Unsupervised Machine Learning Course Final Project

Customer Clustering

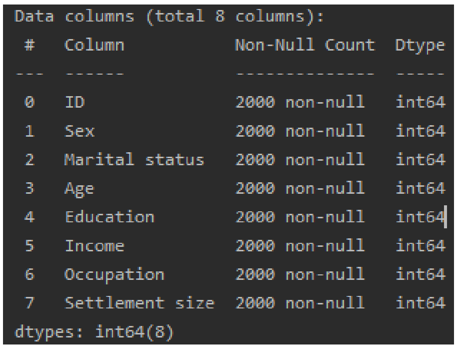
Customer Segmentation is the subdivision of a market into discrete customer groups that share similar characteristics. Customer Segmentation can be a powerful means to identify unsatisfied customer needs.

Data Description

Variable	Data Type	Range	Description
ID	numerical	Integer	Shows a unique identification of a customer
Sex	categorical	{0,1}	0: male 1: female
Martial Status	categorical	{0,1}	0: single 1: non-single (divorced, separated, married, widowed)
Age	numerical	Integer	18: Min value (the lowest age observed in the dataset) 76: Max value
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 data set) 309364: Max value (the highest income observed in the data set)
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	{01,2}	The size of the city that the customer lives in 0: a small city 1: a mid-sized city 2: big city

