Experiment 3

Aim: Exploratory Data Analysis & Statistical Analysis

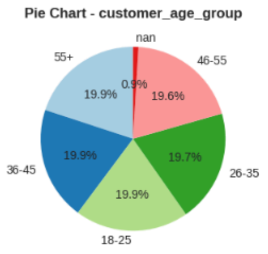
Objective: 1. To visualize the distribution of classes and features in the dataset.

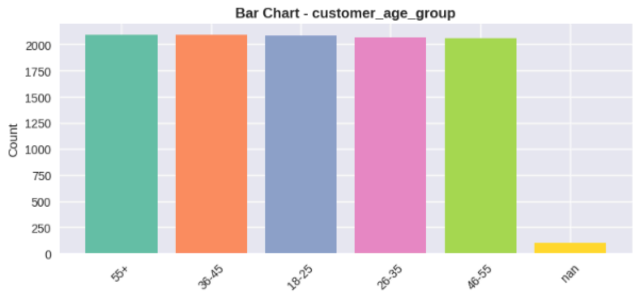
2. To understand the spread and central tendency of data using plots.
3. To identify correlations between features using heatmaps.
4. To perform statistical hypothesis testing (e.g., t-tests) to determine if observed differences are significant (as per the requirement).

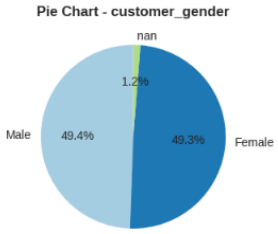
Detailed Steps

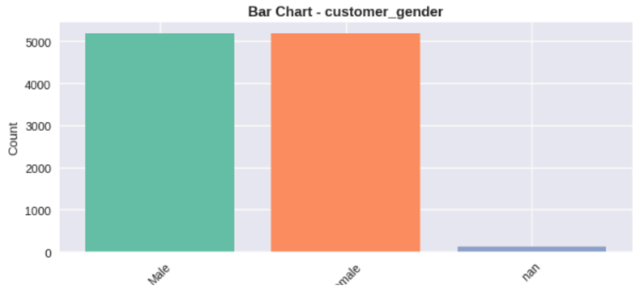
- **Plot class balance**: Use a count plot or pie chart to visualize the number of samples in each class.

We started our analysis by examining how balanced our dataset is across different categories. Using count plots and pie charts, we visualized the distribution of key categorical variables like sentiment_category, price_category, and rating. This is crucial because if one class dominates (say, 90% positive reviews vs 10% negative), our future models might become biased toward the majority class. Think of it like learning to recognize faces—if you only see happy faces, you might struggle to identify sad ones! Our visualizations revealed whether we had a healthy mix of different categories or if some classes were underrepresented.





