Customer Segmentation Modelling with K Mean Clustering and Decision Tree Classifier

Data Science Portfolio

Shalita N. P. Wahyudhie

Table of contents

01

Business Objective

Business Objective

Targeting the right audience can create many business leads such as increasing conversion, improving customer retention, and overall improving revenue. Customer segmentation is a unique and efficient strategy that can help company to find their target audience, resulting in an efficient marketing planning.

Today, when many company rely on digital marketing, and given its effectiveness to target certain demographics, it only make sense for a company to study who their customers are. To achieve this is by using machine learning algorithm.

02

Data Overview

Data Overview

The dataset used in this project is acquired from Kaggle and contains the customer data from a mall. This Dataset contains the information about the customers like Sex, Marital status, Age, Education, Income, Occupation etc. through which we can easily fit our model for better prediction.

The dataset consists of information about the 2,000 individuals from a given area who are customers of a physical 'FMCG' store. All data has been collected through the loyalty cards they use at checkout. The data has been preprocessed and there are no missing values. In addition, the volume of the dataset has been restricted and anonymised to protect the privacy of the customers.

Data Overview

```
<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
```

Data Overview

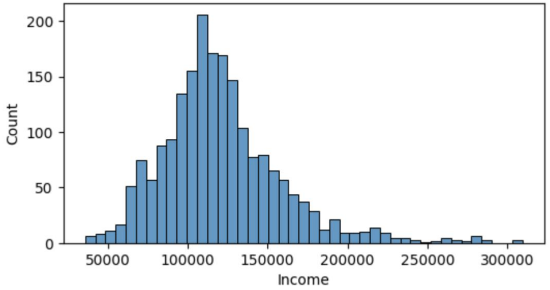
Variable	Data type	Range		Description
ID	numerical	Integer		Shows a unique identificator of a customer.
Sex	categorical	{0,1}		Biological sex (gender) of a customer. In this dataset there are only 2 different options.
			0	male
			1	female
Marital status	categorical	{0,1}		Marital status of a customer.
			0	single
			1	non-single (divorced / separated / married / widowed)
Age	numerical	Integer		The age of the customer in years, calculated as current year minus the year of birth of the customer at the time of creation of the dataset
			18	Min value (the lowest age observed in the dataset)
			76	Max value (the highest age observed in the dataset)
Education	categorical	{0,1,2,3}		Level of education of the customer
			0	other / unknown
			1	high school
			2	university
			3	graduate school
Income	numerical	Real		Self-reported annual income in US dollars of the customer.
			35832	Min value (the lowest income observed in the dataset)
			309364	Max value (the highest income observed in the dataset)
Occupation	categorical	{0,1,2}		Category of occupation of the customer.
			0	unemployed / unskilled
			1	skilled employee / official
			2	management / self-employed / highly qualified employee / officer
Settlement size	categorical	{0,1,2}		The size of the city that the customer lives in.
			0	small city
			1	mid-sized city
			2	big city

