



INTERNATIONAL  
SCHOOL  
VIETNAM NATIONAL UNIVERSITY, HANOI

# CUSTOMER CLUSTERING IN LONDIS SUPERMARKET

Data Mining and Business Analytics



## FINAL REPORT

Group member	Nguyen Thi Ngoc Lan	- 20070943
	Hoang Thi Lan	- 20070942
	Nguyen Thi Thuy Dung	- 20070687
	Tran Huong Quynh	- 20070975
	Tran Thu Hoai	- 20070932
Group	1	
Lecturer	TS. Ha Manh Hung	
Course code	INS2061.01	

Hanoi, December 2023

## Table of Contents

Abstract .....	2
I. Project Proposal .....	3
1.1. Introduction .....	3
1.2. Motivation and Purpose of the research .....	3
1.3. The interesting question(s) to ask using the dataset: .....	4
II. Overview the dataset .....	5
III. Data Cleaning .....	6
3.1. Check missing value .....	6
3.2. Check outlier .....	7
IV. Explore Data Analysis (EDA) .....	8
4.1. Univariate analysis .....	8
4.1.1. Education .....	8
4.1.2. Marital Status .....	8
4.1.3. Children .....	9
4.1.4. Age .....	10
4.1.5. Income .....	11
4.1.6. Product .....	11
4.1.7. Web purchase .....	12
4.2. Bivariate analysis .....	13
4.2.1. Correlation matrix .....	13
4.2.2. Scatter plot .....	14
V. Model .....	14
5.1. K-Mean .....	14
5.2. Gaussian Mixture .....	21
5.3. Hierarchical Clustering .....	23
VI. Conclusion .....	26
VII. Teamwork division .....	27

## **Abstract**

This project is centered on Customer Clustering at Londis Supermarkets, focusing on segmentation to comprehend and cater to diverse customer needs. Utilizing data analysis techniques such as K-Means, Gaussian Mixture Modeling (GMM), and Hierarchical Clustering, the project aims to explore distinct customer groups based on shopping behavior, income, and demographic characteristics. The primary goal is to optimize marketing strategies, enhance shopping experiences, and strengthen customer relationships within a fiercely competitive retail environment.

The research concentrates on data collection and cleansing, followed by robust analysis to identify factors influencing customer behavior. Key inquiries were made concerning spending habits, visit frequency, product diversity, loyalty programs, demographic characteristics, purchase history, and participation in promotions, aiming to understand their impact on customer clustering at Londis Supermarket.

The data exploration phase involves data cleansing, univariate, and bivariate analysis, unveiling valuable insights into customer characteristics and behavior. The subsequent modeling phase employs K-Means, GMM, and Hierarchical Clustering techniques. K-Means clusters customers based on attributes such as birth year, income, and shopping behavior, while GMM elaborates on this clustering, offering a more nuanced perspective of customer groups and their shopping preferences.

Descriptive charts, including scatter plots and bar charts, illustrate the distribution and traits of identified customer groups. Clustering results reveal distinct groups based on income and spending patterns, providing a deeper understanding of customer segments at Londis Supermarket.

Ultimately, the project findings are presented through a comprehensive dashboard synthesizing information gleaned from the analysis. This dashboard is designed to provide an easily comprehensible overview of segmentation and customer behavior, delivering specific recommendations to optimize shopping experiences and fortify customer relationships. This

endeavor aims to equip Londis Supermarket with a competitive advantage in the fiercely competitive retail industry.

## **I. Project Proposal**

### ***1.1. Introduction***



In today's business world, better understanding your customers is not only an advantage but also a necessity for success. Londis Supermarket, in its efforts to meet the increasingly complex wants and needs of customers, especially in the retail sector, has recognized the importance of customer segmentation. This project aims not only to segment customers, but also to provide greater insight into shopping behavior and income, thereby creating a customized marketing strategy and optimizing the shopping experience for each group.

### ***1.2. Motivation and Purpose of the research***

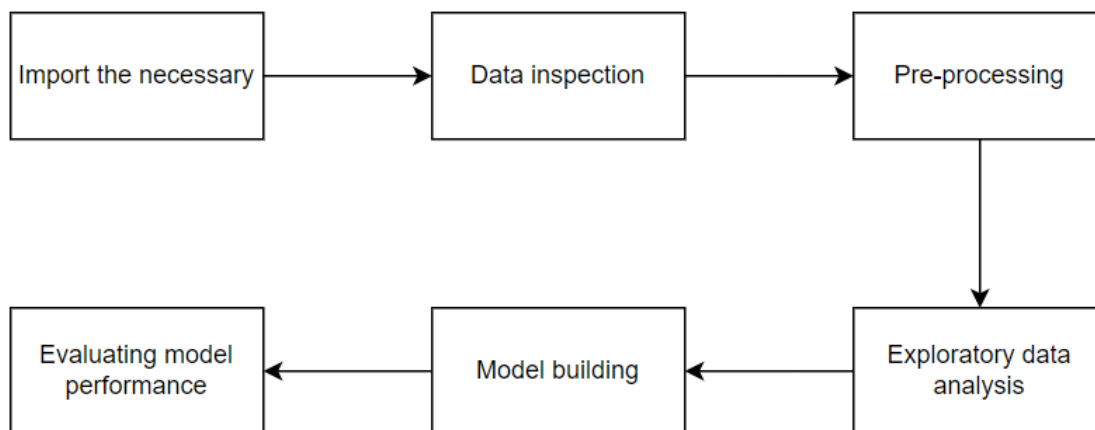
This project aims not only to classify customers but also to delve deeper into their shopping behavior and income. By implementing customer segmentation, Londis Supermarket seeks to craft customized marketing strategies and optimize the shopping experience for each distinct customer group. This initiative is expected to contribute to heightened customer satisfaction while forging a competitive edge in the rapidly evolving retail landscape.

The project commenced with data collection and exploration, wherein the data cleaning process ensured accuracy and completeness. Univariate analysis encompassing education, marital status, age, and income offers detailed insights into the factors influencing customer shopping behavior. The utilization of the K-Means algorithm aids in categorizing customers based on attributes like birth year, income, and website access behavior. Additionally, leveraging the Gaussian Mixture Model (GMM) facilitates the identification of customer clusters characterized by different data, illustrated through the distribution of these clusters via column charts. Results obtained from the GMM model complement and extend information from the

K-Means model, fostering a comprehensive understanding of customer groups and their shopping behavior.

Moreover, the inclusion of Hierarchical Clustering within the modeling phase expands the analysis by identifying hierarchical relationships among data points, further refining the understanding of customer segmentation.

Ultimately, through synthesizing and presenting information via a dashboard, this project aims to create an easily understandable and engaging overview of customer segmentation and behavior modeling. This platform intends to offer actionable recommendations and strategies to enhance the shopping experience and fortify customer relationships, thereby fostering a competitive advantage for Londis Supermarkets within the increasingly competitive retail market.



### Experiment Flowchart

#### *1.3. The interesting question(s) to ask using the dataset:*

1. Can customer spending habits predict their clustering in Londis supermarkets?
2. Is the frequency of visits to Londis a reliable indicator for customer segmentation?
3. Does the variety of products purchased by customers correlate with their clustering in Londis supermarkets?
4. Would membership length or loyalty program participation positively influence customer segmentation in Londis?
5. Are there specific demographic factors that could determine customer clusters in Londis, such as age or household size?





6. Can purchase history and basket size be used to predict customer clustering in Londis supermarket?
7. What factors contribute to a customer being categorized as a high-value shopper in Londis?
8. Is there a correlation between customer engagement with promotional offers and their clustering within Londis supermarket?

## II. Overview the dataset


The dataset, accessible via Business Analysis on Facebook, encompasses crucial information pertinent to customer segmentation within the business domain. It encompasses personal details, transaction amounts, and visit frequencies. Comprising 2240 records with 8 variables, this dataset stands as a comprehensive and valuable resource, facilitating the in-depth analysis necessary for identifying the ideal customer clusters for a company.

AutoSave

Off



Excel

 Search

FileHomeInsertPage LayoutFormulasDataReviewViewAutomateHelp

K23

⌵

⌵

fx

	A	B	C	D	E	F	G	H	I	J			
1		Year	Birth	Education	Marital	Sts	Income	Kid	Home	MntFruits	MntMeatPr	NumWebVisits	Month
2	0	1957	Graduation	Single			58138		0	88	546	7	
3	1	1954	Graduation	Single			46344		1	1	6	5	
4	2	1965	Graduation	Together			71613		0	49	127	4	
5	3	1984	Graduation	Together			26646		1	4	20	6	
6	4	1981	PhD	Married			58293		1	43	118	5	
7	5	1967	Master	Together			62513		0	42	98	6	
8	6	1971	Graduation	Divorced			55635		0	65	164	6	
9	7	1985	PhD	Married			33454		1	10	56	8	
10	8	1974	PhD	Together			30351		1	0	24	9	
11	9	1950	PhD	Together			5648		1	0	6	20	
12	11	1976	Basic	Married			7500		0	16	11	8	
13	12	1959	Graduation	Divorced			63033		0	61	480	2	
14	13	1952	Master	Divorced			59354		1	2	53	6	
15	14	1987	Graduation	Married			17323		0	14	17	8	
16	15	1946	PhD	Single			82800		0	22	115	3	
17	16	1980	Graduation	Married			41850		1	5	19	8	
18	17	1946	Graduation	Together			37760		0	5	38	7	
19	18	1949	Master	Married			76995		0	80	498	5	
20	19	1985	2n Cycle	Single			33812		1	17	19	6	
21	20	1982	Graduation	Married			37040		0	2	73	8	
22	21	1979	Graduation	Married			2447		1	1	1725	1	
23	22	1949	PhD	Married			58607		0	0	86	8	
24	23	1954	PhD	Married			65324		0	0	102	4	
25	24	1951	Graduation	Together			40689		0	3	27	8	
26	25	1969	Graduation	Single			18589		0	4	25	7	

Figure 1: Dataset Excel

The dataset facilitates comprehensive customer understanding for businesses, aiding in product customization aligned with diverse customer needs, behaviors, and preferences. It encompasses customer demographics, education levels from basic to Master's, varying income brackets, and purchasing patterns specifically focusing on fruit and meat items. This information enables targeted product modifications and precise marketing strategies for different customer segments. By clustering ideal customers, this data assists in targeting specific customer groups, optimizing product marketing strategies, and identifying potential high-growth segments, ultimately enhancing businesses' customer-centric approach and product adaptation.

