Chapter 1: Introduction

1.1 Background

In a business, multiple types of customers may buy the product. Companies may have a general understanding of who their customers are. But when in-depth research is made, several interesting insights are found about the customers. By understanding the ideal customers on a deeper level, companies can tailor their products to meet their unique needs, behaviors, and preferences. This analysis will identify different elements of the most promising customer segment. This will help the company to improve their performance.

1.2 Importance

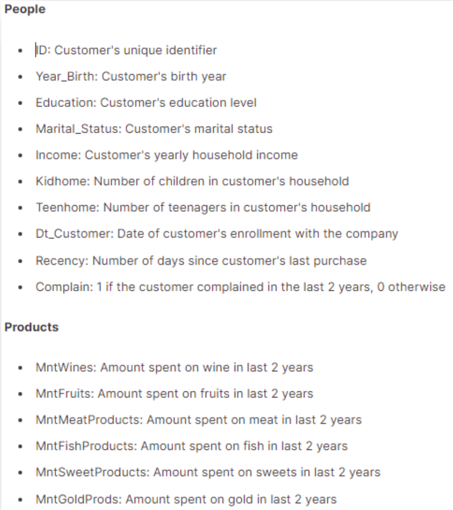
Business leaders in marketing, product development, and finance can gain a significant advantage from this analysis.

1 Managers can create better products by following strategic guidance and evaluating them effectively.
2 Companies in similar industries can also benefit from these findings.
3 With this information, management can outperform competitors and drive business growth.
4 By targeting the right markets, businesses can maximize their sales.

Chapter 2: Dataset Overview

The dataset used in the analysis is sourced from Kaggle. This dataset covers information on customer **characteristics, product, promotion and place**. But this analysis will **only cover characteristics and product**. This focus allows us to delve into the detailed product and customer that would boost the company's performance.

The data dictionary provided on the figure is only for people. And product on which this analysis will be conducted. The People table consists of data related to the characteristics of The customers whereas the products are composed of the Products offered by the company and the amount spent on each.



People

- ID: Customer's unique identifier
- Year_Birth: Customer's birth year
- Education: Customer's education level
- Marital_Status: Customer's marital status
- Income: Customer's yearly household income
- Kidhome: Number of children in customer's household
- Teenhome: Number of teenagers in customer's household
- Dt_Customer: Date of customer's enrollment with the company
- Recency: Number of days since customer's last purchase
- Complain: 1 if the customer complained in the last 2 years, 0 otherwise

Products

- MntWines: Amount spent on wine in last 2 years
- MntFruits: Amount spent on fruits in last 2 years
- MntMeatProducts: Amount spent on meat in last 2 years
- MntFishProducts: Amount spent on fish in last 2 years
- MntSweetProducts: Amount spent on sweets in last 2 years
- MntGoldProds: Amount spent on gold in last 2 years

Figure 1- Data Dictionary

Chapter 3: Data Techniques and Libraries

This analysis effectively a range of libraries and techniques for the data analysis. Some of the key libraries and functions used can be listed as;

