# Analysis and Forecasting of Coffee Sales Trends, Customer Loyalty, and Discount Impact

A Final report for the BDM capstone Project

Submitted by

Name: Sudhanshu Prabhat

Roll number: DS23f1001796

Email: 23f1001796@ds.study.iitm.ac.in

IITM Online BS Degree Program,
Indian Institute of Technology, Madras, Chennai
Tamil Nadu, India, 600036

# Content

# 1 Executive Summary

QahwaCo is a specialty coffee retailing company in Saudi Arabia that offers high-quality beans from several origins to the coffee enthusiasts in Saudi Arabia. As the company continues to grow, it faces challenges in understanding the consumer demand, evaluating the effectiveness of discounts, and planning inventory efficiently.

This study addresses these issues through three objectives:

- Analysing Seasonal Coffee Sales Trends: This study will examine how the sales of each variety of coffee bean fluctuate over time, to identify peak selling season and figure out demand. The goal is to help QahwaCo better anticipate customer demand.
- Exploring Customer Loyalty and Discount Influence: This study is to investigate the purchasing habits of customers and understand what led to repeat orders, while analysing the impact of discounts. This will support QahwaCo in customer retention strategies and scheduling promotional offers.
- Forecasting Coffee Sales for Inventory Planning: This study will utilise sales data, to develop a predictive model to forecast future demand for each coffee variety. These forecasts will aid QahwaCo in making data driven inventory decisions.

This study analyses transactional data from 1 Jan 2023 to 31 Dec 2024 containing key metadata such as date of transaction, unique identification number of the customer, location of the transaction, category of the product, unit price of the product, quantity of the product sold and final sale. Using descriptive statistics, exploratory visualisation, cohort analysis, and local linear time-series forecasting, the study evaluates both behavioural and operational patterns. These methods help uncover how different coffee SKUs perform over time and how customer groups respond to discounts.

The results reveal clear peaks and lows of the products across the time period. Discount shows a noticeable short-term sales lift but also demonstrates an underlying insight that highly regular customers are not affected much by discounts. Forecasting models demonstrate low accuracy because of the highly volatile nature of the product.

Overall, the insights suggest that QahwaCo can enhance profitability by applying discounts more strategically, focusing their inventory to deal with peaks and lows appropriately, and manage their inventory accordingly. These recommendations support more efficient operations, improved customer retention, and better alignment between supply and demand.

# 2 Proof of Originality

The dataset used in this study has been sourced from Kaggle.

Link: https://www.kaggle.com/datasets/halaturkialotaibi/coffee-bean-sales-dataset/data

This dataset contains transactional data from 1 Jan 2023 to 31 Dec 2024. It has a very detailed record of the transaction like the date of transaction and location of the transaction. It explains the details of the product as well like category of the product, unit price of the product and quantity of the product sold. It also provides us with the customer details and customer behaviour like unique identification number of the customer, whether discount used and if used how much.

Colab link for Data Cleaning:
https://colab.research.google.com/drive/1FOUDunei4UAdt6JdnZM34EiNLvb38i0e?usp=sharing

Colab link for Data Analysis:
https://colab.research.google.com/drive/1emGXqvG4qW__FhzKSreCY4U8GHvwN23N?usp=sharing

# 3 Metadata and Descriptive Statistics

## 3.1 Metadata

The dataset used in this study consists of 730 daily transactional records for QahwaCo covering the period 1 January 2023 to 30 December 2024. It provides sales record across multiple cities across Saudi Arabia.

It provides a very detailed record of the transaction like the date of transaction and location of the transaction, the category of the product, unit price of the product, quantity of the product sold, unique identification number of the customer, whether discount used and if used how much. These variables provide a reliable method to understanding customer behaviour, compute product performance, pricing effectiveness, and revenue generation.

This table was used to create 2 more tables to cater to our problem statements. One with the aggregated monthly sales and another with the customer behaviour. These tables will work as the foundation for solving our problem statements.

The dataset has some irregularities. After cleaning the feature engineering and statistical modelling can be implemented fairly easily.

The tables below provides a detailed description of each variable, including its meaning, data type, and observed range or possible values.

**Original Dataset**:

This is the original dataset downloaded from Kaggle. Containing the daily coffee sales recorded across multiple cities in Saudi Arabia from January 2023 and December 2024. Each row contains the details of a customer transaction including date, customer ID, city, product category, specific product, unit price, quantity sold and calculated sales amount. It also has the discount applied to the order and the final sales.

| **Variable Name** | **Description** | **Data Type** | **Range/ Values** |
|---|---|---|---|
| Date | The transaction date for each sales record. | Datetime | 01-01-2023 to 30-12-2024 |
| Customer id | Unique ID representing each customer making a purchase | Integer | 1 -100 |
| City | The city where the transaction occurred | Object (Category) | Riyadh, Jeddah, Dammam, Mecca, Medina, etc |
| Category | The category of the Product. | Object (Category) | “coffee beans” |

| | | | |
|---|---|---|---|
| Product | Specific coffee product purchased | Object (Category) | Brazilian, Colombian, Costa Rica, Guatemala, Ethiopian (5 SKUs) |
| Unit Price | Price per unit (in SAR or Saudi Riyal) of the product purchased. | Integer | 30 – 50 SAR |
| Quantity | Number of units purchased in a single transaction. | Integer | 1 – 50+ units |
| Sales Amount | Gross sales value before any discount (unit_price × quantity) | Integer | 30 – 2000+ SAR |
| Used Discount | Indicates whether a discount was applied. | Boolean | True / False |
| Discount Amount | Amount deducted due to discount for the transaction | Integer | 0 – 200+ SAR |
| Final Sales | Final revenue after subtracting any discounts | Integer | 30 – 2000+ SAR |