### III. Data Cleaning

#### 3.1. Check missing value

```
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2240 entries, 0 to 2239
Data columns (total 9 columns):
#   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   MntFruits            2240 non-null   int64
7   MntMeatProducts      2240 non-null   int64
8   NumWebVisitsMonth    2240 non-null   int64
dtypes: float64(1), int64(6), object(2)
memory usage: 157.6+ KB
```

To understand more about the dataset we will use `data.info()`.

- The dataset consists of 8 columns and 2240 rows
- We have 24 missing values in the 'Income' column.

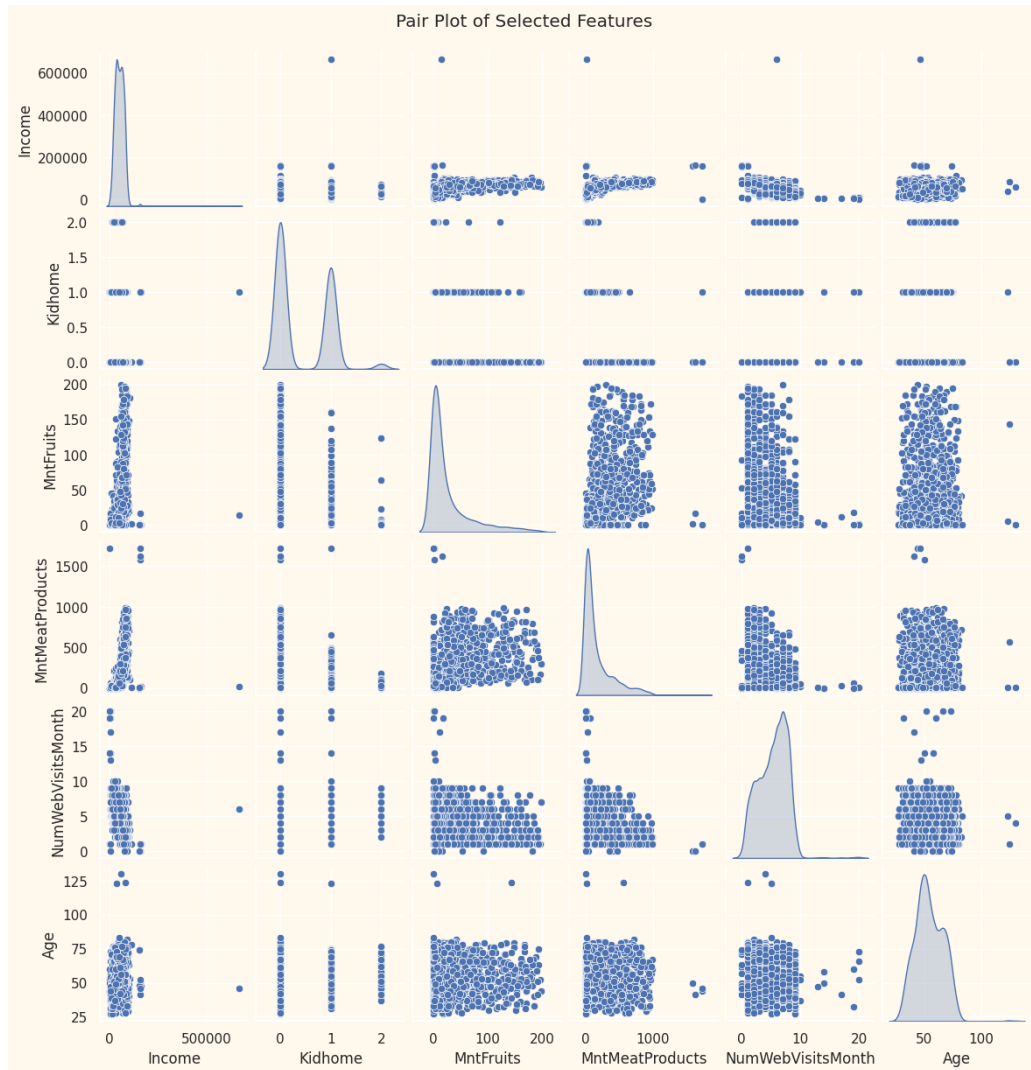
```
data = data.dropna()
print("The total number of data-points after removing the rows with missing values are:", len(data))
```

The total number of data-points after removing the rows with missing values are: 2216

We use `data.dropna` to eliminate missing values.



### 3.2. Check outlier



From the provided code, we use the Seaborn libraries to create a pair plot. It displays an overview of the data and the correlation between data dimensions in pairs on specified columns of data, like 'Income', 'Kidhome', 'MntFruits', 'MntMeatProduct', 'NumWebVisitsMonth' and 'Age'.

Clearly, there are a few outliers in the 'Income' and 'Age' features. We will be deleting the outliers in the data.



```
#Dropping the outliers by setting a cap on Age and income.
data = data[(data["Age"]<90)]
data = data[(data["Income"]<600000)]
print("The total number of data-points after removing the outliers are:", len(data))
```

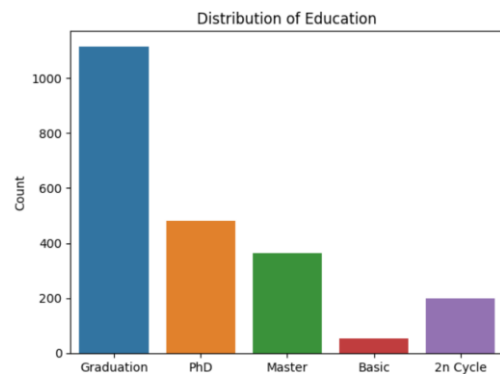
The total number of data-points after removing the outliers are: 2212

## IV. Explore Data Analysis (EDA)

### 4.1. Univariate analysis

#### 4.1.1. Education

```
Total categories in the feature Education:
Graduation    1116
PhD            481
Master        365
2n Cycle      200
Basic          54
Name: Education, dtype: int64
```



Bar charts help compare the distribution of education between groups and find out whether there are significant differences between groups. The analysis show us:

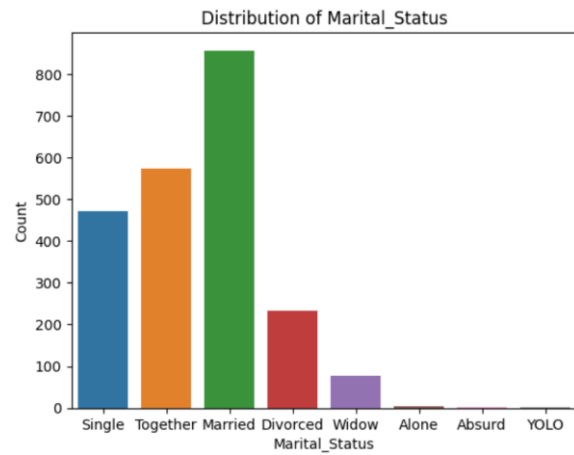
- The number of postgraduates is greater than the number of undergraduate:
  - Graduation: more than 1000
  - PhD: more than 400
  - Master : more than 300
  - 2n Cycle: more than 200
  - Basic: about 100-150

#### 4.1.2. Marital Status

Total categories in the feature Marital\_Status:

Married	857
Together	573
Single	471
Divorced	232
Widow	76
Alone	3
Absurd	2
YOLO	2

Name: Marital\_Status, dtype: int64



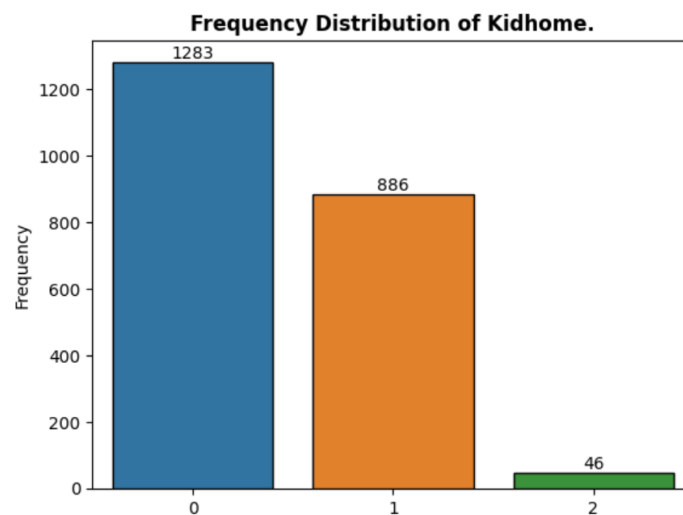
The bar chart shows the distribution of marital status. Through the data analysis, we see:

- People who live together are in a relationship more than people who live alone.