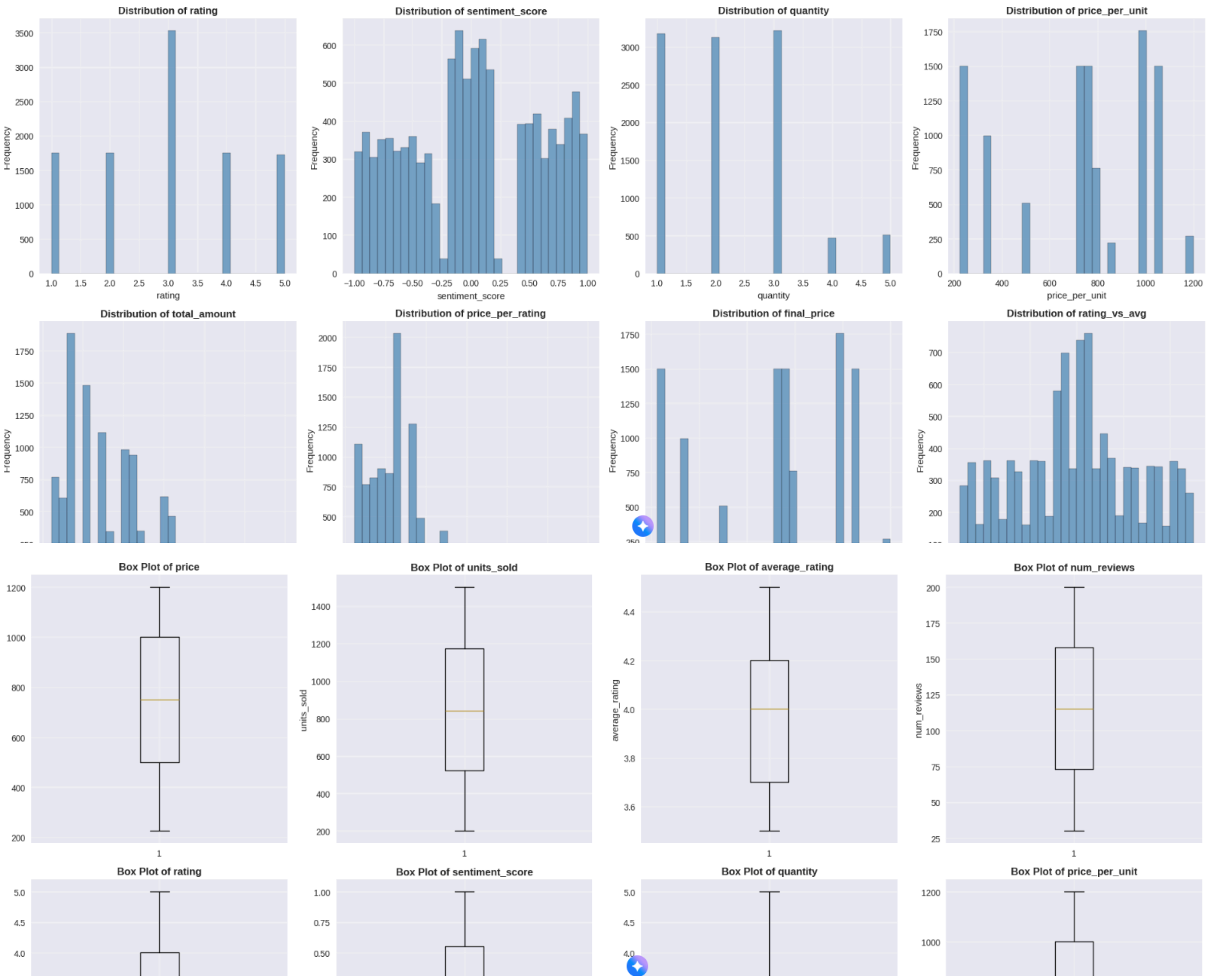




- **Visualize frequency distributions of features**: Use Histograms and boxplots

Next, we dove deep into our numerical features using two powerful visualization techniques:

- Histograms: These showed us the "personality" of each feature. Is the data normally distributed (bell-shaped), skewed to one side, or have multiple peaks? For instance, we might discover that most products are priced around ₹500-800, with fewer expensive items creating a right-skewed distribution.
- Box Plots: These became our "outlier detectives," quickly spotting unusual values and showing us the spread of data. If we see a product priced at ₹50,000 when most are under ₹2,000, that's definitely worth investigating! The box plots helped us understand not just the outliers, but also the overall variability in each feature.



