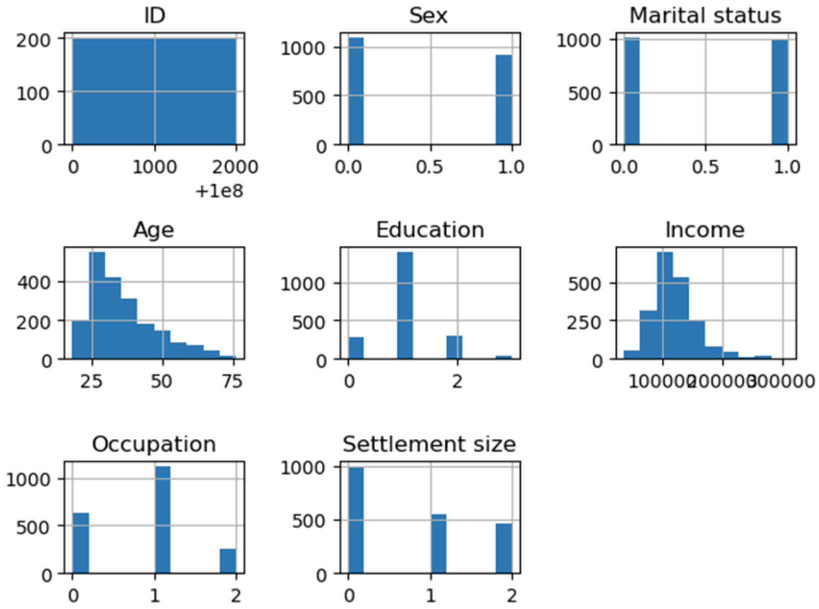
Customer Segmentation Analysis

Introduction

This report focuses on segmenting supermarket customers into distinct target groups using KMeans clustering and agglomerative clustering techniques. The primary objective is to enable businesses to optimize their marketing strategies by gaining more knowledge of different customer segments. The report will encompass several key steps, including exploratory data analysis (EDA), data transformation and scaling, the application and optimization of different clustering techniques, their visual representations, and a detailed breakdown of the resulting customer segments The dataset underpinning this analysis comprises 2000 customer entries, each rich with a variety of attributes such as age, sex, marital status, income, and others.

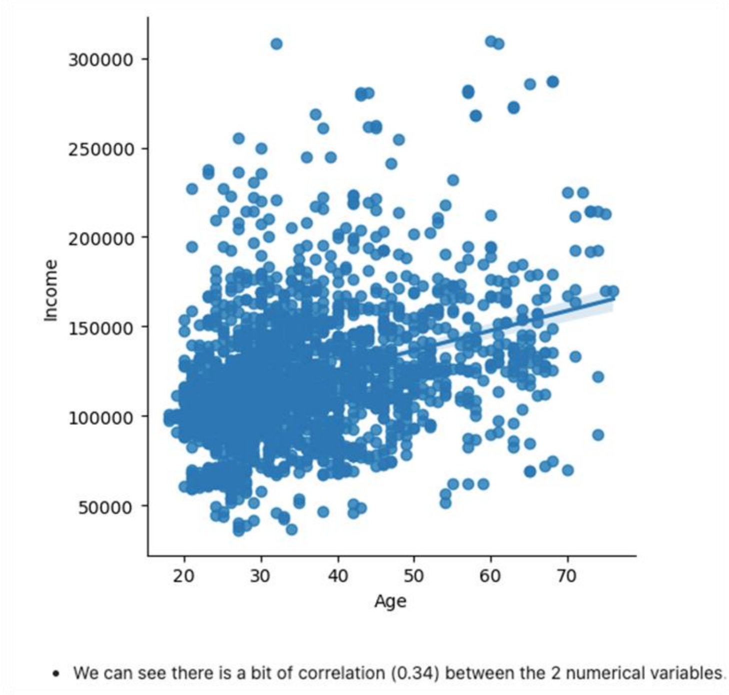
Exploratory Data Analysis



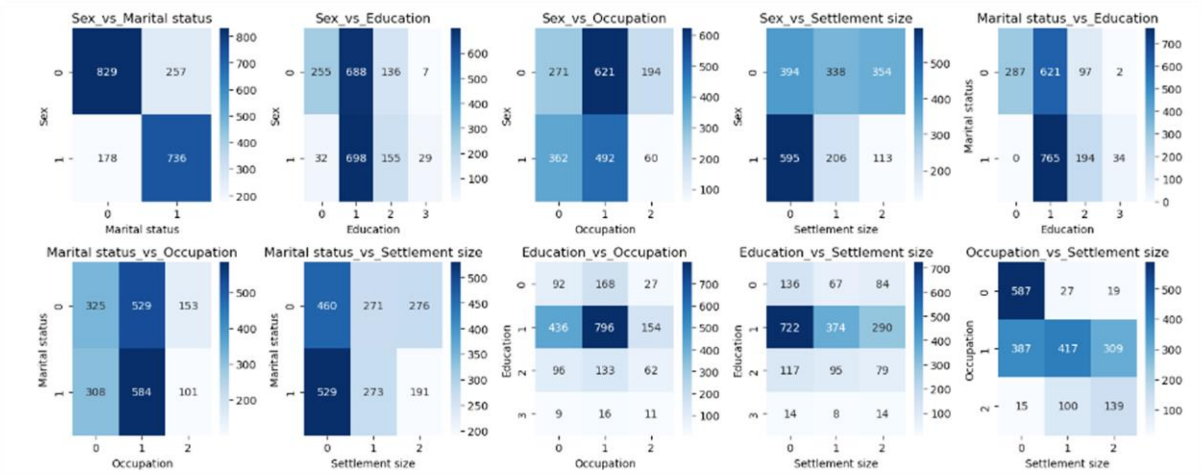
From the above histograms of the variables, we can make certain observations:

- 'ID' should be removed since it's an identifier attribute.
- Gender and marital status are evenly distributed.
- Age and income exhibit right-skewed distributions.
- Majority have high school education.
- Occupation and settlement size show diverse observations.

Relationship between Different Variables



- We can see there is a bit of correlation (0.34) between the 2 numerical variables.



we can see through the visualisations of the contingency tables that there are some interesting relationships between certain categories:

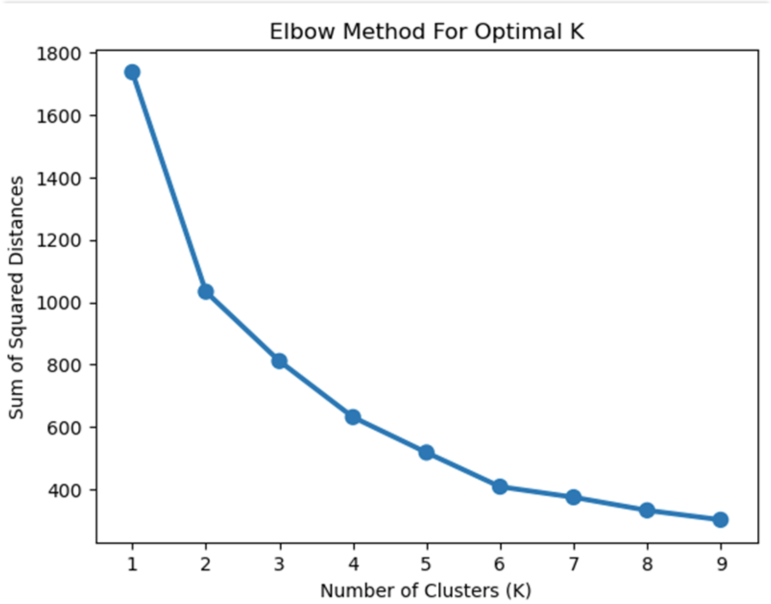
- More females tend live in smaller cities.
- There are more unskilled/unemployed women than skilled/employed and highly qualified men.
- There are more unmarried men than women.
- Married people tend to live in smaller cities.
- There are a lot of High School graduates who work are unemployed/unskilled (0), but there's almost twice as most who are employed and skilled.
- Most customers who have high- school education live in small cities.

- Most unemployed/unskilled customers live in smaller cities, while skilled/ employed customers are evenly distributed across all settlement sizes.

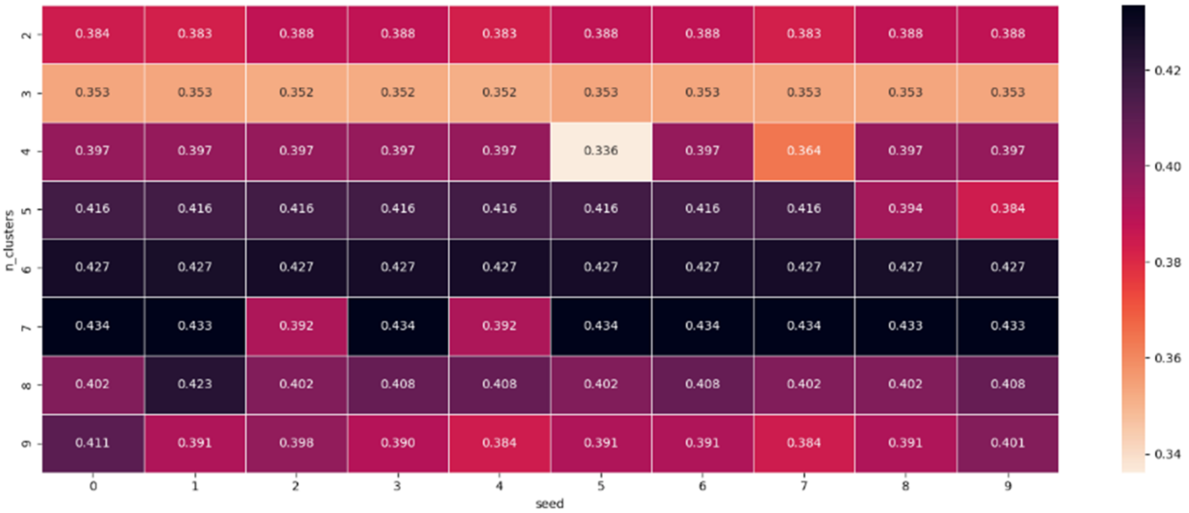
Customer Segmentation

Optimal K

Using elbow method and Silhouette scores we find the optimal number of clusters.