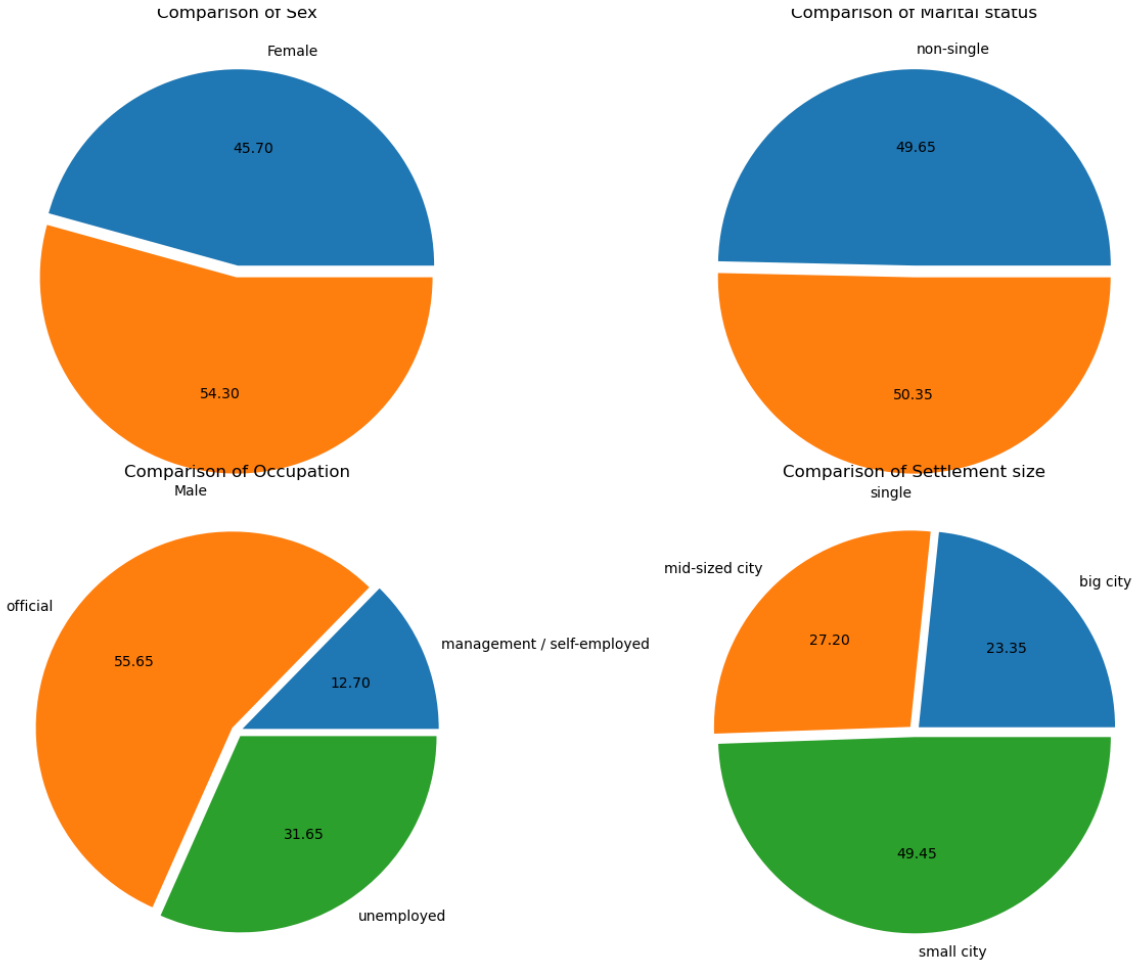
Exploratory Data Analysis

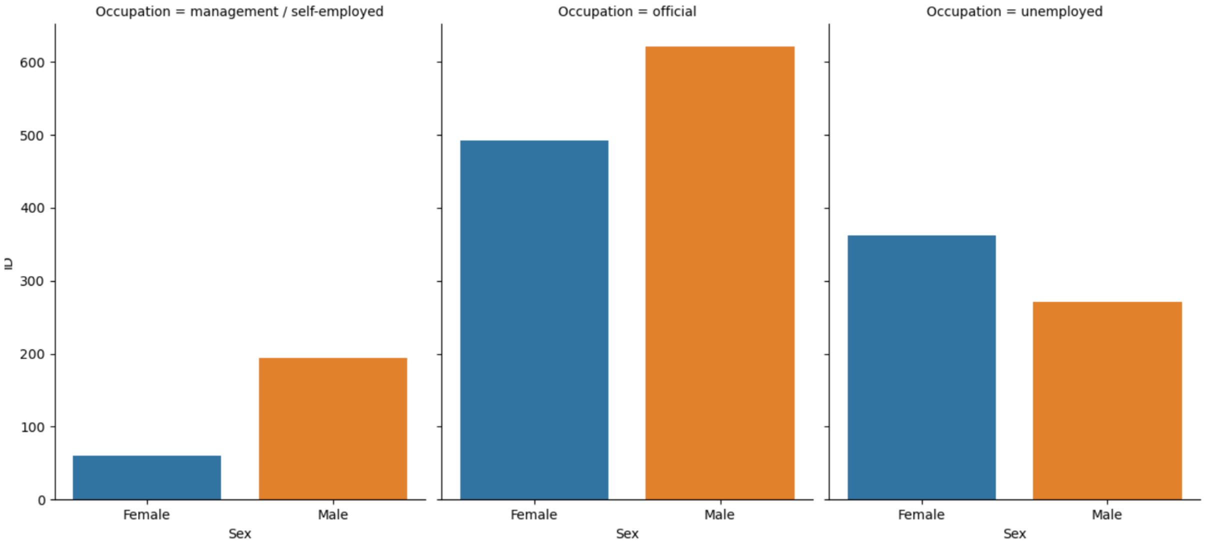
Check rather the data has the non-null data



