COFFEE CHAIN SALES

Data Processing using Python

Anishaa Joseph Reddy
14089318

Contents

Introduction

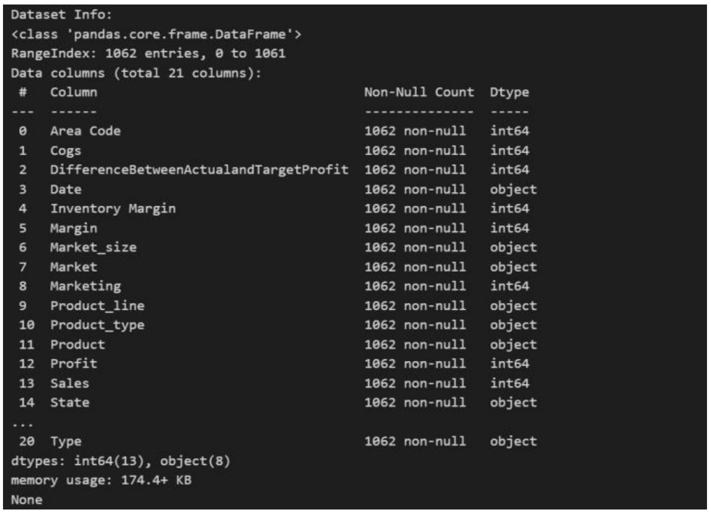
This report presents an analysis of a Coffee Chain Sales dataset using Pandas library, focusing on understanding the characteristics of the data, exploring relationships between variables, and assessing data quality. The dataset contains information about sales, margins, profits, and other relevant metrics for a coffee chain. The objective of this analysis is to gain insights into the performance of the coffee chain and identify key factors influencing sales and profitability.

Dataset

The dataset used for analysis contains information related to sales, profitability, and other relevant metrics for a coffee chain. It comprises a total of 1062 entries, each representing a transaction or observation within the dataset. The dataset consists of 21 columns, each representing a specific attribute or variable. This analysis was done using *info()* function from the Pandas library (Refer Figure 1). The dataset used for analysis can be found at https://www.kaggle.com/datasets/amruthayenikonda/coffee-chain-sales-dataset/data

Some of the key attributes are:

1. Area Code: An integer value representing the area code associated with each transaction.
2. Cogs: An integer value indicating the cost of goods sold (COGS) for each transaction.
3. Date: A timestamp indicating the date and time of each transaction.
4. Margin: An integer value representing the profit percentage for each transaction.
5. Profit: An integer value representing the profit for each transaction.
6. Sales: An integer value representing the sales volume for each transaction.
7. State: A categorical variable indicating the state associated with each transaction.



```
Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1062 entries, 0 to 1061
Data columns (total 21 columns):
 #   Column                               Non-Null Count  Dtype
---  ------                               --------------  -----
 0   Area Code                            1062 non-null   int64
 1   Cogs                                 1062 non-null   int64
 2   DifferenceBetweenActualandTargetProfit  1062 non-null   int64
 3   Date                                 1062 non-null   object
 4   Inventory Margin                     1062 non-null   int64
 5   Margin                               1062 non-null   int64
 6   Market_size                          1062 non-null   object
 7   Market                               1062 non-null   object
 8   Marketing                            1062 non-null   int64
 9   Product_line                         1062 non-null   object
 10  Product_type                         1062 non-null   object
 11  Product                              1062 non-null   object
 12  Profit                               1062 non-null   int64
 13  Sales                                1062 non-null   int64
 14  State                                1062 non-null   object
...
 20  Type                                 1062 non-null   object
dtypes: int64(13), object(8)
memory usage: 174.4+ KB
None
```

