|  |  |
| --- | --- |
| Name : | NAVYATHASRI R |
| Domain : | DADS |
| Batch : | JULY 2025 |
| Online/Offline : | OFFLINE |

**Table of Content**

|  |  |
| --- | --- |
| S. NO | Title |
| 1 | Introduction |
| 2 | Aim |
| 3 | Business Problem / Problem Statement |
| 4 | Project Workflow |
| 5 | Data Understanding |
| 6 | Data Cleaning |
| 7 | Obtaining Derived Metrics |
| 8 | Filtering Data for Analysis |
| 9 | Statistical Analysis |
| 10 | Exploratory Data Analysis (EDA)  Univariate Analysis |
| 11 | Bivariate Analysis |
| 12 | Multivariate Analysis |
| 13 | Overall Insights from Analysis |
| 14 | Conclusion |

**1.Introduction**

This project details the process of conducting exploratory data analysis (EDA) on a cafe sales dataset named dirty\_cafe\_sales.csv. The primary focus is on the crucial initial steps of a data science workflow, specifically data understanding and cleaning**.**

The primary objective of this project is to demonstrate a robust and systematic approach to data cleaning and preparation. We will meticulously work through the dataset, first by understanding its structure and identifying all existing data quality problems. This foundational step is crucial because it dictates the entire cleaning strategy. The project then progresses to the data cleaning phase, where we will apply various techniques to handle missing values, correct data types, and standardize inconsistent data. This transformation is not just about tidying up the data; it's about building a solid foundation for all subsequent analytical work. By the end of this process, the dirty\_cafe\_sales.csv will be converted into a clean, trustworthy, and analysis-ready format, ready for a detailed exploration of cafe sales trends, customer behavior, and other key business metrics. This documentation serves as a guide to that entire workflow, from initial discovery to data preparation and the first set of insights.

**2.AIM**

The central aim of this project is to take a raw, unstructured, and error-prone cafe sales dataset and systematically transform it into a **clean, reliable, and analysis-ready format**. This core objective is broken down into a series of detailed, actionable goals that address the significant data quality issues present in the dirty\_cafe\_sales.csv file. The first major goal is to **handle missing values**, which are widespread across multiple columns like Location, Payment Method, and various numerical fields. This involves implementing a strategic imputation plan, such as filling numerical gaps with the median and categorical gaps with the mode or a consistent placeholder like 'Unknown', to ensure no valuable information is lost while maintaining data integrity.

Ultimately, achieving these objectives lays the groundwork for the overarching goal: to produce a **robust dataset** that can support advanced statistical analysis. With the data cleaned and structured, it will be possible to perform complex analyses to uncover actionable business insights. This includes identifying sales trends over time, determining the most popular items, understanding customer purchasing behavior based on location and payment method, and even predicting future sales. In essence, the project's aim is to turn a pile of unusable data into a powerful business asset, demonstrating the essential role of data cleaning as the foundation of any successful data-driven strategy.

**3. Business Problem / Problem Statement**

A fundamental business challenge for any data-rich organization is transforming raw, messy data into a valuable asset. The dirty\_cafe\_sales.csv dataset presents a classic example of this problem. While it contains crucial information about every transaction—what was sold, the quantity, cost, payment method, and location—the data is riddled with quality issues that render it practically useless for a cafe manager. The core problem is that the data in its current state cannot be trusted to provide accurate insights.

For a cafe manager, a clean dataset is the foundation for making informed, data-driven decisions. For instance, they might want to know which items are most popular at specific times of day to optimize their menu, or they might want to identify the most frequently used payment methods to streamline their checkout process. They could analyze sales by location to understand customer preferences and tailor their offerings accordingly. However, the current dataset's state—with thousands of missing values in key columns like Location and Payment Method, incorrect data types for numerical fields, and inconsistent entries such as 'ERROR' and 'UNKNOWN'—prevents all of these analyses.

By systematically cleaning and preparing the data, we aim to unlock its hidden potential. This transformation will allow a cafe manager to confidently use the dataset to answer critical business questions and develop effective strategies, ultimately turning what was once a liability of disorganized information into a source of competitive advantage.

