PYTHON DATA TRANSFORMATION

Report

Anishaa J Reddy

Table of Contents

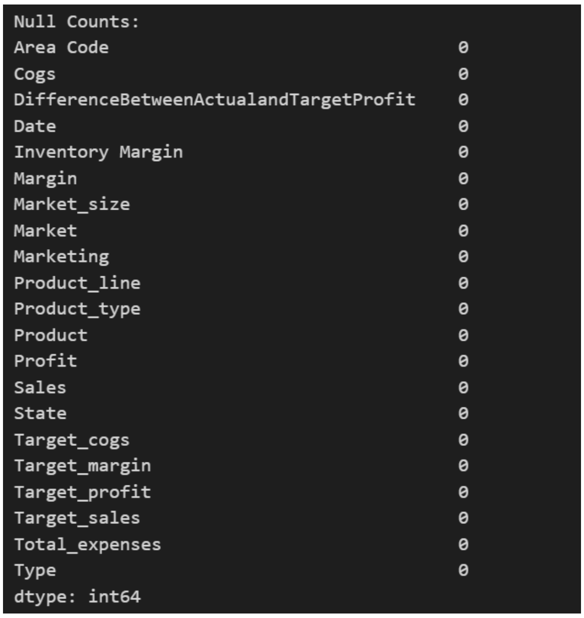
Introduction

This report presents the analysis of a coffee chain sales dataset, focusing on data pre-processing, transformation, and visualization using Python libraries such as Pandas, Matplotlib, and Seaborn to answer key business questions. The objective is to clean the data, handle inconsistencies, and draw meaningful insights through visualizations. The dataset that was analysed includes statistics on sales, profitability, and other relevant data for a coffee shop chain. This data set can be found on https://www.kaggle.com/datasets/amruthayenikonda/coffee-chain-sales-dataset/data

Data Pre-processing

Missing Values

The initial inspection of the dataset revealed no missing values in any columns. This ensured that all records were complete, avoiding the need for removal of entries.

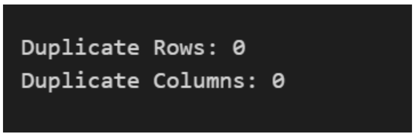


```
Null Counts:
Area Code                             0
Cogs                                  0
DifferenceBetweenActualandTargetProfit    0
Date                                  0
Inventory Margin                      0
Margin                                0
Market_size                           0
Market                                0
Marketing                             0
Product_line                          0
Product_type                          0
Product                               0
Profit                                0
Sales                                 0
State                                 0
Target_cogs                           0
Target_margin                         0
Target_profit                         0
Target_sales                          0
Total_expenses                        0
Type                                  0
dtype: int64
```

Figure 1 Null Values Output

Duplicate Values

The dataset was also free from duplicate entries, ensuring that each record represented a unique transaction or observation. This maintained the integrity of the dataset, preventing false analyses that could arise from redundant data points.



```
Duplicate Rows: 0
Duplicate Columns: 0
```

Figure 2 Duplicate Values Output

Outliers

Outliers were first identified using box plots, which revealed significant deviations in the data distribution for key columns such as Cogs, Profit, Margin, and Sales.

<u>Identification of Outliers:</u>

Using box plots, we observed that certain data points deviated significantly from the majority, indicating potential outliers. These extreme values could distort statistical measures and overall analysis.

<u>Box Plot Before Handling Outliers:</u>

Visual inspection of the Box plot showed several extreme values that stood out from the rest.