**Aggregated Monthly Sales Dataset**:

This dataset is the monthly aggregates sales summary of the original dataset. Each row represents the monthly performance of a product for a specific month. It includes the month of the sales record, the product name, the aggregated sale of the product for the month and the percentage change in sales compared to the previous month.

| Variable Name | Description | Data Type | Range/ Values |
|---|---|---|---|
| year_month | The month for which the sales are aggregated. Represented using the first day of the month. | Datetime | 01-01-2023 to 01-12-2024 |
| product | The specific coffee product for which monthly sales are reported. | Object (Category) | Brazilian, Colombian, Costa Rica, Guatemala, Ethiopian (5 SKUs) |
| monthly_sales | Total sales amount generated by the product in the given month. | Integer | 900 – 13000+ units |
| monthly_pct_change | Percentage change in monthly sales compared to the previous month. | Float | -1 to 4 (percent) |

**Customer behaviour Dataset:**

This dataset is a customer level purchase behaviour dataset. Each row is a aggregated transaction data for each customer id. It includes the number of orders, the total amount spent, the total count of order and the number of repeat orders along with the count of and the average order value for discounted and non-discounted orders for each customer id. This dataset can be used to analyse the loyalty of each customer and the impact of discount.

| Variable Name | Description | Data Type | Range/ Values |
|---|---|---|---|
| customer_id | Unique ID representing each customer making a purchase | Integer | 1 - 100 |

|  |  |  |  |
|---|---|---|---|
| orders_count | Total number of orders placed by the customer. | Integer | 1 – 15 |
| total_spent | Total amount spent by the customer across all orders. | Integer | 1000 – 14000+ SAR |
| avg_order_value | Average value per order for the customer. | Float | 500 – 1600+ SAR |
| discounted_orders | Number of orders where a discount was applied. | Integer | 0 – 11 |
| non_discounted_orders | Number of orders without any discount. | Integer | 0 - 8 |
| avg_order_value_discounted | Average order value for discounted purchases. | Float | 100 – 1700+ SAR |
| avg_order_value_non_discounted | Average order value for non-discounted purchases. | Float | 100 – 1900+ SAR |
| repeat_customer | Indicates whether the customer made more than one purchase. | Boolean | True / False |
| repeat_count | Number of additional purchases beyond the first | Integer | 0 - 14 |

### 3.2 Descriptive Statistics

This section provides a statistical overview of the sales datasets used in this study, highlighting the key characteristics relevant to our problem statements.

**Original Dataset**:

| **Variable Name** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
|---|---|---|---|---|---|---|---|
| Date | 31-12-2023 | Nan | 01-01-2023 | 02-07-2023 | 31-12-2023 | 30-06-2024 | 30-12-2024 |
| Customer id | 51.66 | 29.01 | 1.00 | 27.25 | 52.00 | 77.00 | 100.00 |
| Unit price | 36.79 | 4.95 | 30.00 | 35.00 | 35.00 | 40.00 | 45.00 |
| Quantity | 26.08 | 14.48 | 1.00 | 14.00 | 27.00 | 39.00 | 49.00 |
| Sales amount | 959.92 | 551.28 | 30.00 | 495.00 | 960.00 | 1400.00 | 2205.00 |
| Discount amount | 97.39 | 123.40 | 0.00 | 0.00 | 13.00 | 192.00 | 441.00 |
| Final sales | 862.53 | 509.03 | 24.00 | 448.00 | 840.00 | 1260.00 | 2205.00 |

- The wide range and high std of sales_amount suggest that will contain sudden spikes due to large orders.
- Large max discount suggest that some orders have a huge promotional effect that could be causing repeat orders.
- Difference in std and mean of sales_amount and final_amount suggest that the discount shows a significant difference on overall revenue but might also be the reason behind customer retention.
- Volatility observed in wide range and high std of sales_amount could make the data hard to forecast.

**Monthly Aggregated Dataset**:

| **Variable Name** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
|---|---|---|---|---|---|---|---|
| monthly_ sales | 5291.15 | 2435.93 | 994.00 | 3435.00 | 5220.00 | 6600.50 | 13760.00 |
| monthly_ pct_ change | 0.25 | 0.95 | -0.79 | -0.36 | -0.03 | 0.48 | 3.58 |

- High std of monthly_sales suggests sudden peaks in the sales.

- Monthly_pct_change ranging from -0.79 to 3.58 and a mean of 0.25, this fluctuation suggests volatile growth and could make the forecasting tough.
- The max monthly_pct_change could coincide with a promotional campaign related to discount.
- The negative monthly_pct_change suggest significant drop in monthly sales and could be incorporated in inventory suggestion.

**Customer behaviour dataset**:

| **Variable Name** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
|---|---|---|---|---|---|---|---|
| customer_id | 50.50 | 29.01 | 1.00 | 25.75 | 50.50 | 75.25 | 100.00 |
| orders_count | 7.30 | 2.63 | 1.00 | 5.00 | 7.00 | 9.00 | 15.00 |
| total_spent | 6296.48 | 2586.04 | 1050.00 | 4403.50 | 6161.00 | 7687.25 | 14334.00 |
| avg_order_value | 872.37 | 207.17 | 502.42 | 742.62 | 847.00 | 973.54 | 1682.33 |
| discounted_orders | 3.71 | 1.99 | 0.00 | 2.00 | 4.00 | 5.00 | 11.00 |
| non_discounted _orders | 3.59 | 1.79 | 0.00 | 2.00 | 4.00 | 5.00 | 8.00 |
| avg_order_value _discounted | 772.84 | 272.90 | 160.00 | 588.00 | 772.57 | 904.66 | 1764.00 |
| avg_order_value _non_discounted | 964.78 | 310.11 | 135.00 | 773.75 | 965.83 | 1155.00 | 1932.50 |
| repeat_count | 6.30 | 2.63 | 0.00 | 4.00 | 6.00 | 8.00 | 14.00 |

- Mean of 6.30 and std of 6 of repeat_count suggests that most customers place multiple orders.
- Mean count of discounted and non-discounted orders is close this could suggest similar number of orders with and without discount for each customer.
- Mean of average order value for non-discounted is greater than discounted suggesting that larger orders tend to be paid in full price.