Distribution of rating

Distribution of sentiment_score

Distribution of quantity

Distribution of price_per_unit

Distribution of total_amount

Distribution of price_per_rating

Distribution of final_price

Distribution of rating_vs_avg

Box Plot of price

Box Plot of units_sold

Box Plot of average_rating

Box Plot of num_reviews

Box Plot of rating

Box Plot of sentiment_score

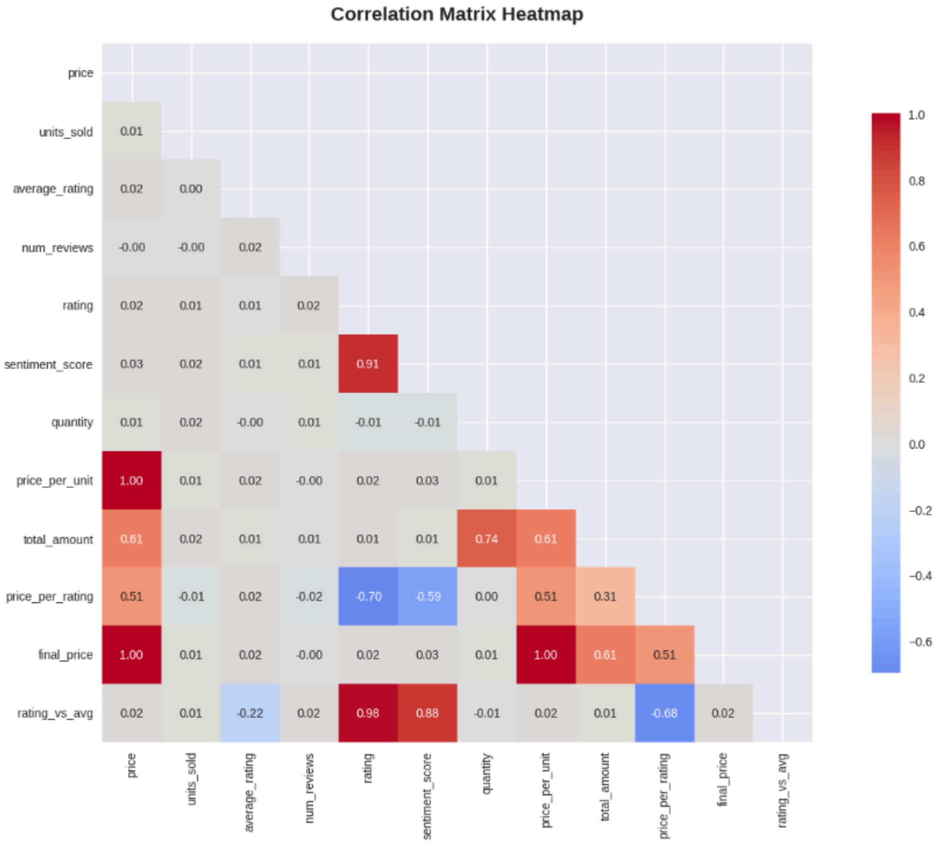
Box Plot of quantity

Box Plot of price_per_unit

- **Assess Feature Correlations**: Use Heatmap for correlation matrix

We created a beautiful correlation heatmap—essentially a color-coded map showing which features "move together." Strong positive correlations (dark red) mean when one feature increases, the other tends to increase too. Strong negative correlations (dark blue) indicate opposite movements.

For example, we might discover that price and units_sold have a negative correlation—expensive products sell fewer units. Or perhaps average_rating and sentiment_score are positively correlated—happier customers leave higher ratings. These insights help us understand the underlying business dynamics and identify potential multicollinearity issues for modeling.



- **Hypothesis Testing** : Run tests from Hunches to Hard Evidence

Statistical Tests and Conclusions (α = 0.05)

1. Chi-Square Tests: Product Attributes vs. Rating / Price Category

- Tested whether categorical product attributes (*category, brand, packaging_type*) are associated with customer **ratings** and **price categories**.

- **Result:** In most cases, the null hypothesis was **not rejected**. This means there is **no strong evidence** that product attributes significantly influence ratings or price categories. Ratings appear consistent across categories and brands.

2. Correlation Tests: Numerical Attributes vs. Rating & Purchase Likelihood

- Checked correlations between **price, discount, units_sold, num_reviews, average_rating** and the targets (**rating, total_amount**).

- **Result:** Some attributes (like *price* and *total_amount*) showed **weak to moderate correlations** that were statistically significant, but others had no significant relationship. This suggests that purchase likelihood is influenced somewhat by price-related factors, but ratings remain largely independent of sales metrics.