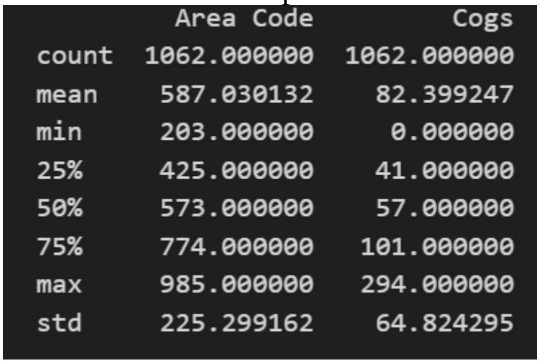
Figure 1 - DataSet Info

Data Processing Outcomes

Statistical Summary

The statistical summary achieved by *describe()* function provides insights into the financial performance and operational dynamics of the coffee chain.

Across all transactions, the average cost of goods sold (COGS) is about $82.40. This shows the average cost that the coffee chain incurs for the products that it offers (Refer Figure 2).

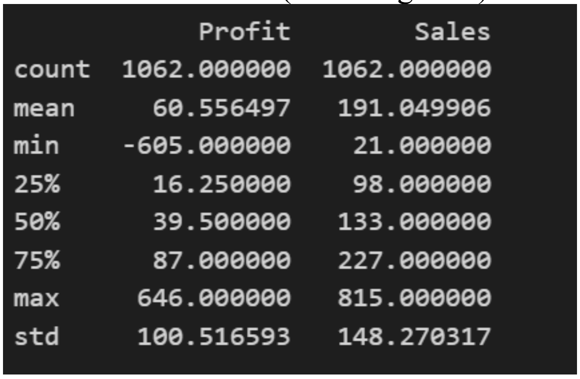


```
         Area Code         Cogs
count  1062.000000  1062.000000
mean    587.030132    82.399247
min     203.000000     0.000000
25%     425.000000    41.000000
50%     573.000000    57.000000
75%     774.000000   101.000000
max     985.000000   294.000000
std     225.299162    64.824295
```

Figure 2 - COGS statistics

An average of $60.56 is the profit made from sales transactions. This shows the average profit the coffee business makes after deducting expenses. The profit numbers, which range from -$605 to $646, show significant variation in profitability between transactions. Negative profit figures indicate situations in which expenses were higher than revenues(Refer Figure 3).

The average sales volume per transaction is $191.05. This represents the average amount of goods the coffee chain sells in a single transaction. The vast range of sales volumes, from $21 to $815, reflects the variety of customer preferences and transaction sizes. The sales volume variability around the average is indicated by the standard deviation, which is roughly $148.27. This variability may be attributed to various reasons, including seasonality, promotional activity, and fluctuations in customer demand (Refer Figure 3).



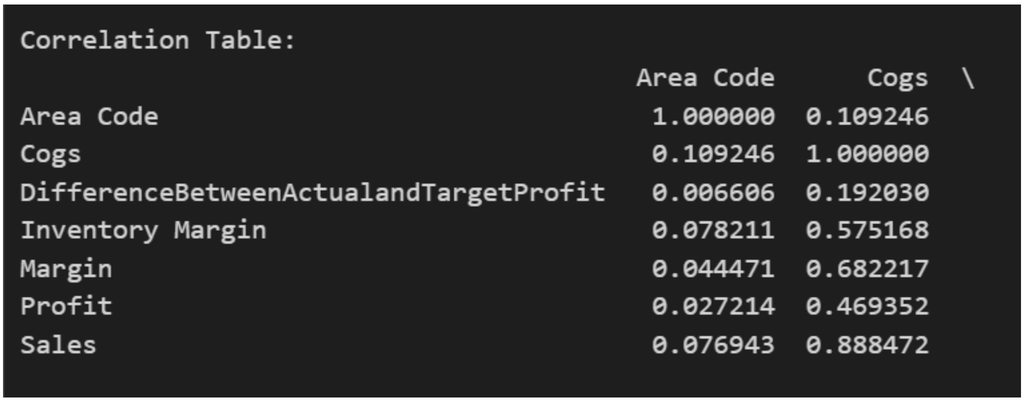
```
            Profit        Sales
count  1062.000000  1062.000000
mean     60.556497   191.049906
min    -605.000000    21.000000
25%      16.250000    98.000000
50%      39.500000   133.000000
75%      87.000000   227.000000
max     646.000000   815.000000
std     100.516593   148.270317
```

Figure 3 - Profit and Sales Statistics

Correlation Analysis

The correlation analysis done using *corr()* on numerical columns examines the relationships between variables in the Coffee Chain Sales dataset.