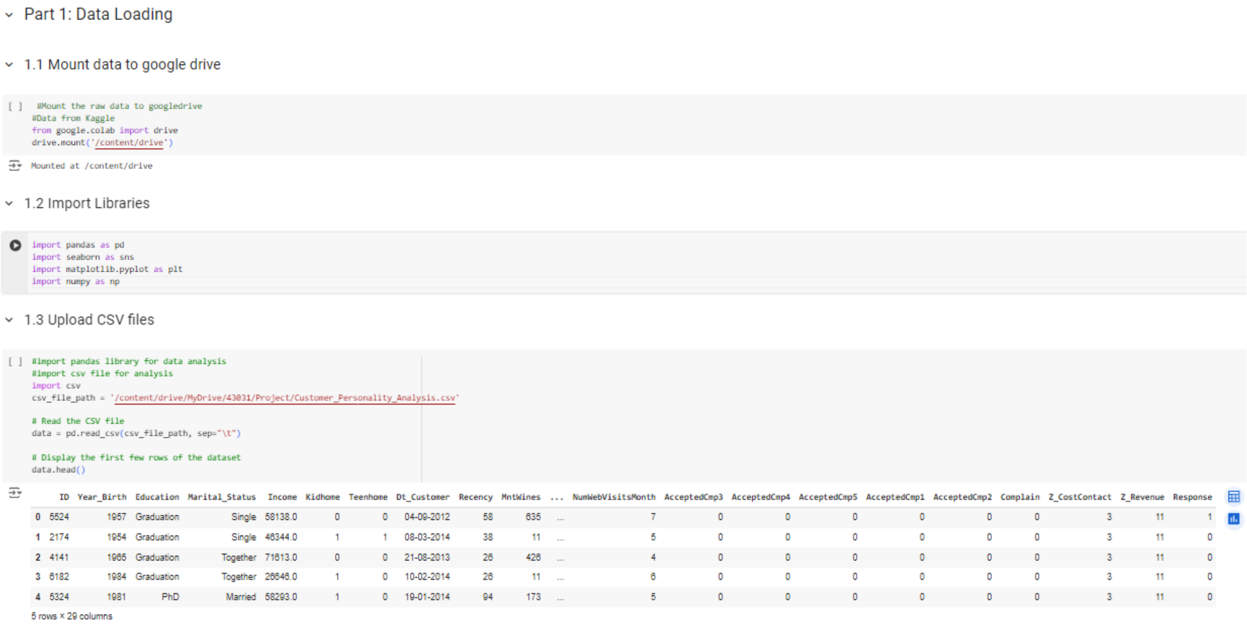
- **Pandas' library** was used for data manipulation and analysis.
- **Numpy** was used for numerical analysis.
- **Matplotlib and Seaborn library** are used for data visualization.
- The cleaning and exploration functions like read function for loading data, info for datatypes and null, heads and tails for viewing data, and various other function to replace headers, check duplicates, identify outliers and so on, is used.
- For analysis, heatmap, line chart, bar graph, scatter plots are used

Chapter 4. Data Analysis

This analysis covers preparing data for analysis. It includes loading, cleaning, verifying, and exploring the data. Each step are discussed further in the following section.

4.1 Data Loading

Data is mounted and accessed on Google drive for analysis. Before cleaning the data, pandas are installed, which can be seen in the screenshot below;



Part 1: Data Loading

1.1 Mount data to google drive

```
[ ]  #Mount the raw data to googledrive
    #Data from Kaggle
    from google.colab import drive
    drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

1.2 Import Libraries

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
```

1.3 Upload CSV files

```
[ ] #import pandas library for data analysis
    #import csv file for analysis
    import csv
    csv_file_path = '/content/drive/MyDrive/43031/Project/Customer_Personality_Analysis.csv'

    # Read the CSV file
    data = pd.read_csv(csv_file_path, sep="\t")

    # Display the first few rows of the dataset
    data.head()
```

```
     ID  Year_Birth  Education  Marital_Status   Income  Kidhome  Teenhome  Dt_Customer  Recency  MntWines  ...  NumWebVisitsMonth  AcceptedCmp3  AcceptedCmp4  AcceptedCmp5  AcceptedCmp1  AcceptedCmp2  Complain  Z_CostContact  Z_Revenue  Response
0  5524        1957  Graduation         Single  58138.0        0         0   04-09-2012       58       635  ...                  7             0             0             0             0             0         0              3         11         1
1  2174        1954  Graduation         Single  46344.0        1         1   08-03-2014       38        11  ...                  5             0             0             0             0             0         0              3         11         0
2  4141        1965  Graduation       Together  71613.0        0         0   21-08-2013       26       426  ...                  4             0             0             0             0             0         0              3         11         0
3  6182        1984  Graduation       Together  26646.0        1         0   10-02-2014       26        11  ...                  6             0             0             0             0             0         0              3         11         0
4  5324        1981         PhD        Married  58293.0        1         0   19-01-2014       94       173  ...                  5             0             0             0             0             0         0              3         11         0
5 rows × 29 columns
```