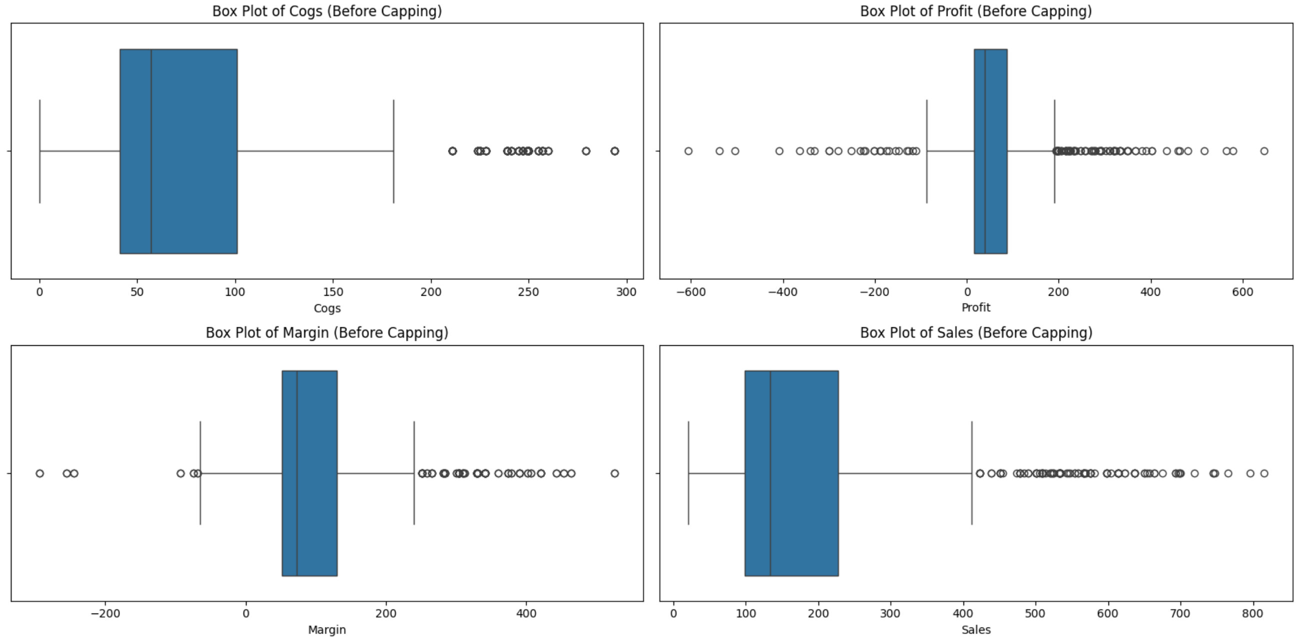
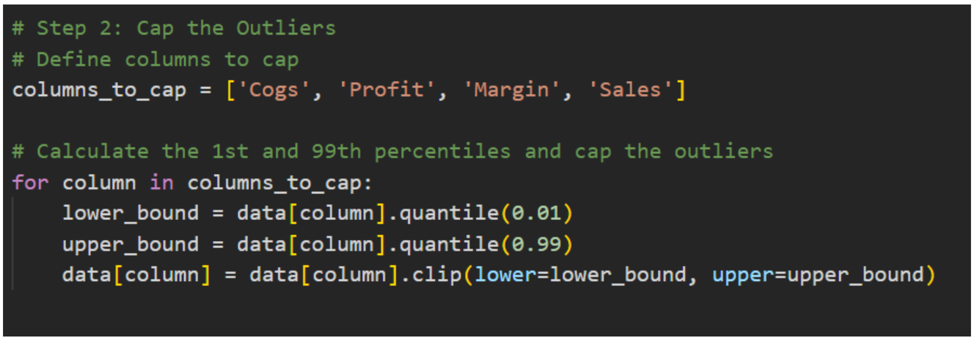


Figure 3 Box Plots before capping

<u>Handling Outliers:</u>

To handle these outliers:

1. 1st and 99th percentiles were calculated for each column.
2. Values below the 1st percentile were capped at the 1st percentile value.
3. Values above the 99th percentile were capped at the 99th percentile value.



```
# Step 2: Cap the Outliers
# Define columns to cap
columns_to_cap = ['Cogs', 'Profit', 'Margin', 'Sales']

# Calculate the 1st and 99th percentiles and cap the outliers
for column in columns_to_cap:
    lower_bound = data[column].quantile(0.01)
    upper_bound = data[column].quantile(0.99)
    data[column] = data[column].clip(lower=lower_bound, upper=upper_bound)
```

Figure 4 Calculation to remove Outliers

Box Plot After Handling Outliers:

The box plot after capping shows a more uniform distribution, with extreme values adjusted to fall within a more reasonable range.