03

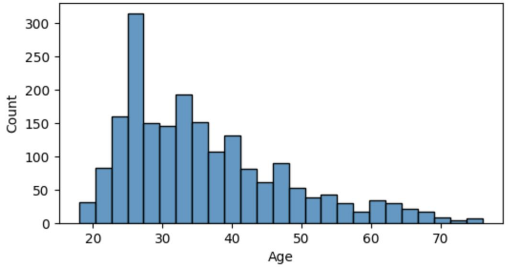
Exploratory Data Analysis

Numerical Features Distribution

Income

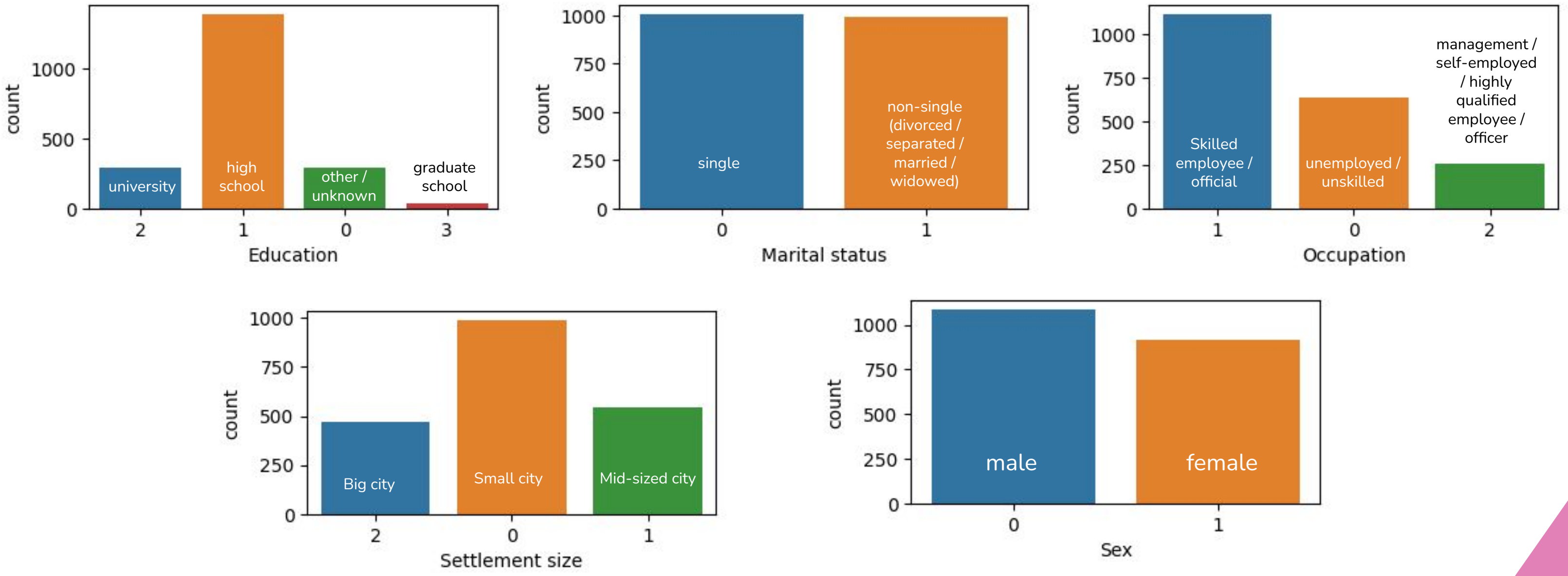


Age



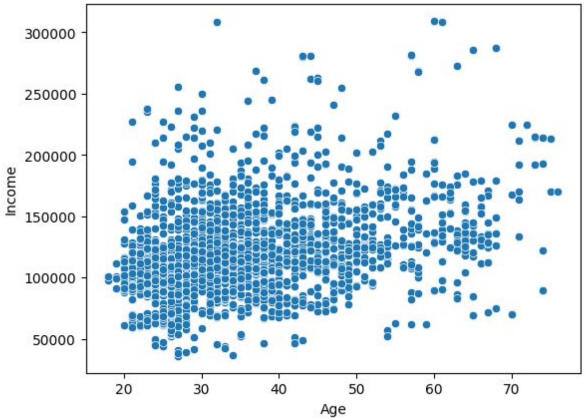
Age feature doesn’t seem to have normal distribution. Income feature’s distribution, however, seems rather normal although it have tail in the higher income side.

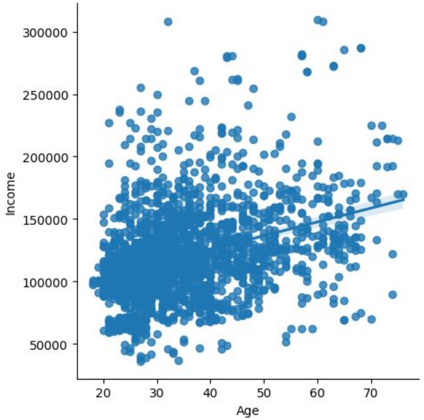
Categorical Features Distribution



The two categories, in Marital status feature have similar amount of data. Male customers count are slightly higher than female customer. Majority customer listed high school as their highest education level. Majority of customer come from small city and majority are skilled / official employee.

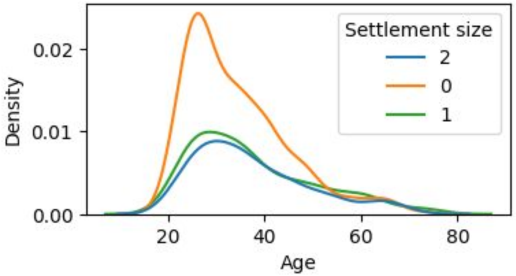
Bivariate Analysis: Age VS Income

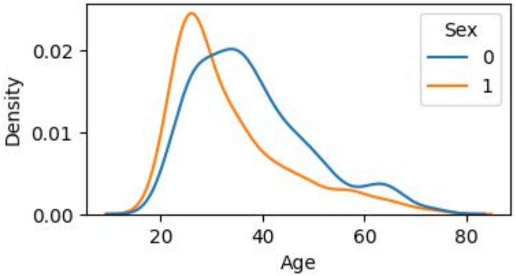


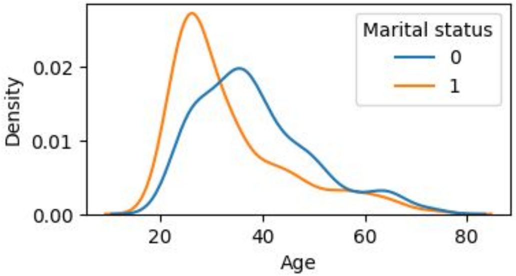


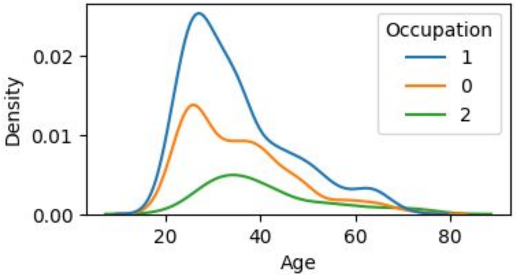
The two numerical features, age and income, have slight positive correlation
(r = 0.341, p-value = 1.6444e-55)

