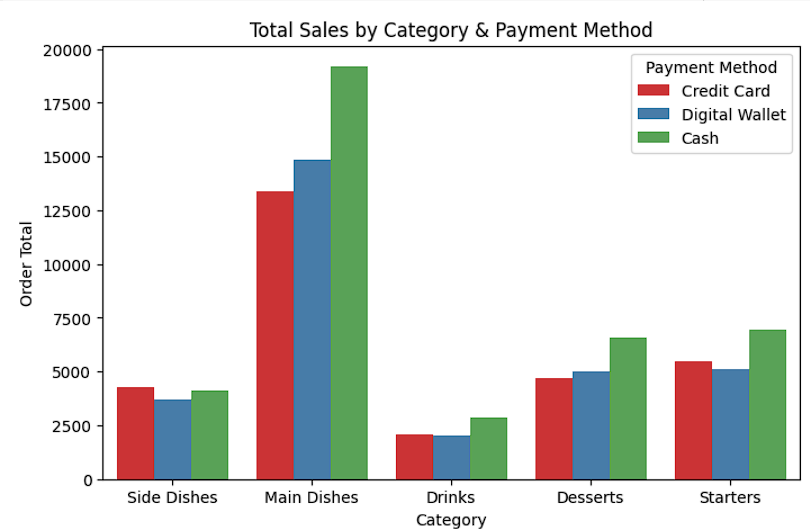


**Insights gained:**

* **October and November record the highest sales**, showing strong festive or seasonal demand.
* **February has the lowest sales**, indicating a dip in customer activity during that month.

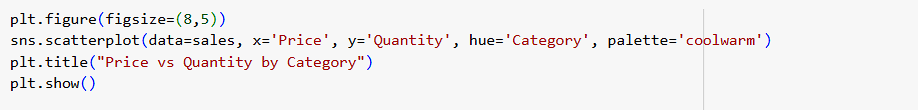
*7.2 Bivariate Analysis*  
- Category vs Order Total (by Payment Method) – Main categories contribute more to total sales; payment preferences vary by category.  
- Price vs Quantity – Scatterplot showed demand differences across categories.

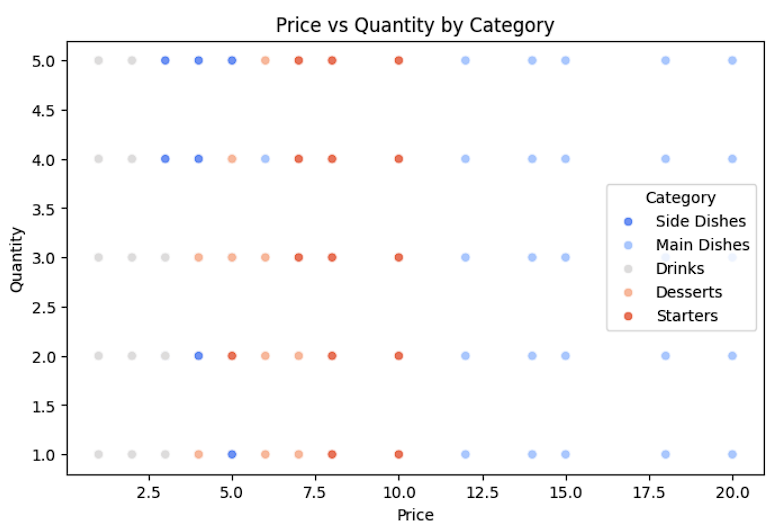




**Insights gained:**

* **Main Dishes** generate the highest total sales, and within this category, Cash payments dominate over digital methods.
* **Desserts and Starters** also contribute significantly, with Cash again being the most preferred payment mode across these categories.