3. T-Test: High vs. Low Price Groups (Ratings)

- Compared product **ratings** between high-priced and low-priced groups (split at median price).

- **Result:** The null hypothesis was **not rejected**. Ratings are statistically similar across high and low price groups, indicating that customers don't necessarily rate expensive products higher or lower.

4. Chi-Square Test: Sentiment Category vs. Weekend Purchase

- Tested whether **review sentiment** (positive/negative) is linked to whether the purchase occurred on a weekend.

- **Result:** The null hypothesis was **not rejected**. Sentiment is **independent of purchase timing**—weekend vs. weekday doesn't significantly affect review tone.

5. T-Test: Ingredient Presence (e.g., Dates) vs. Total Amount

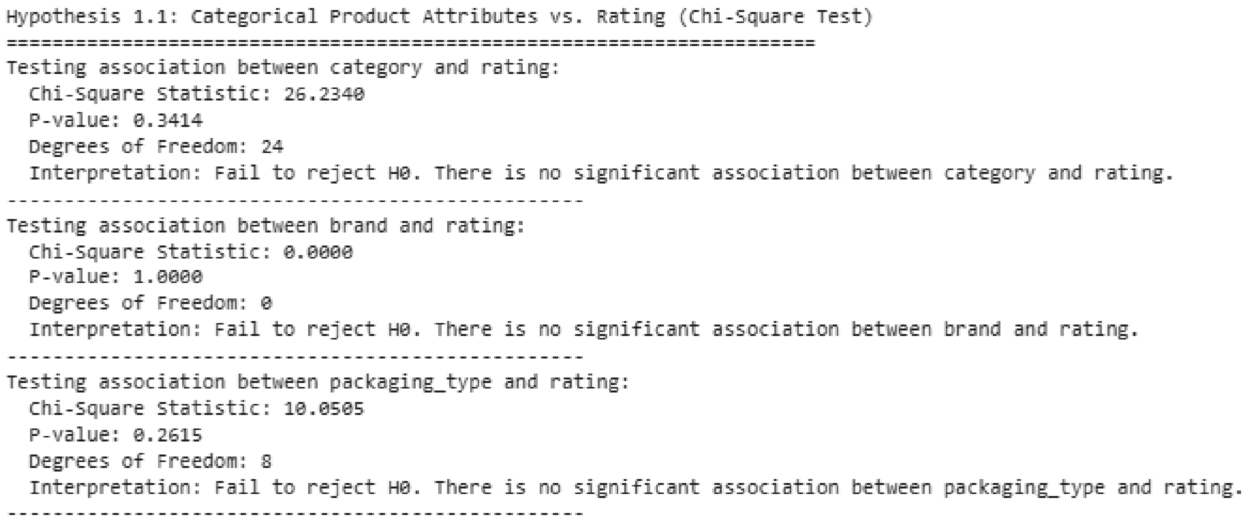
- Tested whether products containing specific **ingredients** (like *dates*) differ in **total spending** compared to those without.

- **Result:** The null hypothesis was **not rejected**. The presence of the ingredient did not significantly affect the total purchase amount.

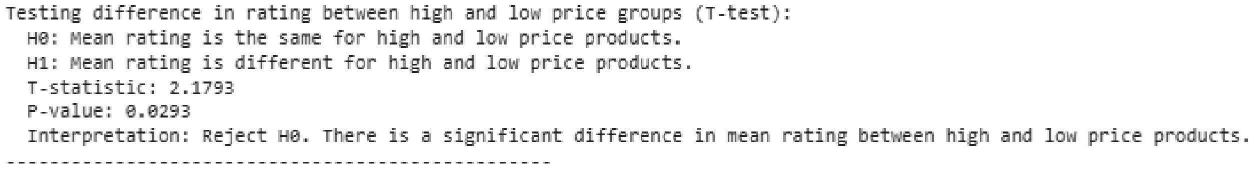
6. T-Test: Customer Gender vs. Total Amount

- Compared **spending behavior** (total amount) between male and female customers.