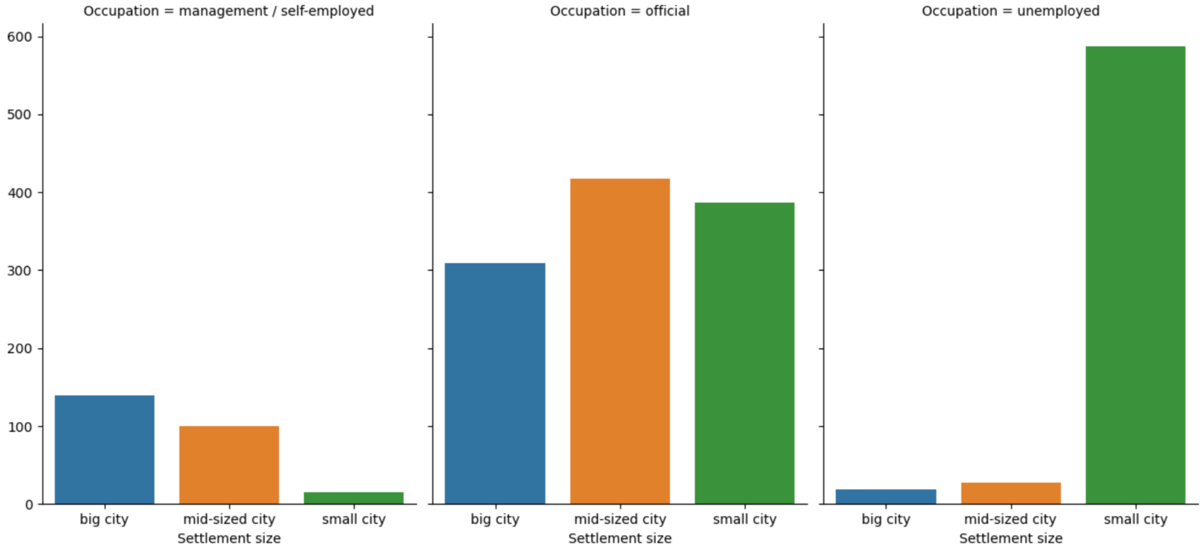
```
Data columns (total 8 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   ID               2000 non-null   int64
 1   Sex              2000 non-null   int64
 2   Marital status   2000 non-null   int64
 3   Age              2000 non-null   int64
 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)
```

As you can see below, the proportion of married/unmarried and male/female are almost the same. And the ‘official’ proportion in the occupation is over 50% and almost half of the people live in the small city.

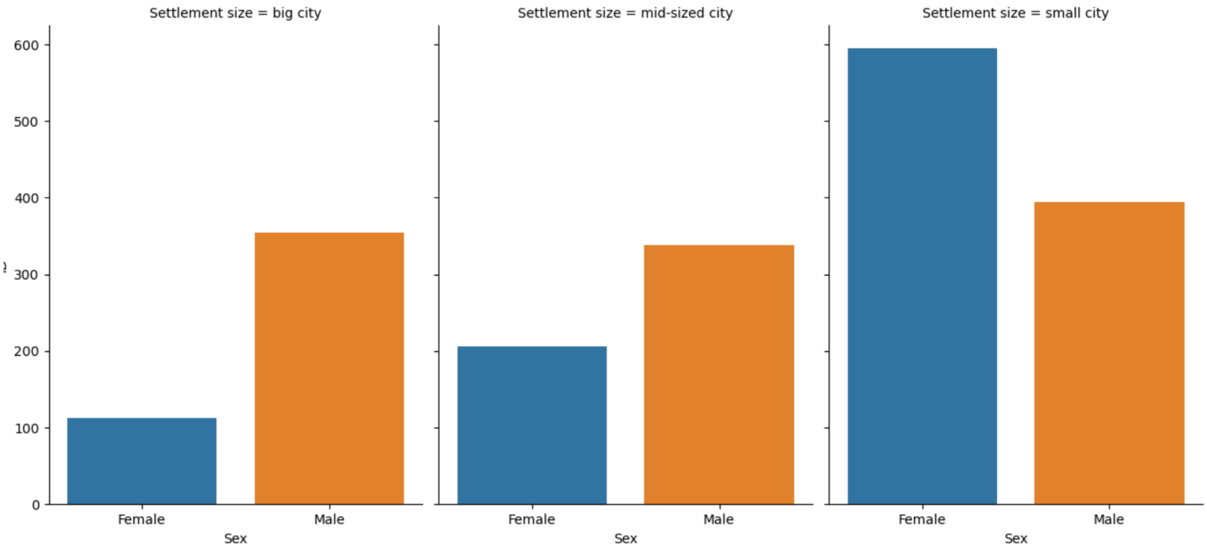




The plot shows that most of the occupations are dominated by males, except the ‘unemployed’ section.