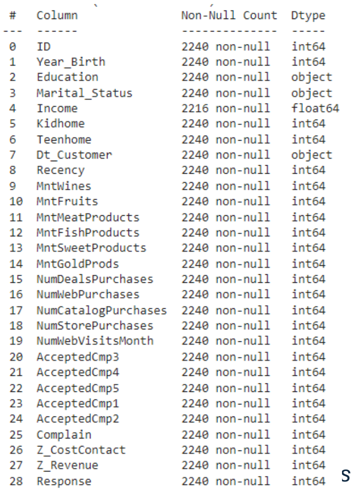
Figure 2- Data Mount and Loading

4.2 Data Cleaning and Verification

4.2.1 Null Values

The function .info is used for null values, and datatype observation in the analysis. There were 24 **missing values on the variable income** present in the data.

To refill the missing value, median is used because medians are less sensitive to outliers.



```
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   ID                   2240 non-null   int64
 1   Year_Birth           2240 non-null   int64
 2   Education            2240 non-null   object
 3   Marital_Status       2240 non-null   object
 4   Income               2216 non-null   float64
 5   Kidhome              2240 non-null   int64
 6   Teenhome             2240 non-null   int64
 7   Dt_Customer          2240 non-null   object
 8   Recency              2240 non-null   int64
 9   MntWines             2240 non-null   int64
 10  MntFruits            2240 non-null   int64
 11  MntMeatProducts      2240 non-null   int64
 12  MntFishProducts      2240 non-null   int64
 13  MntSweetProducts     2240 non-null   int64
 14  MntGoldProds         2240 non-null   int64
 15  NumDealsPurchases    2240 non-null   int64
 16  NumWebPurchases      2240 non-null   int64
 17  NumCatalogPurchases  2240 non-null   int64
 18  NumStorePurchases    2240 non-null   int64
 19  NumWebVisitsMonth    2240 non-null   int64
 20  AcceptedCmp3         2240 non-null   int64
 21  AcceptedCmp4         2240 non-null   int64
 22  AcceptedCmp5         2240 non-null   int64
 23  AcceptedCmp1         2240 non-null   int64
 24  AcceptedCmp2         2240 non-null   int64
 25  Complain             2240 non-null   int64
 26  Z_CostContact        2240 non-null   int64
 27  Z_Revenue            2240 non-null   int64
 28  Response             2240 non-null   int64
```

4.2.2 Data Transformation

- **Change of string** data information to **lowercase** for consistency using *str.**lower*** function
- **Count of Number of Rows and Columns** to verify data.
- **Change of data type** of numerical data to date type using *pd.to_numeric and .astype* function
- **Duplicate Rows and Columns** using. duplicated function. But none was found.
- *.**head*** and *.**tail*** was used to show the **first and last part** of the dataset
- **Outliers**: During data cleaning, some to identify outliers, two box plot was used. Two box plot was used for variables related to spending and other numerical variables separately. This was done for better visualization of the outliers.

  In this way, outliers were observed on almost all variables except kid home, teen home, recency and store purchase.

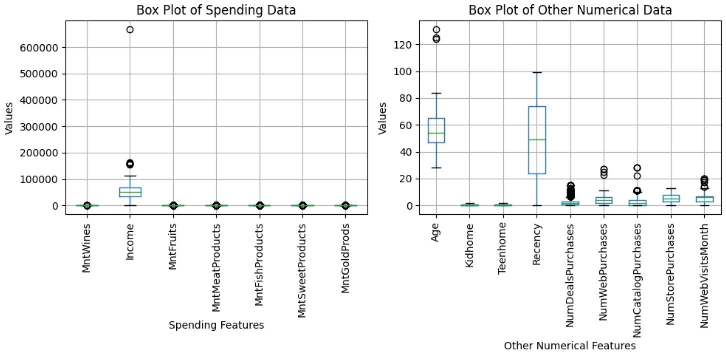
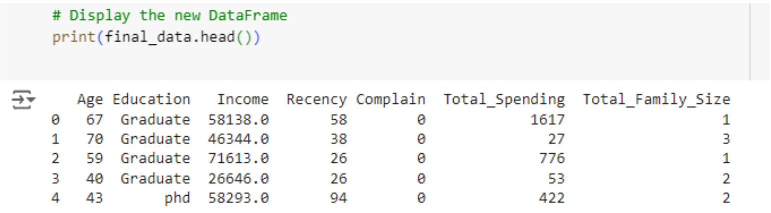