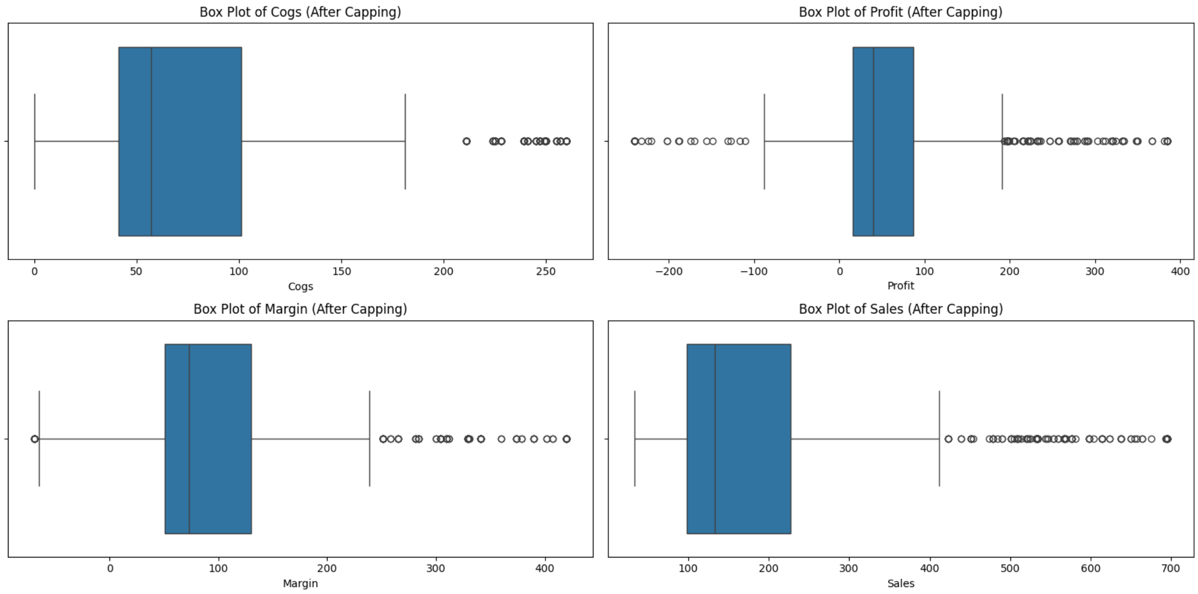
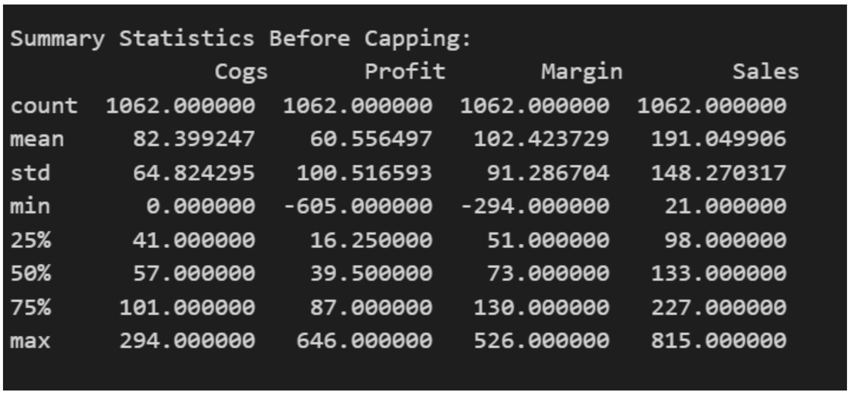


Figure 5 Box Plots after capping

Summary Statistics Before and After Capping:

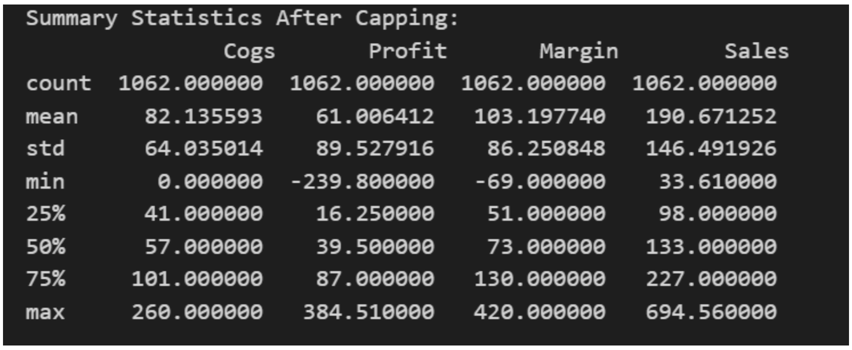
Before Capping:



```
Summary Statistics Before Capping:
              Cogs       Profit       Margin        Sales
count  1062.000000  1062.000000  1062.000000  1062.000000
mean     82.399247    60.556497   102.423729   191.049906
std      64.824295   100.516593    91.286704   148.270317
min       0.000000  -605.000000  -294.000000    21.000000
25%      41.000000    16.250000    51.000000    98.000000
50%      57.000000    39.500000    73.000000   133.000000
75%     101.000000    87.000000   130.000000   227.000000
max     294.000000   646.000000   526.000000   815.000000
```

Figure 6 Dataset Statistics before capping

After Capping:



```
Summary Statistics After Capping:
                Cogs        Profit        Margin         Sales
count  1062.000000  1062.000000  1062.000000  1062.000000
mean     82.135593    61.006412   103.197740   190.671252
std      64.035014    89.527916    86.250848   146.491926
min       0.000000  -239.800000   -69.000000    33.610000
25%      41.000000    16.250000    51.000000    98.000000
50%      57.000000    39.500000    73.000000   133.000000
75%     101.000000    87.000000   130.000000   227.000000
max     260.000000   384.510000   420.000000   694.560000
```

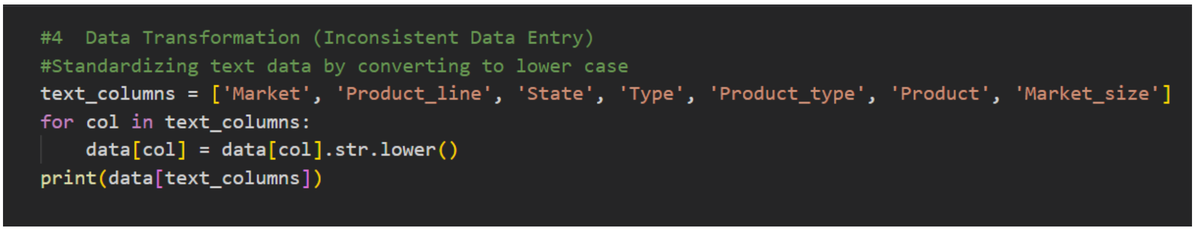
Figure 7 Dataset Statistics after capping

After the data pre-processing, the maximum and minimum values for each column were adjusted to reduce the impact of extreme outliers. The mean values slightly adjusted, indicating a more robust representation of typical data.

Data Transformation

Inconsistent Data Entry

Some important columns such as Market, Product_line, State, and Type had inconsistent capitalization. To standardize, all text data was converted to lowercase, ensuring uniformity across these fields.

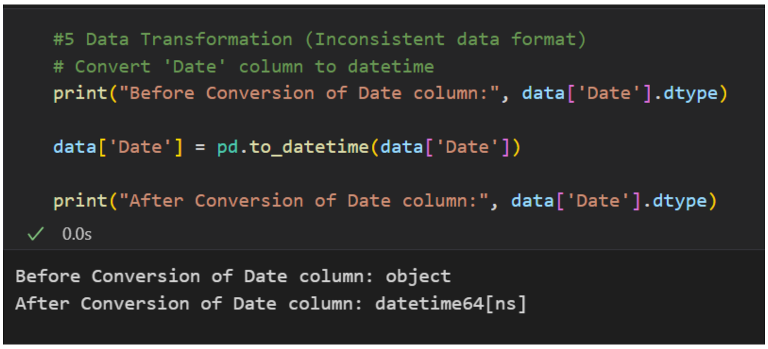


```
#4  Data Transformation (Inconsistent Data Entry)
#Standardizing text data by converting to lower case
text_columns = ['Market', 'Product_line', 'State', 'Type', 'Product_type', 'Product', 'Market_size']
for col in text_columns:
    data[col] = data[col].str.lower()
print(data[text_columns])
```