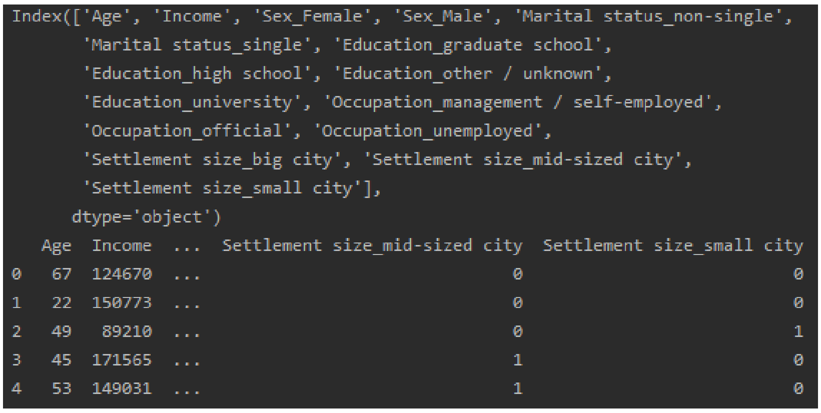
As you can see, people who are managers and self-employed are living in big cities and mid-sized cities. In contrast, the official job is evenly distributed.



By combining the above two plots, we can get an insight that male who lives in the big city is highly to get a job such as management, self-employed and official ones.

Data PreProcessing

In the data, there are several object-type columns (Sex, Marital status, Education. Occupation, Settlement size). For the sake of an efficient training process, I applied one-hot encoding using pd.get_dummies to the above columns. Additionally, the customer ID is deleted due to the fact that ID is not the crucial column for learning.



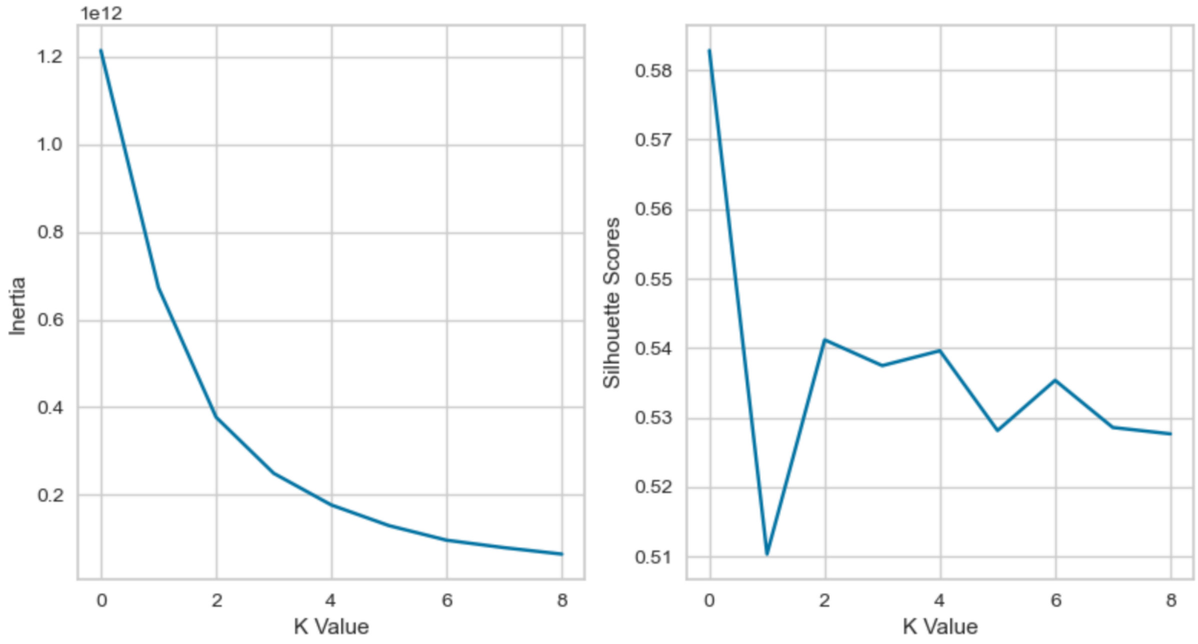
```
Index(['Age', 'Income', 'Sex_Female', 'Sex_Male', 'Marital status_non-single',
       'Marital status_single', 'Education_graduate school',
       'Education_high school', 'Education_other / unknown',
       'Education_university', 'Occupation_management / self-employed',
       'Occupation_official', 'Occupation_unemployed',
       'Settlement size_big city', 'Settlement size_mid-sized city',
       'Settlement size_small city'],
      dtype='object')
   Age  Income  ...  Settlement size_mid-sized city  Settlement size_small city
0   67  124670  ...                               0                           0
1   22  150773  ...                               0                           0
2   49   89210  ...                               0                           1
3   45  171565  ...                               1                           0
4   53  149031  ...                               1                           0
```