The strong positive correlation between sales and margin (0.9394) suggests that as sales increase, the coffee chain achieves higher margins on the products sold. This indicates effective pricing strategies or cost controls that enable the chain to maintain profitability despite increased sales volume (Refer Figure 5).

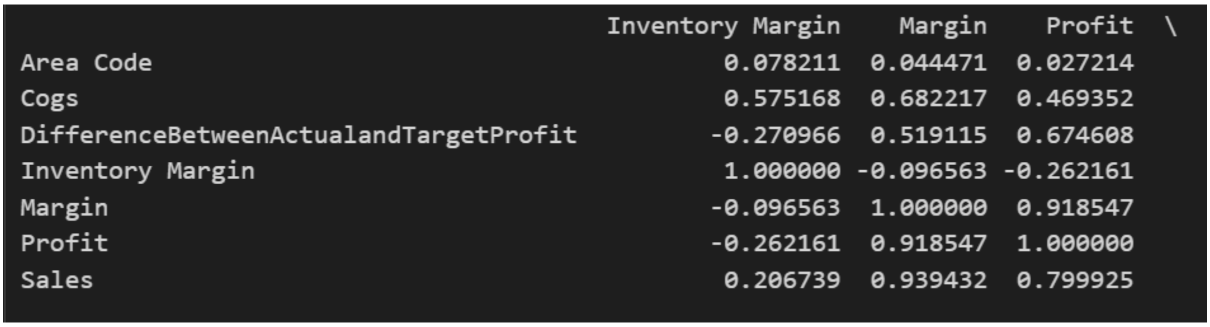


```
Correlation Table:
                                        Area Code      Cogs  \
Area Code                                1.000000  0.109246
Cogs                                     0.109246  1.000000
DifferenceBetweenActualandTargetProfit   0.006606  0.192030
Inventory Margin                         0.078211  0.575168
Margin                                   0.044471  0.682217
Profit                                   0.027214  0.469352
Sales                                    0.076943  0.888472
```

Figure 4 - Correlation between Sales and COGS

Similarly, the strong positive correlation between sales and COGS (0.8885) suggests that higher sales volumes are associated with higher costs of goods sold. This may reflect increased production or procurement costs to meet demand, highlighting the importance of cost management and efficiency in operations (Refer Figure 4).

The moderate positive correlation between sales and profit (0.7999) indicates that higher sales volumes contribute positively to profitability, although to a lesser extent than the relationship between sales and margin or COGS. This suggests that while increased sales drive higher profits, other factors such as operating expenses or overhead costs may also influence profitability (Refer Figure 5).



