CUSTOMER SEGMENTATION ANALYSIS

Findings and Recommendations

Smriti Pradhananga
Student ID: 46188290

MACQUARIE
University

Table of Contents

Introduction 2
Exploratory Data Analysis 3
Numeric EDA 4
Categorical EDA 5
Customer Segmentation 6
1. K-means clustering 6
Choosing an optimal number of customer segments/clusters 6
2. Hierarchical clustering 9
Difference between two clusterings 10
Recommendations 11
Conclusion 12

Introduction

This report entitled “Customer Segmentation Analysis: Findings and Recommendations” has the goals to examine the data collected by loyalty cards and divide the customers into proper segments. The primary objective of this report is to attain detailed insights into the various types of customers that shop in the supermarket.

Customer segmentation is the process of determining comparable segments in terms of one or more certain qualities. This classification aims to maximise the value of each customer to your organisation by optimising marketing to each category and ensuring that individual clients gain the most pertinent and appropriate communications.

This paper analyses the 2000 data points provided by loyalty cards in the supermarket. These data points include customer information on demographic characteristics such as sex, marital status, age, education, income, occupation, and settlement size and ID.

Exploratory Data Analysis

The supermarket's dataset collected through loyalty cards contains 2000 customer information, including unique ID, sex, marital status, age, education, income, occupation and settlement size.

Variable	Data type	Details
ID	Integer	Unique identificator of a customer.
Sex	Categorical	0: male, 1: female
Marital status	Categorical	0: single, 1: non-single (divorced / separated / married / widowed)
Age	Numerical	Age of customer
Education	Categorical	0: other/ unknown, 1: high school, 2: university, 3: graduate school
Income	Numerical	Annual income
Occupation	Categorical	0: unemployed/ unskilled, 1: skilled/ official, 2: management / self-employed / highly qualified employee / officer
Settlement size	Categorical	0: small city, 1: mid-sized city, 2: big city