# 4. Analytical Approach and Methodology

This project was done using python as it has many readily available packages. The workflow consists of cleaning, feature engineering, EDA and modelling.4.1 Data ingestion

**4.1 Data Ingestion:**

- Importing the necessary libraries:
  i. NumPy for numerical operations and array manipulation pandas for loading
  ii. Pandas for loading, cleaning and transforming the dataset.
  iii. Matplotlib to create visualisations such as line charts and bar plot.
  iv. Seaborn to create more enhanced visualisations.
  v. Math for mathematical functions.
  vi. Scikitlearn for machine learning models.
- Reading the raw csv file using pd.read_csv().
- Initial description of the dataset using .info(), .describe() and .nunique().

**4.2 Cleaning:**

- Date parsing: The date column had inconsistent format of date that required to be converted to datetime using pd.to_datetime().
- Type coercion: Numeric field like unit_price, quantity, sales_amount, discount_amount, final_sales were coerced to numeric using pd.to_numeric() and .astype(str) for string objects.
- Standardising of column names to deal with ‘ ‘ and ’-‘ in column names.
- Deduplication: Duplicate values were searched for using .duplicated() and none were found.
- Sorting the dataset in chronological order of date.

**4.3 Feature Engineering**

- Time based features: year_month was extracted from date for creating the monthly sales analysis dataset using .dt.to_period('M').dt.to_timestamp().
- Monthly aggregation: for the monthly sales trend analysis daily sales were aggregated monthly to get monthly_sales and their percentage difference was taken from previous month to get monthly_pct_change.
- Monthly sales dataset: year_month, product, monthly_sales and monthly_pct_change were combined to form the monthly sales dataset.
- Customer loyalty feature: grouped by customer_id and aggregated customer behaviour features like order_count, total_spent, avg_order_value, repeat_customer etc.
- Discount features: discount features of discount_orders, non_dicount_orders, avg_discount_order_value and avg_non_discount_order_value are created.

- Aggregating discount features to the customer behaviour dataset for discount impact analysis.

### 4.4 Descriptive Statistics

- Summarising statistics to compute the mean, median, std, min, max, 25%, and 75% of the datasets using .describe().
- This provides a statistical understanding of the dataset.

### 4.5 Time Series Analysis

Logic:

- To identify the peaks and lows in the demand of the SKUs.
- To identify if the data shows some trend in the sales.

Method:

- The statistical properties of the monthly sales of each product are found and studied.
- The line charts for the products from monthly_sales are plotted and studied.
- The seasonal trends are also plotted on a heatmap for clear peaks.
- The monthly percent change for each product is plotted and studied.
- The insights of both the visualisations are used to make inventory improvisation decisions.

### 4.6 Discount and Loyalty Analysis

Logic:

- To summarise customer behaviour, customer level features were created to understand general buying pattern.
- To calculate the impact of discount on retention, the proportion of repeat buyers and how often they used discount was calculated.
- To measure how discount dependence changes with loyalty, discount ratio (= discounted orders / total orders) was compared to repeat counts.

Method:

- Ratio of repeat customers is calculated from the customer behaviour dataset.
- Ratio of repeat customers who have used discount is calculated.
- Visualisation for repeat customers who have used discount vs who have never used discount is plotted.
- Visualisation for number of repeat orders against the average discount ratio (ratio of orders that had discount) is plotted for those customers to analyse if regular customers relied on discounts.

### 4.7 Forecasting

Logic:

- Training a forecasting model to help QahwaCo predict the sales.

- Provide alternatives if the sales can not be forecasted.

Method:

- The data was split into training and test data.
- A model object was created.
- The training data was used to train the model.
- The sales for the test data were predicted.
- The predicted values were compared to the actual values using mape (Mean Absolute Percentage error).

### 4.8 City Level analysis

Logic:

- To understand purchasing power, the average order value of the cities was computed. Higher AOV represents tendency to buy expensive products or large orders.
- To determine price sensitivity for using discount, discount usage rate and average discount amount was computed. High discount usage means more sensitive to price.
- Identify cities with high sales to highlight key markets.

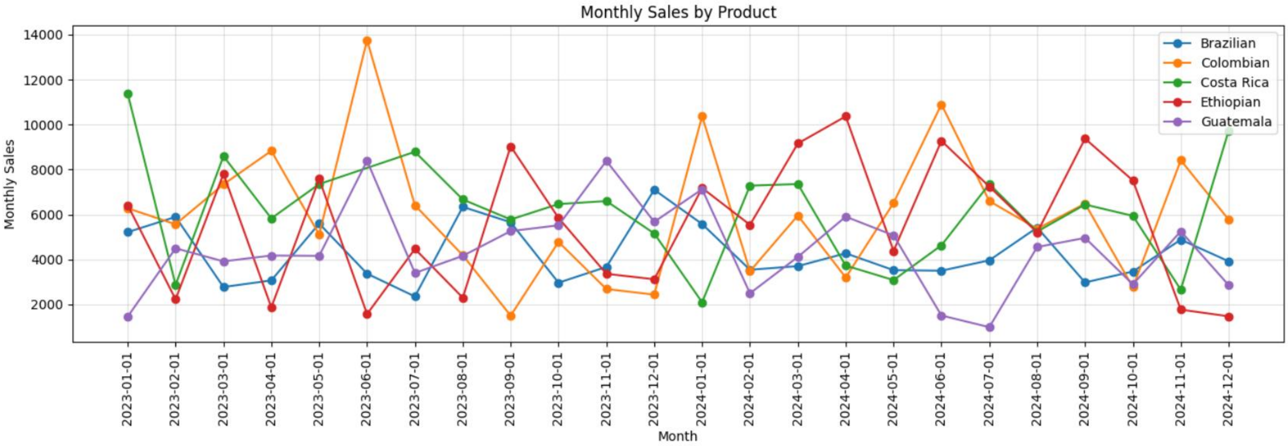
Method:

- The average order value for the cities was calculated and used to rank the cities.
- The discount usage rate and average discount amount is calculated for all the cities.
- Total revenue and revenue share percent is calculated and plotted in a pie chart.