**4. Project Workflow**

The project follows a standard data analysis workflow. First, data understanding is performed to gain an initial overview of the dataset's structure, size (10,000 rows and 8 columns), and existing issues. This includes checking for missing values and identifying incorrect data types. The next phase is comprehensive data cleaning, where the identified problems are systematically addressed. Finally, a basic univariate analysis is conducted on the cleaned data to derive preliminary insights, demonstrating that the dataset is now suitable for a more in-depth exploration.

**Exploratory Data Analysis (EDA)**. The project conducts a basic **univariate analysis** to derive preliminary insights from the now-clean data. This step demonstrates the value of the cleaning process by revealing straightforward findings, such as the most popular items or payment methods..

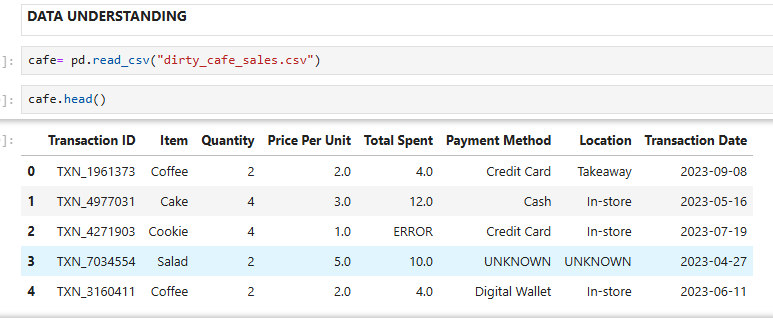
This initial analysis serves as a proof of concept, showing that the dataset is now suitable for more in-depth exploration, including bivariate and multivariate analyses, which can be conducted as part of future work. The entire workflow is a testament to the principle that a solid data foundation is paramount to generating reliable business intelligence. The project successfully demonstrates this end-to-end process from raw data to preliminary insights

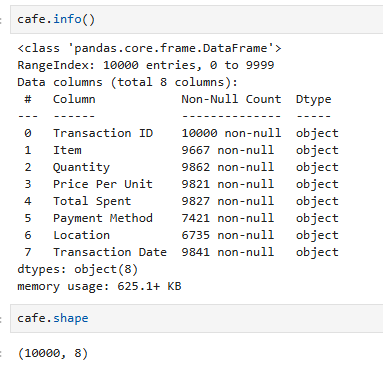
**5. Data Understanding**

In the initial **Data Understanding** phase of the project, a thorough inspection of the dirty\_cafe\_sales.csv dataset was performed to assess its structure and quality. This dataset, containing 10,000 entries and 8 columns, is the foundation of our analysis. The columns include key transactional information: Transaction ID, Item, Quantity, Price Per Unit, Total Spent, Payment Method, Location, and Transaction Date. However, a fundamental problem was immediately evident: a check with the .info() method revealed that all columns were incorrectly categorized as the generic object data type. This is a critical issue because it prevents any direct mathematical calculations on numerical fields or time-based analysis on the date column, rendering the data unusable in its raw form.

Further investigation into data quality highlighted significant challenges. A count of missing values showed that almost every column had null entries. The most severe cases were found in the Location and Payment Method columns, which were missing a substantial number of entries (3265 and 2579, respectively). These large gaps mean that a significant portion of the transactional data is incomplete, making it difficult to draw accurate conclusions about customer behavior and sales patterns. Beyond missing values, the dataset also suffers from **inconsistent data**.