Table 1: Variable and its description

Table 1 shows how the values of these variables have been modified for analysis purposes.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 8 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   ID               2000 non-null   int64
 1   Sex              2000 non-null   int64
 2   Marital status   2000 non-null   int64
 3   Age              2000 non-null   int64
 4   Education        2000 non-null   int64
 5   Income           2000 non-null   int64
 6   Occupation       2000 non-null   int64
 7   Settlement size  2000 non-null   int64
dtypes: int64(8)
memory usage: 125.1 KB
```

Figure 1: Check for null values

From Figure 1, we can see that there are no missing values and the values integer type only.

Numeric EDA

	AGE	INCOME
MEDIAN	33	115548.5
MIN	18	35832
MAX	76	309364

Table 2: Descriptive statistics of numeric variables

The minimum age on the dataset is 18 while maximum is 76. The minimum income that a customer earns is $35832 while maximum is $309364. The medians are 33 and $115548.5 for age and income respectively.

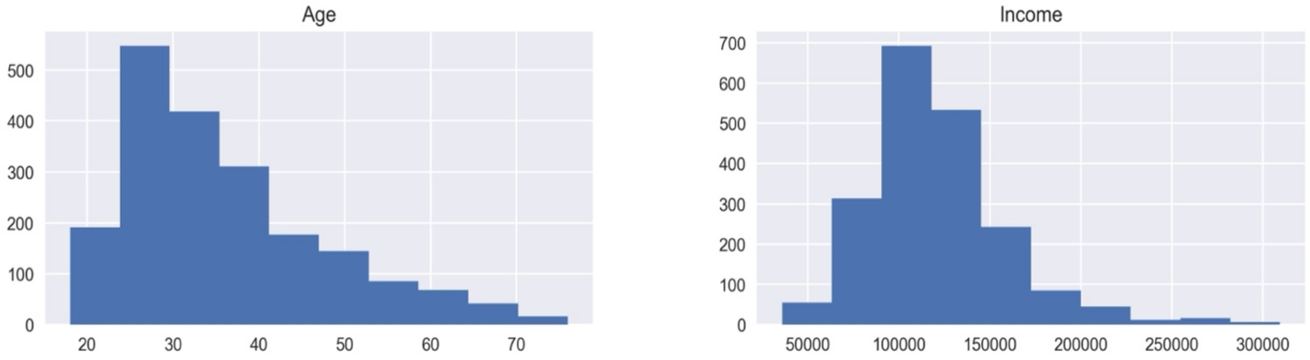

Figure 2: Histogram for Age and Income