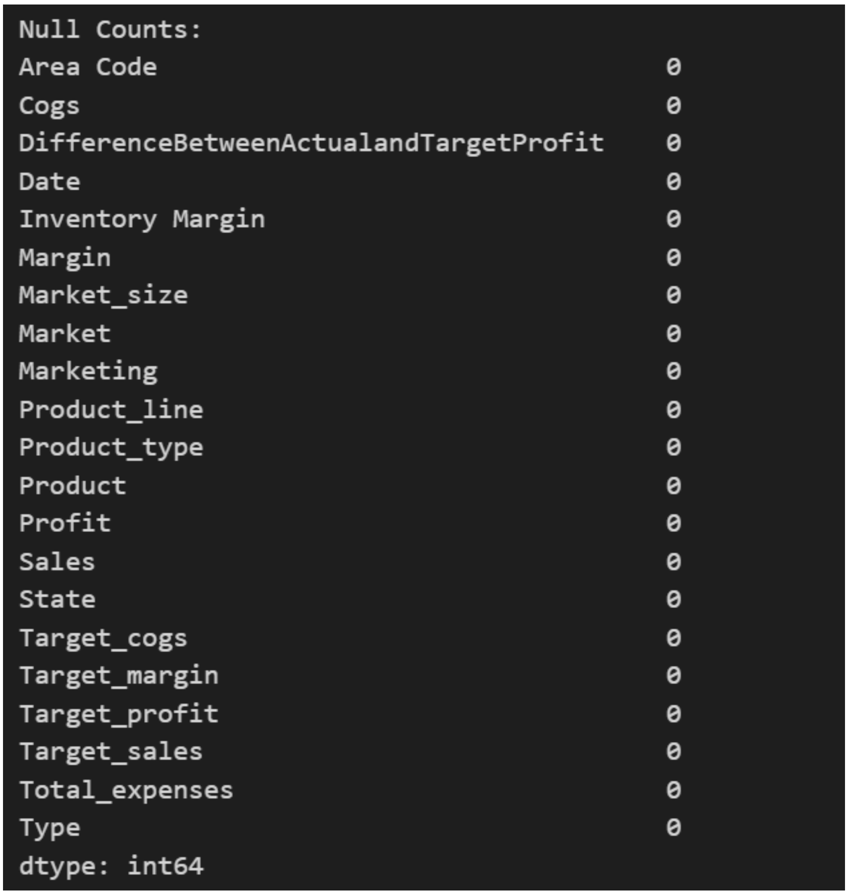
```
                                        Inventory Margin    Margin    Profit  \
Area Code                                       0.078211  0.044471  0.027214
Cogs                                            0.575168  0.682217  0.469352
DifferenceBetweenActualandTargetProfit         -0.270966  0.519115  0.674608
Inventory Margin                                1.000000 -0.096563 -0.262161
Margin                                         -0.096563  1.000000  0.918547
Profit                                         -0.262161  0.918547  1.000000
Sales                                           0.206739  0.939432  0.799925
```

Figure 5 - Correlation between Sales and Margin, Profit

Data Quality

Missing Values

Finding and fixing missing variables that could compromise the accuracy and dependability of the analysis is essential after analysing the dataset. There are no missing values in any of the columns, suggesting that the dataset is rather clean. The *info()* function's output makes this clear by showing that each column has a non-null count equal to the total number of entries (1062). Not only this but also *isnull()* functions return 0 further confirming the results.