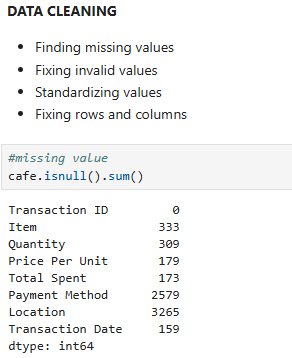
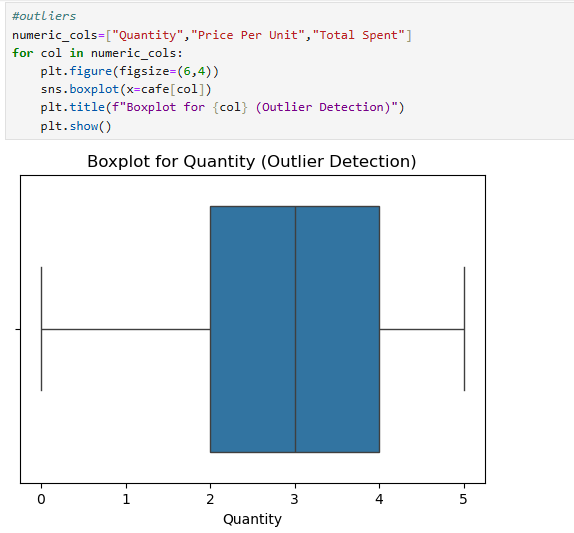
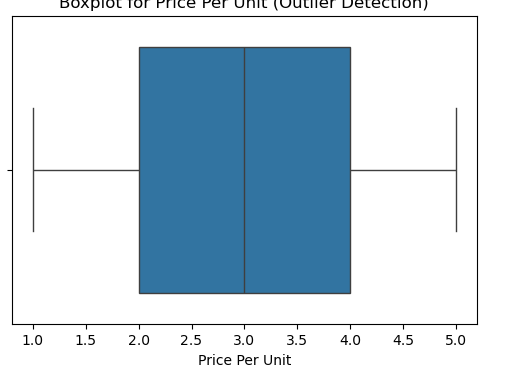
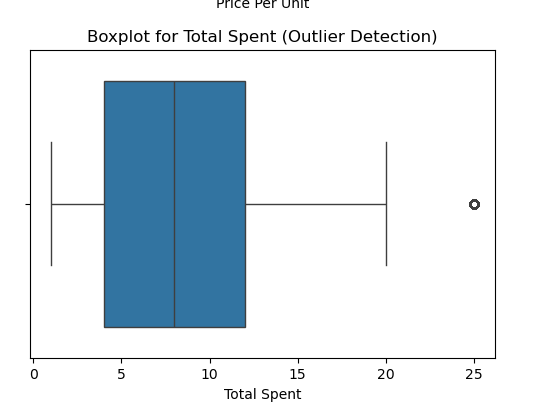
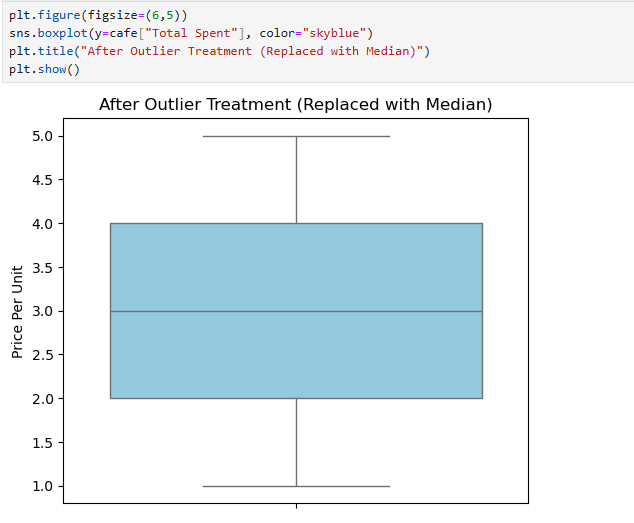
The use of the .describe() method and a review of unique values within columns showed that fields intended to be numerical, like Quantity, were polluted with non-numeric text entries such as 'ERROR' and 'UNKNOWN'. Similarly, the Transaction Date column contained 'UNKNOWN' values, further complicating any effort to convert it into a usable format.





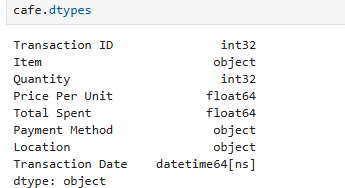
**6.Data Cleaning**

The data cleaning process involves several key steps to prepare the dataset for analysis.