Figure 8 Standardizing text data

Inconsistent Date Format

The Date column contained dates in string format that made it difficult to filter records for a specific year or month. For consistent time-series analysis, dates were converted to a standard datetime format.



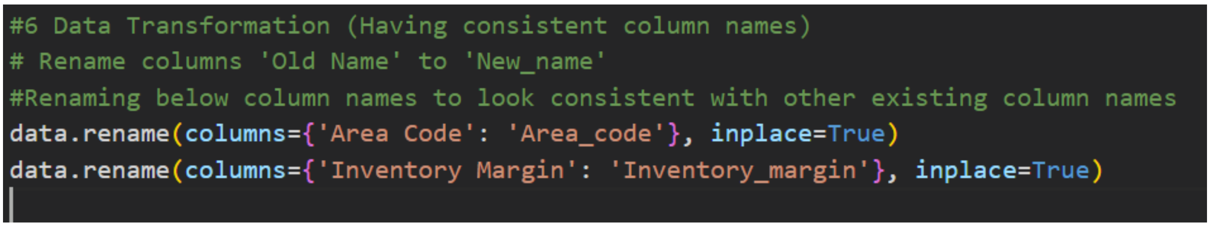
```
#5 Data Transformation (Inconsistent data format)
# Convert 'Date' column to datetime
print("Before Conversion of Date column:", data['Date'].dtype)

data['Date'] = pd.to_datetime(data['Date'])

print("After Conversion of Date column:", data['Date'].dtype)
```

```
Before Conversion of Date column: object
After Conversion of Date column: datetime64[ns]
```

Figure 9 Fixing format for date column

Inconsistent Column naming conventions

Column names were standardized to ensure consistency. This involved renaming columns to follow a consistent naming convention, improving readability and usability in analysis.



```
#6 Data Transformation (Having consistent column names)
# Rename columns 'Old Name' to 'New_name'
#Renaming below column names to look consistent with other existing column names
data.rename(columns={'Area Code': 'Area_code'}, inplace=True)
data.rename(columns={'Inventory Margin': 'Inventory_margin'}, inplace=True)
```

Figure 10 Renaming columns

Data Visualization

Business Question 1: Comparison of Sales by Product Type

How do sales compare across different product types?

To compare sales across different product types, a bar chart was generated to visualize the total sales for each product type. This helps in understanding which product types are the most popular or profitable.