- We can observe the clusters forming an elbow at around 5-6 clusters. The curve is not highly distinct, and we could probably use more than 5 clusters as well, but more or less variance explained by each cluster flattens down after 4-5 clusters.
- We'll further use Silhouette scores to validate our assumption.



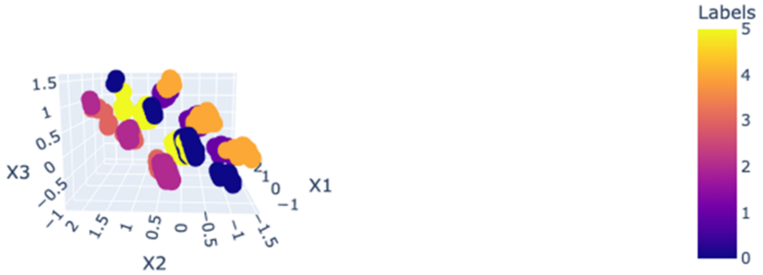
n_clusters \ seed	0	1	2	3	4	5	6	7	8	9
2	0.384	0.383	0.388	0.388	0.383	0.388	0.388	0.383	0.388	0.388
3	0.353	0.353	0.352	0.352	0.352	0.353	0.353	0.353	0.353	0.353
4	0.397	0.397	0.397	0.397	0.397	0.336	0.397	0.364	0.397	0.397
5	0.416	0.416	0.416	0.416	0.416	0.416	0.416	0.416	0.394	0.384
6	0.427	0.427	0.427	0.427	0.427	0.427	0.427	0.427	0.427	0.427
7	0.434	0.433	0.392	0.434	0.392	0.434	0.434	0.434	0.433	0.433
8	0.402	0.423	0.402	0.408	0.408	0.402	0.408	0.402	0.402	0.408
9	0.411	0.391	0.398	0.390	0.384	0.391	0.391	0.384	0.391	0.401

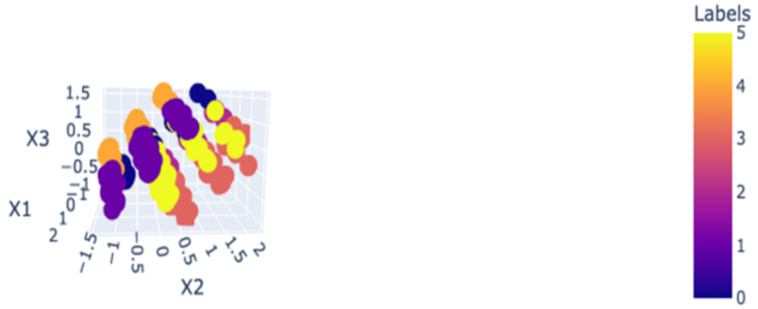
From the heatmap we can see that Silhouette score peaks at 6 or 7 clusters, which is in line with the assumption we made through the elbow method. We'll get cohesive clusters where each data point is more like its cluster compared to others if we use 6 or 7 clusters.

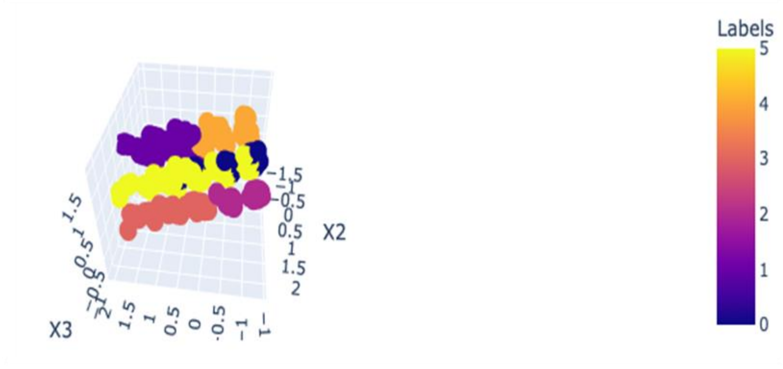
Visualising Kmeans with the help of PCA

Since its difficult to visualise Kmeans with a lot variables, we use PCA to reduce the data to 3 Principal Components , which helps us visualise the clusters in 3 dimensions.

Note: The optimal number of Principal Components that explain the variance might be different, this is purely meant for Visual imaging.