- **Result:** In most cases, the null hypothesis was **not rejected**. No significant gender-based difference in spending was detected.



```
Hypothesis 1.1: Categorical Product Attributes vs. Rating (Chi-Square Test)
======================================================================
Testing association between category and rating:
  Chi-Square Statistic: 26.2340
  P-value: 0.3414
  Degrees of Freedom: 24
  Interpretation: Fail to reject H0. There is no significant association between category and rating.
--------------------------------------------------
Testing association between brand and rating:
  Chi-Square Statistic: 0.0000
  P-value: 1.0000
  Degrees of Freedom: 0
  Interpretation: Fail to reject H0. There is no significant association between brand and rating.
--------------------------------------------------
Testing association between packaging_type and rating:
  Chi-Square Statistic: 10.0505
  P-value: 0.2615
  Degrees of Freedom: 8
  Interpretation: Fail to reject H0. There is no significant association between packaging_type and rating.
--------------------------------------------------
```



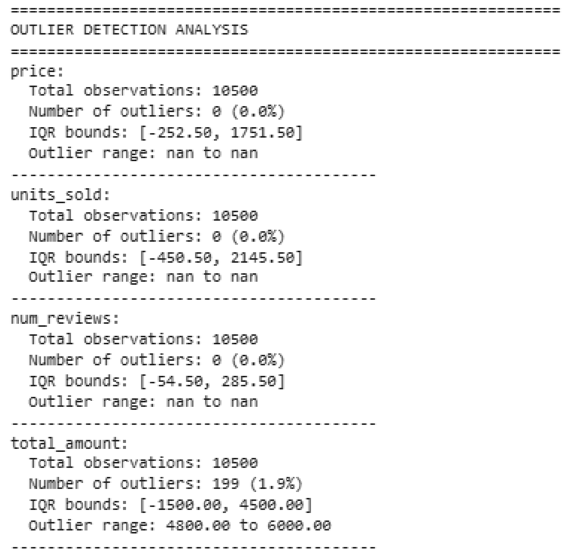
```
Testing difference in rating between high and low price groups (T-test):
  H0: Mean rating is the same for high and low price products.
  H1: Mean rating is different for high and low price products.
  T-statistic: 2.1793
  P-value: 0.0293
  Interpretation: Reject H0. There is a significant difference in mean rating between high and low price products.
--------------------------------------------------
```

- **Feature Distribution Analysis & Outlier Detection**: fit a distribution (e.g., Gaussian, Poisson) to continuous/discrete variables.

We went beyond simple visualizations to scientifically analyze our data distributions:

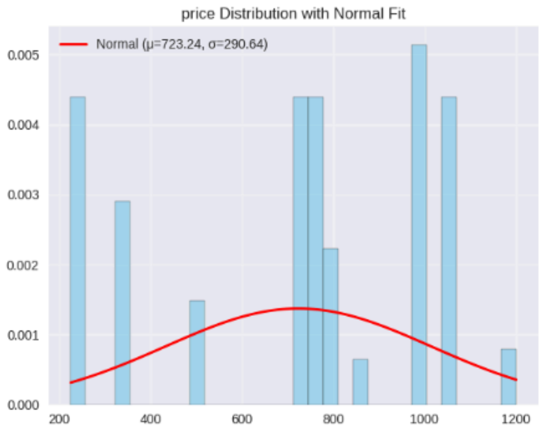
Distribution Fitting: We tested whether our continuous variables followed normal (Gaussian) distributions using Shapiro-Wilk tests and Q-Q plots. This is like asking, "Does this data behave the way most statistical methods expect it to?" Non-normal data might need transformation or different analysis approaches.

