COFFEE SHOP SALES

****

**About Dataset**

The dataset coffee shop sales analysis aims to uncover insights to enhance business performance. By examining sales trends, product performance, and store-level variations, the study identifies peak periods, high-demand products, and location-based performance differences. It evaluates pricing impacts, customer purchasing behavior, and inventory optimization opportunities. Additionally, the analysis explores correlations between transaction factors, detects anomalies, and offers recommendations to improve efficiency, ultimately providing actionable insights for strategic decision-making.

# Objective

The objective of this dataset analysis is to derive actionable insights to enhance the coffee shop's operational and strategic decision-making. This includes identifying peak sales periods, evaluating product performance across categories, and analyzing store-level variations to pinpoint high-demand products and top-performing locations. By exploring customer purchasing behavior and the impact of pricing on sales, the analysis aims to optimize inventory management and promotional strategies. Additionally, it seeks to uncover correlations between transaction factors, detect anomalies, and provide data-driven recommendations to improve overall business efficiency and performance.

# Overview of Dataset

* Total Rows : 149116
* Total Columns : 16

**Table Representation of Dataset :**

|  |  |  |
| --- | --- | --- |
| Field | Type | Null |
| transaction\_id | Int | NO |
| transaction\_date | Datetime | NO |
| transaction\_time | datetime | NO |
| transaction\_qty | Int | NO |
| store\_id | Int | NO |
| store\_  location | Object | NO |
| product\_id | Int | NO |
| unit\_price | Float | NO |
| product\_  category | Object | NO |
| product\_  type | Object | NO |
| product\_  detail | Object | NO |
| year | Int | NO |
| month | Int | NO |
| unit\_price\_  Indian | float | NO |
| price\_  category | Object | NO |
| total\_sales | float | NO |

# Data Cleaning

Data cleaning is a process that ensures the accuracy and reliability of the dataset. It includes:

1. Eliminating Null Values :

Check if any null value exist or not :

df. isna (). sum ()

2. Eliminate the Duplicates :

Check if any duplicates exists or not :

df. duplicated ()

Check if any duplicates exists or not :

df. duplicated (). sum ()

Then delete the duplicates :

df. drop\_duplicates ( inplace = True)

Recheck whether any duplicates is still there :

df. duplicated (). sum ()

Check the new shape of the dataset :

df. shape

Here the dataset didn’t contain any null values. So no need to eliminate any null values or fill any values using fillna(). Also that it didn’t contain any duplicate values. So that, after all these operations

This dataset was found to be contain 149116 rows and 16 columns.

3. Data type Conversion :

Check column datatypes using :

df. dtypes

Convert datetype of transactions\_date to date\_time using :

df[’ transaction\_date ’] = pd. to\_datetime ( df[’ transaction\_date ’],

format =’% m -% d -% Y ’)

Convert Datetype of transaction\_time to date\_time using :

df[’ transaction\_time ’] = pd. to\_datetime ( df[’ transaction\_time ’],

format =’% H -% M :% S’)

Here the datatype of columns ‘transaction\_date’ and ‘transcation\_time’ were changed to datetime.

There were no other columns to change its datetype.

# Feature Engineering

Feature engineering is performed to create new columns that enhance the analysis of the dataset.

1. Creating Unit Price in India :

df[‘unit\_price’\_india] = df[‘unit\_price’]\*84

here the unit\_price is in dollars, so that we create a new column in which the that price is inr form.

# 

1. Creating Year of sales :

df['year']=df['transaction\_date'].dt.year

1. Creating the Month of sales :

df['month']=df['transaction\_date'].dt.month

Here there were 2 columns named year and month extracted from ‘transaction\_date’ to check the sales of year and month respectively.

1. Creating the Total Sales :

df[‘total\_sales’] = df[‘transaction\_qty] \* df[‘unit\_price\_indian’]

Total sale of a coffee shop can be found by multiplying quantity to the price.