Specifically:

- Married 857
- Together 573
- Single 471
- Divorced 232
- Widow 76
- Alone 3
- Absurd 2
- YOLO 2

#### 4.1.3. Children



Analysis show that families with children (1 or 2) have fewer children than families without children

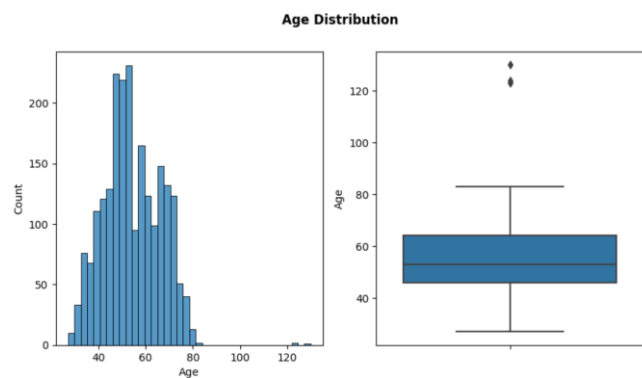
- One child: 886
- 2 child: 46
- No child: 1283

#### 4.1.4. Age

```

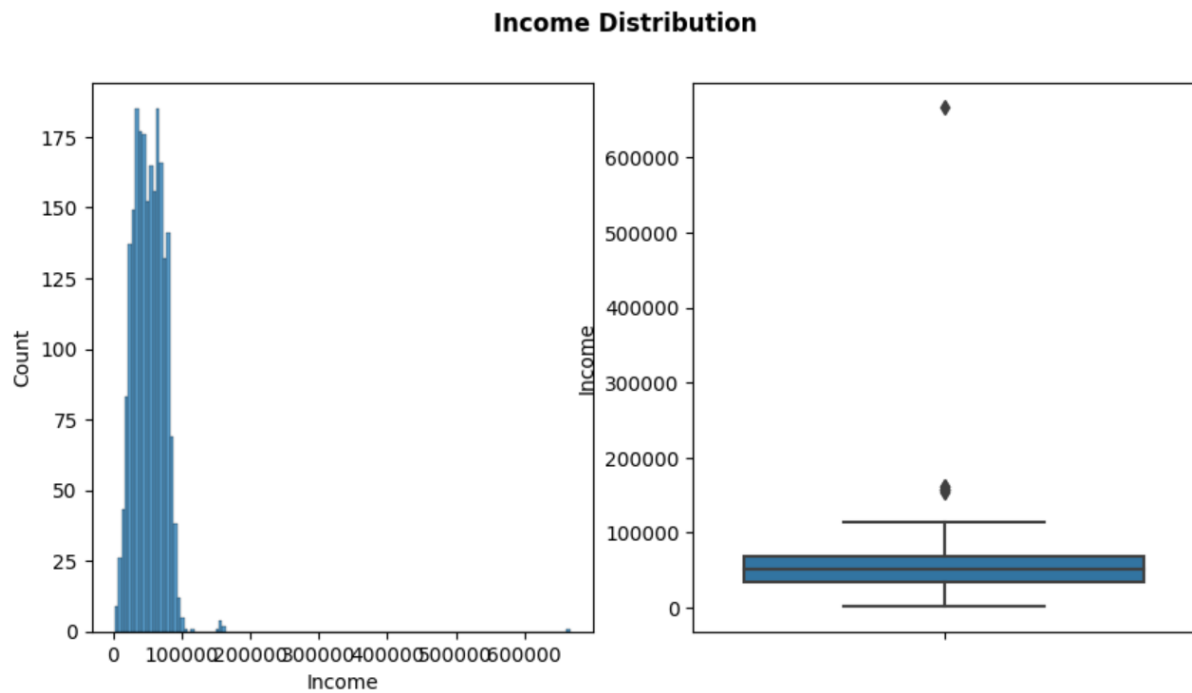
0      66
1      69
2      58
3      39
4      42
..
2235   56
2236   77
2237   42
2238   67
2239   69
Name: Age, Length: 2216, dtype: int64

```



The histograms all help to better understand the distribution of age data in the dataset. The box plot shows the distribution of the data for each quartile and identifies outliers. In summary, after analysis, we see the maximum age ranges from 30 - 60 and there are exceptions such as over 120.

#### 4.1.5. Income

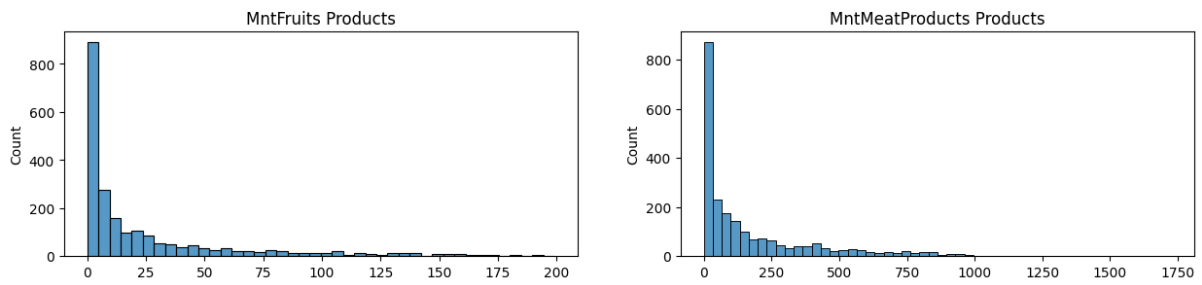


The graph shows the distribution of income in each group. We can see that income ranges from about 25000 - 75000 and there are exceptions like above 150000.

#### 4.1.6. Product

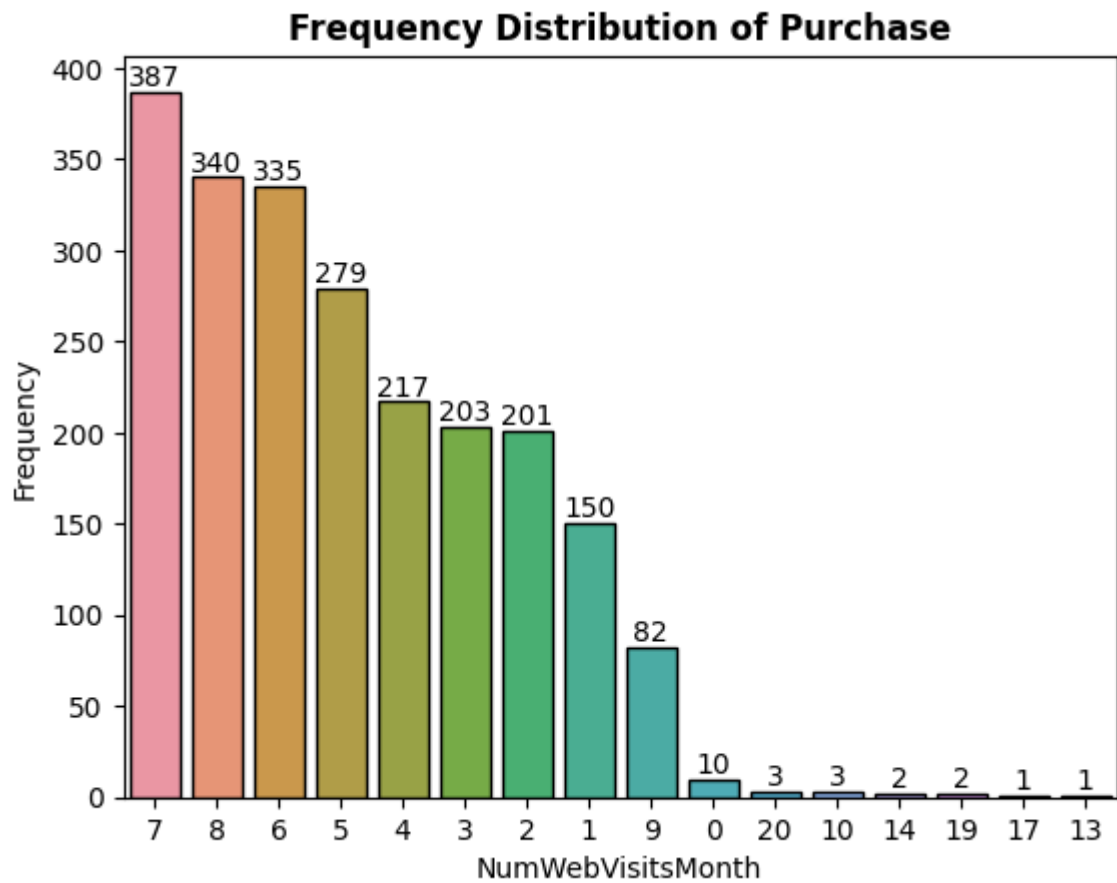
	MntFruits	MntMeatProducts		
<b>count</b>	2216.000000	2216.000000	MntFruits	58405
<b>mean</b>	26.356047	166.995939	MntMeatProducts	370063
<b>std</b>	39.793917	224.283273	dtype: int64	
<b>min</b>	0.000000	0.000000		
<b>25%</b>	2.000000	16.000000		
<b>50%</b>	8.000000	68.000000		
<b>75%</b>	33.000000	232.250000		
<b>max</b>	199.000000	1725.000000		