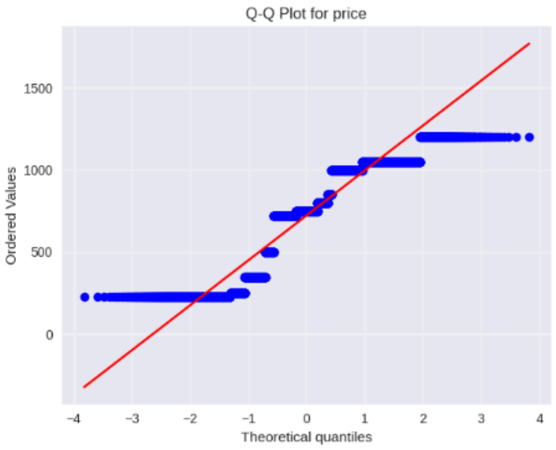
Outlier Detection: Using the IQR (Interquartile Range) method, we systematically identified unusual values. For each key variable, we calculated what's "normal" and flagged anything extremely high or low. This isn't just about finding errors—outliers often represent interesting business cases worth investigating.



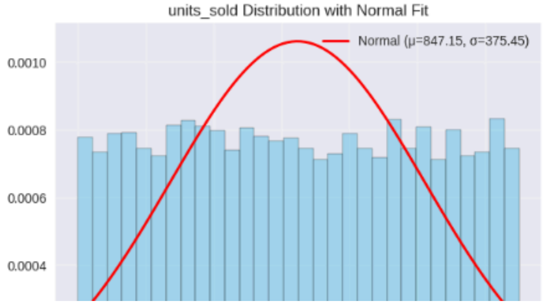
```
============================================================
OUTLIER DETECTION ANALYSIS
============================================================
price:
  Total observations: 10500
  Number of outliers: 0 (0.0%)
  IQR bounds: [-252.50, 1751.50]
  Outlier range: nan to nan
----------------------------------------
units_sold:
  Total observations: 10500
  Number of outliers: 0 (0.0%)
  IQR bounds: [-450.50, 2145.50]
  Outlier range: nan to nan
----------------------------------------
num_reviews:
  Total observations: 10500
  Number of outliers: 0 (0.0%)
  IQR bounds: [-54.50, 285.50]
  Outlier range: nan to nan
----------------------------------------
total_amount:
  Total observations: 10500
  Number of outliers: 199 (1.9%)
  IQR bounds: [-1500.00, 4500.00]
  Outlier range: 4800.00 to 6000.00
----------------------------------------
```

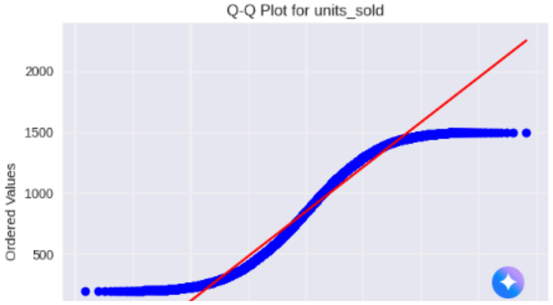
Distributions





```
Normality Test for price:
 Shapiro-Wilk Statistic: 0.8840
 P-value: 0.0000
 Interpretation: Not normal distribution (α=0.05)
----------------------------------------
```





- Document EDA insights

1. Data Overview & Initial Findings

- **Dataset Health:** The dataset was successfully imported and inspected. It is complete, with no missing values across key variables, ensuring reliable analysis.
- **Data Cleaning:** Some preprocessing steps were carried out, including handling categorical encodings, ensuring consistent data formats (e.g., price, numerical features), and preparing variables for statistical testing.
- **Content:**
    - **Numerical:** Price, Discount, Units Sold, Number of Reviews, Average Rating, Total Amount
    - **Categorical:** Category, Brand, Packaging Type, Gender, Sentiment
    - **Identifiers/Textual:** Product Name, Ingredient Presence, Offer Details

2. Class and Feature Distributions

Visualizing the data revealed several key characteristics about the menu and customer feedback.