# 5. Results and Findings

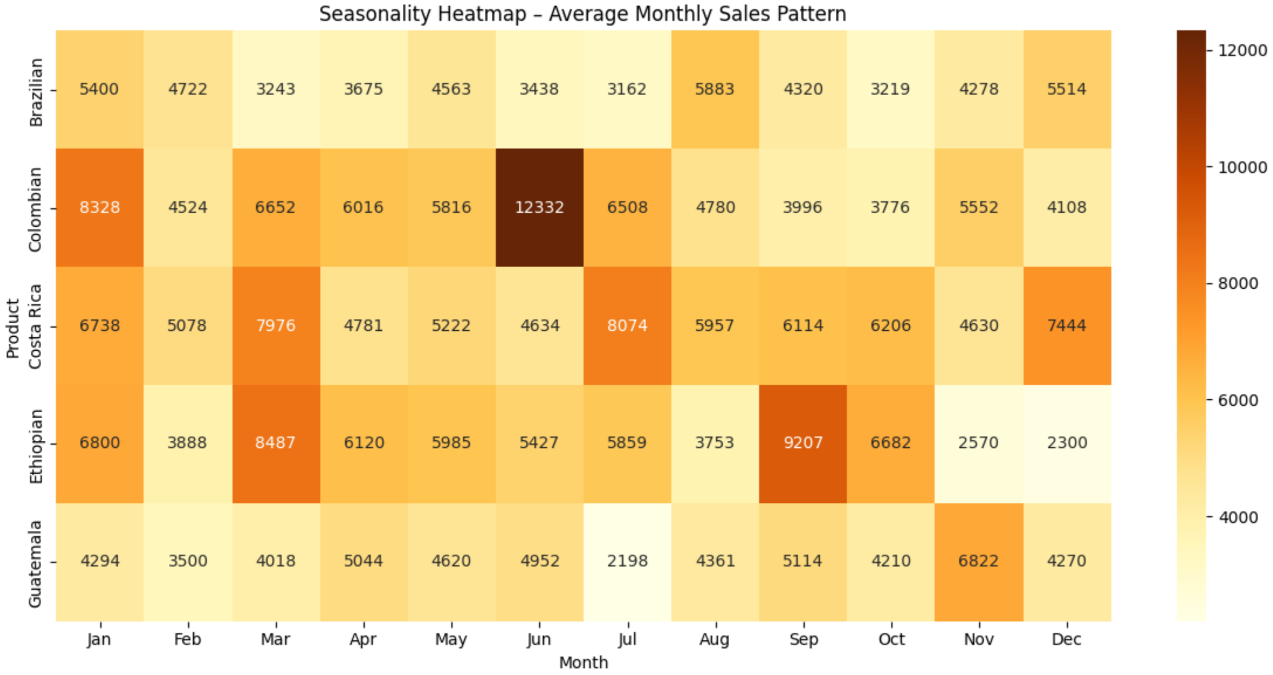
## 5.1 Trend Analysis of Products

To analyse the trends, a line chart and seasonal heatmap were used to plot the monthly sales. The line chart helps in revealing the peak or low trends, repeating seasonal peaks and volatility.



Visual 5.1.1 Multi Line Chart of the monthly sales

The heatmap summarises the average monthly sales for each product in a single plot, it visualises the seasonal highs and lows more clearly.



Visual 5.1.2 Seasonal Heatmap of monthly sales

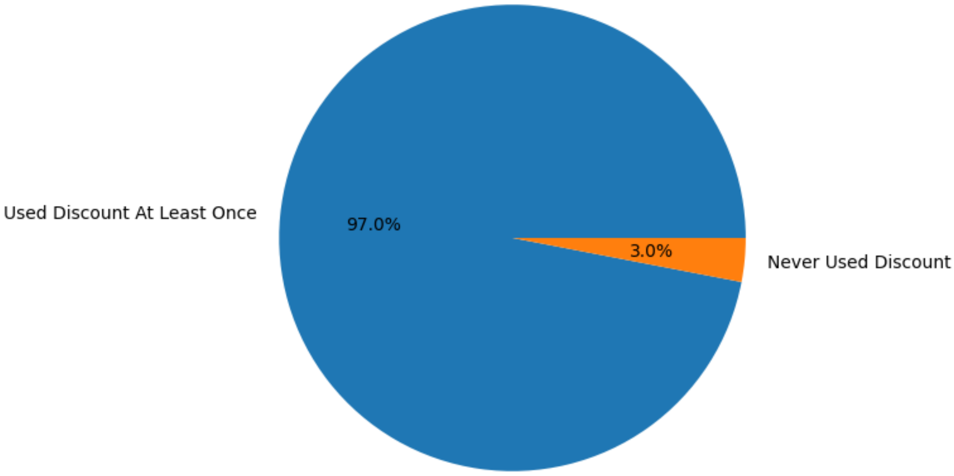
Overall Insights:

| **Coffee Variety** | **Peak Demand** | **Low Demand** | **Remark** |
|---|---|---|---|
| **Brazilian** | December and August | March and April | There seems to be some trend with high demand in the latter half. |
| **Colombian** | June, January and March | February and July | Demand is very volatile seems to be influenced by promotions. |
| **Costa Rica** | March, July and December | February and May | Demand is very volatile seems to be influenced by promotions. |
| **Ethiopian** | March and September | January and June | Demand is very volatile seems to be influenced by promotions. |
| **Guatemala** | June and August | January, February and November | Demand is very volatile seems to be influenced by promotions. |

- June-August show consistently higher sales for most coffee verities.
- November-December generally has secondary peaks across most coffee varieties.
- January- February show lower sales across most coffee varieties.
- Brazilian has the most stable demand with visible seasonality.