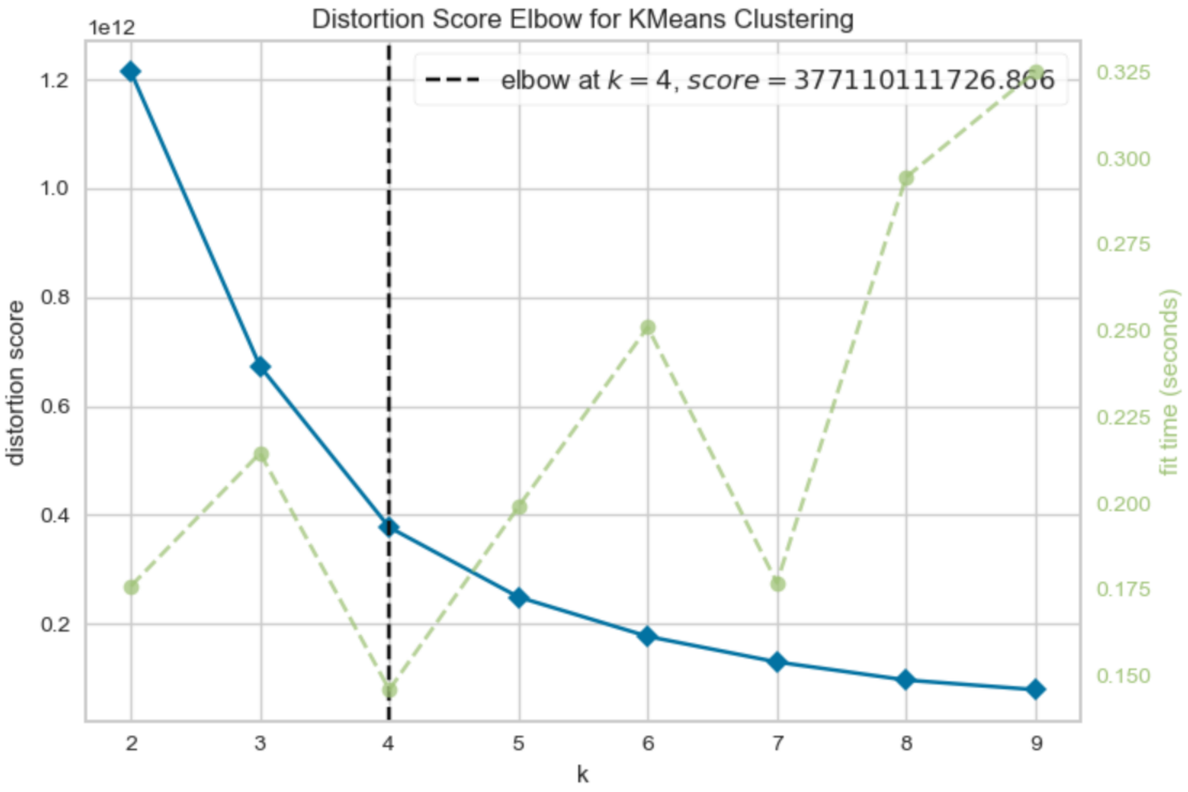
Unsupervised Models

Before getting started, I measured and contrasted each of the algorithms through silhouette scores.