Histograms above show a clearer distribution of age and income. The X-axis represents age and income, while Y-axis, number of customers. The highest count is for people who are between the age of 25-30 and for income is between 70,000 to 100,000. Both graphs are right skewed.

Categorical EDA

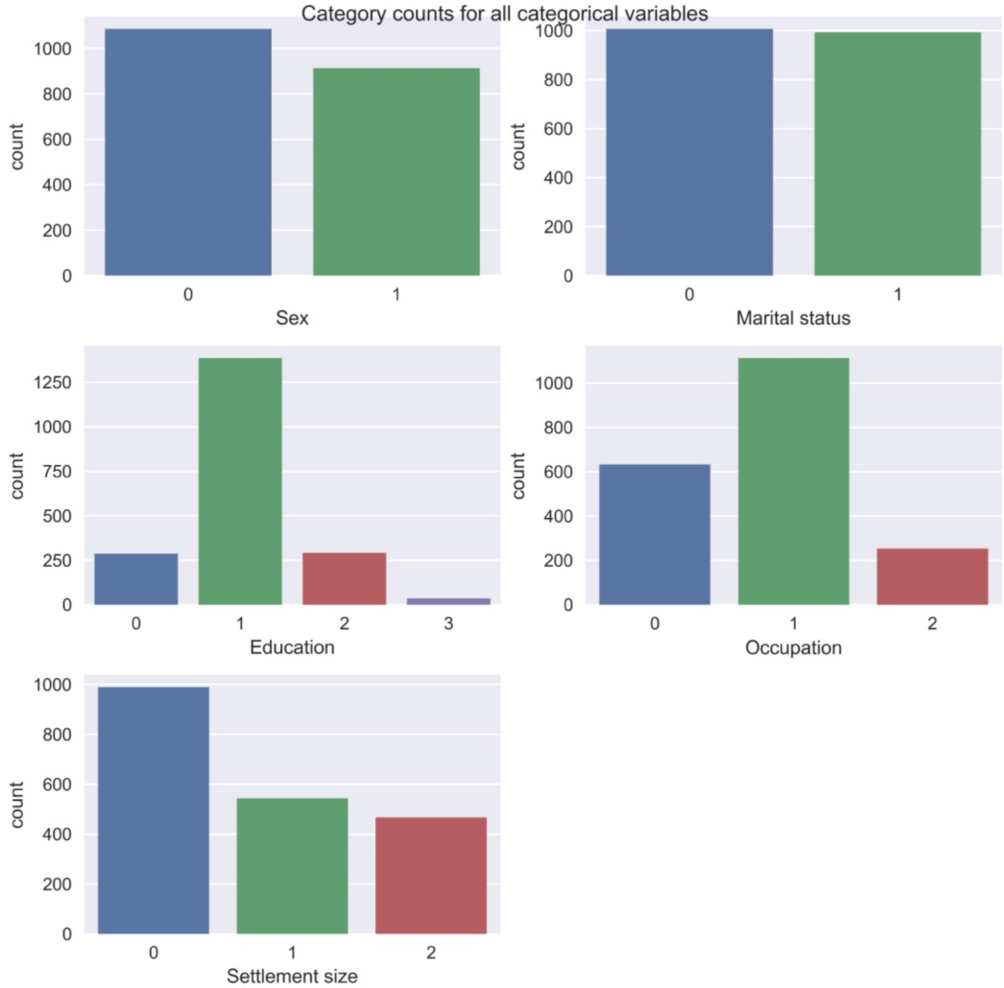

Figure 1: Category counts for categorical variables

From figure3 and table 1, the number of males is 1086 and females is 914. Similarly, individuals that are single are 1007 and non-single 993. Most level of education from the customers are in high-school(1386), 291 are university graduates, 287 are other and 36 are graduates. Furthermore, 1113 customers are skilled employees, 633 are unskilled/unemployed and 254 highly qualified employees. Lastly, 989 customers live in small, 544 in mid-sized and 467 in large cities.

Customer Segmentation

1. K-means clustering

K-means attempts to group similar types of items into clusters. It detects similarities between items and groups them into segments.

Choosing an optimal number of segments

Elbow method works on k-means clustering, which involves computing the sum of squared errors(SSE) for each value of k(from Figure 6, k is from 1-10) over a range of data points.

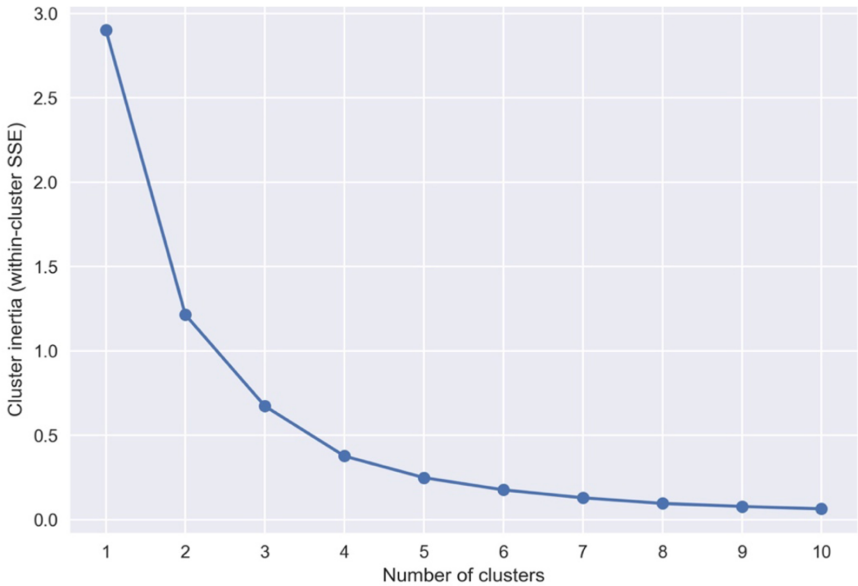

Figure 5: Elbow Method to determine number of clusters(k)