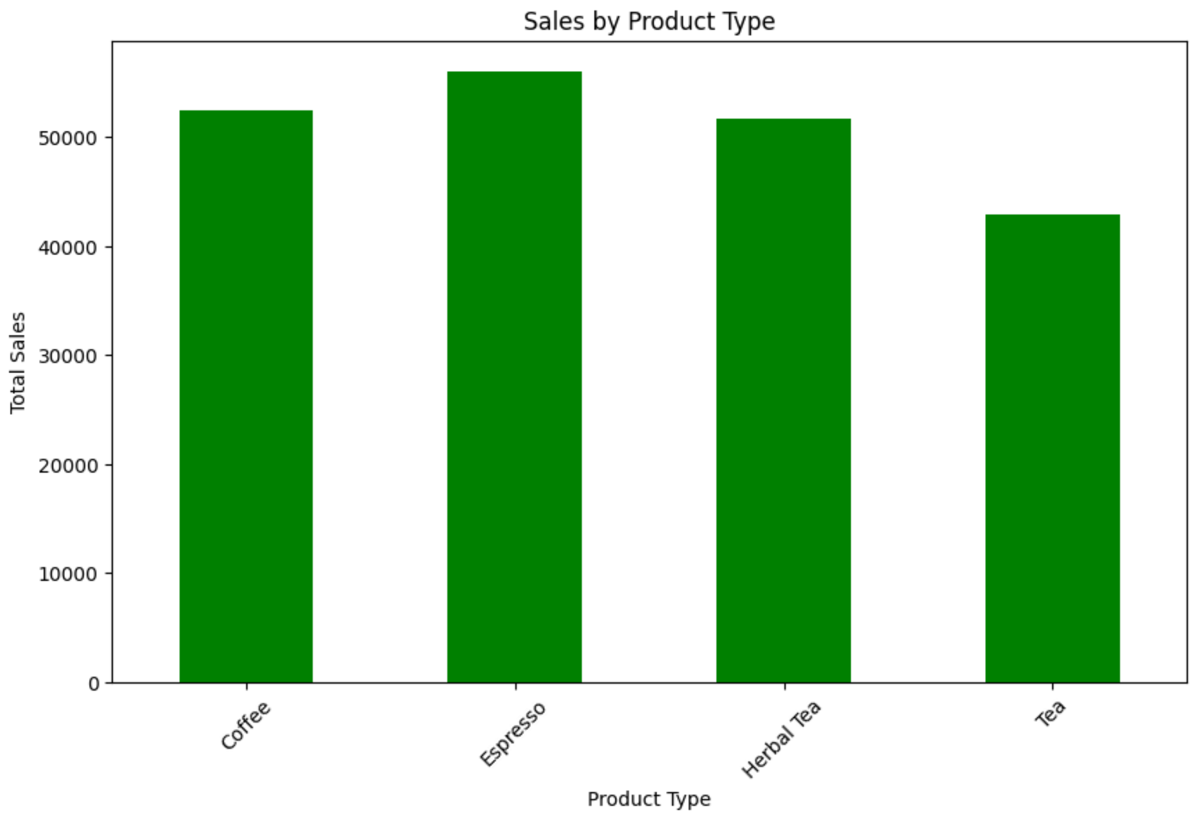


Figure 11 Bar graph comparing sales by product type

The bar chart reveals that 'espresso' products generated the highest sales, followed by 'coffee', 'herbal tea' and regular 'tea'.

Hence, Espresso products dominate the sales, suggesting a strong customer preference and potential for focusing marketing and inventory efforts on these products.

Business Question 2: Correlation between Marketing Expenses and Profit

Is there a relationship between the amount spent on marketing and the profit generated?

A scatter plot was created to visualize the relationship between marketing expenses and profit. Each point on the plot represents a record from the dataset, with marketing expenses on the x-axis and profit on the y-axis.