Figure 4- Box Plot for Outlier

This is the final dataset called final_data showcasing the first four rows set after cleaning. Using this dataset further data exploration is conducted.



```
# Display the new DataFrame
print(final_data.head())
```

```
   Age Education   Income  Recency  Complain  Total_Spending  Total_Family_Size
0   67  Graduate  58138.0       58         0            1617                  1
1   70  Graduate  46344.0       38         0              27                  3
2   59  Graduate  71613.0       26         0             776                  1
3   40  Graduate  26646.0       26         0              53                  2
4   43       phd  58293.0       94         0             422                  2
```

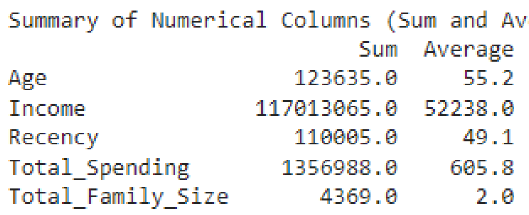
Figure 5- Clean Dataset (First Four Rows)

4.3 Exploratory Analysis

Three major analysis is done in this part, first the sum and average of each numerical values, second the statistical summary of each value, and third the correlation analysis.

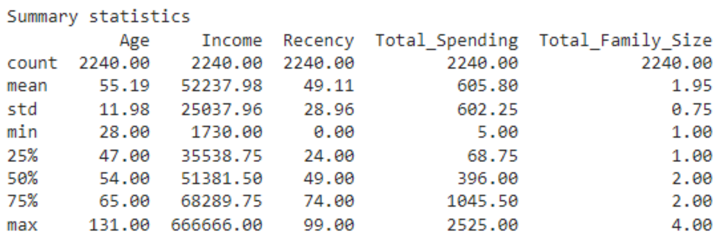
- Sum and average of numerical variables using .sum and .mean function
- .describe was used for summary statistics.
- .corr was used for correlation analysis

However, to enhance the exploratory analysis, visualization is used, which provides us with the following output



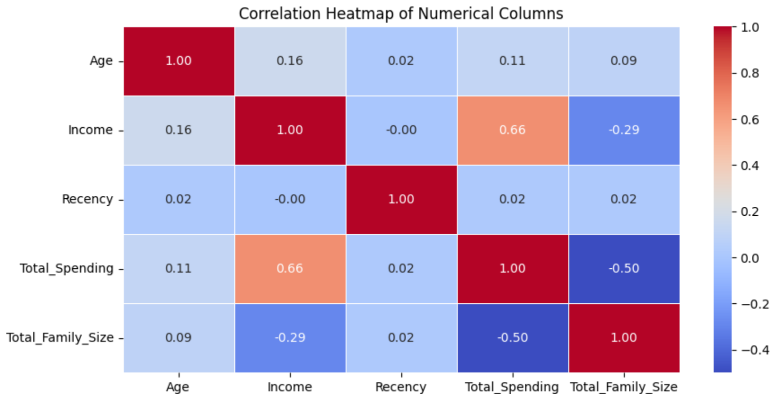
```
Summary of Numerical Columns (Sum and Average):
                         Sum   Average
Age                 123635.0      55.2
Income           117013065.0   52238.0
Recency             110005.0      49.1
Total_Spending     1356988.0     605.8
Total_Family_Size     4369.0       2.0
```