### Products Distribution



The histograms all help to better understand the distribution of product data in the dataset. This will show distribution of product data by MntFruits Products and MntMeatProducts. Analysis show that the level of meat purchases is higher than the level of fruit purchases

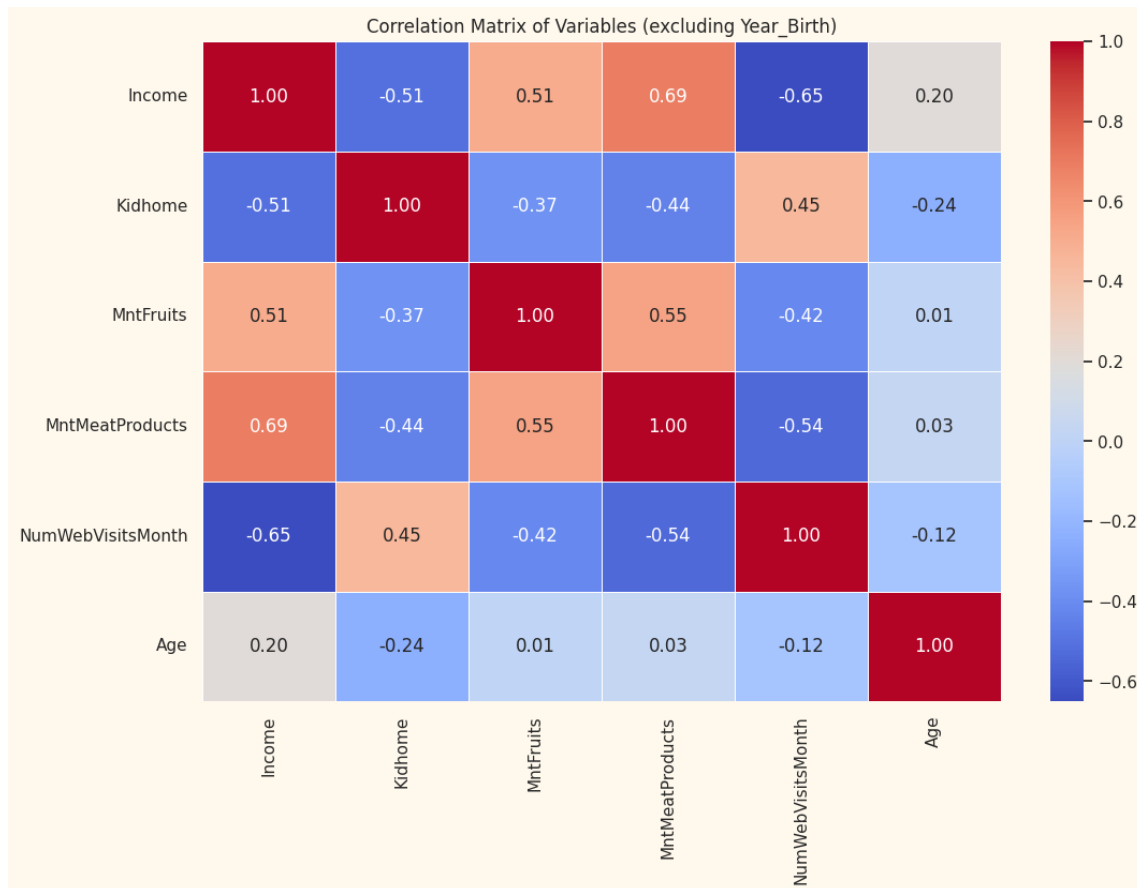
#### 4.1.7. Web purchase



The bar chart shows the distribution of purchase. We can see that the highest level of online purchases is 2-7 times per month.

## 4.2. Bivariate analysis

### 4.2.1. Correlation matrix



The graph visualizes the correlation between the variables in the data set. Correlation values are displayed in colors and numbers, making it easier to see the degree of correlation between pairs of variables.

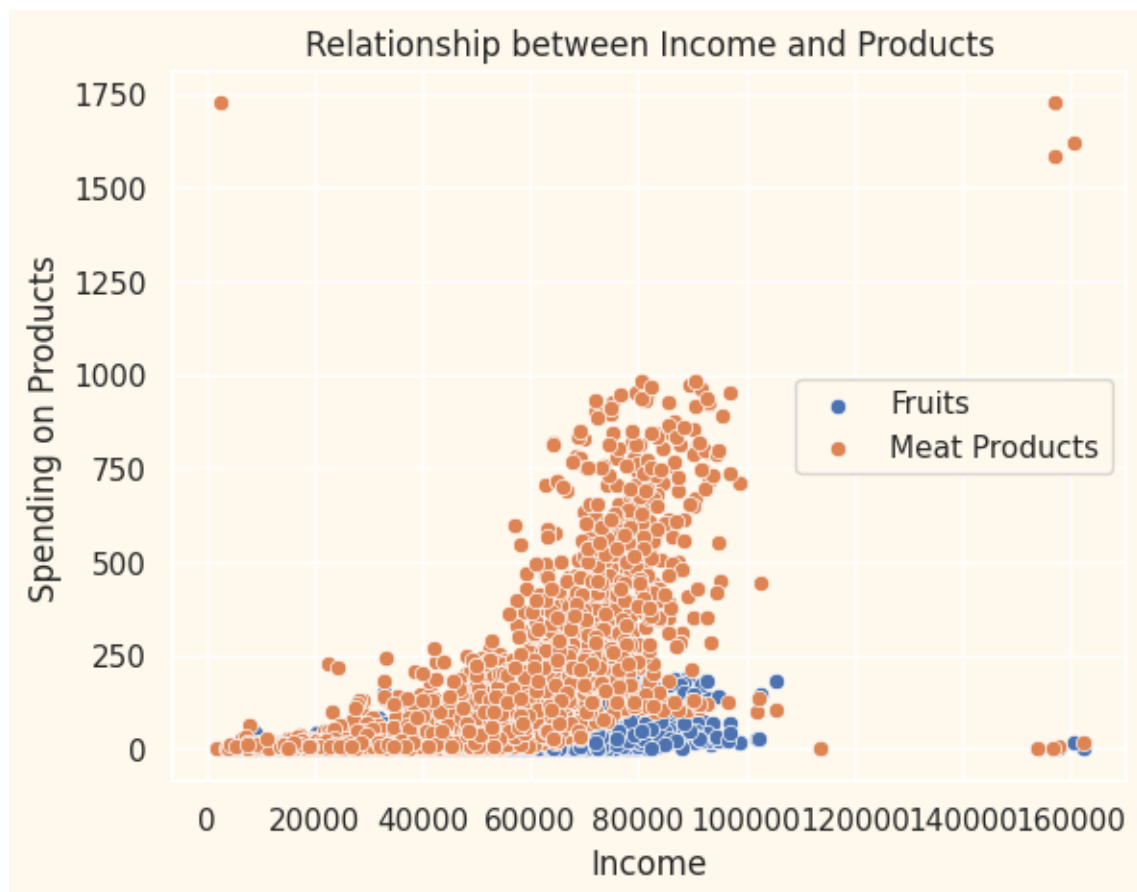
Income has some positive correlation with the MntMeatProducts and MntFruits. But income is highly correlated with MntMeatProducts. This shows that customers who spend more money on meat or fruit often also have a higher total spend.

Another positive correlation is MntMeatProducts and MntFruits.

The final positive correlation is between NumWebVisitsMonth and Kidhome. Families with children are often less likely to shop online.

We also have a negative correlation between NumWebVisitsMonth and income as previously seen.

#### 4.2.2. Scatter plot



The scatter plot points to reflect the density of points when they overlap. This graph helps us to see the dispersion of the data points and the relationship between income and product (fruits and meats).

On the scatter plot, there is a clear trend that customers with high incomes tend to spend more on meat products (MntMeatProducts). However, there are also some customers with low income but high spending on meat. The relationship is not absolute, but it can be seen that there is a correlation between income and expenditure on meat products.

## V. Model

### 5.1. K-Mean

```
selected_features = ['Year_Birth', 'Income', 'Kidhome', 'MntFruits', 'MntMeatProducts', 'NumWebVisitsMonth']

scaler = StandardScaler()
scaled_data = scaler.fit_transform(data_cleaned[selected_features])

num_clusters = 3
```



Select features including 'Year\_Birth', 'Income', 'Kidhome', 'MntFruits', 'MntMeatProducts', and 'NumWebVisitsMonth'.

Normalizing data helps bring features to the same range of mean 0 and standard deviation

1. This can improve the performance of some K-Means machine learning models

Data Transformation: normalize selected data (selected\_features) from DataFrame data\_cleaned. The result is a NumPy array containing normalized data.

Number of Clusters: 3 clusters

```
kmeans = KMeans(n_clusters=num_clusters, random_state=42)
kmeans.fit(scaled_data)
```

/usr/local/lib/python3.10/dist-packages/sklearn/cluster/\_kmeans.py:100: UserWarning: Initializing KMeans with n\_clusters=3 and random\_state=42. This may lead to non-deterministic results. To ensure reproducibility, please set the random\_state parameter to a fixed value.