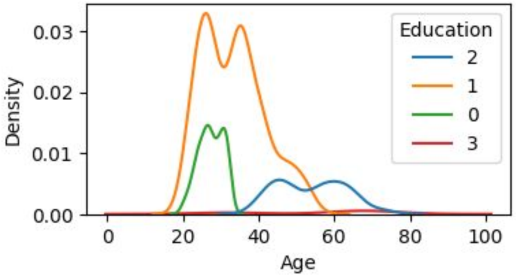
Bivariate Distributions with Age





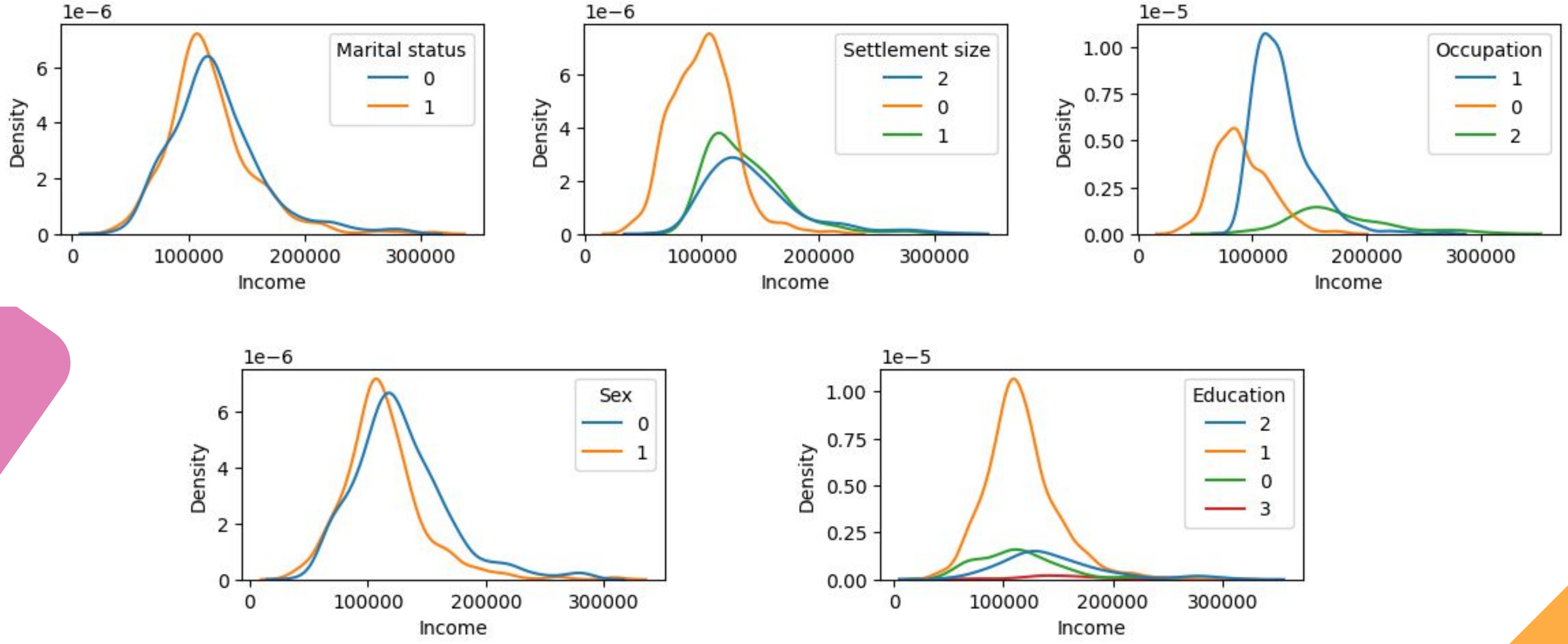






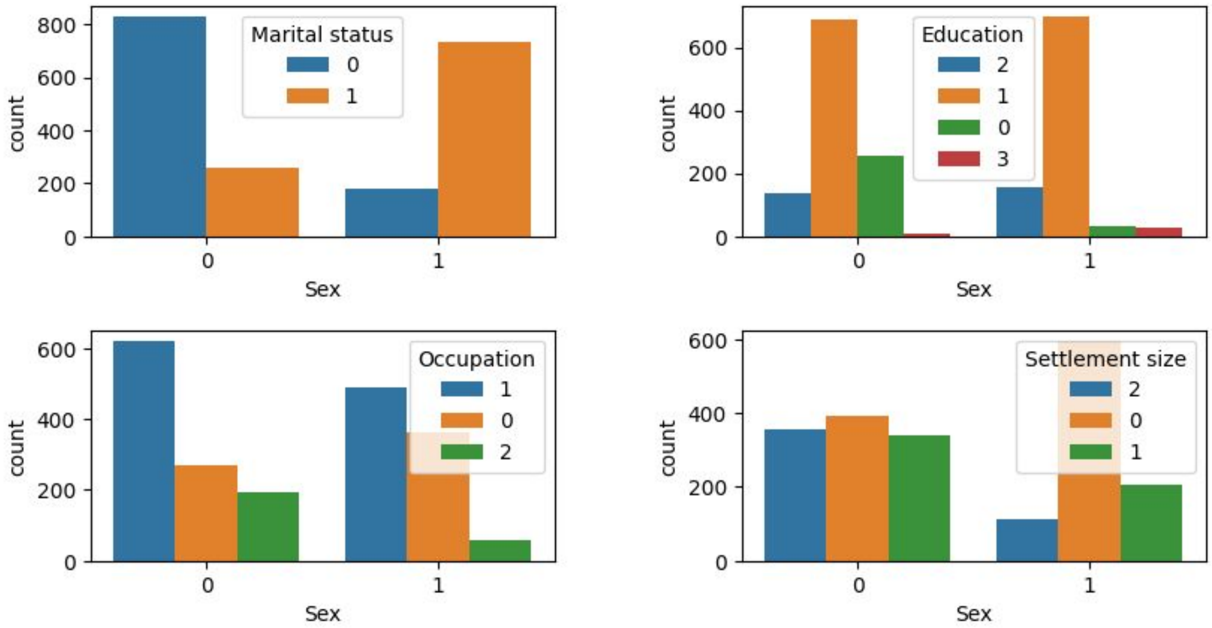
From graphs above, there seem to be significant variation in age for each education categories.