Figure 7-Sum and Average of Metrics



```
Summary statistics
           Age      Income  Recency  Total_Spending  Total_Family_Size
count  2240.00     2240.00  2240.00         2240.00            2240.00
mean     55.19    52237.98    49.11          605.80               1.95
std      11.98    25037.96    28.96          602.25               0.75
min      28.00     1730.00     0.00            5.00               1.00
25%      47.00    35538.75    24.00           68.75               1.00
50%      54.00    51381.50    49.00          396.00               2.00
75%      65.00    68289.75    74.00         1045.50               2.00
max     131.00   666666.00    99.00         2525.00               4.00
```

Figure 6 Summary Statistics



Correlation Heatmap of Numerical Columns

	Age	Income	Recency	Total_Spending	Total_Family_Size
Age	1.00	0.16	0.02	0.11	0.09
Income	0.16	1.00	-0.00	0.66	-0.29
Recency	0.02	-0.00	1.00	0.02	0.02
Total_Spending	0.11	0.66	0.02	1.00	-0.50
Total_Family_Size	0.09	-0.29	0.02	-0.50	1.00

Figure 8 Correlation Heatmap

4.4 Findings

4.4.1 Spending by Category

The customer spends most on wine products followed by meat products and then gold products. Customers mostly buy wine products, followed by meat products which are the significant sales compared to other sales.

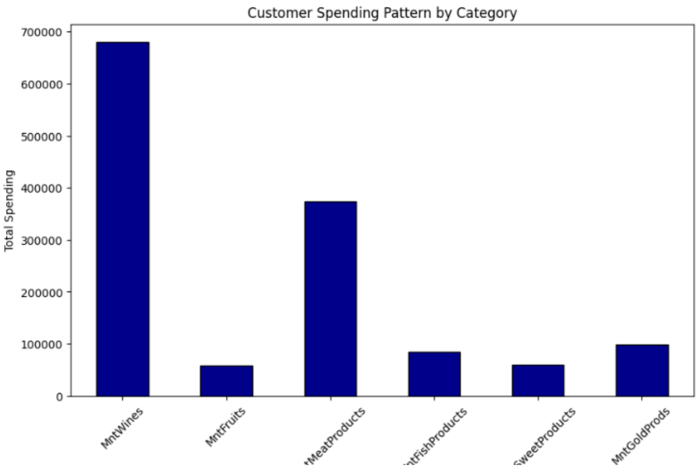
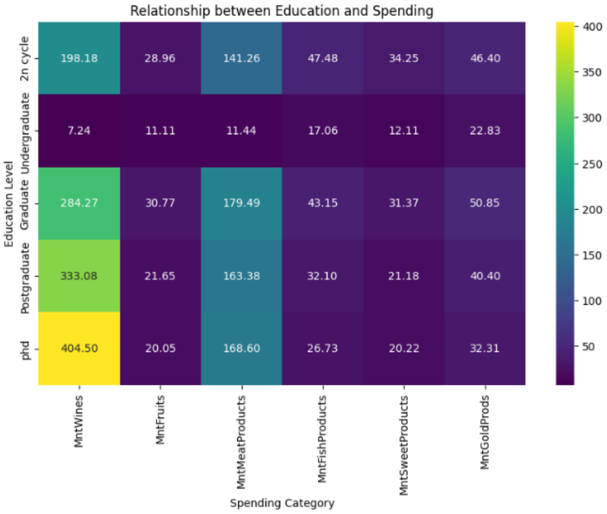


Figure 9- Customer Spending Pattern by Category

4.4.2 Education Affect Spending?

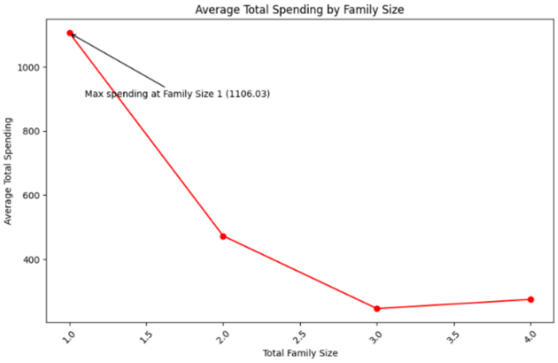


PhD students spend the most on wine products and then on meat products. Undergraduates are an exceptional case of spenders because they mostly buy gold products and are the lowest consumer of wine products.