KMeans

KMeans(n\_clusters=3, random\_state=42)

Create a K-Means model object with the number of clusters determined by num\_clusters (in this case, 3 clusters) and random\_state to ensure consistency of results when rerunning the model.

kmeans.fit(scaled\_data) uses scaled\_data normalized data to train the K-Means model. This process classifies each data point into one of the clusters based on correlation and distance between points.

The results of the K-Means model will show the K-Means model that was created with KMeans(n\_clusters=3, random\_state=42)

```

cluster_labels = kmeans.labels_

data_cleaned['Cluster'] = cluster_labels

<ipython-input-24-acb5a72e0f2a>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable
data_cleaned['Cluster'] = cluster_labels

print(kmeans.cluster_centers_)

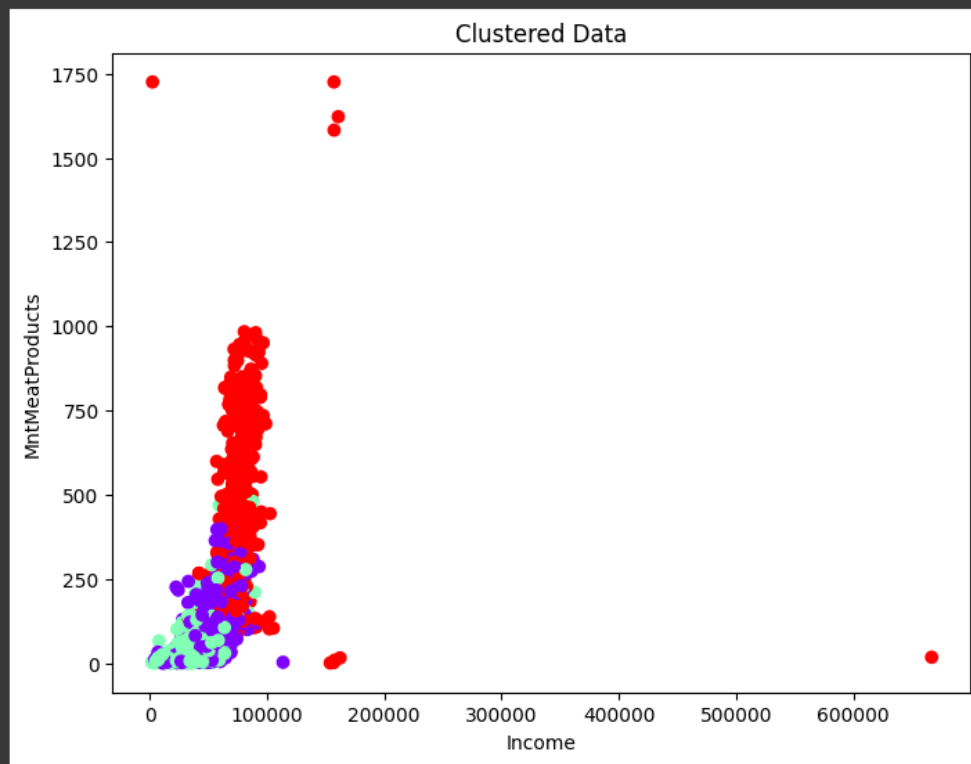
[[-0.45779805  0.05239425 -0.79809986 -0.24993708 -0.26849841  0.01859107]
 [ 0.36233642 -0.62779618  1.05783248 -0.49772987 -0.57393502  0.6056458 ]
 [ 0.00463519  1.03215511 -0.74249706  1.22758626  1.38786735 -1.09281512]]

```

The line of code `cluster_labels = kmeans.labels_` is used to extract the labels of the clusters that the K-Means model has classified the data into

Add a new column named 'Cluster' to the `data_cleaned` DataFrame. This column will contain the cluster labels that the K-Means model has classified for each data point and print out the center of each cluster that the K-Means model has identified.

```
plt.figure(figsize=(8, 6))
plt.scatter(data_cleaned['Income'], data_cleaned['MntMeatProducts'], c=data_cleaned['Cluster'], cmap='rainbow')
plt.xlabel('Income')
plt.ylabel('MntMeatProducts')
plt.title('Clustered Data')
plt.show()
```



The red cluster includes data points with low income and low meat spending.

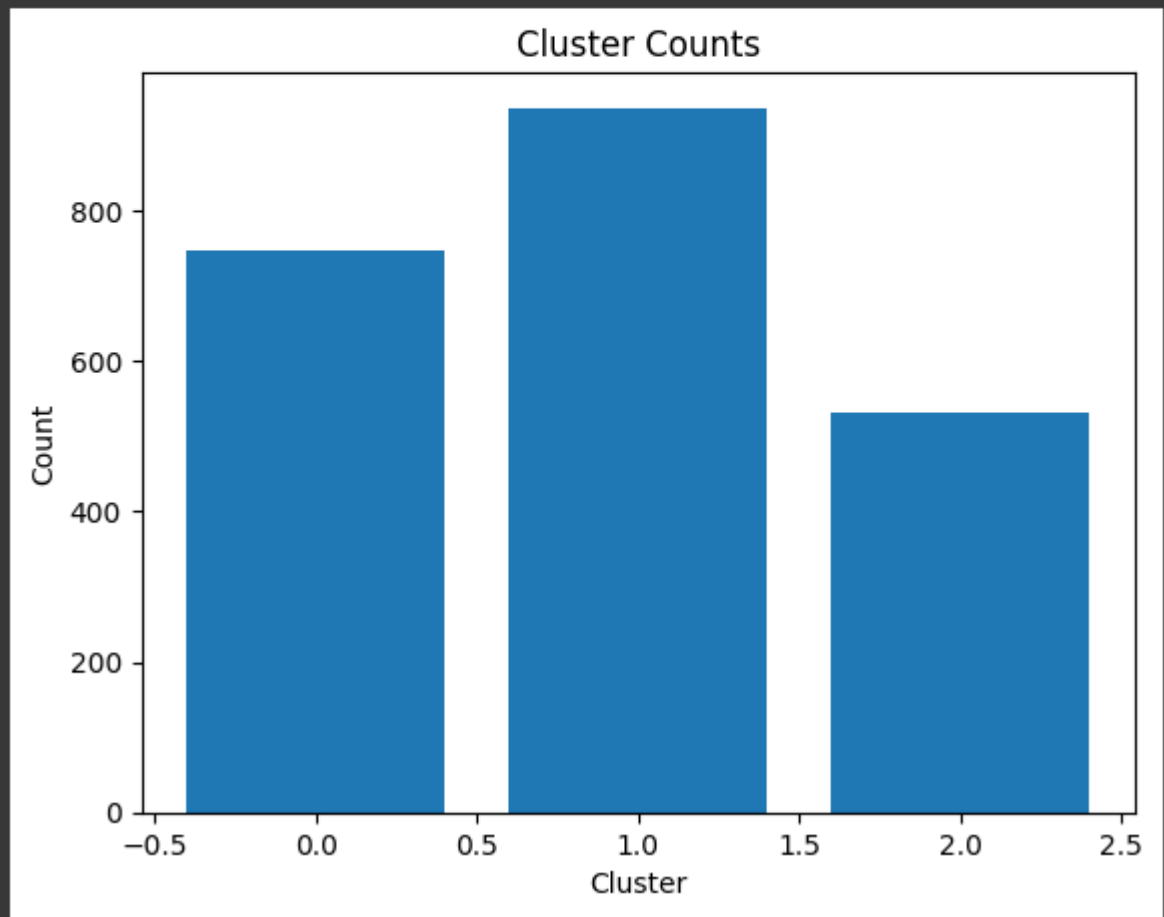
The blue cluster includes data points with average income and average meat spending.

The purple cluster includes data points with high income and high meat spending.

This code creates a scatter plot of the clustered data, with the points colored by cluster. The plot shows the relationship between income and the amount of money spent on meat products. The plot shows that there is a positive correlation between income and the amount of money spent on meat products. This means that there are groups of people who have similar income and spending patterns.

```
cluster_counts = data_cleaned['Cluster'].value_counts()
```

```
plt.bar(cluster_counts.index, cluster_counts.values)  
plt.xlabel('Cluster')  
plt.ylabel('Count')  
plt.title('Cluster Counts')  
plt.show()
```



Cluster 0: data points with low income and low meat spending.

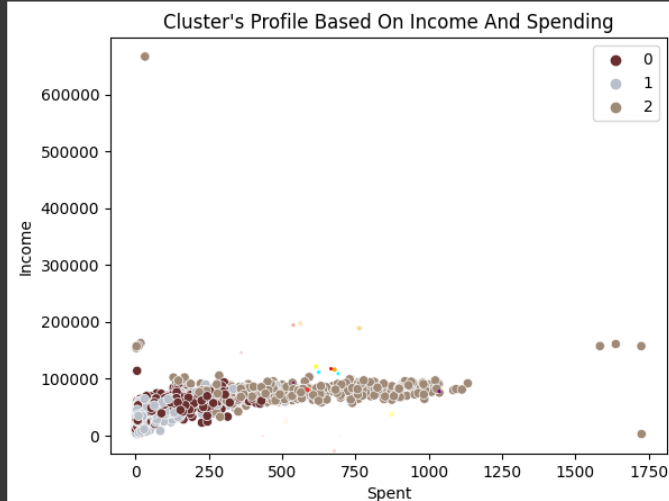
Cluster 1: data points with average income and average meat expenditure.