Bivariate Distributions with Income



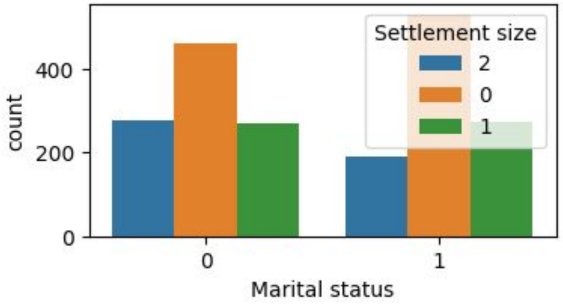
From graphs above, there seem to be significant variation in income for each occupation categories. The income also for customer who lived in small city compared to mid-sized and big city.

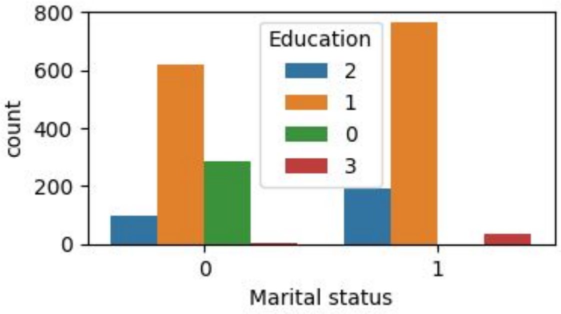
Categorical Variable: Sex

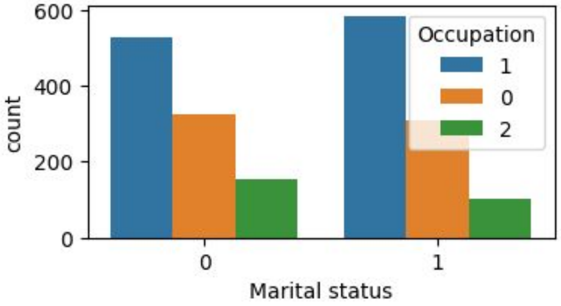


The distribution of male and female customer in their marital status is different. Male customer also has more unknown education status than female customers. Female customers majority lives in small city, compared to male customers whose settlement distribution is uniform.

Categorical Variable: Marital Status

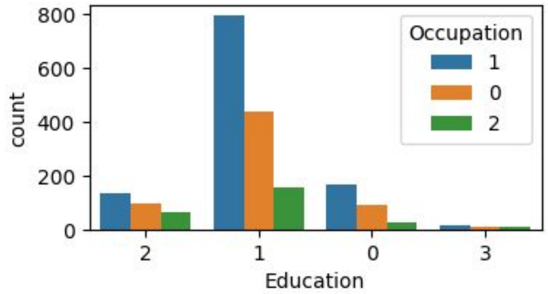


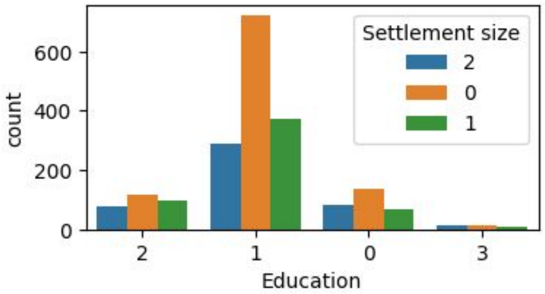


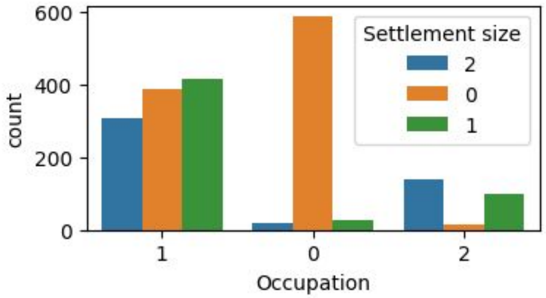


Single customers listed more unknown education status than the non-single customers.