We can observe that as SSE value starts to decline as the number of clusters rises. After k=4, there is not much to gain as SSE is very low. Thus, we choose 4 as an optimal number of clusters.

label_km	Age(Median)	Income(Median)	Sex(Mode)	Marital status(Mode)	Education(Mode)	Occupation(Mode)	Settlement size(mode)	Count
Emergent	38	152267.0	Male	Single	High School	Skilled Employee/Official	Mid-sized City	455
Standard shoppers	32	114262.0	Female	Non-single	High School	Skilled Employee/Official	Small City	954
Well-Established	43	214732.0	Male	Single	University	Management/Self-Employed/Highly Qualified Employee/Officer	Big City	105
Working Class	29	81838.5	Female	Non-single	High School	Unemployed/Unskilled	Small City	486

Table 4: Clusters/Segments with K-means

Table 4 shows the clusters as per k-means.

The numerical and categorical variables in table 4 show the median and mode. Thus, it is important to know that as we are categorising them, not all customers should be single female for instance.

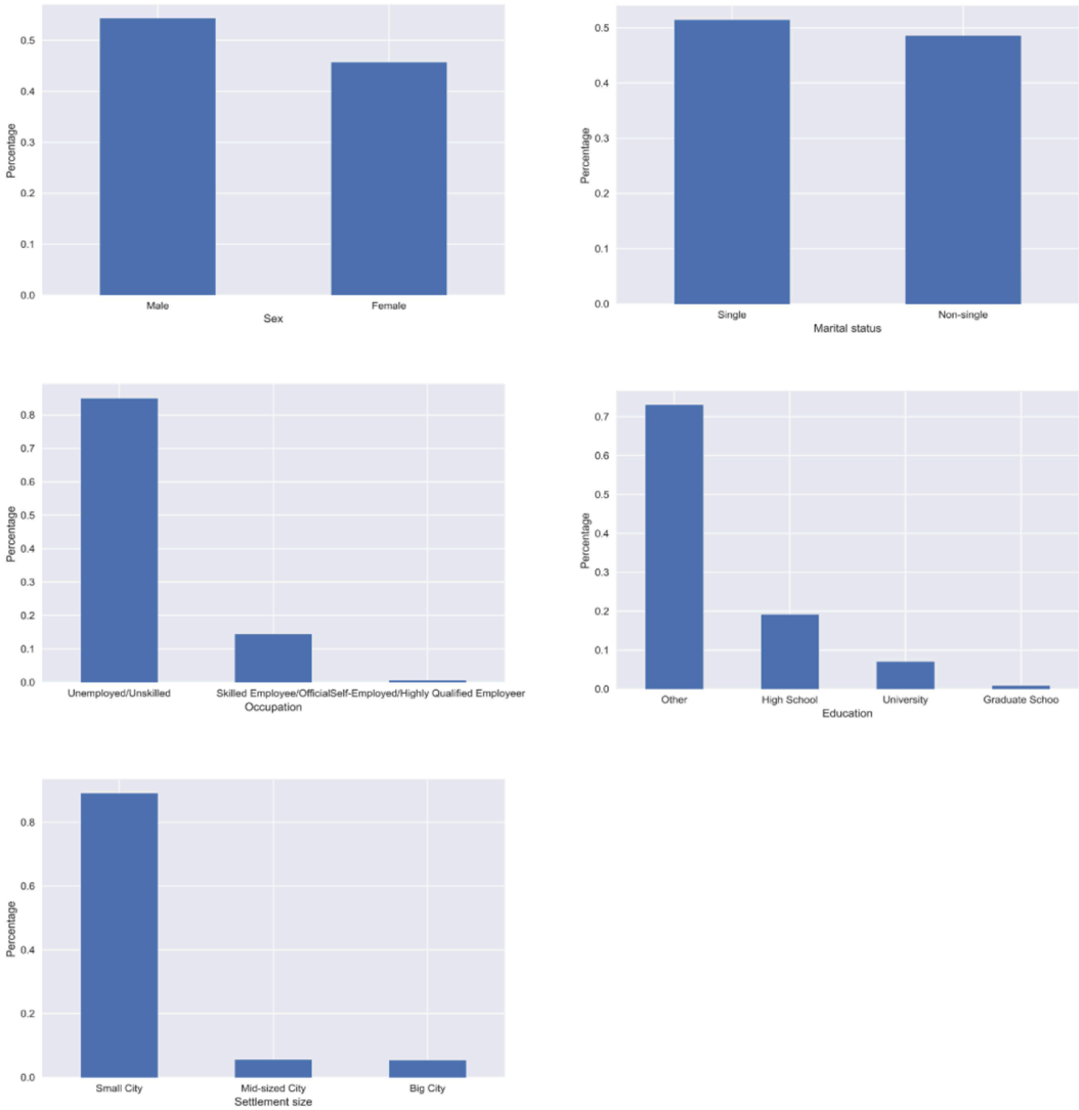

Figure 6: Frequencies of categorical variables in Working-class segment

This graph illustrates the percentage frequency of categorical variables in Working class segment. Sex and marital status have almost equal distribution in categories which implies the cluster variables could change.