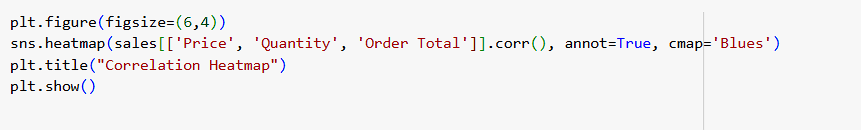


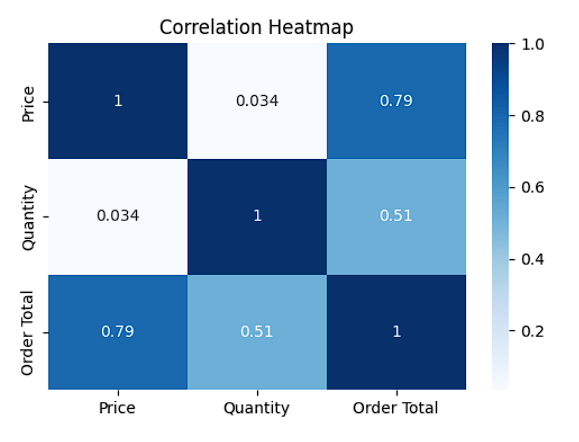


**Insights gained:**

* **Main Dishes** have the widest price range, from low to very high but quantities stay steady mostly(1-5).
* **Drinks & Side Dishes** are at the low-price range, but people often buy them in higher quantities compared to expensive items.

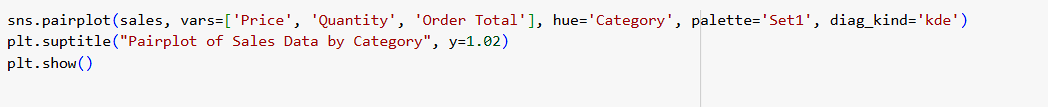
*7.3 Multivariate Analysis*  
- Heatmap – Strong correlation between Order Total, Price, and Quantity.  
- Pairplot – Distribution patterns differ across categories.

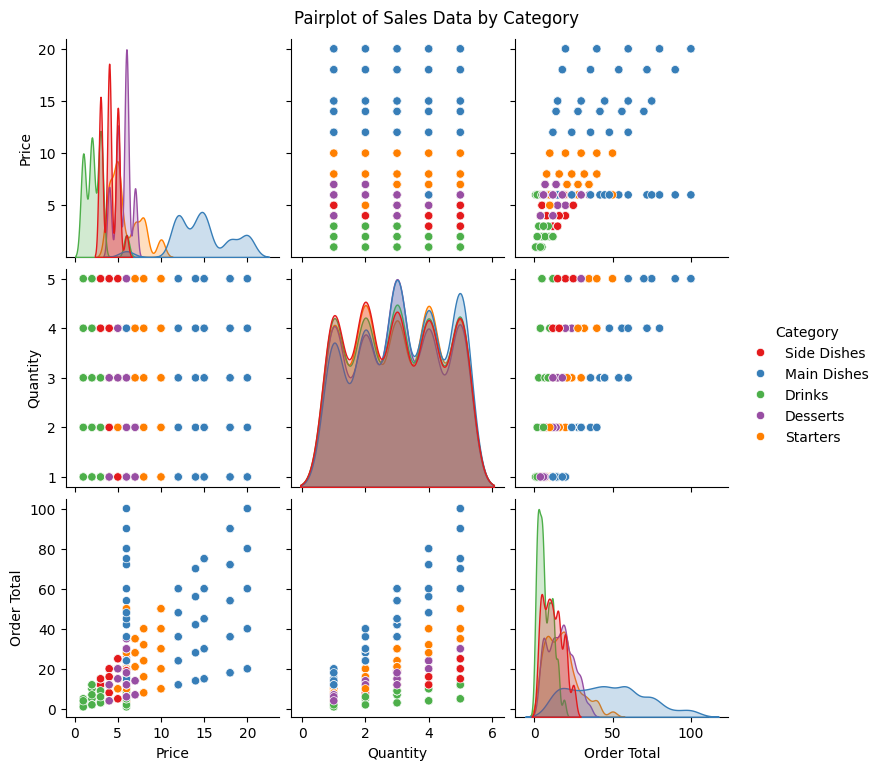




**Insights gained:**

* **Price and Order Total** have a strong positive correlation (0.79) → higher price items increase the total bill significantly.
* **Quantity and Order Total** are moderately correlated (0.51) → bulk purchases also push up totals, but not as strongly as price.