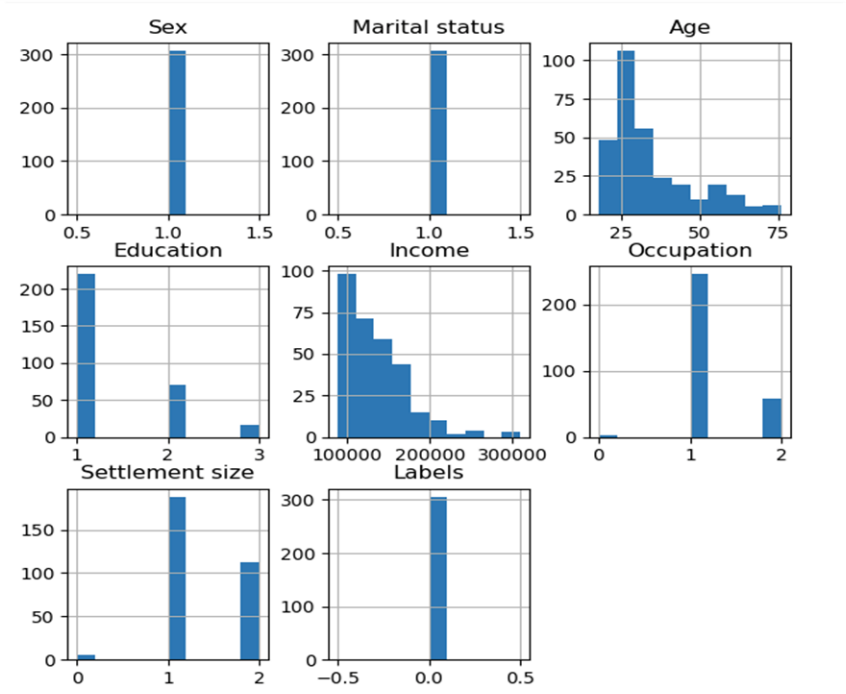


- As we can see the clusters are distinct and separated well.

Key Segments

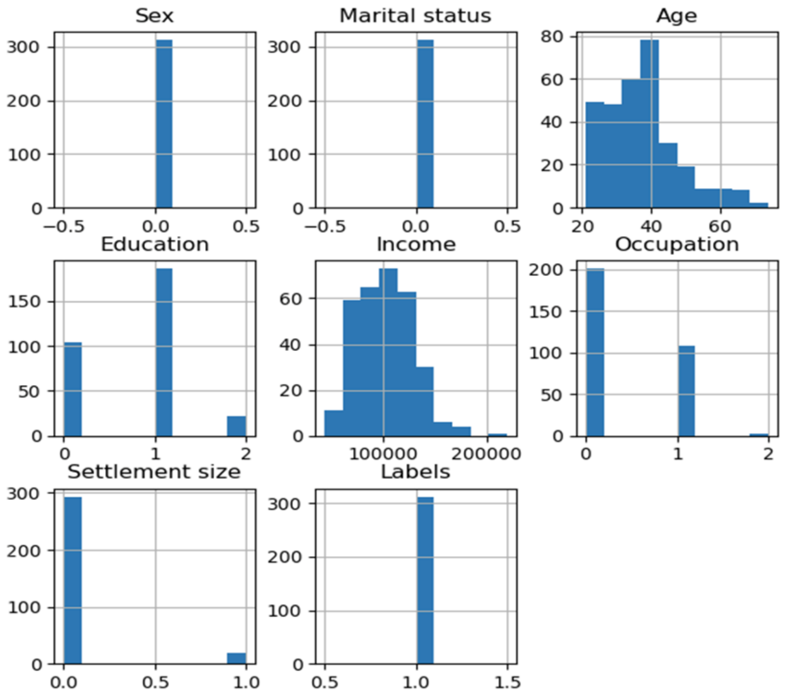
Cluster 0:

This cluster is represented by married women. They're mostly High School graduates, lie in the low-income strata and are mostly skilled or highly skilled. They mostly live in medium or big sized cities.



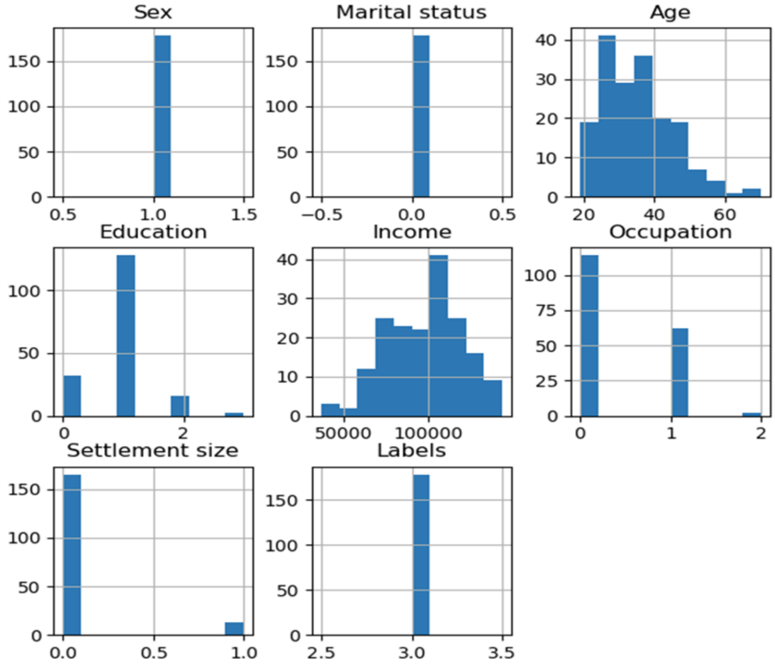
Cluster 1 :

This cluster is represented by single males, most are high school graduates, earn a bit more than the previous cluster and live in medium to large sized cities.