Categorical Variable





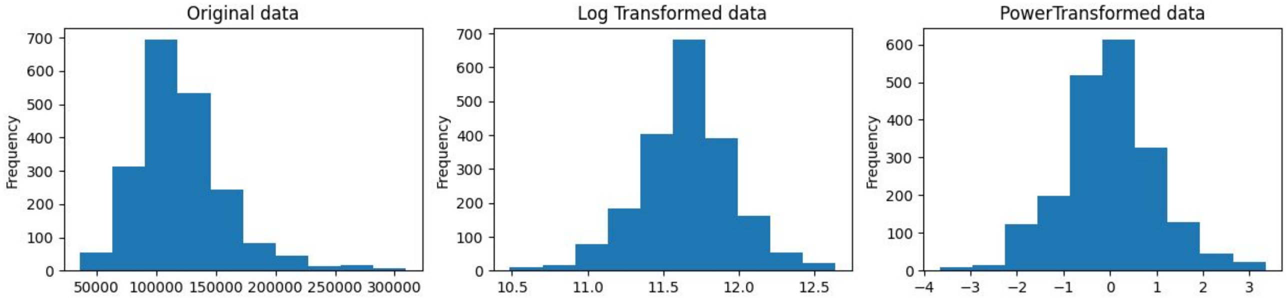


04

Feature Engineering

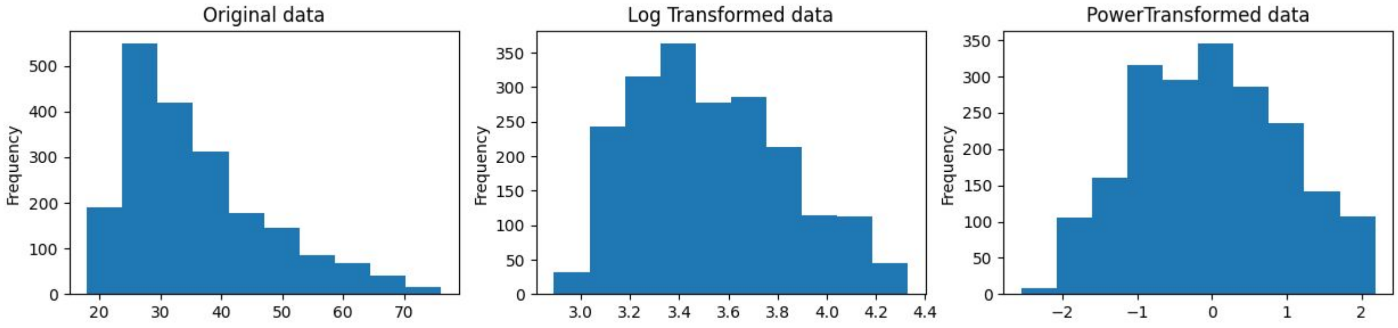
Feature Transforming: Income

To prepare Income feature for modelling, we test its normality and obtain the result that p-value for the null hypothesis of Income being normally distributed is 2.50e-98. This conclude that the Income feature is not normally distributed. Further, we do methods of feature transforming such as log transformation and power transform to normalize Income feature.



Feature Transforming: Age

The p-value for the null hypothesis of the Age feature being Normally distributed is 3.343e-56. This conclude that Age is not normally distributed. We then do some feature transforming function such as log transform and power transform.



Feature Transforming & Scaling

Income

	statistic	pvalue
Original data	449.473326	2.500964e-98
Log transform	32.357037	9.413664e-08
PowerTransformer	27.859212	8.921730e-07

Age

	statistic	pvalue
Original data	255.475892	3.342834e-56
Log transform	111.094201	7.519703e-25
PowerTransformer	161.196197	9.924088e-36

After log and power transformation we see that these cannot transform Age and Income feature to be normally distributed. Even so, we chose to still do the transformation, power transformation for Income feature, and log transformation for Age feature.

Scaling All Features

Feature scaling : MinMaxScaler

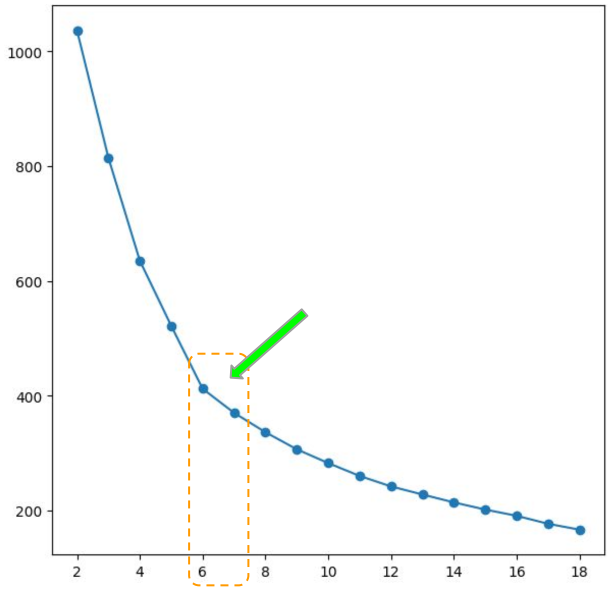
```
[ ]  from sklearn.preprocessing import MinMaxScaler

     scaler = MinMaxScaler()
     X = scaler.fit_transform(customer_transformed)
```

We then scaled all feature with MinMaxScaler so all features have the same weight to the model. MinMaxScaler scales and translates each feature individually such that it is in the given range on the training set, e.g. between zero and one.

