Customer Segmentation

Problem Statement:

In the competitive retail landscape, understanding customer behavior and preferences is crucial for the success of shopping malls. The aim of this project is to perform comprehensive customer segmentation and behavior analysis using data collected from a shopping mall. By segmenting customers into distinct groups based on their characteristics and behaviors, the mall can tailor its marketing strategies, optimize resource allocation, and enhance the overall shopping experience for its customers.

Key Objective:

1. **Customer Segmentation:** Divide the mall's customer base into distinct segments/clusters based on their demographic attributes (age, gender, income), purchase history, visit frequency, and spending patterns.

2. **Behavior Analysis:** Explore the shopping behavior of each customer segment, identifying trends, preferences, and patterns. This analysis could include peak shopping times, popular stores, common purchase categories, and more.

3. **Customer Profiling:** Develop detailed profiles for each customer segment, providing insights into their preferences, needs, and motivations. This information can guide targeted marketing efforts.

About the Data:

The dataset consists of information about the purchasing behaviour of **2,000 individuals** from a given area when entering a physical 'FMCG' store. All data has been collected through **the loyalty cards they use at checkout.** In addition, the volume of the dataset has been **restricted and anonymised** to protect the **privacy** of the customers.

Variable description including type, range and full description:

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. 0 = male / 1 = female
Marital status	categorical	{0,1}	Marital status of a customer. 0 = single / 1 = non-single
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 (Min. age = 18 / Max. age = 78)
Education	categorical	{0,1,2,3}	Level of education of the customer. 0=no education / 1=high-school / 2=university / 3=graduate
Income	numerical	Real	Self-reported annual income in US dollars of the customer.
Occupation	categorical	{0,1,2}	Category of occupation of the customer. 0=unemployed / 1=employee/oficial / 2=management or self-employed
Settlement size	categorical	{0,1,2}	The size of the city that the customer lives in. 0=small / 1=mid-size / 2=big

Methodology:

1. Data Preprocessing: Clean and preprocess the collected data, handle missing values, and standardize variables.

2. Feature Selection/Engineering: Identify relevant features for segmentation and analysis. Create new features if needed.

3. Customer Segmentation: Apply clustering algorithms (e.g., K-means, hierarchical clustering) to group customers based on various attributes.

4. Behavior Analysis: Analyze spending patterns, visit frequency, peak shopping times, and preferred stores/product categories for each segment.

5. Customer Profiling: Create detailed profiles for each segment, describing their demographics, preferences, and behavior.

Expected Outcomes:

1. Clear understanding of the mall's customer segments and their unique characteristics.

2. Insights into shopping behavior patterns and preferences.

3. Personalized recommendations to enhance the shopping experience.

4. Improved customer retention through churn prediction and targeted strategies.

5. Enhanced marketing campaigns and resource allocation strategies.

By successfully completing this project, the shopping mall can not only optimize its operations and marketing efforts but also create a more personalized and enjoyable shopping experience for its customers.

Glimpse of the dataset:

	ID	Sex	Marital status	Age	Education	Income	Occupation	Settlement size
320	100000321	0	1	28	1	165190	2	0
425	100000426	0	1	21	1	115494	1	2
1354	100001355	1	1	19	1	99519	1	1
1902	100001903	0	0	33	0	117308	1	0
1710	100001711	0	0	21	0	59255	0	0

General info on the dataset:

We'll first run a quick statistical analysis of the data. We will check for null values, get the general sense of the data and the types, to see if they reflect what's shown in the table above. We use:

- The .info() method on the DataFrame to understand the data types in each column. We notice there are no missing values and that all columns have numerical values in them.
- The .describe() method on the DataFrame shows summary info on the regular statistics you may apply to categorical/numerical data.

```
<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
```

	count	unique	top	freq
Sex	2000	2	0	1086
Marital status	2000	2	0	1007
Education	2000	4	1	1386
Occupation	2000	3	1	1113
Settlement size	2000	3	0	989

	count	mean	std	min	25%	50%	75%	max
ID	2000.0	1.000010e+08	577.494589	100000001.0	1.000005e+08	100001000.5	1.000015e+08	100002000.0
Age	2000.0	3.590900e+01	11.719402	18.0	2.700000e+01	33.0	4.200000e+01	76.0
Income	2000.0	1.209544e+05	38108.824679	35832.0	9.766325e+04	115548.5	1.380722e+05	309364.0

From info, we seen there is **no missing** value present in the data.