2. Hierarchical clustering

Hierarchical clustering is the process of dividing data into segments based on some measure of similarity, determining how they're alike and different, and narrowing the data even further.

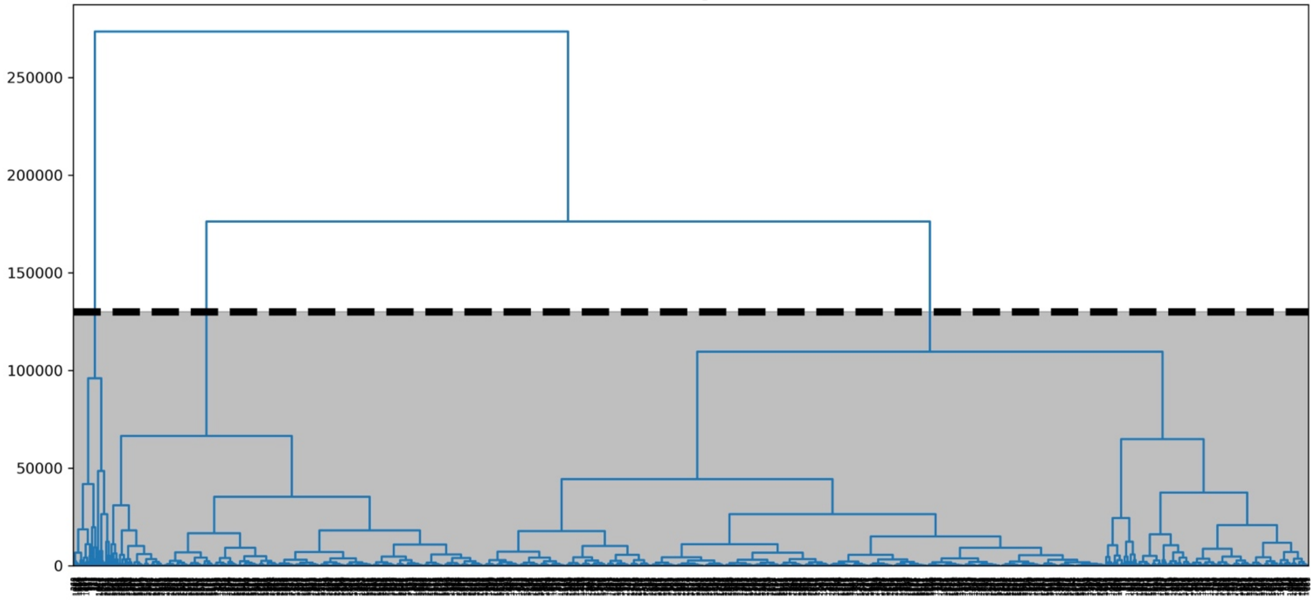

Figure 7: Dendrogram

From the above dendrogram we see that the dataset can be divided into 4 clusters in the shaded region.

label_hc	Age(Median)	Income(Median)	Sex(Mode)	Marital status(Mode)	Education(Mode)	Occupation(Mode)	Settlement size(mode)	Count
Emergent	39.0	164590.0	Male	Single	High School	Skilled Employee/Official	Mid-sized City	328
Standard shoppers	33.0	120160.0	Male	Single	High School	Skilled Employee/Official	Small City	1009
Well-Established	43.5	235538.5	Male	Single	University	Management/Self-Employed/Highly Qualified Employee/Officer	Big City	60
Working Class	28.0	86015.0	Female	Non-single	High School	Unemployed/Unskilled	Small City	603

Table 5: Segments with Hierarchical clustering

Both clustering produces almost identical clusters as per table 4 and table 5.

Difference between two clusterings

Clusters provided by these two algorithms won't be the same. However, most of the observations ought to belong in the same clusters.