04

Modelling

K Means for Clustering: Optimal Cluster Number



K means is one of the most popular clustering algorithms. The goal of K means is to group data points into distinct non-overlapping subgroups. One of the major application of K means clustering is segmentation of customers to get a better understanding of them which in turn could be used to increase the revenue of the company.

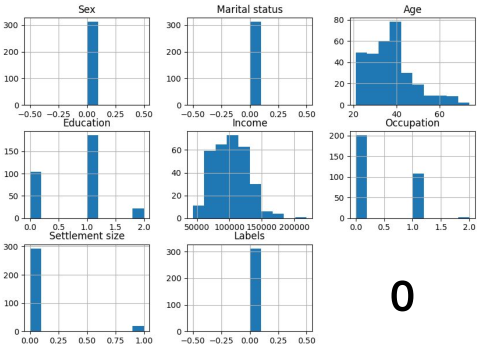
In this project, we use K means algorithm to analyze the mall customers segments. First we estimate the best number of cluster to test using the Elbow Method.

Silhouette Score

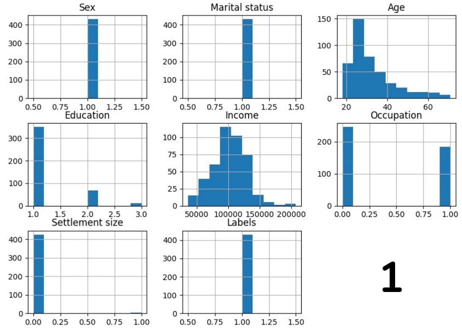
The best value is 1 and values near 0 indicate overlapping clusters. From the heatmap, we obtained a slight spike in silhouette score for 6-7 clusters, which is relevant with the Elbow Method result. Further, we decided to use 6 clusters to better understand the mall's customers.

With 14 clusters or above, our model would overfit and it also wouldn't be as efficient to have 14 or more clusters of customers in marketing context.