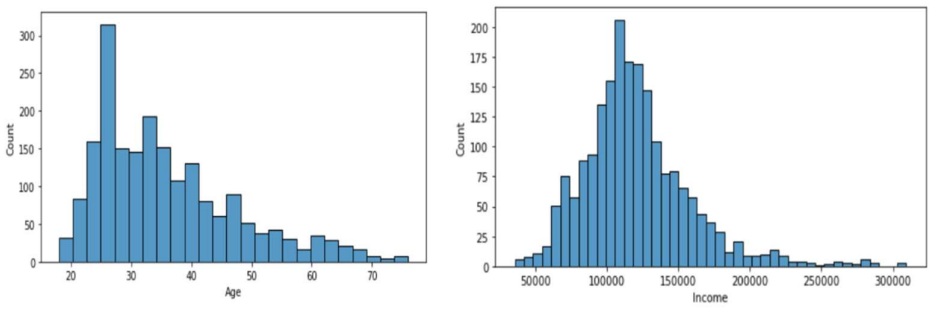
EDA

We start by taking a look at the distributions of the numerical variables Age and Income.

- The variable 'Age' has a heavy right skew, generated because of the lower limit at zero of the variable. If we are using K-Means clustering, there will be no need to normalize the feature, but we may have to do so for other models.

- The feature 'Income' has the same right skew problem as the 'Age' feature. We'll have to be wary of this depending on the model we select.

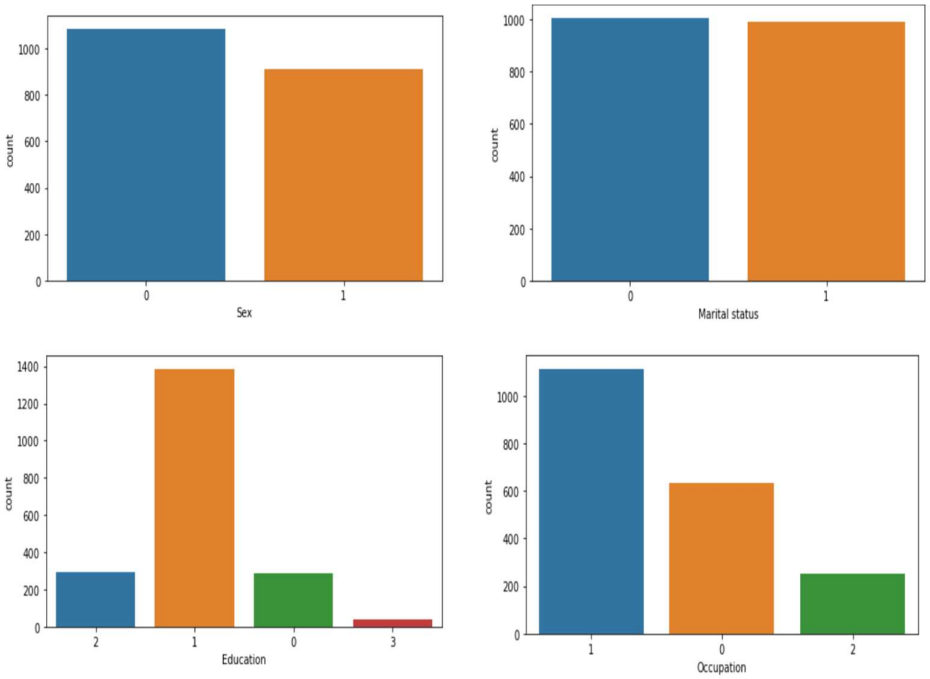
Numerical Variables Distribution

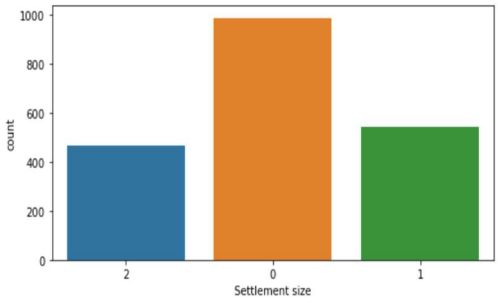


Categorical Variables Distribution

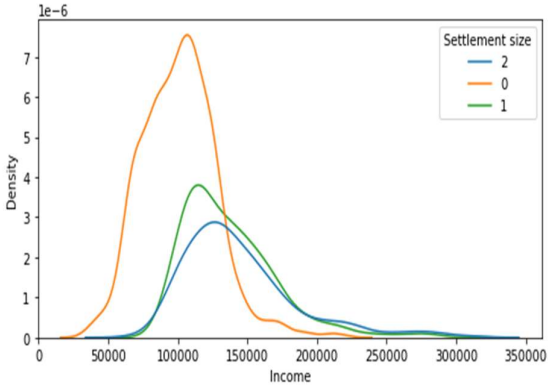
Next, we take a look at categorical variables. Running univariant countplots, we reach the following conclusions:

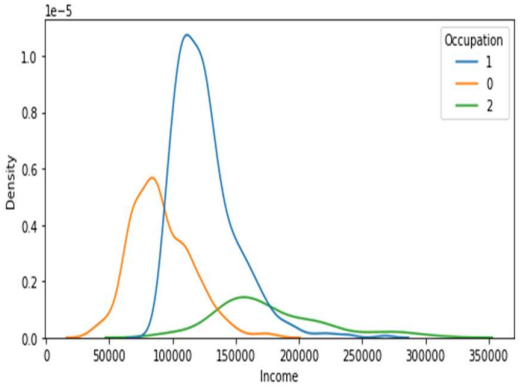
- Customer genders are quite balanced
- Marital status is also balanced
- Most customers have high school level education. There are only 1,8% graduate students. We could join both categories ('2' and '3') in a category called university & over
- Both the 'Occupation' and 'Settlement size' features seem to have a larger enough quantity of instances of each category.

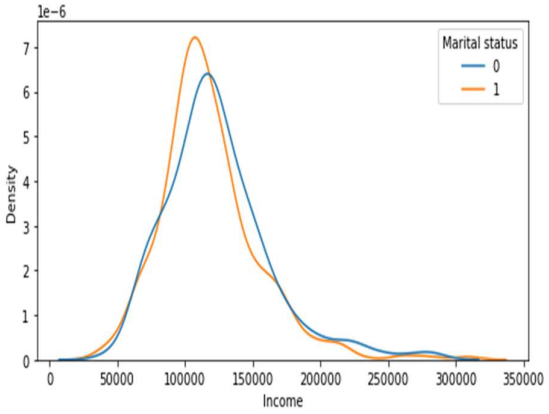




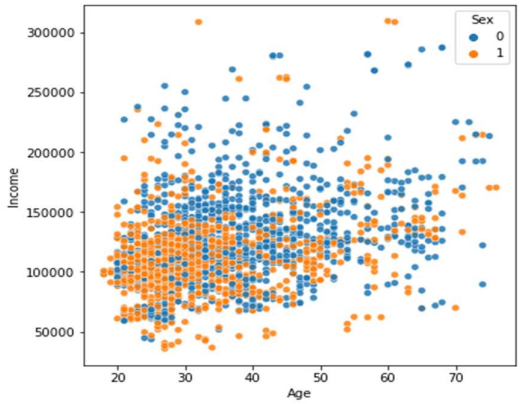
Categorical vs Numerical analysis:



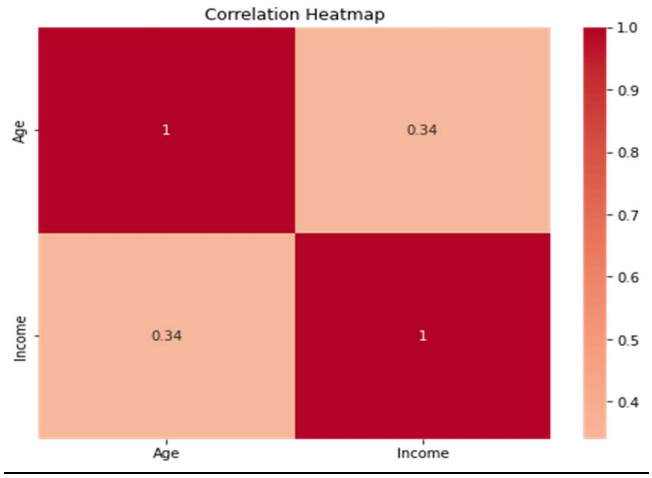




Multivariate analysis:



Correlation between Age and Income:



Comment:

First, we notice that some numerical features (Age and Income) have a right-skewed normal distribution. We will have to correct that for the model to perform correctly, since it assumes normality in our features. Most likely, a **log transform** will correct this skew.