1. KMeans Clustering

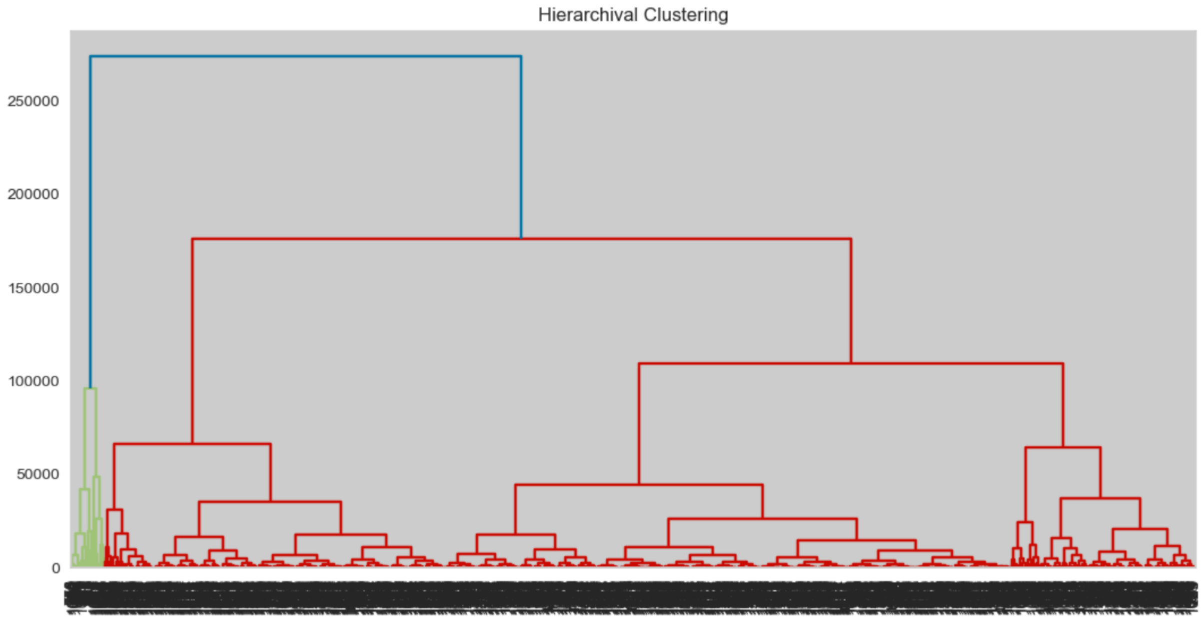
To find the best number of clusters, I used inertia, silhouette scorers, and KMeans Elbow visualization.





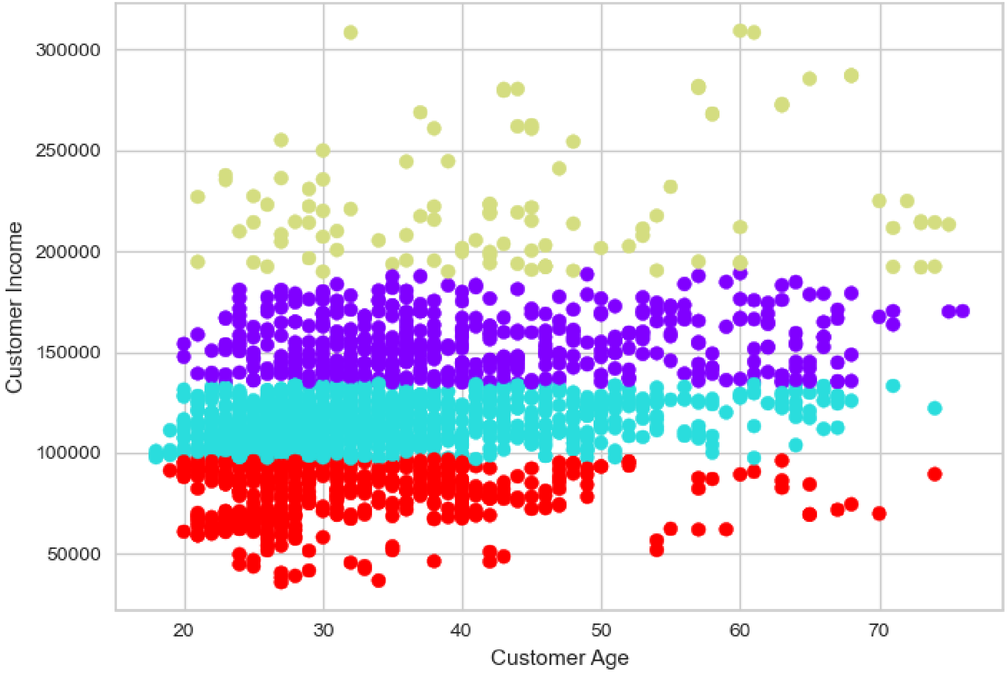
The silhouette score at k=4 is 0.53026 which is quite good.