Cluster 2: data points with high income and high meat spending.

```
pal = ["#682F2F", "#B9C0C9", "#9F8A78", "#F3AB60"]
```

```
pl = sns.scatterplot(data = data_cleaned, x=data_cleaned["Spent"], y=data_cleaned["Income"], hue=data_cleaned["Cluster"], palette= pal)
pl.set_title("Cluster's Profile Based On Income And Spending")
plt.legend()
plt.show()
```

```
<ipython-input-31-42342c541cd4>:3: UserWarning: The palette list has more values (4) than needed (3), which may not be intended.
pl = sns.scatterplot(data = data_cleaned, x=data_cleaned["Spent"], y=data_cleaned["Income"], hue=data_cleaned["Cluster"], palette= pal)
```

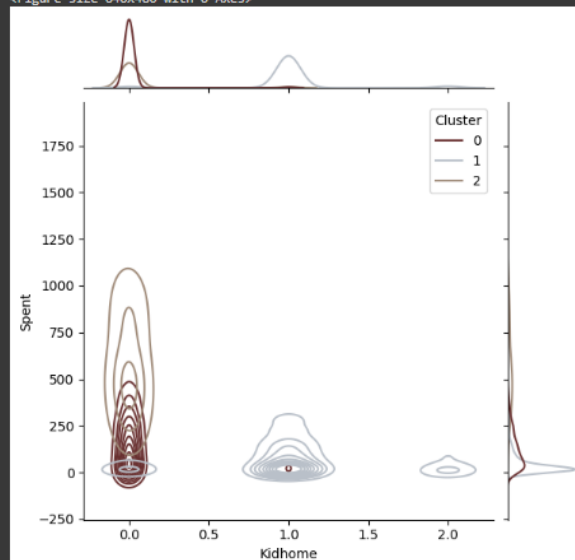


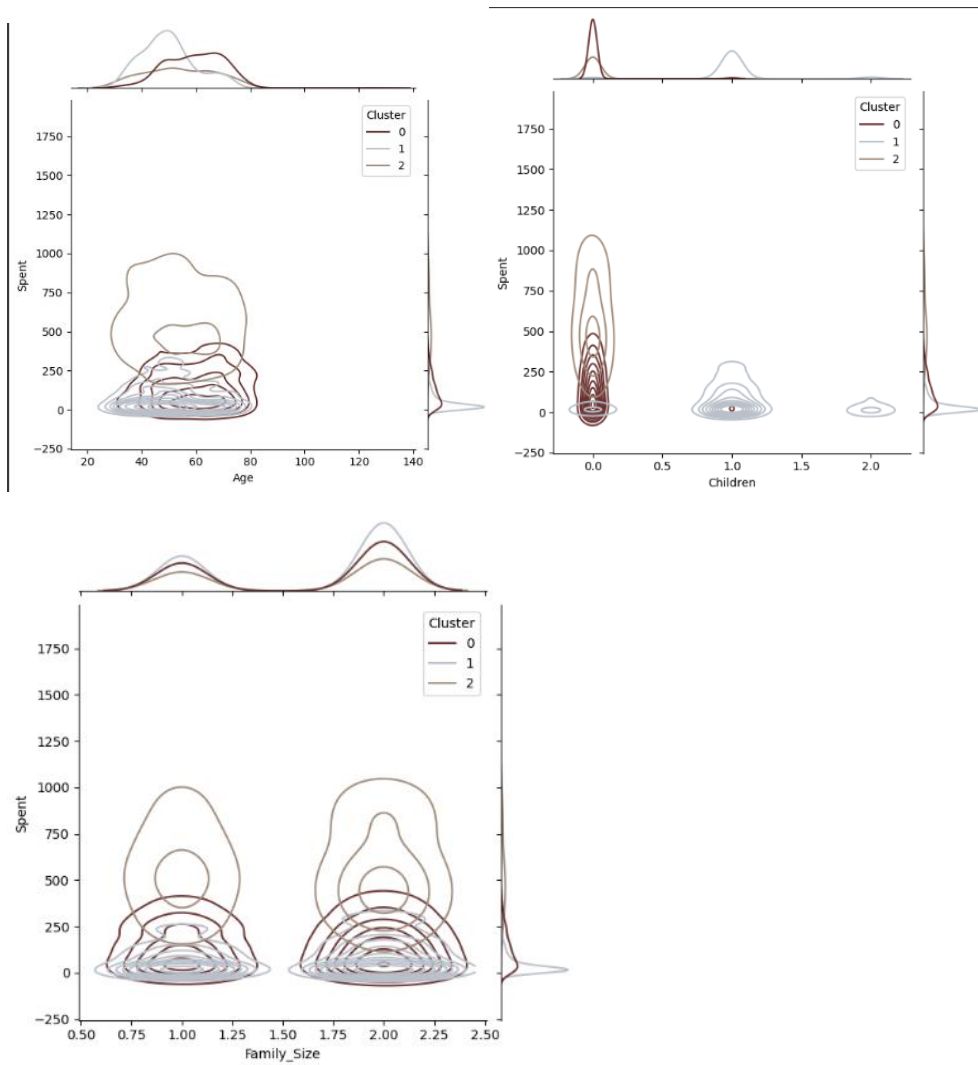
```
import seaborn as sns
import matplotlib.pyplot as plt
```

```
Personal = ["Kidhome", "Age", "Children", "Family_Size", "Education", "Living_With"]
```

```
for i in Personal:
    # Kiểm tra xem biến có phải là kiểu số hoặc ngày tháng không
    if data_cleaned[i].dtype in ['int64', 'float64']:
        plt.figure()
        sns.jointplot(x=data_cleaned[i], y=data_cleaned["Spent"], hue=data_cleaned["Cluster"], kind="kde", palette=pal)
        plt.show()
    else:
        print(f"Ignoring {i} as it is not a numeric variable.")
```

```
func(x=self.x, y=self.y, **kwargs)
/usr/local/lib/python3.10/dist-packages/seaborn/axisgrid.py:1877: UserWarning: The palette list has more values (4) than needed (3), which may not be intended.
func(x=self.x, ax=self.ax_marg_x, **kwargs)
/usr/local/lib/python3.10/dist-packages/seaborn/axisgrid.py:1883: UserWarning: The palette list has more values (4) than needed (3), which may not be intended.
func(y=self.y, ax=self.ax_marg_y, **kwargs)
<Figure size 640x480 with 0 Axes>
```





## 5.2. Gaussian Mixture

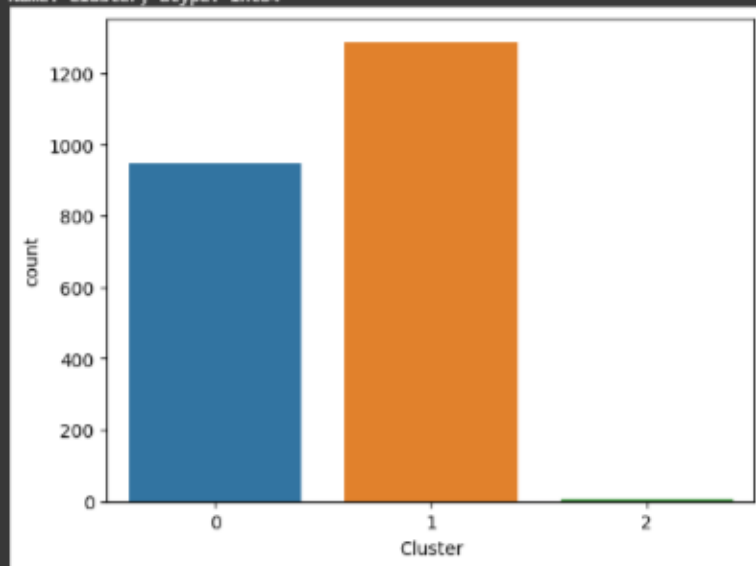
```
# Sử dụng GaussianMixture để phân cụm dữ liệu
gmm = GaussianMixture(n_components=3, random_state=42)
labels = gmm.fit_predict(selected_data_scaled)

# Thêm nhãn cụm vào DataFrame gốc
data['Cluster'] = labels

# Hiển thị thông tin về kích thước của mỗi cụm
print(data['Cluster'].value_counts())

# Hiển thị phân phối của các cụm
sns.countplot(x='Cluster', data=data)
plt.show()
```