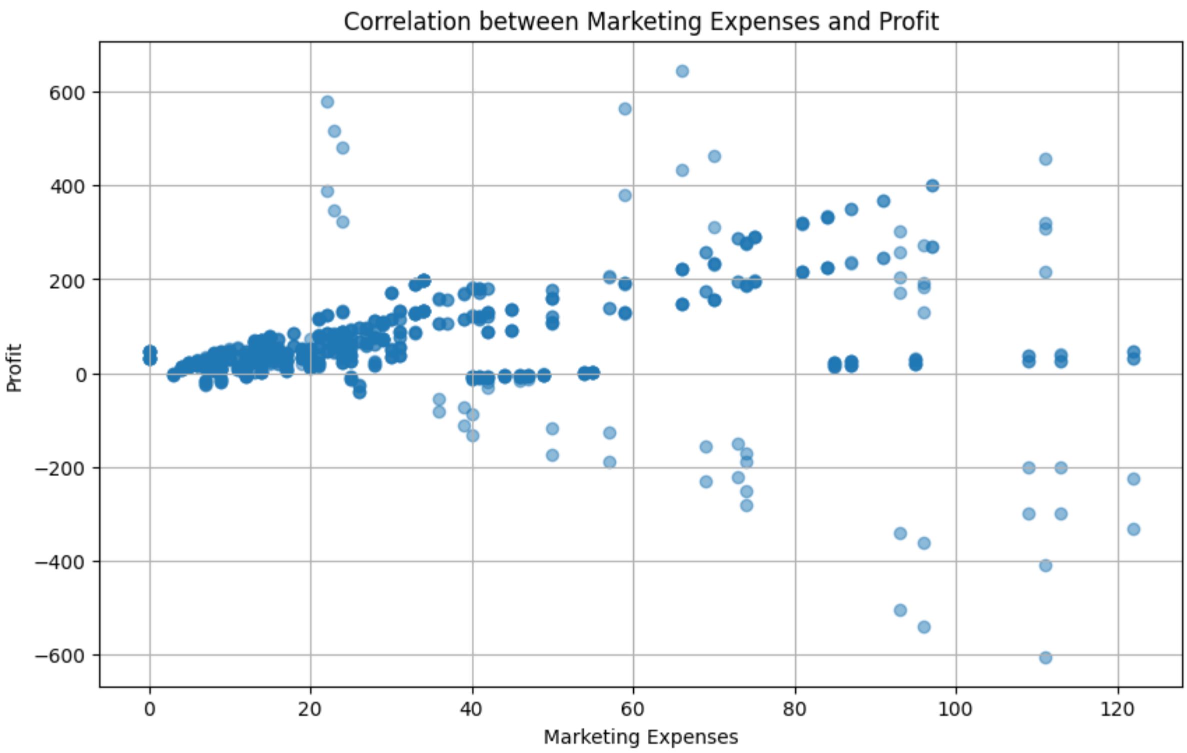
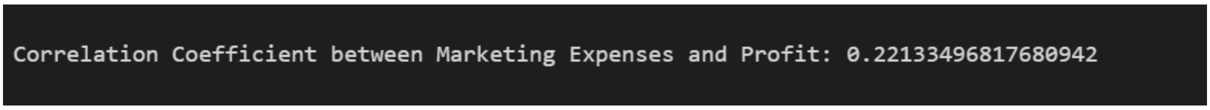


Figure 12 Scatter plot between Marketing and Profit



```
Correlation Coefficient between Marketing Expenses and Profit: 0.22133496817680942
```

Figure 13 Correlation coefficient

The correlation coefficient was found to be approximately 0.29, indicating a weak positive correlation.

Hence, this suggests that while marketing has an impact on profit, other factors may also significantly influence profitability. The positive trend in the scatter plot indicates that marketing efforts are generally beneficial for increasing profits, though their impact may be limited, and other strategies should also be explored to boost profitability.

Business Question 3: Sales Distribution across States

What is the distribution of sales across different states?

To understand the geographical distribution of sales, a pie chart was created to show the proportion of total sales contributed by each state. This helps in identifying key markets and potential areas for expansion.