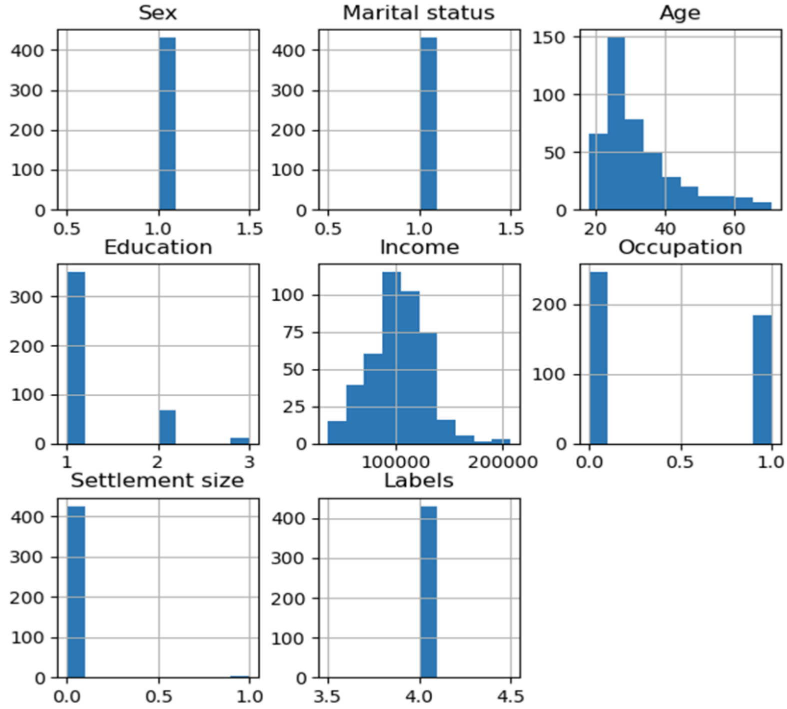
Cluster 3:

This cluster represents by single women , who have mostly graduated from High- School, medium income and mostly unemployed or work as officials, and the majority of them live in smaller cities, and are under the age of 40.

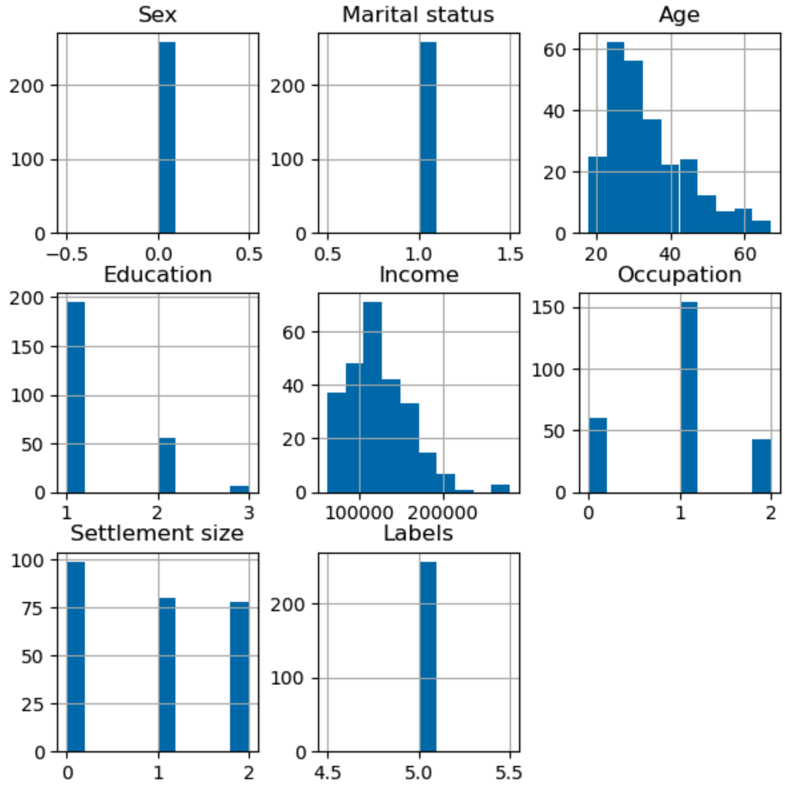


Cluster 4:

This cluster is represented by married females, with high school education, medium income bracket, who live in smaller cities, and are either unemployed or self-employed.