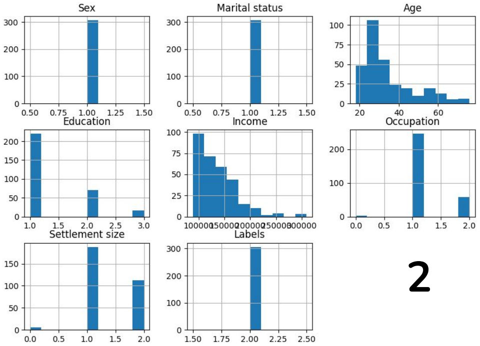
Customer Clusters



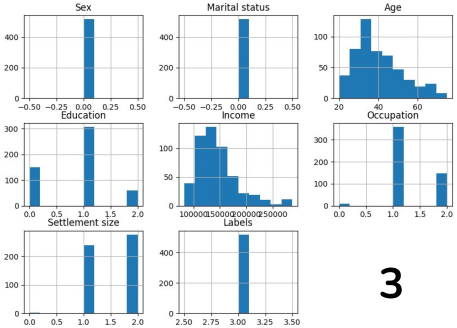
0



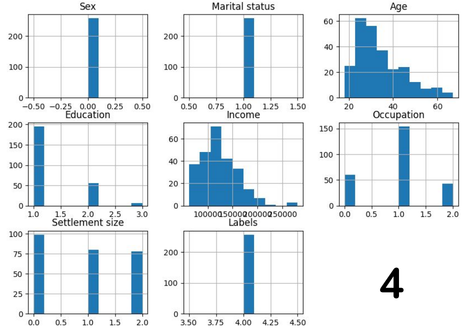
1



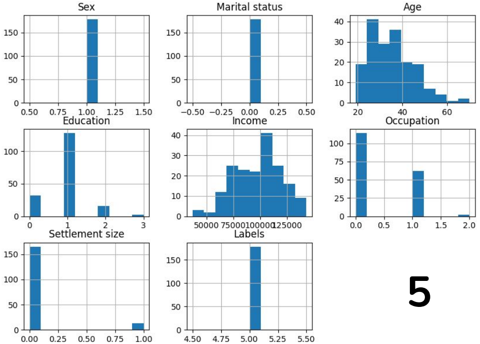
2



3

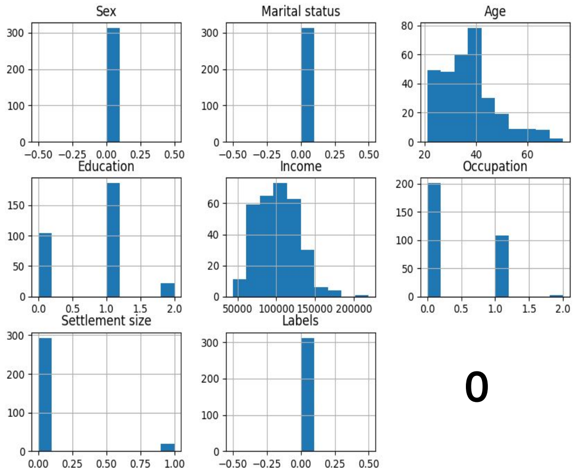


4



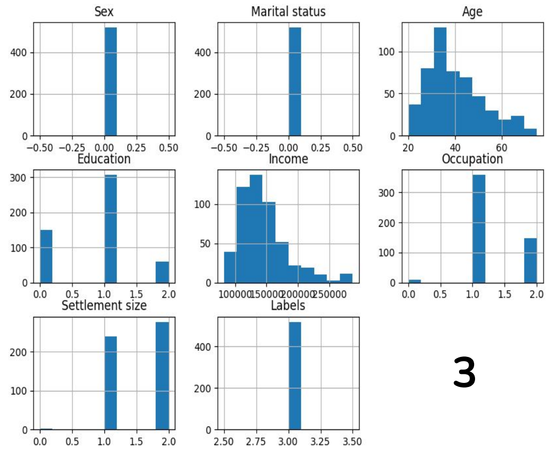
5

Customer Clusters: Single Male



0

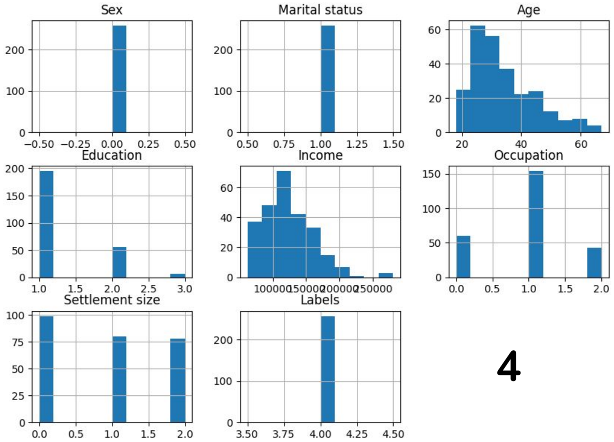
- lower income
- younger
- mostly from small city
- majority are unskilled or unemployed



3

- higher income
- older
- from mid-sized to big city
- more skilled and highly qualified worker

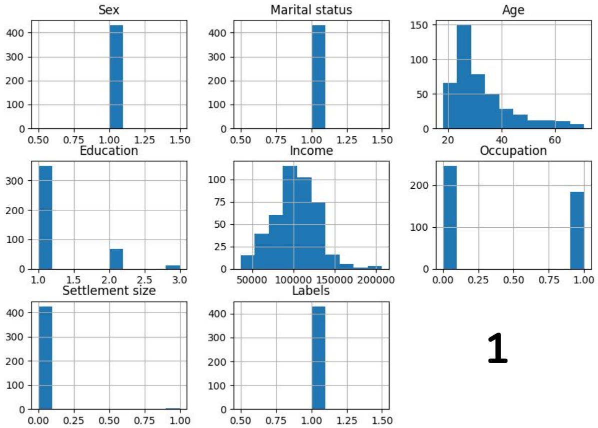
Customer Clusters: Non-single Male



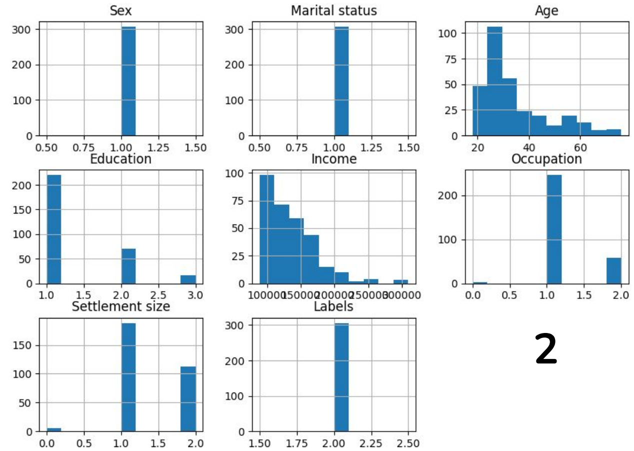
4

- higher income than Cluster 0
- mostly from small city
- majority are skilled employee

Customer Clusters: Non-single Female



1

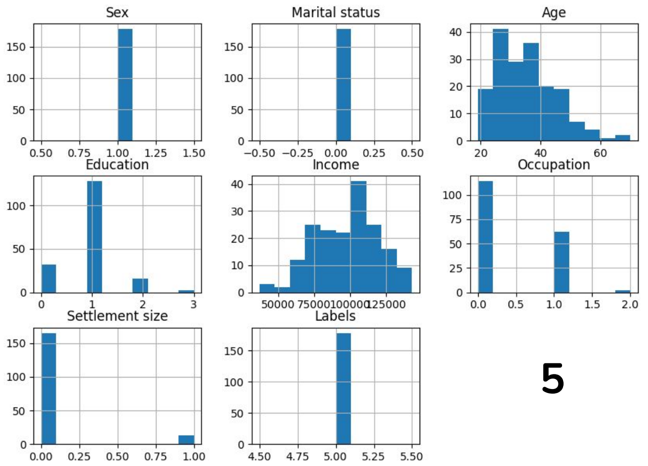


2

- lower income
- mostly from small city
- majority are unskilled or unemployed

- higher income
- from mid-sized to big city
- more skilled and highly qualified worker