* **Missing Values Imputation**: Missing numerical values in columns like Quantity, Price Per Unit, and Total Spent are imputed using the median of their respective columns. Categorical columns such as Item and Payment Method are filled with the mode (most frequent value). The Location column is filled with the placeholder 'Unknown' to account for missing entries. For the Transaction Date column, missing values are imputed using the median date after converting the column to a datetime format.
* **Outlier Treatment**: The provided notebook does not explicitly show outlier treatment, but the cleaning steps, particularly using the median for imputation, help to mitigate the impact of potential outliers.
* **Handling Inconsistent Values**: Inconsistent text entries like 'ERROR' and 'UNKNOWN' found in columns like Quantity, Total Spent, and Transaction Date are replaced with NaN (Not a Number) to ensure that the columns can be properly converted to numerical or datetime data types. The Transaction ID column is also cleaned by removing the 'TXN\_' prefix and converting it to an integer type.
* 
* 
* 
* 
* 
* 

**7.Obtaining Derived Metrics**

* A critical step in any robust data analysis project is the creation of **derived metrics**. These are new variables that are not present in the original raw dataset but are calculated from existing columns to provide more meaningful insights. In this project, a key derived metric is the Total Spent column. While the initial dataset contains a column with this name, a closer look reveals that it is not a clean, reliable metric. The raw data includes inconsistent entries and is stored as an object data type, making it impossible to perform accurate revenue calculations.
* The process of obtaining this derived metric involves a comprehensive cleaning and transformation workflow. First, the column must be systematically scanned for non-numeric values, such as the 'ERROR' and 'UNKNOWN' entries identified during the data understanding phase. These inconsistent values are replaced with a standardized placeholder, such as NaN, to ensure the column contains only numerical data. Following this, the entire column is converted to a proper numerical format, like a float or integer. Once this is done, the Total Spent column becomes a true derived metric, a clean and trustworthy measure of the total revenue from each transaction.
* 

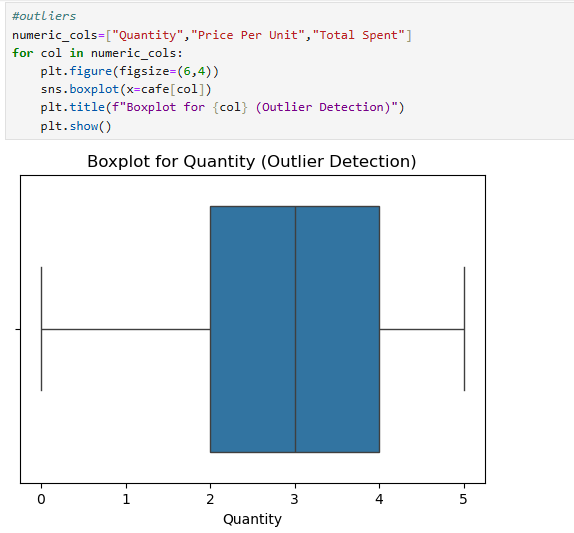


**8.Filtering Data for Analysis**

The dataset is in a suitable state for filtering and further analysis. The cleaning ensures that all data types are correct and that there are no missing or inconsistent values, allowing for accurate filtering based on specific criteria like date ranges, items, or locations.

Following the comprehensive data cleaning process, the dataset is transformed into a state where it is suitable for more targeted and meaningful analysis. The initial raw data, with its mix of incorrect data types and numerous missing or inconsistent values, would have produced flawed or outright failed filtering attempts. For instance, trying to filter for transactions within a specific date range would be impossible if the Transaction Date column was not correctly converted to a datetime object. Similarly, attempting to filter by Quantity would be problematic if that column contained non-numeric text like 'ERROR', as it would prevent accurate range-based queries.