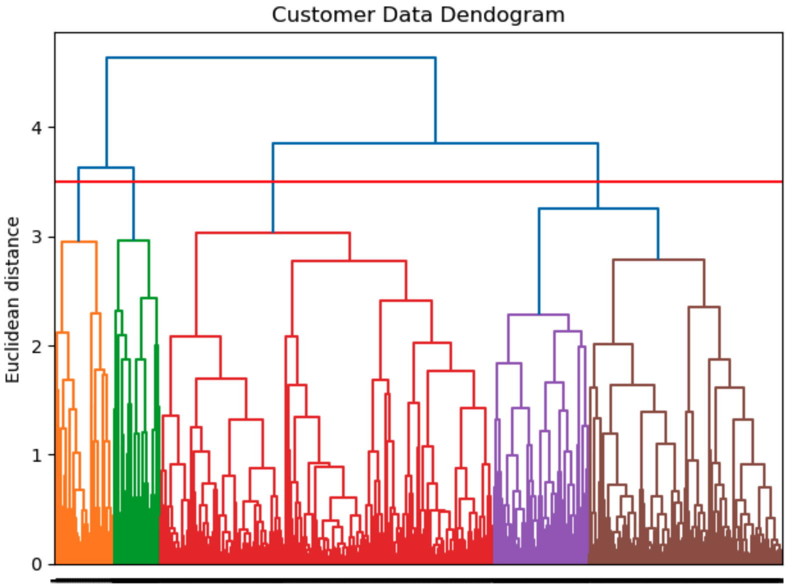


Cluster 5
This cluster represents married males , who're mostly unskilled or are officials , have medium income, are mostly officials , who live in small to large sized cities.



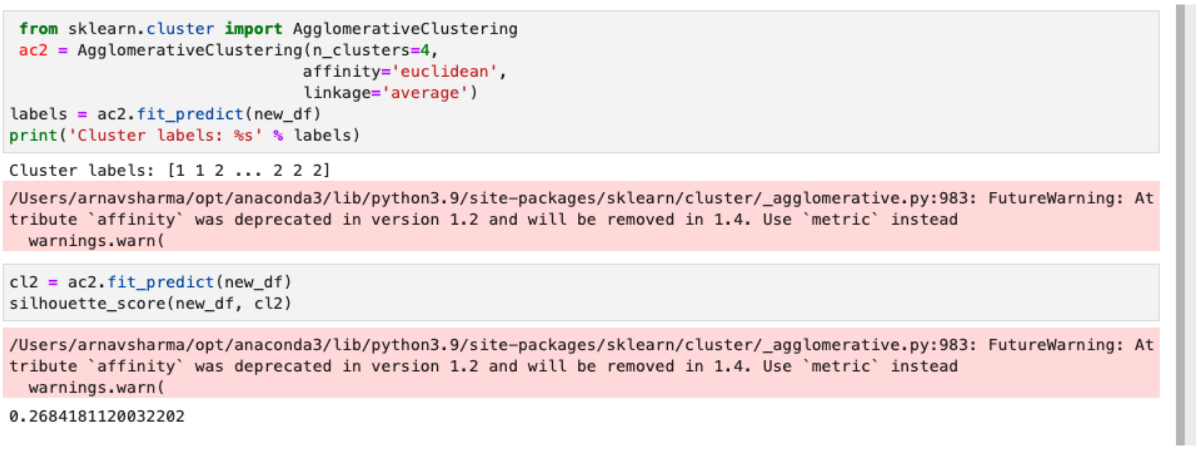
Hierarchical Clustering

We plot the dendrogram after creating the distance matrix.



From the above dendrogram we can conclude that 4-5 clusters are appropriate for the data.

After pruning the number of clusters and try different combinations of parameters, the following model rendered us the highest silhouette scores.



```
from sklearn.cluster import AgglomerativeClustering
ac2 = AgglomerativeClustering(n_clusters=4,
                              affinity='euclidean',
                              linkage='average')
labels = ac2.fit_predict(new_df)
print('Cluster labels: %s' % labels)
```

```
Cluster labels: [1 1 2 ... 2 2 2]
/Users/arnavsharma/opt/anaconda3/lib/python3.9/site-packages/sklearn/cluster/_agglomerative.py:983: FutureWarning: At
tribute `affinity` was deprecated in version 1.2 and will be removed in 1.4. Use `metric` instead
  warnings.warn(
```

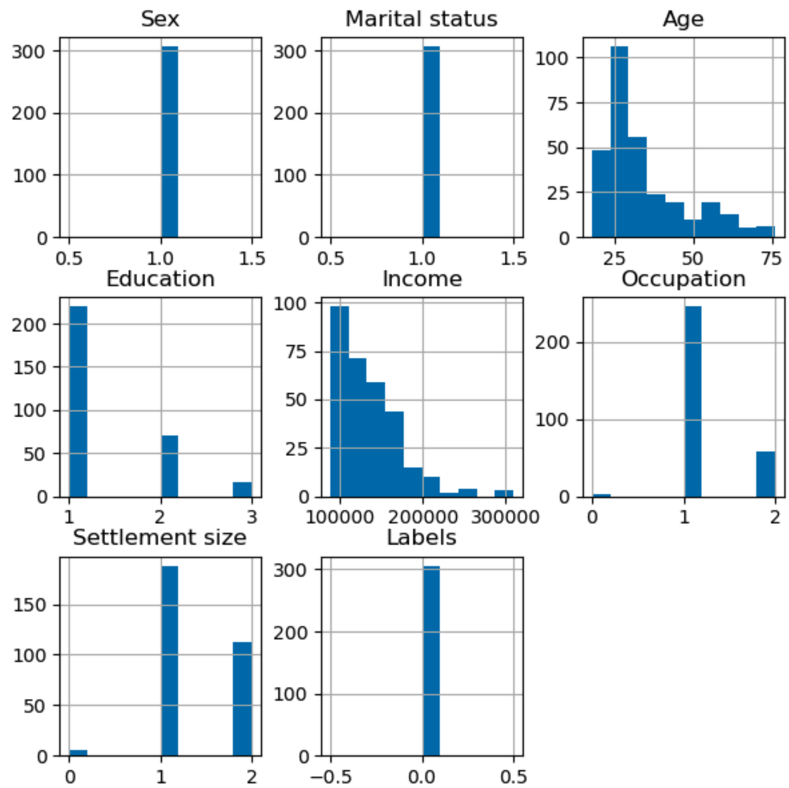
```
cl2 = ac2.fit_predict(new_df)
silhouette_score(new_df, cl2)
```

```
/Users/arnavsharma/opt/anaconda3/lib/python3.9/site-packages/sklearn/cluster/_agglomerative.py:983: FutureWarning: At
tribute `affinity` was deprecated in version 1.2 and will be removed in 1.4. Use `metric` instead
  warnings.warn(
0.2684181120032202
```