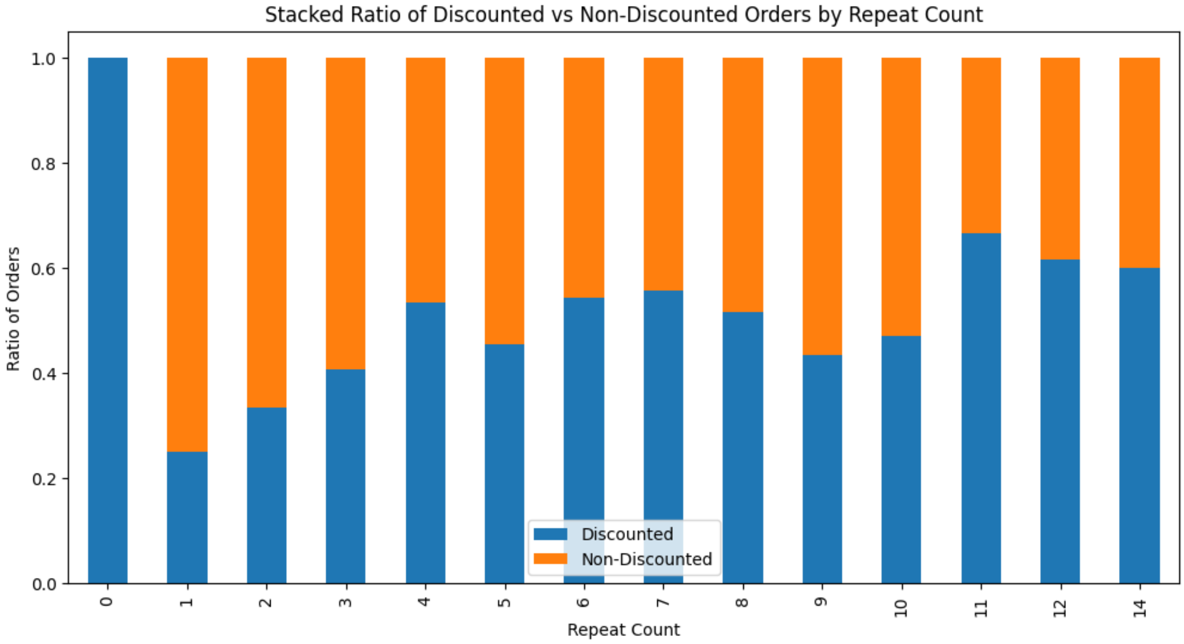
### 5.2 Customer Loyalty and Discount Impact

To represent the impact of discount on getting repeat customers, the ratio of repeat customers that had used discount ever was found.



Visual 5.2.1 Pie chart of repeat customers that used discount ever vs customers who never used discount

To compare the impact of the discount on different repeat order levels, a stacked bar chart was used, it visualises this more clearly.



Visual 5.2.2 Stacked ratio of discounted vs non discounted orders by repeat count

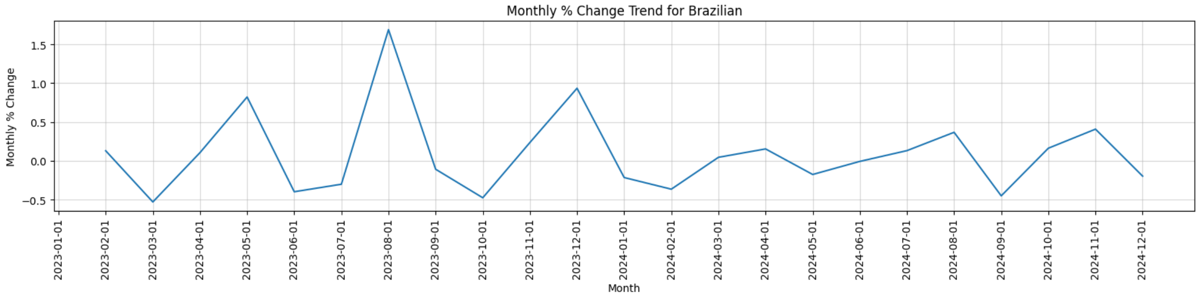
Overall Insights:

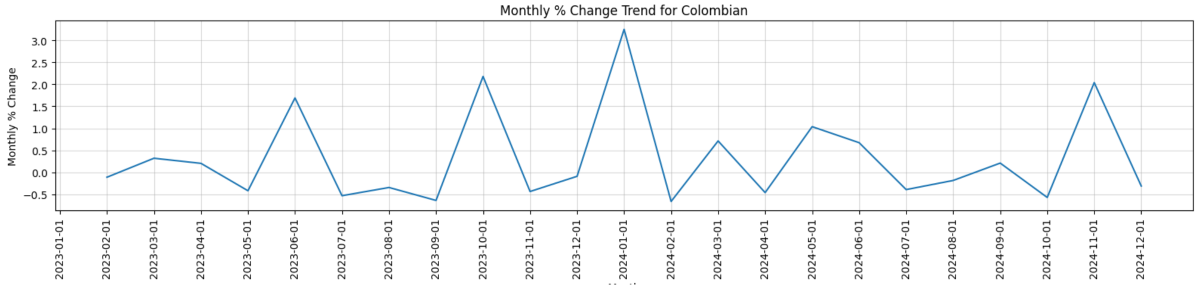
- Majority of the customers of QahwaCo tend to reorder, 99% of all customers were repeat buyers.
- Discount strongly influences the orders after the first purchase, as 97% of repeat customers had used discount at least once and the ratio of discount orders shows positive relation in the first four orders.