# Univariate Analysis and Visualization

Outliers check using Boxplot

Import matplotlib.pyplot as plt

# then drawing boxplot of numerical columns

for i in num1:

  plt.figure(figsize=(5,3))

  plt.boxplot(df[i])

  plt.title(f'Boxplot of {i}')

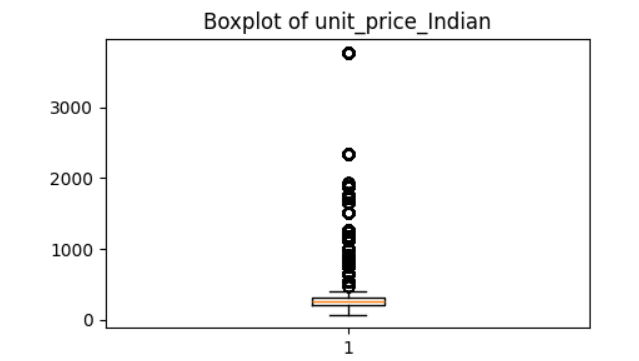
  plt.show()

A diagram of a box plot

Description automatically generated

Boxplot of Transaction Quantity

In this figure, we can see that there contains outliers, we can’t delete this outliers that the transactions quantities contins upto 8 except 7.



Boxplot of Unit Price in India

Here also we can see an outlier. But from df[‘unit\_price\_indian’].max(), we can see that the maximum price is above 3000.

Histogram of Numerical columns

For drawing histogram –

for i in num1:

  plt.hist(df[i],bins=5,color='green')

  plt.title(f'Histplot of {i}')

  plt.show()

bA green rectangular bar graph

Description automatically generated

Histplot of Transaction Quantity

From this figure, we can see that most of the quantities that customers were ordered are 1 to 2 . only least amount of customers were ordered more than 2 quantities.

A green rectangular bar graph

Description automatically generated

Histplot of Unit Price Indian

It’s the plot of unit price Indian and we can see that around 100 to 800 priced things were the one that mostly sold out. And from 1600 to 2300 were least sold out.

Countplot of categorical columns

For drawing Countplot

Import seaborn as sns

for i in obj1:

  plt.figure(figsize=(20,10))

  sns.countplot(x=df[i])

  plt.title(f'Countplot of {i}')

  plt.xticks(rotation=45)

  plt.show()

A blue rectangles with white text

Description automatically generated

Barplot of Store locaton

A screenshot of a computer screen

Description automatically generated

From this figure we can assure that the sales where high in Hells Kitchen and Astoria. Lower Manhattan is lower than both when comparing, but not too low.

A graph with blue squares

Description automatically generated

Barplot of Product Category

From this figure, it is evident that Coffee stands as the best-selling product. Tea also exhibits a high sale rate. Package Chocolate got least sale rate.

A graph of blue bars

Description automatically generated with medium confidence

Barplot of Product type

we can see that Gourmet Brewed Coffee got more sale, where Green beans and Green Tea the least sold product type.

**Bivariate Analysis and Visualization**

Heatmap of the Numerical columns

# For drawing heatmap

var1=df.corr(numeric\_only=True)

plt.title("Heatmap")

sns.heatmap(var1,annot=True)

plt.show()

A screenshot of a computer

Description automatically generated

# From this we can see that, between unit\_price and total\_sales:,there is a strong positive correlation (0.69). This implies that as the unit price increases, the total sales value also tends to increase.

# Also transaction\_qty and total\_sales shows a moderate positive correlation (0.36). This indicates that higher transaction quantities contribute somewhat to total sales, but it is not the strongest factor.