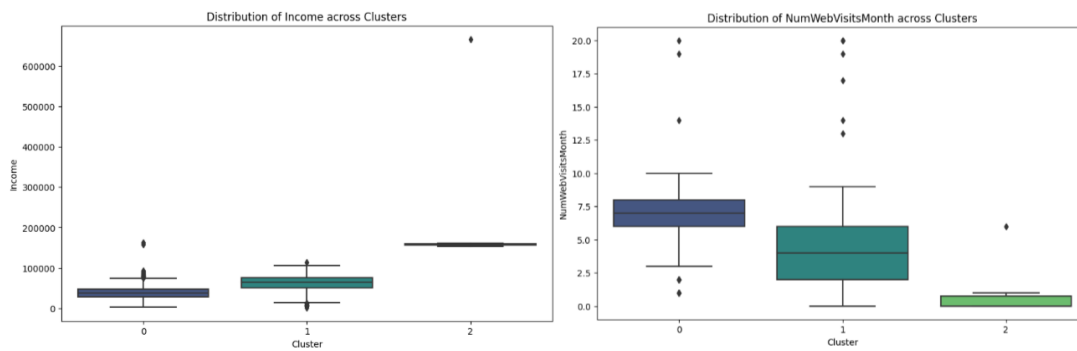
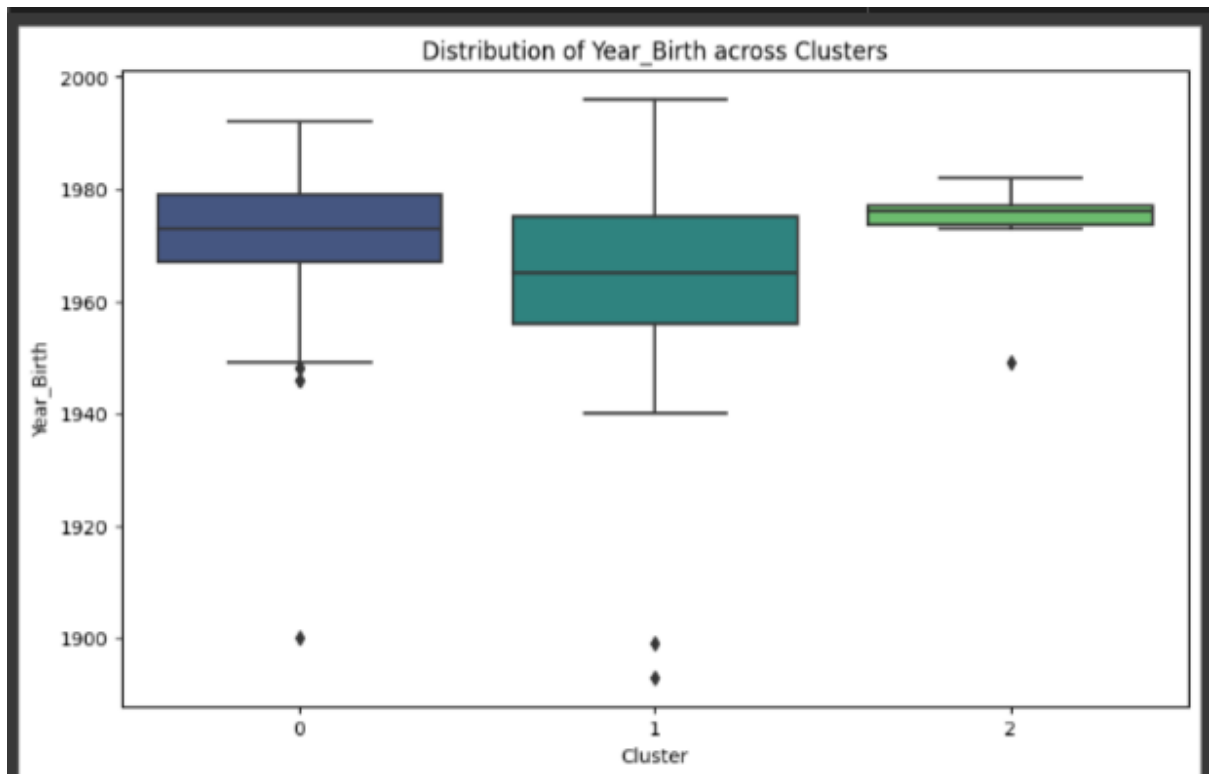
```
1    1288
0     946
2         6
Name: Cluster, dtype: int64
```

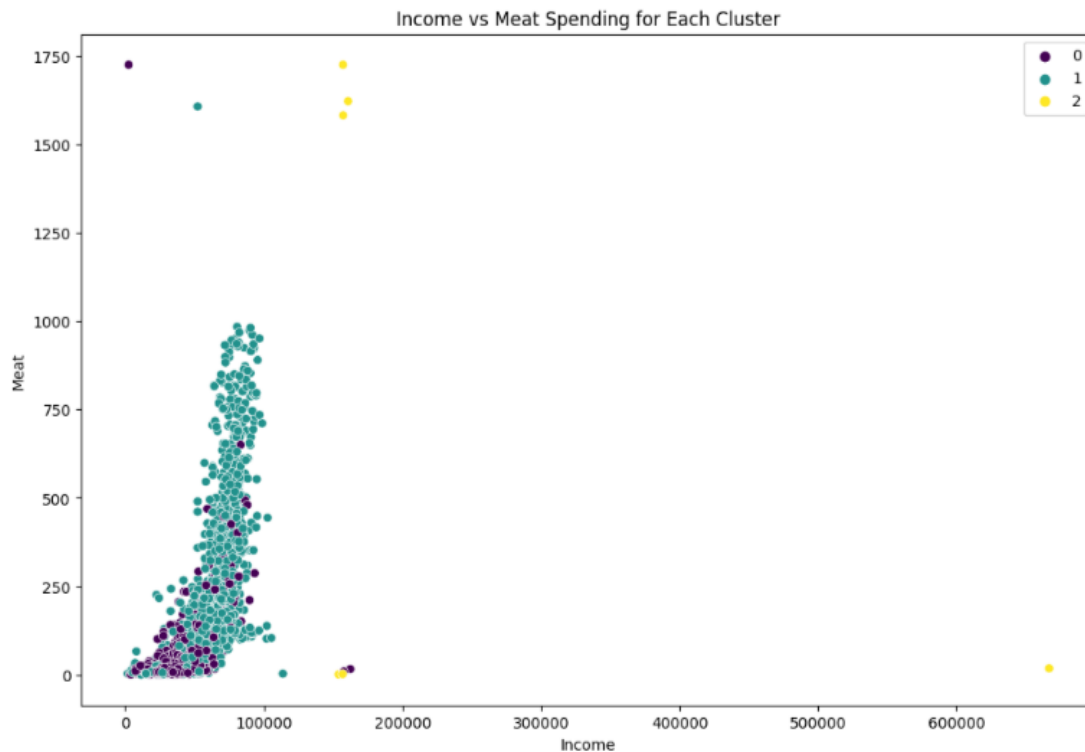


Use GaussianMixture to cluster the data into 3 clusters.

Column chart (countplot) to show the number of data points in each cluster.







#### Cluster 0:

Customers in this cluster have an average income, with an average number of children and monthly web visits.

They consume a fairly steady amount of fruit and meat compared to the other two clusters.

#### Cluster 1:

This cluster contains a large number of customers with a high average income, no or few children, and low to moderate monthly web visits.

Customers in this cluster consume a wider variety of fruits and meat than cluster 0, especially in meat consumption.

#### Cluster 2:

This is a cluster with a very small number of customers, but very high income.

For this cluster, there is a greater focus on high education, and they tend to consume fewer fruits but more meat.

This cluster also has high variability in monthly web visits.

### 5.3. Hierarchical Clustering

The process of customer segmentation at Londis Supermarket commences by selecting crucial features and eliminating data rows with missing values. Subsequently, the data undergoes standardization to ensure a consistent value range. Using the Ward linkage method, a

hierarchical clustering tree is constructed and visualized through a dendrogram. This aids in identifying customer clusters based on their shopping behavior. This approach allows Londis Supermarket to gain deeper insights into customer profiles and categorize them into groups with similar purchasing tendencies.

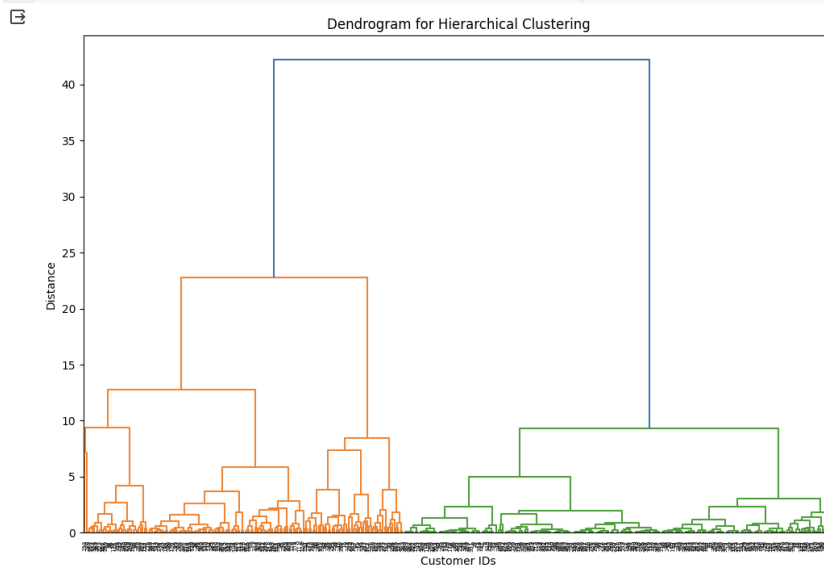
```
[89] import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy.cluster.hierarchy import linkage, dendrogram, fcluster
from sklearn.preprocessing import StandardScaler
import seaborn as sns

[90] # Xóa các hàng chứa giá trị NaN và chọn các đặc trưng quan trọng
selected_features = ['Income', 'MntFruits', 'MntMeatProducts']
data_cleaned = data[selected_features].dropna()

[91] # Chuẩn hóa dữ liệu
scaler = StandardScaler()
scaled_data = scaler.fit_transform(data_cleaned)

[59] # Sử dụng phương pháp liên kết ward để xây dựng cây phân cấp
linkage_matrix = linkage(scaled_data, method='ward')

# Vẽ biểu đồ dendrogram
plt.figure(figsize=(12, 8))
dendrogram(linkage_matrix, orientation='top', labels=data_cleaned.index, distance_sort='descending', show_leaf_counts=True)
plt.title('Dendrogram for Hierarchical Clustering')
plt.xlabel('Customer IDs')
plt.ylabel('Distance')
plt.show()
```



```
[66] # Xác định số lượng cụm dựa trên đồ thị dendrogram
num_clusters = 3 # Chọn số cụm dựa trên cắt cây trên dendrogram
clusters = fcluster(linkage_matrix, num_clusters, criterion='maxclust')
```