Key Segments

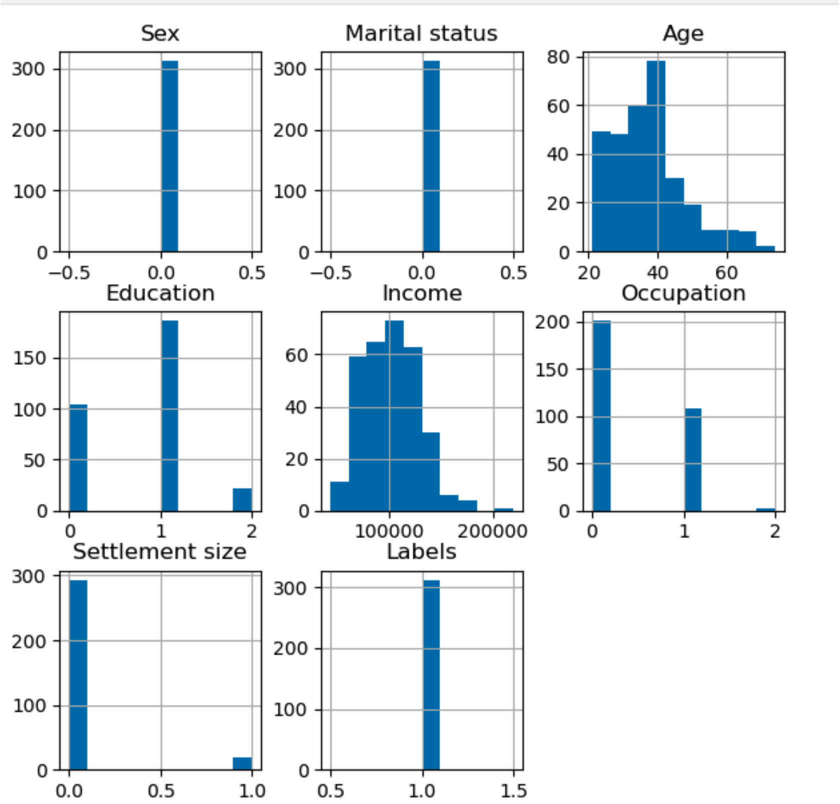
Cluster 0

This cluster is marked by married women, usually uneducated, low-income strata, are employed and live in medium to big cities.



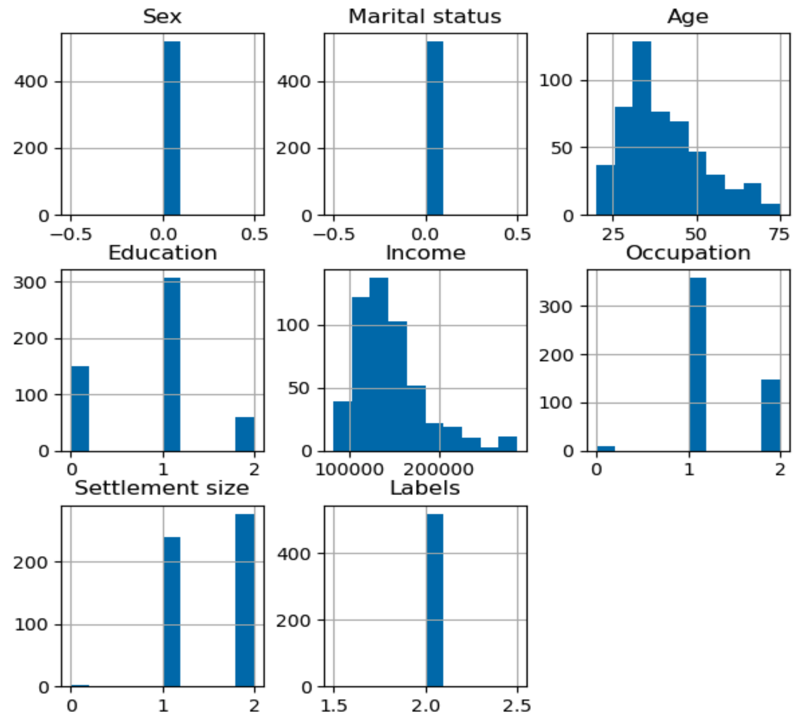
Cluster 1

This cluster is represented by unmarried men, mostly educated, in lower to middle income bracket and employed, mostly living in small cities.