```
((df1['label_km']) == (df1['label_hc'])).value_counts()

True     1666
False     334
dtype: int64
```

Figure 8: Segment differentiation in two clustering

We can see from the above figure that out of total, 1666(83.3%) customers are labelled as same by both clusterings and, 334(16.7%) are misplaced. There is less misplacement thus our segmentation is valid.

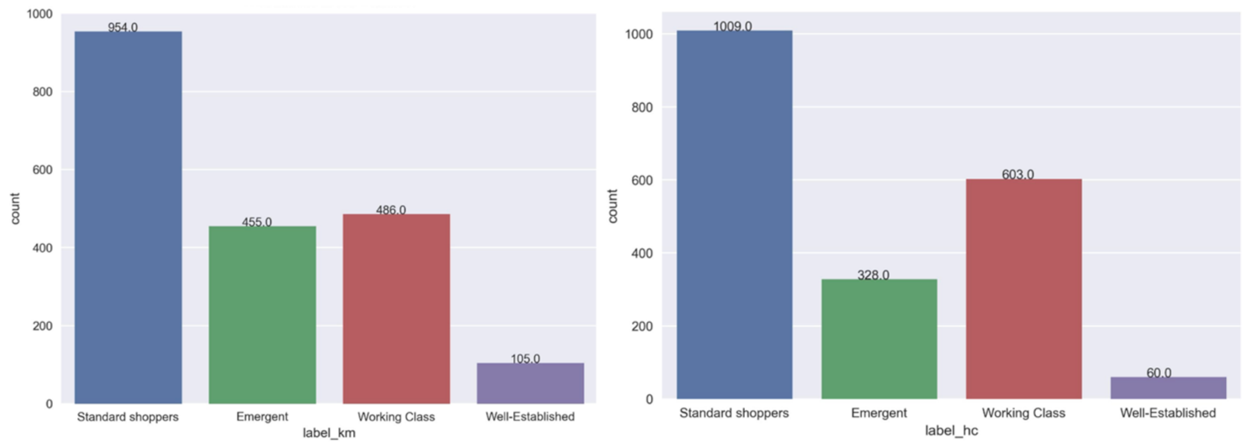

Figure 9: Distribution of clusters

Major differences are:

1. In k-means, standard shoppers are mostly non-single females while the latter is mostly single males. This was mentioned in Figure 6.
2. Distribution in clusters(Figure 9).

Recommendations

From table 4 and 5, the following is the segment and the recommendation:

SEGMENT	DESCRIPTION	MARKETING TECHNIQUE
Emergent	Single male high school graduates who are skilled employees with more than average income and live in mid-sized cities.	This segment consists of people who regard grocery shopping as a chore that should be completed as quickly and efficiently as possible. They're drawn to potentially time-saving technological innovations such as online shopping and self-checkout/scanning.
Standard shoppers	Non-single/Single high school male/female graduates who are skilled employees with less than average income and live in small city.	Most customers are from this segment, so company should focus on retaining them. The supermarket should push frequent shopper programs and offer exclusive deals.
Well-established	Single male university graduates who are highly skilled employees with highest income and live in big cities.	Email marketing is an effective way to influence this segment to purchase products. The company should focus on organic produce, and specialty cheeses.
Working class	Non-single high school female graduates who are unskilled/unemployed employees with lowest income and live in small city.	The company must persuade them to buy groceries. They may also purchase other items that high earning segments do not purchase from supermarkets. A person earning $10,000/month, for example, would not buy socks from Woolworths, whereas a person earning $1000/month might.

Conclusion

By now we know that clustering is a fantastic application for marketing. Upon the completion of customer segmentation analysis, we found four clusters: emergent(low career and experience, high education), working class(low career, education), well-established(high career, education) and standard shoppers(high career, low education, and experience).

I would like to recommend online shopping to Emergent, frequent shopper programs to Standard shoppers, Email-marketing to well-established and other-than-grocery items to Working class.