Coffee Bean Sales Analysis

Doğa Tatlı - 31149

01 – Data Preparation

The dataset was cleaned and standardized to ensure consistency and readiness for analysis.

Key steps:

- Removed missing and duplicate entries.
- Standardized column names and data formats.
- Converted the Date column to a datetime object.
- Created derived fields such as:
    - Coffee_Bean_Type from product names,
    - Bean_Category based on Unit Price (Premium vs. Standard),
    - Price_Category mapped from price levels (Budget, Standard, Premium, Luxury),
    - Month extracted from date for seasonality analysis.

These transformations ensured the dataset was suitable for both exploratory analysis and machine learning workflows.

02 – Exploratory Data Analysis (EDA)

This phase explored the structure, trends, and relationships in the dataset.

Highlights:

- Identified dominant coffee bean types and pricing tiers.
- Found seasonality in monthly sales and quantities.
- Observed that Used_Discount correlated with higher quantities sold.
- City-level breakdowns revealed regional differences in demand.
- Set the stage for statistical validation and feature selection.

EDA laid the analytical foundation by revealing patterns that informed hypothesis testing and modeling decisions.

03 – Hypothesis Testing

Statistical techniques were used to confirm or reject assumptions suggested by EDA.

Examples:

- **T-tests:** Confirmed discounts significantly increased purchase quantity ($p < 0.001$).
- **ANOVA:** Showed that different coffee bean types led to significantly different sales volumes.
- **Chi-square tests:** Validated associations between categorical variables like City and Bean_Category.

These tests provided statistical support for EDA insights and improved confidence in feature importance for modeling.

04 – Visualizations (Enhanced)

Twelve detailed and interpretable visualizations were created to support findings and modeling strategy.

Key Visuals:

1. Distribution of Coffee Bean Types
2. Price Category Distribution
3. Scatter Plot with Regression: Price vs Quantity
4. Boxplot: Quantity by Bean Category

5. Monthly Sales Trend with Regression
6. Discount Usage Pie Chart
7. Quantity by City (Top 10)
8. Average Unit Price by City
9. Total Sales by Coffee Bean Type
10. Correlation Heatmap of Numeric Variables
11. Monthly Quantity by Bean Category
12. Boxplot: Final Sales by Price Category

These plots provided visual confirmation of patterns in product popularity, pricing behavior, regional variation, and seasonal effects.

05 – Predictive Goal Definition

A regression problem was defined to transition from analysis to prediction.

Predictive Objective:

Estimate the number of units (Quantity) sold per transaction.

Type: Supervised Learning – Regression

Target Variable: Quantity

Selected Features:

- Unit Price (continuous)
- Used_Discount (binary)
- City (categorical)
- Coffee_Bean_Type (categorical)
- Price_Category and Bean_Category (derived categorizations)
- Month (captures seasonality)

Supporting Visuals:

- Histogram: Distribution of Quantity
- Boxplots: Quantity by Price Category, Quantity by City
- Barplot: Average Quantity by Coffee Bean Type
- Heatmap: Feature correlations

The notebook successfully outlined a complete ML task with validated features and transitioned the project into the modeling phase.

06 – Summary of Findings

6.1 Bean Type and Price Relationship

- Premium beans (Ethiopian and Colombian) have significantly higher prices.
- Despite price, high sales volumes suggest strong perceived value.
- T-tests confirm pricing significantly impacts purchasing quantity ($p < 0.05$).

6.2 Price Sensitivity Analysis

- Moderate negative correlation ($r \approx -0.32$, $p < 0.001$) between price and quantity.
- Regression confirmed price as a significant predictor.
- Price elasticity is most visible between $35–$40.
- Ethiopian bean buyers show lower sensitivity, suggesting luxury positioning potential.

6.3 Regional Patterns

- Chi-square tests and ANOVA show that preferences and pricing tolerance vary significantly by city.
- Tailored regional strategies are likely more effective.

6.4 Discount Impact Analysis

- Discounts significantly boost purchase volume ($p < 0.001$).
- Discount effects vary by product and price tier.
- Premium items respond differently to discount strategies.

6.5 Time-Based Patterns

- Monthly trends reveal seasonality with peaks near holidays and mid-year.
- These trends are useful for seasonal marketing and stocking.

07 – Conclusion and Next Steps

The analysis confirms that customer purchasing behavior is influenced by:

- Price and discount usage
- Product type and category
- Regional preferences
- Seasonality

Next Steps:

1. **Machine Learning Models**
   Predict purchase quantity using selected features.
2. **Customer Segmentation**
   Group consumers based on preferences and behavior.
3. **Recommendation Systems**
   Suggest products or discounts based on past behavior.
4. **Dynamic Pricing Strategies**
   Adjust prices in real time based on demand and elasticity.