2. Agglomerative Clustering



The silhouette score of the Agglomerative clustering is 0.51026 which is similar to KMeans clustering.

The silhouette score of the KMeans algorithms is better than the Agglomerative clustering, so I choose the final model as the KMeans algorithm.



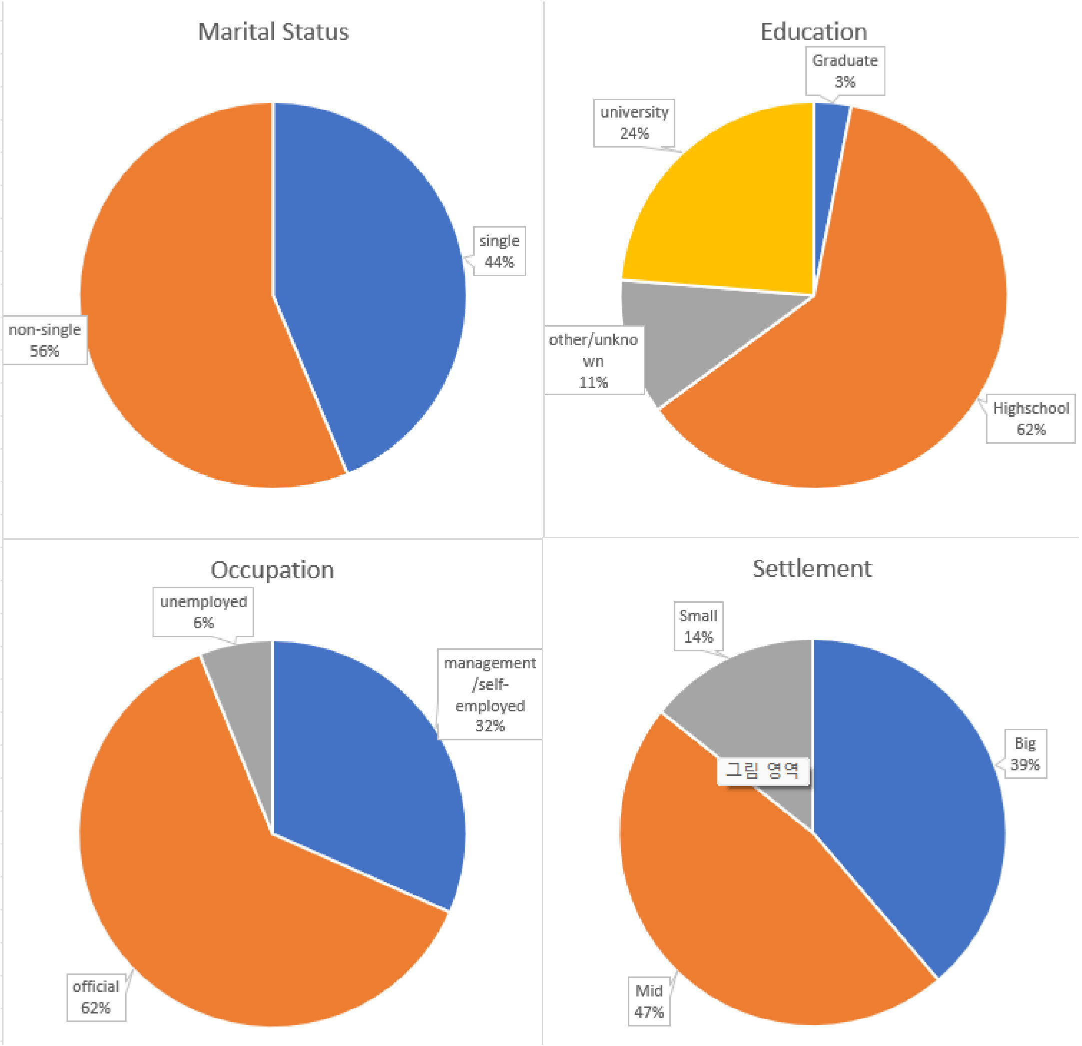
As you can see, 4 clusters are easily shown by the age-income scatter plot. The other columns are divided through one-hot encoding, so it is better to watch the age income scatter plot to see the well-divided clusters.