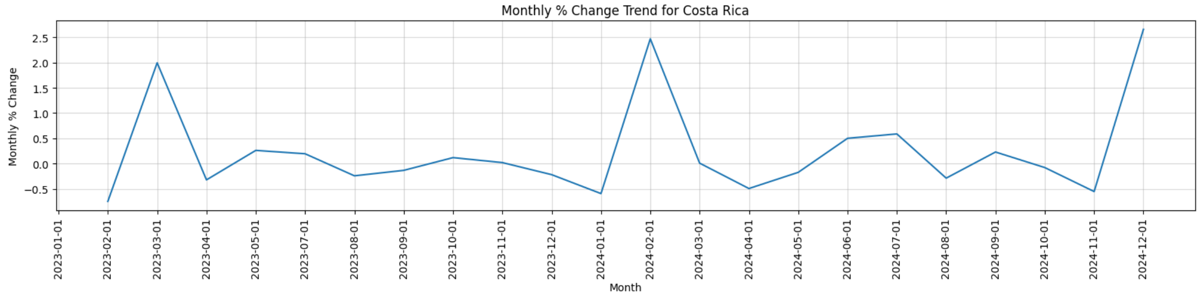
- Over time the most loyal customers became less dependent on discount, customers with 7+ orders rely less on discount.

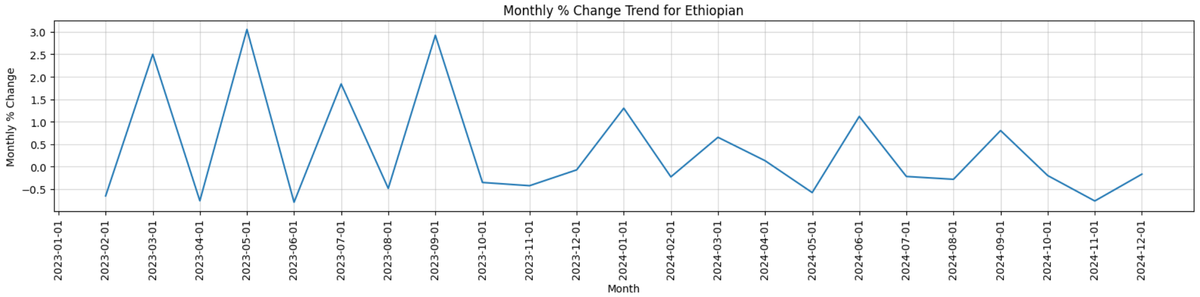
### 5.3 Forecasting

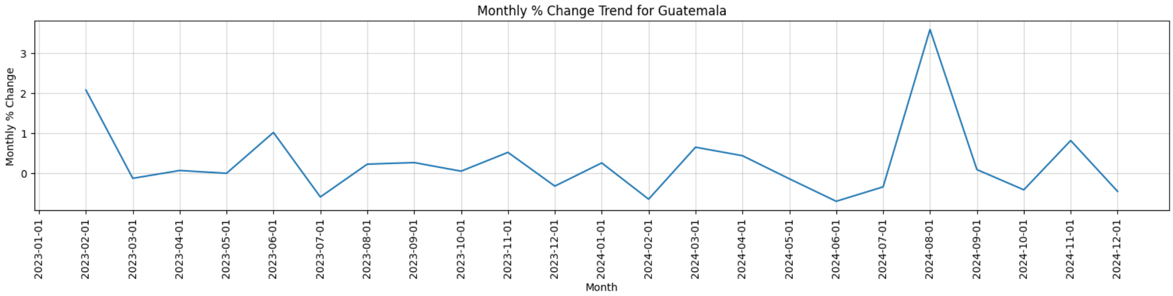
The monthly percentage change line chart was used to visualise the month on month volatility of the products.











Visual 5.3.1 The monthly percent change for each product

The forecasting revealed low accuracy due to high volatility of the dataset.

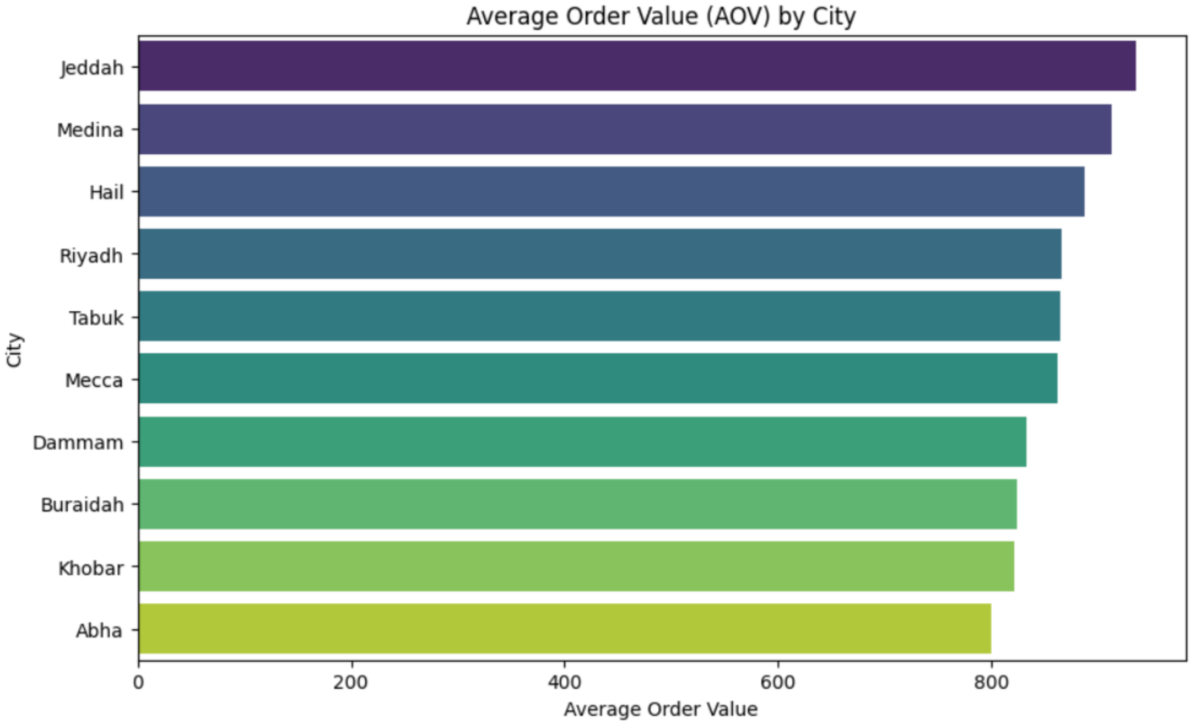
| **Product** | **Mape** |
|---|---|
| Brazillian | 11.650572826639388 |
| Colombian | 50.843704399520576 |
| Costa Rica | 58.37013979418412 |
| Ethiopian | 468.1829785007408 |
| Guatemala | 164.07015370690215 |