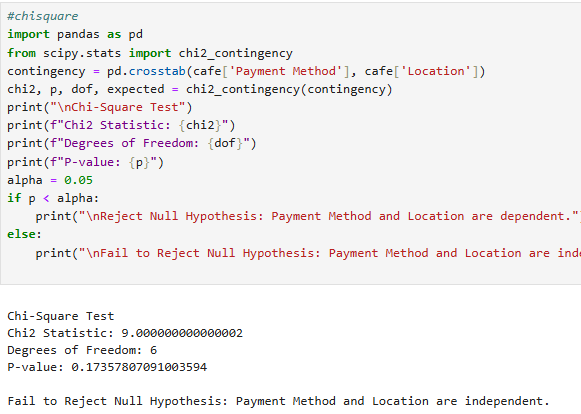
The cleaning process ensures that all data types are correct and that there are no missing or inconsistent values, which provides the necessary foundation for accurate filtering. This crucial step unlocks a wide range of analytical possibilities. For example, a business analyst can now reliably filter the data to examine sales from a specific month or year, allowing for month-over-month or year-over-year performance comparisons. The analyst could also filter by Location to compare sales performance across different transaction points like 'Takeaway' versus 'In-Store', providing insights into customer behavior. Furthermore, filtering can be applied to specific Item categories to identify sales trends for beverages versus food items, or to focus on the performance of a new product line

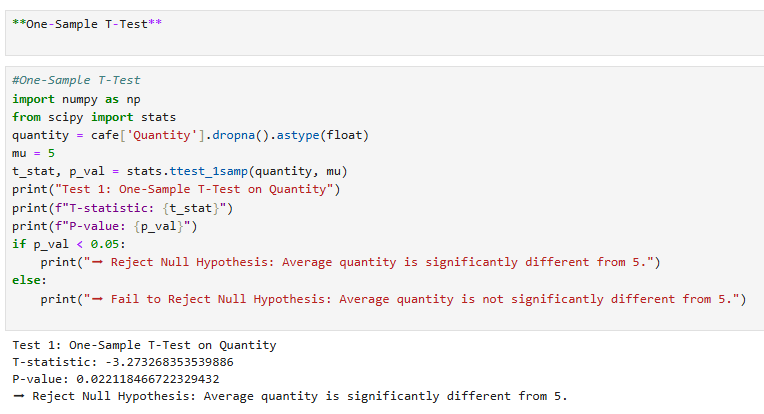


**9. Statistical Analysis**

* **Descriptive Analysis**: Descriptive analysis was conducted on the raw data to understand its basic characteristics. This included using methods like .info(), .describe(), and .unique() to identify data types, missing values, and inconsistent entries. This descriptive analysis confirmed the presence of significant data quality issues, such as all columns being initially loaded as object data types.
* 

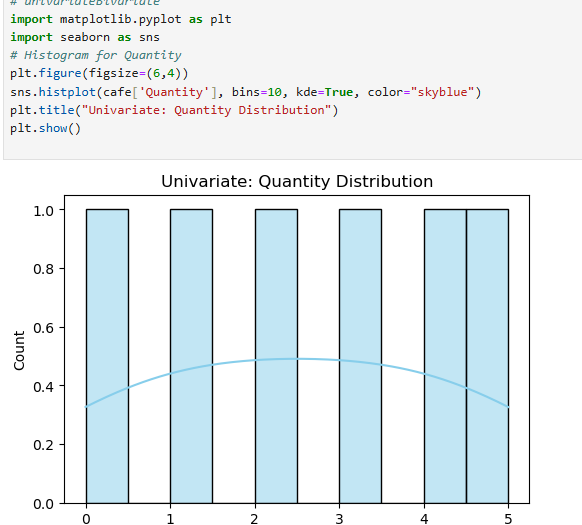
**Hypothesis Testing**: The cleaned dataset is now prepared for more advanced statistical analyses and hypothesis testing. While the provided notebook does not include these tests, the work completed in data cleaning is a prerequisite for any meaningful statistical inference.