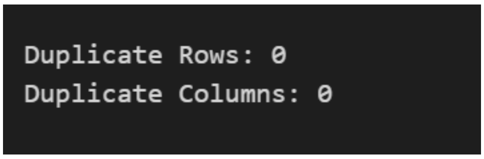


```
Null Counts:
Area Code                              0
Cogs                                   0
DifferenceBetweenActualandTargetProfit 0
Date                                   0
Inventory Margin                       0
Margin                                 0
Market_size                            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 4 - Null Values

Duplicate Entries

Duplicate rows have the potential to distort analytical findings and cause errors in assumptions made from the data. Fortunately, the dataset had no duplicate entries which was concluded by using *duplicated.count()* function on rows/columns. This implies that all of the dataset's entries are unique from one another and do not contain exact duplicates of one another.

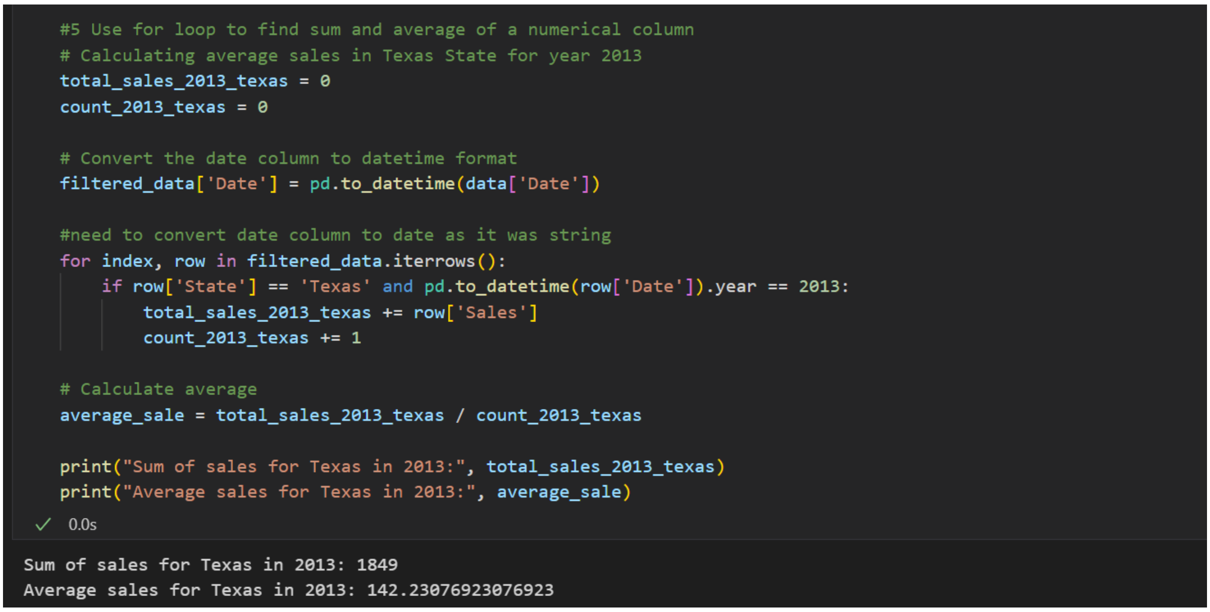


```
Duplicate Rows: 0
Duplicate Columns: 0
```

Figure 7 - Duplicate entries

Challenges

TimeStamp Handling: To extract significant insights or trends over time, additional processing may be necessary for the "Date" column, which is stored as an object datatype. For time-series analysis and trend identification, it is essential to ensure that timestamps are handled correctly, including parsing and conversion to datetime format. For example, in Figure 8 the “Date” column had to be converted to a datetime format using *to_datetime()* as it was not possible to extract sales for a particular year (in this case year 2013) due to the column being a string.



```
#5 Use for loop to find sum and average of a numerical column
# Calculating average sales in Texas State for year 2013
total_sales_2013_texas = 0
count_2013_texas = 0

# Convert the date column to datetime format
filtered_data['Date'] = pd.to_datetime(data['Date'])

#need to convert date column to date as it was string
for index, row in filtered_data.iterrows():
    if row['State'] == 'Texas' and pd.to_datetime(row['Date']).year == 2013:
        total_sales_2013_texas += row['Sales']
        count_2013_texas += 1

# Calculate average
average_sale = total_sales_2013_texas / count_2013_texas

print("Sum of sales for Texas in 2013:", total_sales_2013_texas)
print("Average sales for Texas in 2013:", average_sale)
```

✓ 0.0s

```
Sum of sales for Texas in 2013: 1849
Average sales for Texas in 2013: 142.23076923076923
```

Figure 8 - Fixing Incorrect Data Type for "Date" column

Conclusion

To wrap up, using Python's Pandas library, we analysed the coffee chain sales dataset and discovered valuable insights. We found connections between sales and profitability, thanks to statistical summaries and correlation analysis. Although the dataset was clean overall, maintaining consistent data entry is vital. The analysis done using Python, provides useful information for the coffee chain industry to plan ahead and streamline operations.