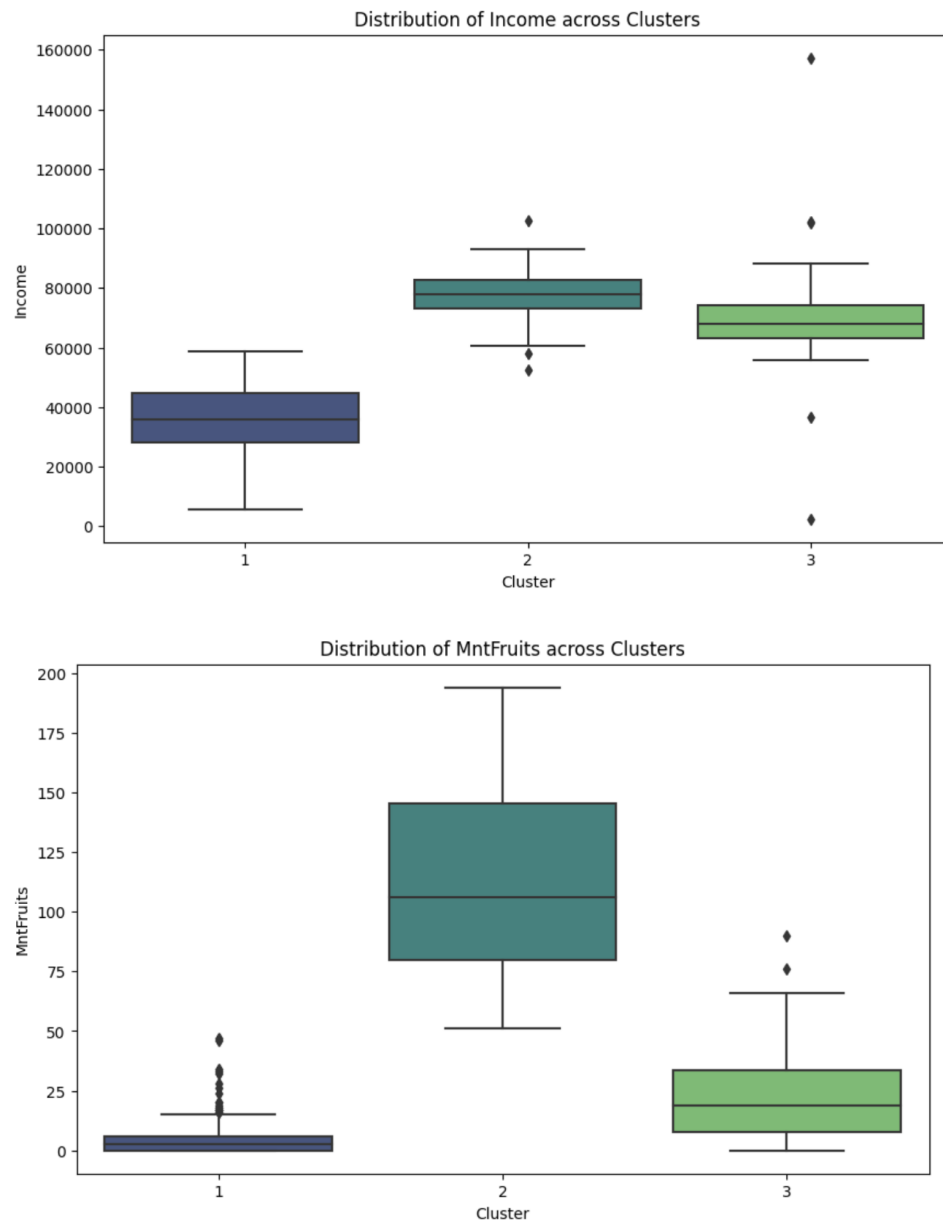
```
[67] # Thêm nhãn cụm vào DataFrame
data_cleaned['Cluster'] = clusters
```

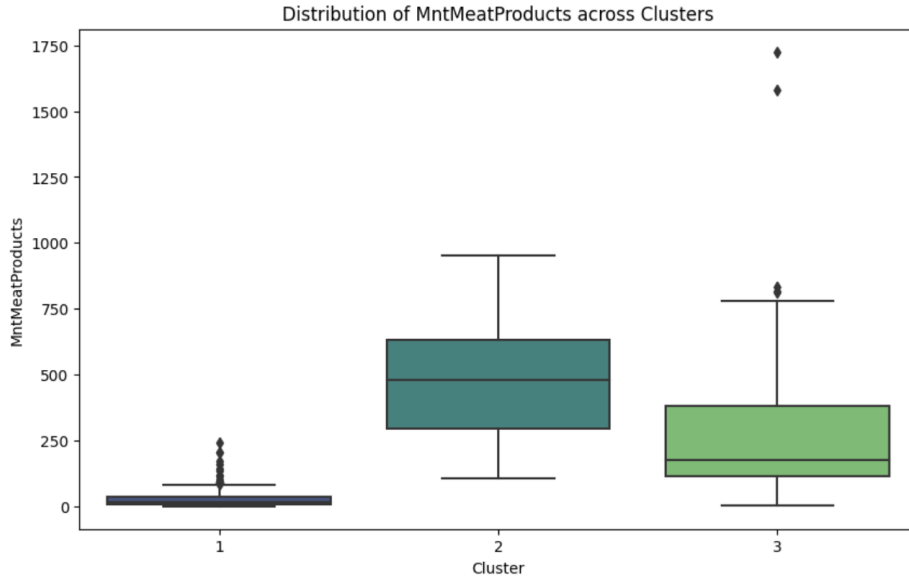
```
[68] # Hiển thị thông tin về kích thước của mỗi cụm
print(data_cleaned['Cluster'].value_counts())
```

```
1    222
3    114
2     51
Name: Cluster, dtype: int64
```

```
[69] # Biểu đồ phân phối của các đặc trưng theo cụm
for feature in selected_features:
    plt.figure(figsize=(10, 6))
    sns.boxplot(x='Cluster', y=feature, data=data_cleaned, palette='viridis')
    plt.xlabel('Cluster')
    plt.ylabel(feature)
    plt.title(f'Distribution of {feature} across Clusters')
    plt.show()
```

The process of customer classification at Londis Supermarket using the Hierarchical Clustering method. Initially, essential features are selected, and the data is standardized. Subsequently, the number of clusters is determined based on the dendrogram. Each customer is assigned a cluster label, and a distribution plot for important features within each cluster is generated. This procedure enhances Londis Supermarket's understanding of customer groups and optimizes business strategies.





## VI. Conclusion

The Customer Clustering Project at Londis Supermarket has opened new avenues, aiming not only to identify customer groups but also to gain profound insights into their shopping behavior and income. By amalgamating data analysis methods such as K-Means, Gaussian Mixture Model, and Hierarchical Clustering, we have acquired a multifaceted perspective on the market and customer segments.

Integrating algorithms like Gaussian Mixture Model (GMM) and K-Means aids Londis Supermarkets in more effective customer segmentation and analysis. GMM provides detailed information about customer groups, helping understand the characteristics and shopping behaviors of each group, thereby refining marketing strategies and tailored shopping experiences for each segment.

K-Means assists in grouping customers based on similar shopping patterns. Combined with GMM and Hierarchical Clustering, we gain a deeper understanding of customer behaviors and preferences, refining marketing strategies accordingly. This allows us to personalize products, advertisements, and shopping experiences for specific customer groups. As a result, Londis Supermarkets deliver a more personalized, enjoyable shopping experience, enhancing customer satisfaction and loyalty.

Ultimately, the profound comprehension of customers and the intricate modeling of shopping behavior not only enable Londis to develop astute marketing strategies but also foster stronger interactions and connections with customers. This not only fosters a competitive advantage but also serves as the cornerstone for sustainable development in an increasingly competitive retail market.

## VII. Teamwork division

No.	Student Name	Student ID	Tasks	Contribution
1	Nguyễn Thị Ngọc Lan	20070943	Explore Data Analysis	20%
2	Hoàng Thị Lan	20070942	Explore Data Analysis	20%
3	Nguyễn Thị Thùy Dung	20070687	Project proposal Overview the dataset Conclusion	20%
4	Trần Hương Quỳnh	20070975	Model	20%
5	Trần Thu Hoài	20070932	Model	20%