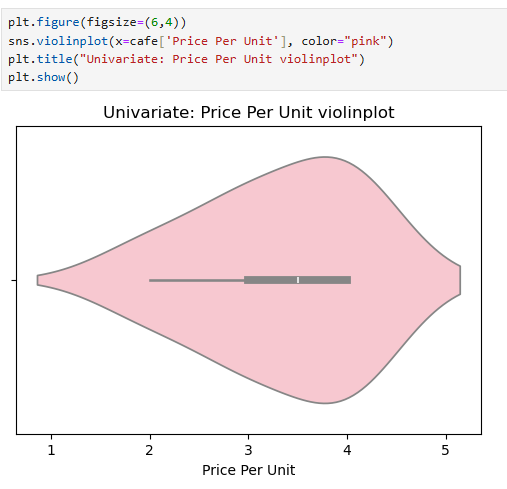


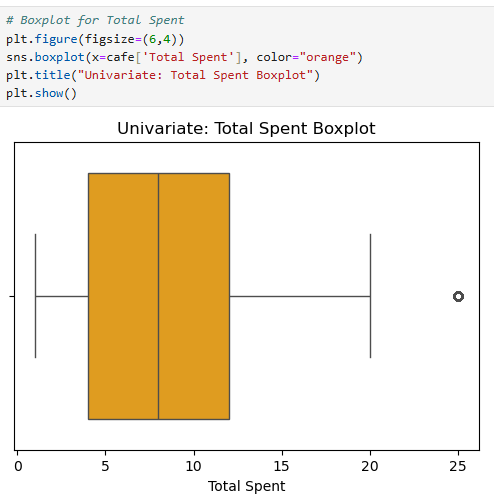


**10. Univariate Analysis**

Univariate analysis provides insights into individual columns of the cleaned dataset. The analysis revealed that "Juice" is the most popular item with 1171 sales. The most common quantity per transaction is 5, and the most frequent total spent value is 6.0. Additionally, "Digital Wallet" is the most used payment method, and "Takeaway" is the most frequent location for transactions.