Second, about the data itself, we noticed that there is a small correlation between age and income, as expected. People in smaller cities have lower income in the dataset. Income is higher as occupation feature is higher. Non-singles (married, divorced, widowed or separated) tend to have higher income than single people. When they are older, males tend to have higher income than females. Most unemployed people and married people in the dataset live in small cities. There are more unemployed women than men in the dataset.

Feature Scaling (MinMaxScaler);

Since there are categorical features scaled between 0 and 1 ('Sex' feature), we will use scikit learn's MinMaxScaler to scale out data between 0 and 1. This is done so that the scale is equivalent to that of the 'Sex' feature (all values will be between zero and one). This is important for the K-Means clustering model, as it uses distance as a measure of similarity. Therefore, if we don't scale the features, we won't get accurate clusters.

Clustering Models

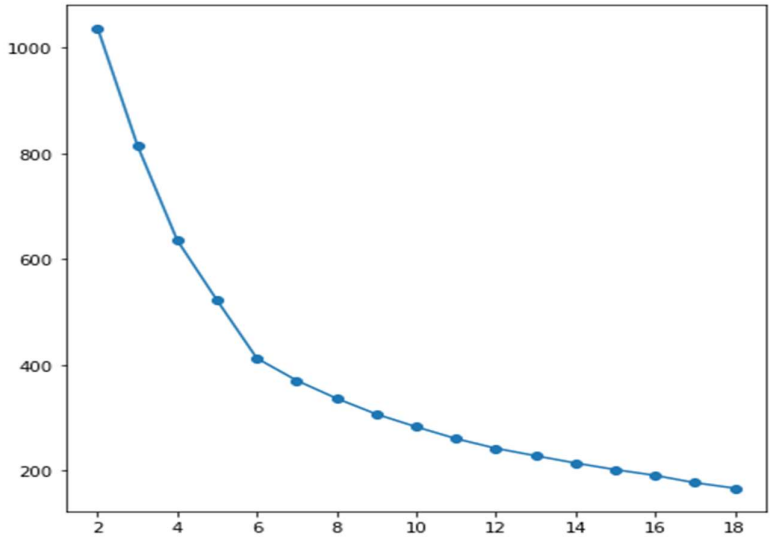
K-Means Clustering

K-Means is a clustering algorithm based on distance to determine the similarity of different points. It creates clusters by assigning points to the cluster nearest to them.

Selecting the correct number of clusters

There are several ways to select the correct number of clusters, but they are all based in the amount of business insight we get from each cluster. It isn't useful to find a lot of clusters if we cannot interpret them, or gain nothing by separating them into different categories. The number of clusters is an input to the model, but we cannot determine the number of clusters beforehand. So a good approximation is using the elbow method and selecting the number of groups that reduces a metric to a considerable amount; more wouldn't add much information and less would mean the metric could still be improved significantly.

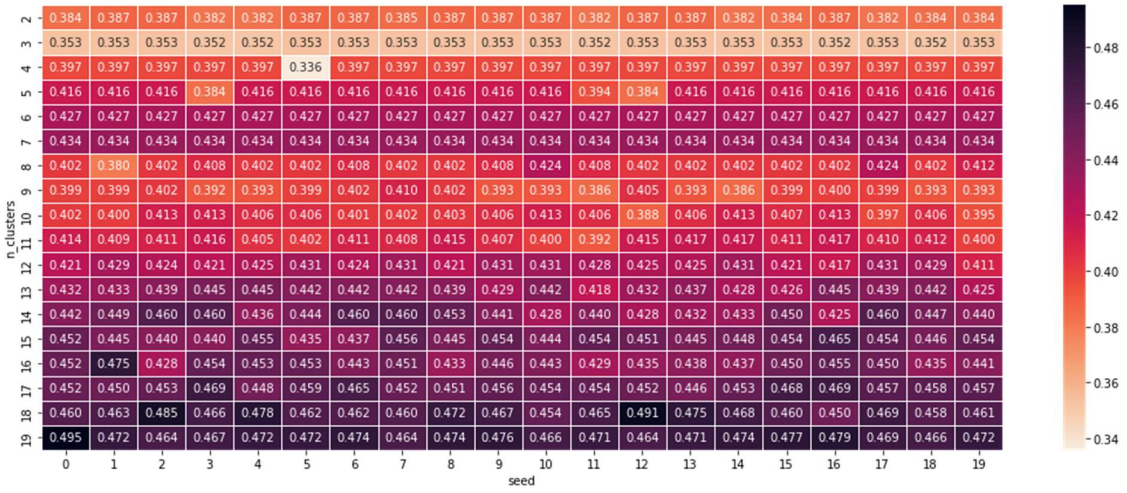
Elbow method:



When running the loop shown above, we notice that the elbow happens around 6-7 clusters, which would be a good approximation. The curve isn't very clear and you could also say that 12 is a good number as well, but you should understand that 12 clusters is generally too much, so we would rather lose some information about the groups our customers belong to, than gaining more accuracy in the clustering used.

Even though we have a somewhat convincing result above, we will use the Silhouette scores to see if we can gain more insight on how many clusters should we use. See the procedure below:

Silhouette scores:



From the graph shown above, we see that there is a spike in the Silhouette score for 7 clusters, which is coherent with the results we got from the elbow method. This, results show that choosing either 6 or 7 clusters should result in somewhat separated groups, which is what we are looking for. We see again that more clusters would reduce the metrics even further, but interpretation for such a small quantity of features with so many clusters would be really hard.

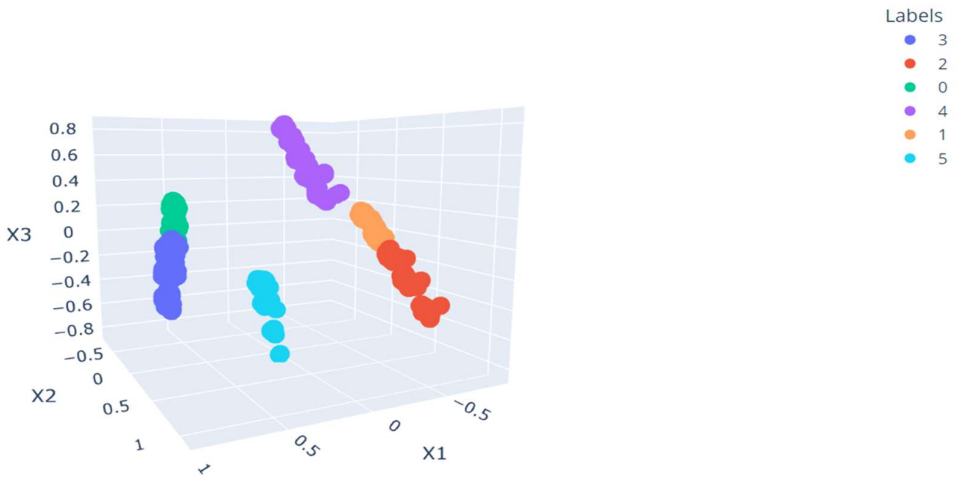
We decide to create 6 and 7 clusters and use our business understanding to determine which classification provides more insights about the customers. Another option would be to select 14 or more, which is unrealistic given the size of the dataset.

Principal Component analysis:

Earlier we have seen that our data has multicollinearity (Age and Income has correlation) and also 7 dimensional (without ID feature). So, we need to apply PCA to address those problems. For better visualization we need to reduce dimension to 3 or 2 or 1. I decide to take first three principle components only.

```
pca = PCA(n_components=3, random_state=42)
X_pca = pca.fit_transform(X)
```

K-means Clustering with 6 clusters:



Visualization

We have already clustered the data into 6 distinct groups and done PCA to get 3 features out of the 7 we originally had. It is always important to remember that using PCA inherently means a loss of information, so the projections of the data in the new features X1, X2 and X3 can have some overlapping points, but in reality, when using K-Means clustering the border points are clearly defined.

Summarized Clusters:

Cluster 0 --> Single + Small city + Male + Employed (employee/self-employed)

Cluster 1 --> Non-single + Female + small city + unemployed or employees (there are very little cases of medium/big city + unemployed so we discard them)

Cluster 2 --> Non-single + Female + medium/large city + management/self-employed (there are very little cases of small city + self-employed so we discard them)

Cluster 3 --> Single + Medium/big city + Male + Employed (employee or self-employed)

Cluster 4 --> Non-single + Males

Cluster 5 --> Single + Small city + Female (there are very little cases of medium/big city + Female so we discard them)