Coffee Bean Sales Analysis – EDA & Hypothesis Testing Report

Doğa Tatlı - 31149

This report presents a complete analysis of customer purchasing behavior regarding coffee bean products. The analysis was conducted through four well-organized and modular Jupyter notebooks: data preparation, exploratory data analysis (EDA), hypothesis testing, and visualizations. These notebooks built upon one another to uncover trends, correlations, and statistically significant patterns related to pricing, discount usage, and regional preferences.

01 – Data Preparation

This notebook initiates the analysis pipeline by importing and preparing the raw dataset for further exploration. The dataset includes key attributes such as transaction date, customer ID, city, product name, unit price, quantity, and discount information.

The primary goals in this notebook were:

- **Data Formatting:** The Date column was successfully converted into a proper datetime format, enabling the creation of derived features like Year, Month, Day, and Day_of_week. These features allow for time-based trend analysis in later stages.
- **Categorical Transformation:**
    - A Price_Category feature was manually constructed based on the Unit Price, separating beans into four classes: Budget ($30), Standard ($35), Premium ($40), and Luxury ($45).
    - A new feature Coffee_Bean_Type was derived from the original product column to unify naming conventions.
    - The beans were also grouped into Bean_Category (Standard vs Premium) using a logical threshold on the price.

These enrichments transformed the original raw data into a more structured and analysis-ready form, which is essential for running valid statistical tests and generating relevant visualizations later on.

02 – Exploratory Data Analysis (EDA)

This notebook focused on summarizing the structure and distribution of the dataset, both visually and numerically.

Key Tasks:

- **Univariate Distributions:**
  Histograms were plotted for Unit Price, Quantity, Sales Amount, and Final Sales. These showed positively skewed distributions, especially for purchase quantities and revenue, indicating the presence of a few high-value transactions.
- **Categorical Counts:**
  Count plots revealed the popularity of different bean types, with Costa Rica and Ethiopian beans appearing most frequently. Sales counts by city suggested that Riyadh and Jeddah were the most active markets.
- **Temporal Patterns:**
  Monthly sales trends showed clear seasonal variations, with peaks in particular months likely influenced by holidays or cultural consumption habits.
- **Correlation Analysis:**
  A heatmap revealed a moderate correlation between unit price and final sales, and a

stronger correlation between quantity and sales. These relationships provided the basis for the formal statistical tests carried out in the next phase.

By the end of this notebook, a solid understanding of the data's shape, skewness, and composition was achieved. These insights guided the hypotheses tested in the next notebook.

03 – Hypothesis Testing

This notebook validated several data-driven hypotheses using formal statistical tests. It went beyond visualization to quantify relationships and differences.

1. Bean Type and Purchase Quantity (T-test):

A T-test compared the average quantities purchased for premium (Ethiopian and Colombian) vs. standard beans. The result ($p < 0.05$) confirmed a statistically significant difference, showing that premium beans are purchased in different quantities, influenced by their higher price.

2. Price Sensitivity (Correlation & Regression):

Pearson correlation and simple linear regression were used to investigate the relationship between Unit Price and Quantity. The correlation was moderately negative and statistically significant ($r \approx -0.32$, $p < 0.001$), indicating that as prices increase, quantity tends to decrease. Regression confirmed that price significantly predicts quantity, validating economic expectations.

3. Regional Preferences (Chi-Square & ANOVA):

A Chi-square test on a contingency table of cities and bean types revealed a highly significant result ($p < 0.001$), indicating that coffee preferences vary greatly between cities. ANOVA confirmed statistically significant differences in average prices paid by city, supporting region-specific strategies.

4. Discount Effectiveness (T-test):

A comparison of quantities purchased with vs. without discounts showed a statistically significant difference ($p < 0.001$). Discounts effectively increase purchasing volume, supporting their use in targeted marketing strategies.

All hypotheses were statistically validated, offering strong evidence for behavioral trends in the dataset.

04 – Visualizations

The final notebook focused on high-quality visual outputs that communicate the findings to non-technical stakeholders. It complemented the numerical analysis by enabling visual intuition.

Key Visuals Included:

- **Boxplots & Countplots:** Comparing quantity distributions across price categories, bean types, and cities.
- **Scatter Plots & Regressions:** Showing the relationship between price and quantity by bean type.
- **Heatmaps:** Mapping regional preferences for bean types as percentages.
- **Bar Charts:** Highlighting discount usage across product and price categories.

Each figure was saved in a visualizations/ folder with proper resolution and labels. These visuals not only support but also **validate** the statistical insights generated in the previous stages.

5. Summary of Findings

Based on the combined analysis:

5.1 Bean Type and Price Relationship

- Premium beans (Ethiopian and Colombian) are sold at significantly higher prices than standard ones.
- Despite the price, their high sales suggest consumers are willing to pay more for perceived quality.
- A T-test confirmed that pricing affects purchasing quantity ($p < 0.05$).

5.2 Price Sensitivity Analysis

- A moderate negative correlation ($r \approx -0.32$, $p < 0.001$) indicates that higher prices generally lead to lower purchase quantities.
- Regression results further confirmed price as a significant predictor.
- Price elasticity is most prominent between $35–$40.
- Consumers of Ethiopian beans showed less price sensitivity, suggesting potential for luxury positioning.

5.3 Regional Patterns

- Chi-square tests ($p < 0.001$) and ANOVA results revealed that coffee preferences and price willingness vary by city.
- Targeted regional pricing and marketing are likely to improve effectiveness.

5.4 Discount Impact Analysis

- Discounts significantly increased average purchase quantities ($p < 0.001$).
- Discount usage patterns varied across product types and price tiers.
- Premium products showed a different discount behavior compared to standard ones, suggesting discount strategies should be segmented.

5.5 Time-based Patterns

- Monthly trends show clear seasonality, with peaks during holidays and certain mid-year periods.
- These insights are valuable for seasonal planning and promotional campaigns.

6. Conclusion and Next Steps

The findings strongly support the hypothesis that customer behavior is influenced by multiple factors, including price, product type, discounts, region, and time.

Next Steps:

1. **Machine Learning Models** – Predict purchase quantity using price, location, and discounts.
2. **Customer Segmentation** – Identify and profile consumer clusters based on preferences.
3. **Recommendation Systems** – Suggest optimal beans or discounts for different users.
4. **Dynamic Pricing Strategies** – Optimize pricing in real-time for revenue maximization.

These initiatives will help coffee retailers operate more efficiently, maximize revenue, and tailor offerings to consumer behavior using data-driven strategies.