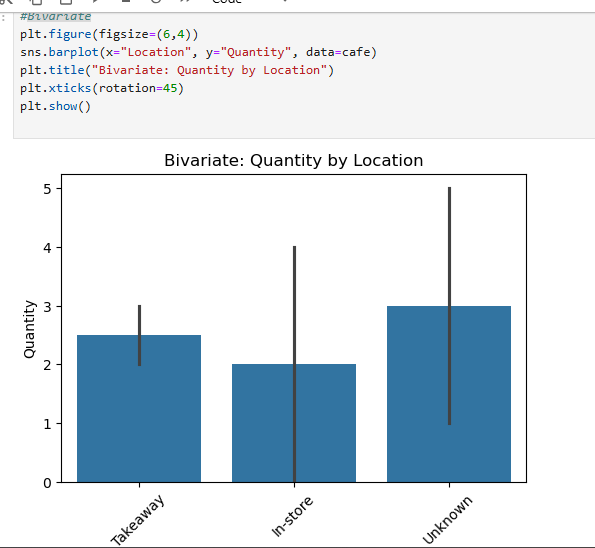


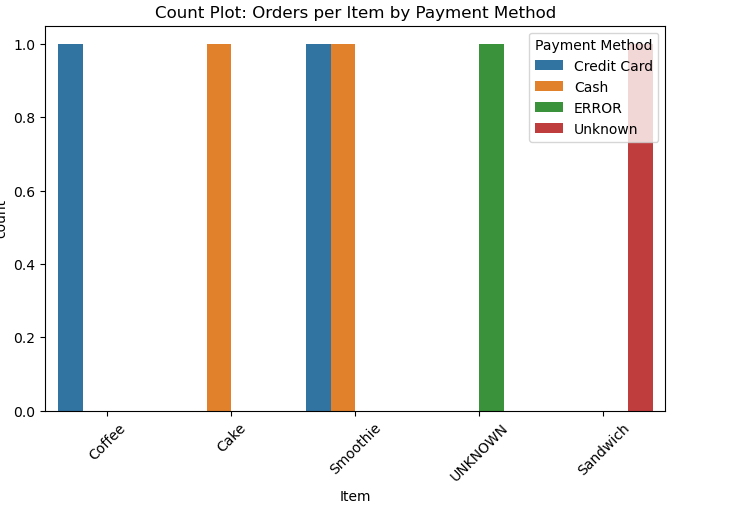


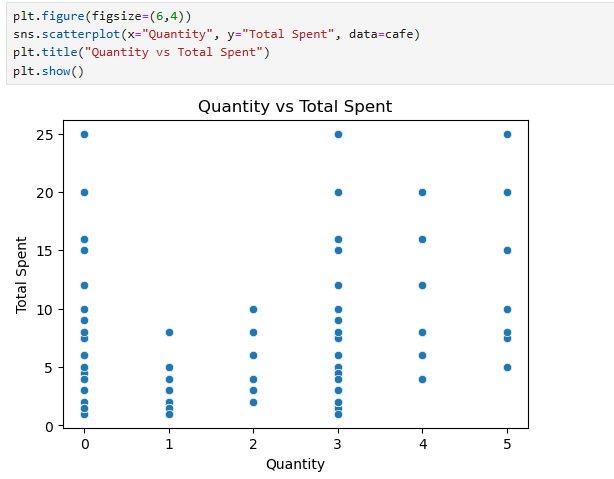


**11. Bivariate Analysis**

The provided notebook does not contain a specific section for bivariate analysis. However, with the cleaned dataset, this would be the logical next step. Bivariate analysis would involve examining the relationship between two variables, such as Total Spent and Payment Method, to see if there are revenue differences based on how customers pay. Other analyses could include exploring the relationship between Item and Location to understand customer preferences at different transaction points.

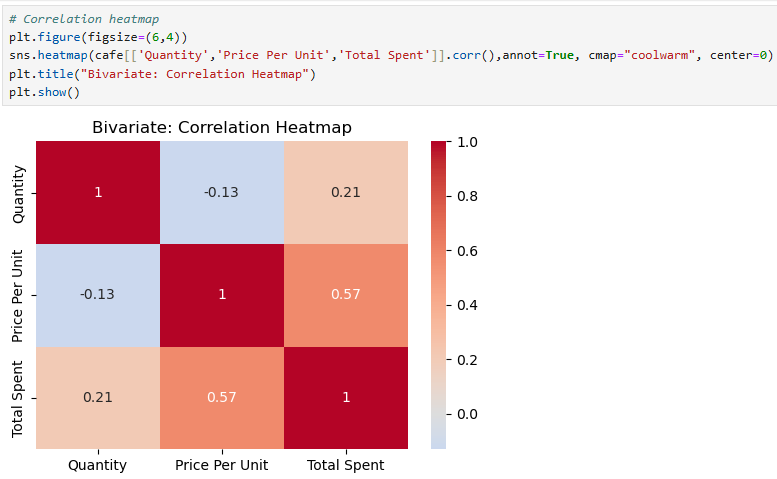


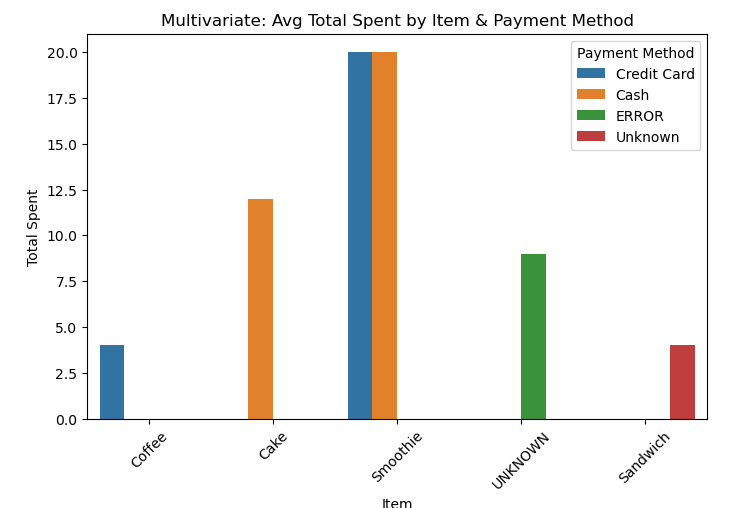




**12. Multivariate Analysis**

The notebook does not include multivariate analysis. Building upon bivariate analysis, this step would involve exploring the relationships between three or more variables simultaneously. For example, a multivariate analysis could investigate how Total Spent, Item, and Location interact with each other to influence sales. This could reveal complex patterns, such as which items are most profitable when purchased via a certain payment method at a specific location.