# 

Lineplots

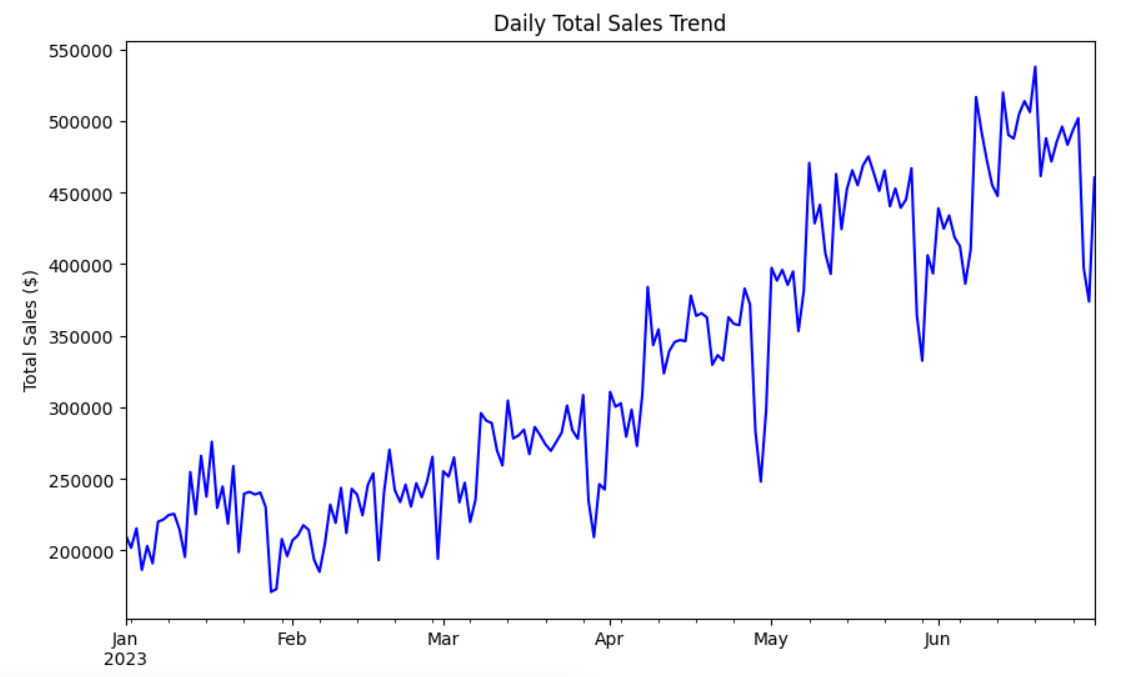
A graph of a line

Description automatically generated

# Lineplot of Transaction qty and unit price indian

From this figure, we can see the Unit Price Indian is relatively low and steady for transaction quantities between 1 and 4. Starting from Transaction\_qty = 5, there is a sharp increase in the unit price, peaking significantly at Transaction\_qty = 8. We can also assume that higher transaction quantities might correspond to higher-priced premium products, driving the increase in unit price.

# 



# Line of Tota sales and Date

The graph shows a general upward trend in daily total sales throughout the period from January to June 2023. There are fluctuations, but the overall direction is positive.

Also there seem to be some seasonal variations in sales. For instance, there's a noticeable dip in sales around April, which might be due to factors like holidays or seasonal changes in consumer behavior.

The graph indicates that the highest sales were recorded in June 2023. There's a sharp increase in sales towards the end of June, suggesting a significant peak during this period. The peak in June could be attributed to various factors like successful marketing campaigns, end-of-season sales, or holiday promotions that drove increased consumer interest and spending.

The graph also shows a noticeable dip in sales around April 2023. This suggests that sales were relatively lower during this month compared to other months in the given period. The dip in April might be related to factors like holidays, seasonal changes in consumer behavior, or economic factors that influenced purchasing power.

A graph of a bar chart

Description automatically generated with medium confidence

# Barplot of the Top 5 product by Revenue

This chart shows the top 5 products that have generated the most revenue for the business.

Product ‘Sustainably Grown Organic Lg’ generates significantly more revenue than the other products.

Customers may prefer this product due to its features, quality, or brand reputation. Or This product might be priced higher than the others, leading to higher revenue per unit sold.

The remaining 4 products have similar revenue levels.

A graph of blue bars

Description automatically generated

# Barplot of Product and revenue

This chart shows the revenue generated by different products sold by a business.

Barista Espresso is the top revenue-generating product It has the highest bar on the chart, indicating it brings in the most money. Customers may prefer this product due to its taste, quality, or brand reputation.

As we can see that Organic chocolate, Green Tea and Green beans have least revenue.

Pie Plots

A pie chart of different colored circles

Description automatically generated

Coffee is the most popular product category, accounting for 39% of the total count. This suggests that coffee is the most frequently purchased or consumed product.

Tea follows closely behind with 30% of the total count, indicating a significant portion of customers prefer tea.