Overall Insights:

- These varieties lack any consistent trend, meaning statistical forecasting cannot reliably predict their future values.
- Volatile beans cannot rely on forecasting, instead safety stock buffer, short review cycles and fast replenishment techniques need to be implemented.

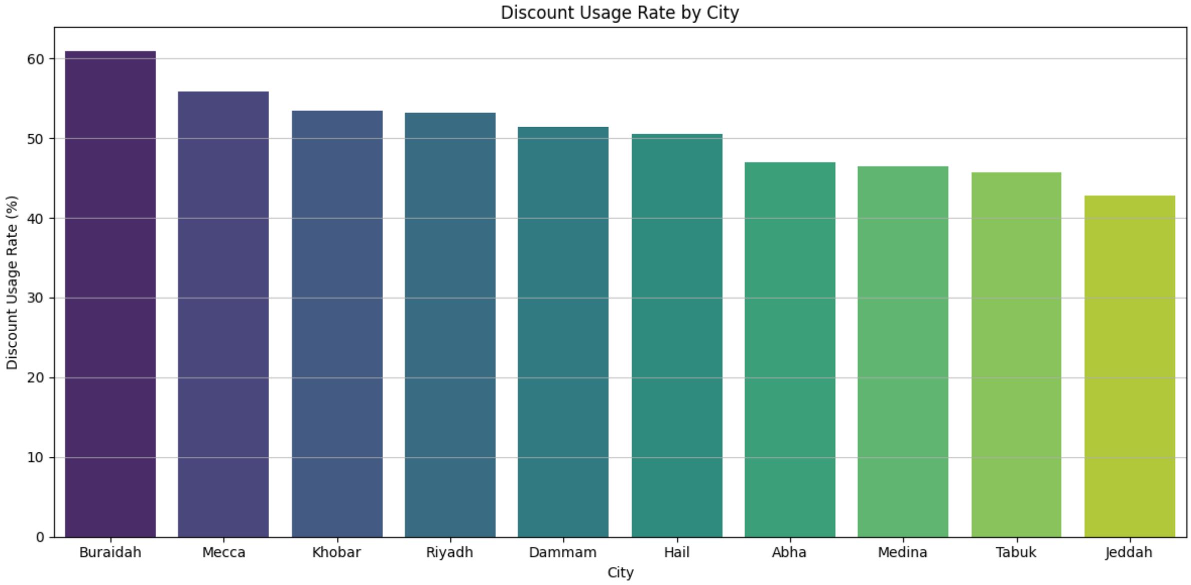
### 5.4 City level analysis

To analyse the sales performance of various cities their average order value was calculated and visualised using a bar plot as it gives an uncluttered ranking of Average Order Value.



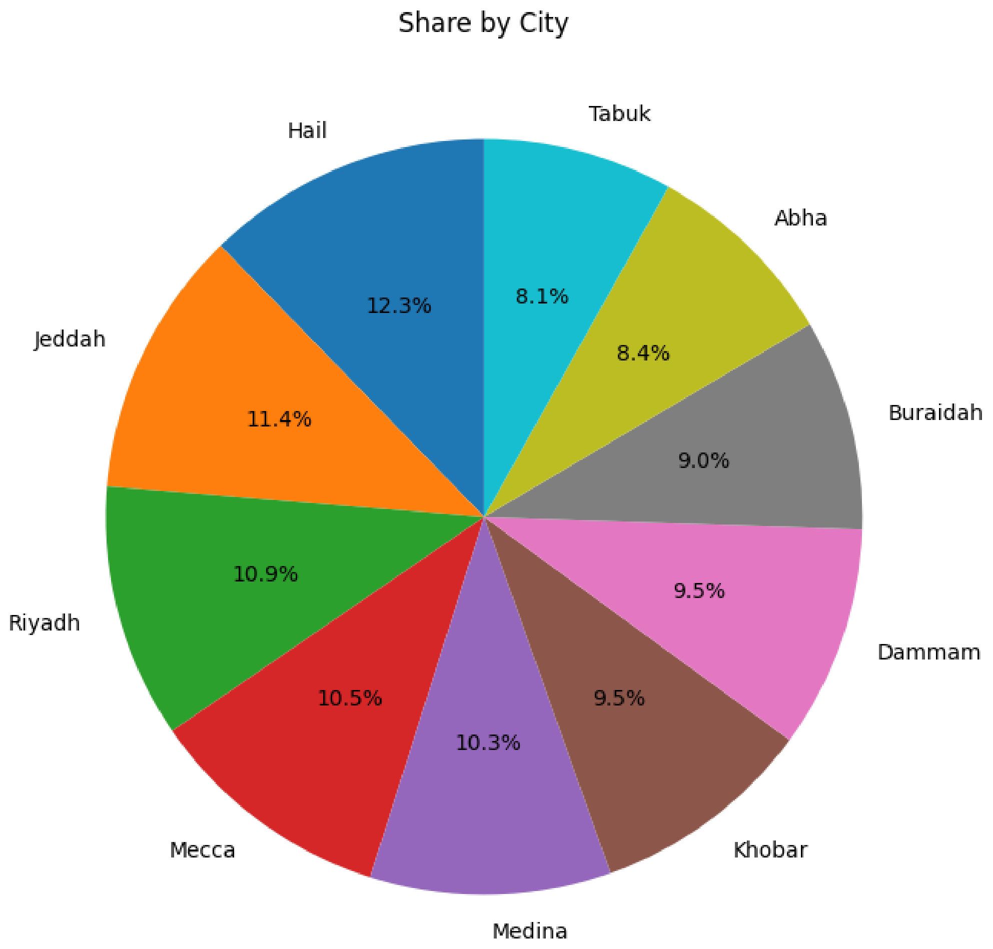
Visual 5.4.1 Bar chart of Average Order Value