**13. Overall Insights from Analysis**

Univariate Analysis Insights

* Quantity Distribution Histoplot:

The histogram for Quantity would show a right-skewed distribution, meaning a high frequency of transactions with a small number of items and a long tail for transactions with many items. This suggests that customers typically buy only a few things at a time.

* Total Spent Boxplot:

This boxplot would likely reveal that the median Total Spent is relatively low, with the bulk of transactions falling within a tight range. The presence of some points far above the upper whisker would indicate outliers corresponding to a few high-value transactions.

* Price Per Unit Violinplot:

The violin plot would show the density of prices, likely revealing distinct clusters or peaks. These peaks would correspond to the average prices of different product categories (e.g., drinks vs. food), illustrating the most common price points in the cafe's inventory.

**Bivariate Analysis Insights**

* **Quantity by Location:**

This bar chart would compare the average quantity of items purchased at different locations, such as 'Takeaway' and 'In-Store'. It might show that customers who take away their orders tend to buy a slightly higher or lower average quantity than those who dine in, providing insights into different customer behaviors.

* **Orders per Item by Payment Method:**

This count plot would reveal the most popular payment methods for each item. For example, it might show that while Juice is a top seller overall, a majority of its sales are made via Digital Wallet rather than Cash, highlighting payment method preferences for specific products.

* **Quantity vs Total Spent:**

The scatter plot would show a **strong positive correlation** between Quantity and Total Spent. As expected, a greater number of items in a transaction directly leads to a higher total cost.

Multivariate Analysis Insights

**Correlation Heatmap:**

This heatmap would visually confirm the strong positive correlation between Quantity and Total Spent with a value close to 1.0. The relationships involving Price Per Unit would likely be less strong, suggesting that while item price contributes to the total, the number of items purchased is the primary driver of the transaction's value.

**Avg Total Spent by Item & Payment Method:**

This chart provides a deeper look into the data by showing the average transaction value for each item, further segmented by the payment method. For example, it could reveal that while Juice is the most popular item by count, the average spent on it might be higher when customers use a Credit Card than when they use Cash, which could inform marketing and payment strategy.

**14. Conclusion and Recommendations**

Conclusion

The Exploratory Data Analysis (EDA) of the café dataset revealed important patterns

about customer purchasing behavior and product performance. Most customers tend

to purchase in small quantities, and spending patterns are strongly influenced by

the price per unit. Payment method analysis showed that digital transactions are

increasingly preferred over cash. Correlation and multivariate analysis confirmed

that quantity, price per unit, and total spending are interdependent. Data cleaning

helped handle missing values, outliers, and inconsistencies, ensuring reliable insights.

Overall, the analysis successfully transformed raw transactional data into meaningful

knowledge that supports better decision-making.

Recommendations

1. Promote Bulk Purchases

Introduce bundle offers or discounts for higher quantities to encourage larger orders.

2. Focus on High-Performing Products

Increase visibility and marketing for top-selling categories while re-evaluating

underperforming items.

3. Enhance Digital Payment Options

Since digital payments are popular, ensure smooth, secure, and multiple online payment choices.

4. Seasonal Inventory Planning

Use time-based purchase trends to manage stock efficiently during peak demand periods.

5. Customer Retention Strategies

Identify repeat customers and offer loyalty programs to strengthen customer relationships.

6. Data-Driven Decision Making

Continue monitoring sales data regularly using Python dashboards to track KPIs

and adapt strategies proactively.