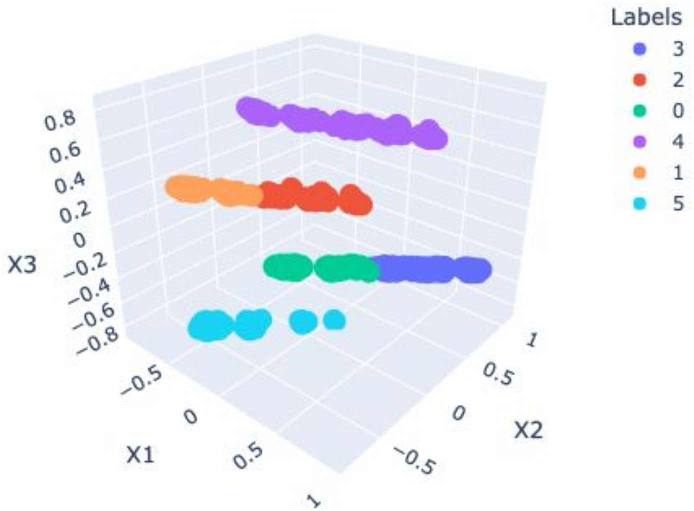
Customer Clusters: Single Female



- higher income than Cluster 1
- mostly from small city
- majority are skilled employee

Evaluate Clustering with PCA

From the PCA result, we conclude that the model predicted customer clusters nicely, with no bad overlapping between clusters. The elbow method and silhouette score also showed coherent result indicating that the model is suitable to segment the mall's customer.



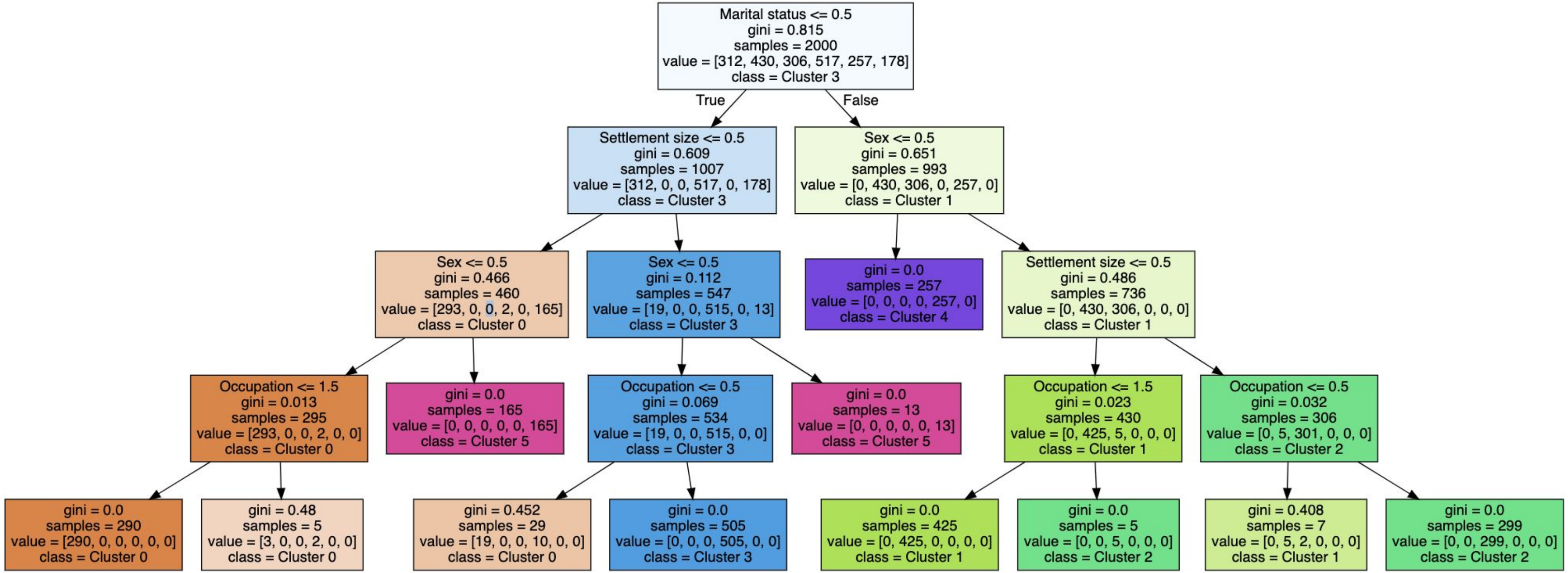
Decision tree as a method to interpret clusters

```
▾ DecisionTreeClassifier
DecisionTreeClassifier(max_depth=4, min_samples_leaf=5, random_state=42)
```

We further build a decision tree model as a way to be able to interpret clusters. The result showed that the model is extremely accurate at predicting the customer groups. Hence, we can expect the split point to be accurate as well. We can proceed with the interpretation of the model using this technique.

	precision	recall	f1-score	support
0	0.96	1.00	0.98	312
1	1.00	1.00	1.00	430
2	1.00	0.99	1.00	306
3	1.00	0.98	0.99	517
4	1.00	1.00	1.00	257
5	1.00	1.00	1.00	178
accuracy			0.99	2000
macro avg	0.99	1.00	0.99	2000
weighted avg	0.99	0.99	0.99	2000

Decision tree as a method to interpret clusters



Thanks!

Feel free to reach out,

salsa.wahyudhie@gmail.com
+62 811 311 2210
LinkedIn: Shalita Nafisah Putri Wahyudhie