Interestingly, customers with basic level of education are the primary type of customer for fish products compared to other education levels.

Figure 10- Heatmap- Education Affect Spending ?

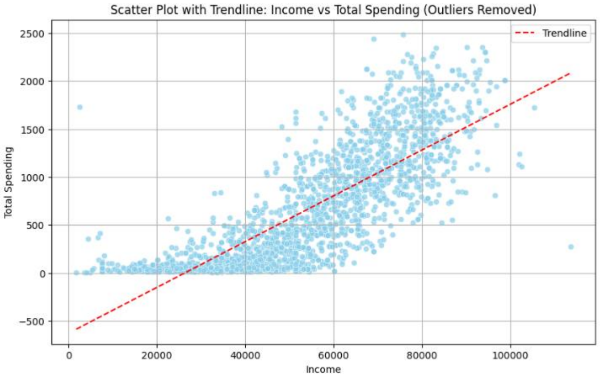
4.4.3 Family Size Affects Spending?



Single individuals spend more on items compared to family.

Figure 11- Family Size Affects Spending- Line Chart

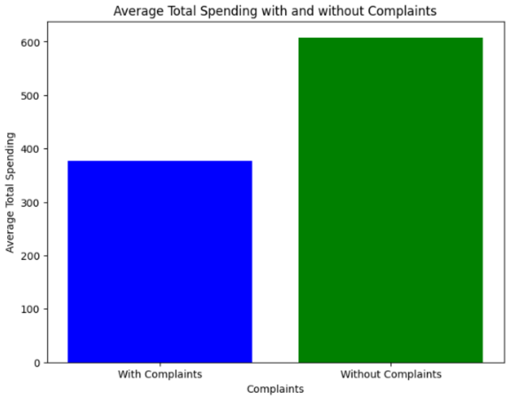
4.4.4 Relationship between Income and Spending



More income leads to more spending. And the relationship is linear.

Figure 12- Scatter Plot- Income vs Spending

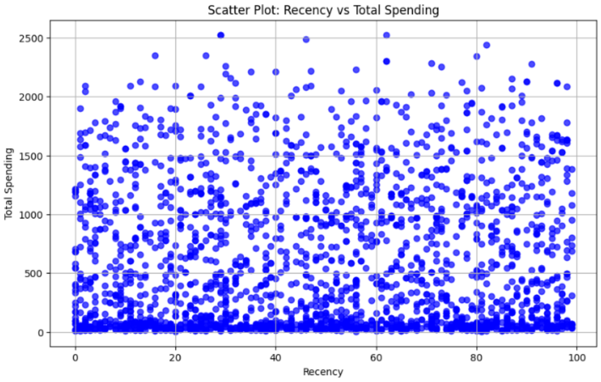
4.4.5 Complaints Affect Spending?



Customers with complaints spends less on items

Figure 13- Complaints Affect Spending ?

4.4.6 Recency Affects Spending?



Recency has no relationship with spending.

Figure 14- Recency Affect Spending?

Chapter 5. Conclusion

During the analysis, important steps such as data cleaning, verification, and exploration were conducted. Here are some key findings from the analysis process:

- Missing income filled with median value in the dataset.
- Outliers were detected and removed for better analysis.
- A table showcasing summary statistics, correlation matrix, and the sum and average of metrics was created during the exploration phase.
- The following variables were important in for sales
    - Family Size
    - Education Level
    - Complaints
    - Income

These findings provide a solid foundation for further analysis of the data. These can also help in building predictive modelling.