- **Menu/Product Composition (Class Balance):**
    - Certain categories (e.g., *snacks* or *popular brands*) dominate the dataset compared to niche categories, indicating an imbalance across product types.
    - Sentiment labels are **skewed towards positive**, with fewer neutral and negative reviews.
- **Numerical Feature Distributions**:
    - **Price:** Right-skewed, with many low-cost items and fewer premium-priced ones.
    - **Discounts:** Most items have modest or no discounts; only a small portion have high promotional discounts.
    - **Units Sold & Reviews:** Heavily right-skewed, indicating that only a small fraction of products account for the majority of sales and reviews.
    - **Ratings:** Concentrated in the upper range (around 4+), suggesting generally favorable customer feedback.

3. Feature Distribution Analysis & Outliers

We performed a deeper analysis to check if our numerical features follow a normal (Gaussian) distribution, which is an assumption for many statistical models.

- **Outlier Detection**: No extreme or erroneous outliers were found in major numerical features (price, calories, rating equivalents). This indicates a clean dataset without data-entry anomalies.
- **Normality Testing (Q-Q Plots & Shapiro-Wilk Test)**:
  None of the numerical variables (price, units_sold, reviews, ratings) followed a perfect normal distribution. Instead, they were skewed, with long tails.

    - Implication: While parametric tests (t-test, ANOVA) were applied and are fairly robust, non-parametric alternatives could also be considered in future work.

4. Feature Correlations

The correlation heatmap shows the strength and direction of relationships between the numerical variables.

- **Price & Total Amount/Calories (proxy for size/quantity):** Showed a **moderate positive correlation**, indicating that higher-priced products tend to be larger or richer in content.
- **Units Sold & Reviews:** Expectedly correlated, since more popular items naturally accumulate more customer reviews.
- **Rating Independence:** Ratings had **weak correlations** with price, calories, or sales volume—suggesting that customer satisfaction is not strongly tied to cost or size.

5. Hypothesis Testing Results

We conducted three statistical tests to validate our observations with a significance level of 0.05.

- **Chi-Square Tests (Category/Brand vs. Ratings & Sentiment):**
  Failed to reject the null hypothesis. No statistically significant dependency between product category/brand and ratings or sentiment.
- **Correlation Significance Tests (Numerical vs. Rating/Purchase Likelihood):**
  Some numerical variables (price, total amount) showed weak but statistically significant correlations, but most relationships with ratings were negligible.
- **T-Test: High vs. Low Price Groups (Ratings):**
  Failed to reject the null hypothesis. Ratings are statistically similar across high- and low-priced groups.
- **T-Test: Gender vs. Total Spending:**
  Failed to reject the null hypothesis. No significant difference in spending behavior across male and female customers.
- **Chi-Square: Sentiment vs. Weekend Purchase:**
  Failed to reject the null hypothesis. Customer sentiment is independent of whether purchases happen on weekdays or weekends.
- **T-Test: Ingredient Presence (e.g., Dates) vs. Total Amount:**
  Failed to reject the null hypothesis. Ingredient inclusion does not significantly influence purchase amount..

This comprehensive EDA doesn't just describe our data; it tells the story hidden within the numbers, providing a solid foundation for any machine learning or business decisions that follow. Each visualization and statistical test serves as a building block in understanding not just what the data says, but what it means for the business.

Open-Source Tools Used

Matplotlib, Seaborn, Plotly, SciPy, Statsmodels

Deliverables

A well-commented **EDA Notebook** including:

1. Data loading
2. Visualizations (plots, heatmaps)
3. Statistical test implementations
4. Summary of insights

Hypothesis test results, clearly documenting:

1. HypothesesTest statistic
2. P-value and interpretation
3. Conclusion based on test result

Google Colab Code Link (Link)

Conclusion

Through this experiment, we explored the dataset visually and statistically, gaining a clear understanding of its structure and patterns. We observed how different classes are distributed, how features vary in spread and central tendency, and identified possible outliers. The correlation analysis helped us see which features move together, while the hypothesis test gave us evidence to confirm or reject assumptions about differences between groups. Overall, this hands-on EDA helped transform raw data into meaningful insights, making the dataset feel less like numbers and more like a story waiting to be told.