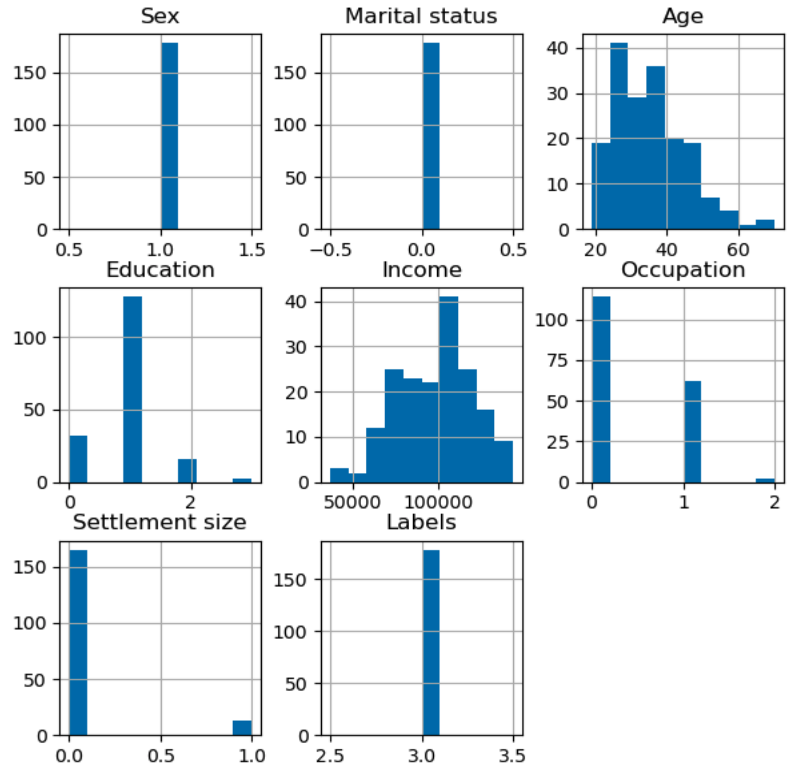


Cluster 2

This cluster is represented by unmarried men living in medium to big cities who are educated, earn medium income and works as officials, officers etc.



Cluster 3



This cluster is marked by unmarried women who’re educated, earn high salaries, are employed and live in small cities.

Comparison

KMeans method ranks higher in silhouette scores than hierarchal, indicating it will have better and more Cohesive clusters with its datapoints staying within its boundaries. The clusters in KMeans are more detailed and much more precise than hierarchal clustering.

Recommendations

Based on the customer segmentation results from both K-Means and Hierarchical clustering, we can tailor marketing techniques to effectively target each cluster:

Cluster 0 (K-Means): Focus marketing efforts on married women with lower education, offering products and promotions suitable for their budget. Emphasize convenience and affordability for those residing in medium to big cities.

Cluster 1 (K-Means): Target single males who are high school graduates and earn relatively more. Create promotions that appeal to their specific interests and living in medium to large cities.

Cluster 3 (K-Means): Concentrate on single women with higher education and income, offering premium products and tailored services. Highlight the convenience and options for those in smaller cities.

Cluster 4 (K-Means): Tailor marketing for married females with medium income who live in smaller cities. Focus on their specific needs and budget-conscious solutions.

Cluster 5 (K-Means): Develop marketing strategies for married males with various job roles, highlighting products and services suitable for medium income levels. Consider the preferences of those living in small to large cities.

Hierarchical Clustering yields four distinctive customer segments. Focus on:

1. Cluster 0: Target budget-conscious married women in medium to big cities with promotions emphasizing convenience.

2. Cluster 1: Appeal to unmarried men in small cities with budget-friendly offers and tailored campaigns.

3. Cluster 2: Engage educated, medium-income unmarried men in medium to large cities with personalized promotions.

4. Cluster 3: Attract affluent, educated unmarried women in smaller cities with premium products and exclusive offers.

Note: We don’t use age as an indicator in our analysis, since its distribution is similar throughout all the clusters in both the techniques.

Conclusion

In conclusion, our customer segmentation analysis, utilizing KMeans and agglomerative clustering on a dataset of 2000 entries, has successfully identified distinct customer segments. From married women in low-income brackets to single males with higher incomes, each segment exhibits unique characteristics. These insights enable businesses to tailor marketing strategies effectively.

The comparison between KMeans and agglomerative clustering favours KMeans for its higher silhouette scores, indicating more cohesive clusters. The strategic implications are substantial, allowing businesses to allocate resources efficiently and enhance customer engagement.

While age wasn't a differentiator, attributes like education, income, and marital status played pivotal roles. Continuous adaptation of these segments will be crucial as market dynamics evolve.