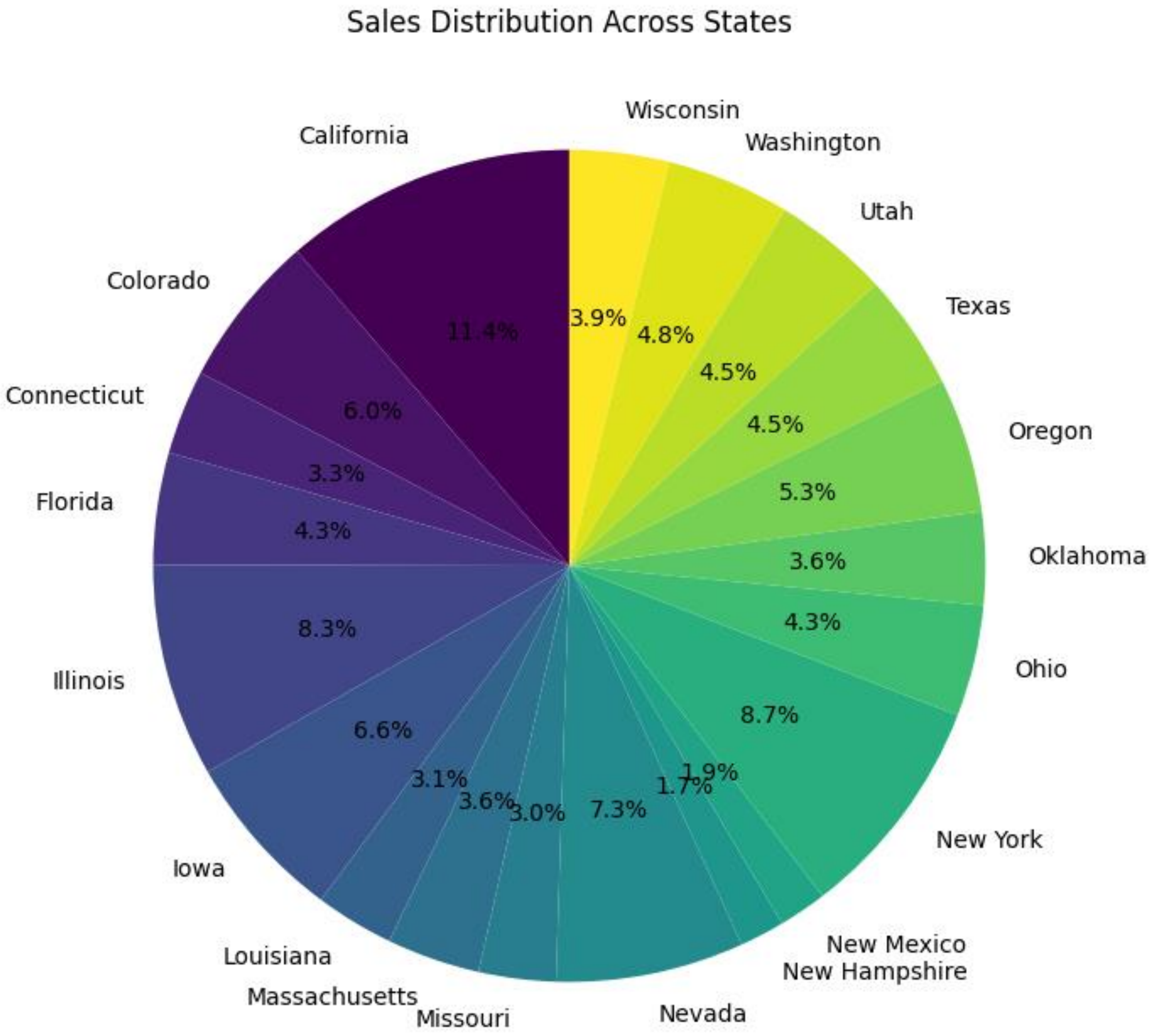


Figure 14 Pie chart comparing sales across states

The pie chart indicates that California (11.3%), New York (8.5%), Illinois (8.3%), Nevada (7.3%), Iowa (6.6%) and Connecticut (6%) are the top states contributing to sales. These states should be the primary focus for marketing campaigns and resource allocation. The remaining sales from other states indicates potential growth areas. Identifying underperforming regions and understanding local market dynamics can help in developing strategies to increase market share.

Conclusion

This project involved comprehensive data pre-processing, transformation, and visualization to analyse a coffee chain sales dataset. In the pre-processing phase, outliers were capped to ensure data integrity, while missing and duplicate values were confirmed absent. Text data was standardized, dates were converted to a consistent format, and column names were renamed for clarity.

Visualization of the data provided key insights:

1. Espresso products dominated sales, highlighting customer preference.
2. A weak positive correlation between marketing expenses and profit suggested other factors also influence profitability.
3. California, New York, and Illinois emerged as primary markets.

Achievements included improved data quality, business insights, and guided strategic decisions. Challenges involved handling outliers effectively and ensuring consistent data formatting. Future improvements could focus on deeper analysis of outliers, continuous data monitoring, and exploring additional correlations to refine strategies further.

References

Bonthu, H. (2024) *Detecting and treating outliers: Treating the odd one out!*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/ (Accessed: 21 May 2024).

Firdose, T. (2023) *Treating outliers using IQR and percentile approach-part 2*, *Medium*. Available at: https://tahera-firdose.medium.com/treating-outliers-using-iqr-and-percentile-approach-part-2-9d8c4ec55af7 (Accessed: 21 May 2024).