Bakery items make up 15% of the total count, suggesting a good demand for bakery products.

The remaining categories, such as Drinking Chocolate, Flavors, Coffee Beans, Loose Tea, Packaged Chocolate, and Baked Goods, have relatively smaller shares of the total count.

Possible Interpretations :

The high percentage of coffee and tea indicates a general preference for these beverages among the customers. The availability and variety of coffee and tea products might be contributing to their popularity.

A pie chart with different colored circles

Description automatically generated

# Pie Plot of Revenue by Product Category

Coffee and Tea together account for a significant portion of the revenue, with Coffee contributing 38.6% and Tea contributing 28.1%. This indicates that these two categories are the primary revenue drivers for the business.

Bakery and Drinking Chocolate follow with 11.8% and 10.4% of the revenue, respectively, suggesting they are also important contributors to the overall revenue

The remaining categories, such as Flavors, Coffee Beans, Packaged Chocolate, and Loose Tea, have smaller shares of the total revenue.

A screenshot of a graph

Description automatically generated

A pie chart with numbers and a number on it

Description automatically generated

# Pie Plot of Revenue Distribution by Month

From this, we can see that June has the largest share of the revenue, accounting for 23.8%. This indicates that Month 6 is the most profitable month for the business.

Months 1, 3, and 5 also have significant shares. These months contribute 11.7%, 14.1%, and 22.4% to the total revenue, respectively, suggesting they are also important revenue drivers.

Months 2 and 4 have the smallest shares of the revenue, contributing 10.9% and 17.0%, respectively.

# Overall Conclusion and Analysis Summary

Analysis Summary

{'Sales Trends (Total Sales by Date)': transaction\_date

2023-01-01 210688.80

2023-01-02 201881.40

2023-01-03 215460.00

2023-01-04 186488.40

2023-01-05 203183.40

...

2023-06-26 493575.60

2023-06-27 501954.60

2023-06-28 397227.60

2023-06-29 373863.00

2023-06-30 460430.88

Name: total\_sales, Length: 181, dtype: float64,

'Top Performing Products (Quantity and Revenue)': total\_quantity total\_revenue

product\_detail product\_category

Sustainably Grown Organic Lg Drinking Chocolate 4453 1776747.0

Dark chocolate Lg Drinking Chocolate 4668 1764504.0

Latte Rg Coffee 4497 1605429.0

Cappuccino Lg Coffee 4151 1481907.0

Morning Sunrise Chai Lg Tea 4346 1460256.0

... ... ...

Lemon Grass Loose Tea 152 114273.6

Guatemalan Sustainably Grown Coffee beans 134 112560.0

Spicy Eye Opener Chai Loose Tea 122 112215.6

Earl Grey Loose Tea 142 106755.6

Dark chocolate Packaged Chocolate 118 63436.8

[80 rows x 2 columns],

'Store Performance (Transactions and Revenue)': total\_transactions total\_revenue

store\_id store\_location

8 Hell's Kitchen 50735 19866938.28

3 Astoria 50599 19508488.44

5 Lower Manhattan 47782 19324809.00,

'Category Breakdown (Quantity and Revenue)': total\_quantity total\_revenue

product\_category

Coffee 89250 22676005.80

Tea 69737 16498099.80

Bakery 23214 6914513.76

Drinking Chocolate 17457 6082944.00

Coffee beans 1828 3367161.00

Branded 776 1142988.00

Loose Tea 1210 941942.40

Flavours 10511 706339.20

Packaged Chocolate 487 370241.76}

* **Year**, Sales exclusively from 2023.
* **Month**, Most sales within a six-month period
* **Transaction quantity**, Most sales were singles followed by 2 lots

From the wide range of services from the three stores :

* + Coffee stands as the best-selling product.
  + Tea also exhibits a high sales rate.
* **Top Revenue Product Types :**
  + **Barista Espresso** with a revenue of **90000**
  + **Brewed Chai tea** with a revenue of **70000**
  + **Hot chocolate** with a revenue of **70000**
  + **Gourmet brewed coffee** with a revenue of **70000**
  + **Brewed Black tea** with a revenue of **45000**