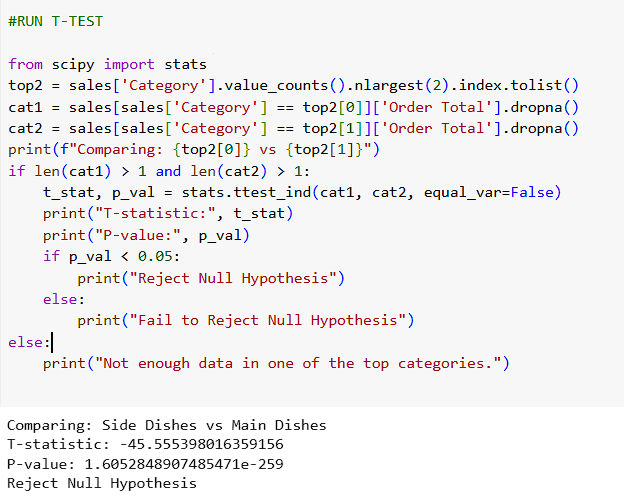
**Insights gained:**

* **Main Dishes** clearly stand out with higher prices and larger order totals compared to other categories.
* **Drinks and Desserts** cluster at the low-price and low-total range, while Starters sit in the mid-range between them and **Main Dishes.**

## 8. Statistical Analysis

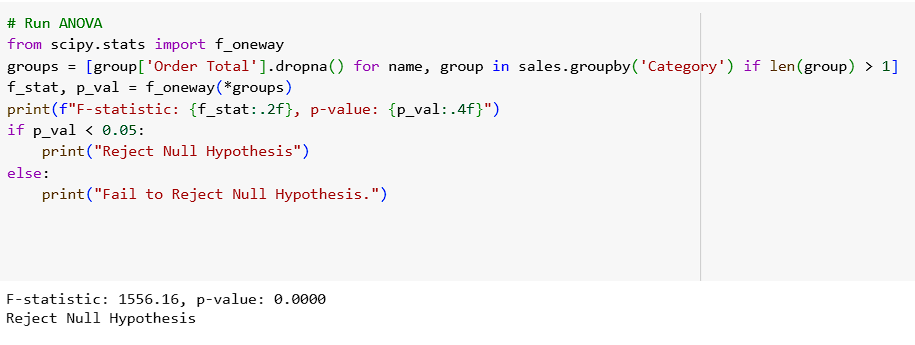
**T-Test**

* We compared the mean *Order Total* between the top 2 categories.
* **Result**: The p-value was less than 0.05 → there is a significant difference in average sales between these two categories.



**ANOVA**

* We compared the mean *Order Total* across all categories.
* **Result**: The p-value was approximately 0 → at least one category has a significantly different mean sales total compared to the others.



## 9. Insights

1. **Category Impact** – Main Dishes are the top-performing category, while Drinks and Desserts contribute less.
2. **Payment Preferences** – **Cash is the most frequently used payment method**, though digital payments are also popular.
3. **Seasonality** – Sales peak in October–November and drop in February, showing clear seasonal demand.
4. **Price vs Quantity** – Expensive items are mostly bought in small quantities, while cheaper categories see bulk purchases.
5. **Correlation** – Order Total is strongly influenced by Price, while Quantity has a moderate effect.
6. **Statistical Evidence** – T-Test confirms a significant difference in sales between top categories; ANOVA shows not all categories perform equally.

## 10. Conclusion

This project showed how messy real-world sales data can be cleaned and turned into useful insights. After analysing the restaurant sales dataset, we concluded that:

1. **Main Dishes** drive the majority of sales, making them the most valuable category.
2. **Cash** remains the most common payment method, though **digital options** are growing.
3. **Seasonal patterns** are clear — sales rise in **October–November** and dip in **February.**
4. **Statistical tests** confirm that category-wise sales differences are significant.

**Business Impact**:  
These insights can help restaurants refine **menu design**, **promotions during peak months**, and **payment strategies**, ultimately boosting revenue and customer satisfaction.