The dependence of cities on discount was calculated using discount usage rate for each city.



Visual 5.4.2 Bar chart of discount usage rate for each city.

To prioritise the markets, they were ranked on the revenue share of the total revenue.



Visual 5.4.3 Pie chart of revenue share for each city.

Overall insights:

- High AOV does not mean high discount usages rate, Jeddah has the highest AOV 935.68 but lowest discount usage rate 42.85% indicating low reliance on discount.
- High discount dependency does not guarantee higher sales. Buraidah has the highest discount usage rate 60.87% but has only 9.03% of share.
- Hail, Jeddah and Riyadh have the highest share in sales, accounting for over a third of the total sales.

# 6. Interpretation of results and recommendations

**Problem Statement 1**: Analysis of Coffee Sales Trend.

Insightful interpretation:

- The line chart and heatmap for Brazilian beans demonstrates clear seasonal peaks and lows.
- Other varieties do not have as clear seasonality but do show irregular high amplitude peaks and lows.
- Aggregated statistics confirm that Brazilian has repeatable month on month pattern suitable for seasonal inventory planning.

Actionable recommendation:

- Adopt seasonal forecasting to plan the inventory for Brazilian bean and higher safety stock for peak seasons. For other beans instead of seasonal planning, treat them as intermittent demand SKUs with a higher safety stock in general.
- Schedule promotional offers for each variety just before their expected peaks to capture demand, especially during June-August (overall peak).
- For volatile SKUs implement flexible replenishment windows.

Impact:

- Reduced stockouts during peak seasons by higher demand capture and better customer experience.
- Reduced inventory cost by optimised inventory management.
- Better return over investment for promotions by aligning them with expected demand window.

**Problem Statement 2**: Analysis of customer loyalty and impact of discount.

Insightful interpretation:

- 99% of repeat buyers indicates very strong customer loyalty.
- 97% of repeat buyers being discount users represents high conversion rate of first-time customers due to discount.
- First-time buyers rely heavily on discounts, and discount shows positive relation for first few orders. However, discount usage stabilises around 40% to 60% after that, showing that loyal customers rely less on promotions.

Actionable recommendation:

- Use introductory discount to target first-time buyers. Using a new registration discount or a voucher.

- Promote bundles to introduce multiple products at early stage to capture specific tastes.
- Reduce discount spend on loyal customers, instead offer loyalty perks like early access to new products or goodies.
- Customers with low discount ratio should be offered quality-based marketing due to less reliance on discount.
- Customers with high discount ratio should receive small frequent promotions and promotional bundles to acquire taste and convert to lower discount ratio.

Impact:

- Higher customer retention with lower promotional cost.
- Higher Return over investment on promotions.
- Improved customer relations and loyalty.

**Problem Statement 3**: Forecasting coffee sales.

Insightful interpretation:

- Brazilian bean shows forecastable sales pattern with a reliable accuracy.
- Other coffee beans have high volatility in pattern making forecasting sales unreliable.

Actionable recommendations:

- Brazilian coffee can be forecasted with confidence to manage the inventory.
- For volatile products implement other inventory strategies:
- Safety stock buffer: Holding additional inventory proportional to volatility.
- Shorter review cycle to reduce risk of stockout.

Impact:

- Improved inventory management with reduced cost of inventory and wastage.
- Reduced over stocking and lower stockout risk.
- Operational flexibility and faster pivoting with shorter review cycles.

**Problem Statement 4**: Analysing the sales across various cities.

Insightful interpretation:

- Jeddah, Medina and Hail are cities with higher AOV, representing higher value baskets. Buraidah, Khobar and Abha are cities with lower AOV, representing lower value baskets (may be price sensitive too).
- High AOV does depend on high discount usage rate.
- High sales does not depend on high discount usage rate.

- Cities with high sales have a strong AOV and do not rely on high discount. Natural demand, product preference or purchasing power are the sales driving factors in these cities.

Actionable recommendation:

- Focusing premium or standard pricing on cities with high AOV and low to moderate discount dependency.
- Focusing promotions on cities with high discount dependency.
- Prioritising inventory allocation based on share of the city.

Impact:

- Improved revenue by reducing promotions in naturally strong markets.
- Improved resource allocation and inventory management.
- City specific strategies help capture market more efficiently.
- Reduced stockout in high demand areas.