Comparison between CLusters

Description of divided clusters

- Cluster 1

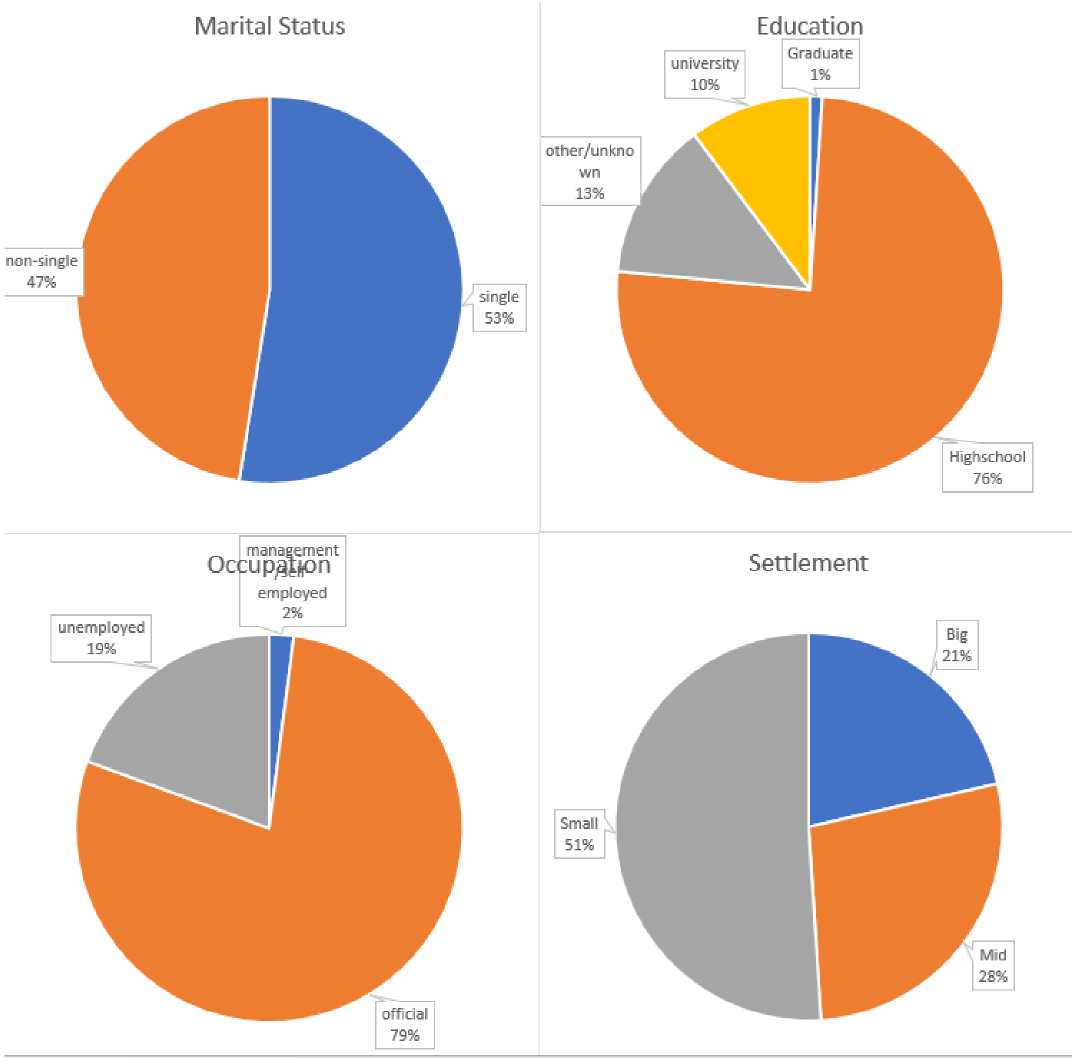
Average Age: 40 / Average Income: 154237



As you can see Cluster 1 people are usually married, and their last status of education is in high school or university. And they almost have official jobs and live in a big or mid-sized city.

- Cluster 2

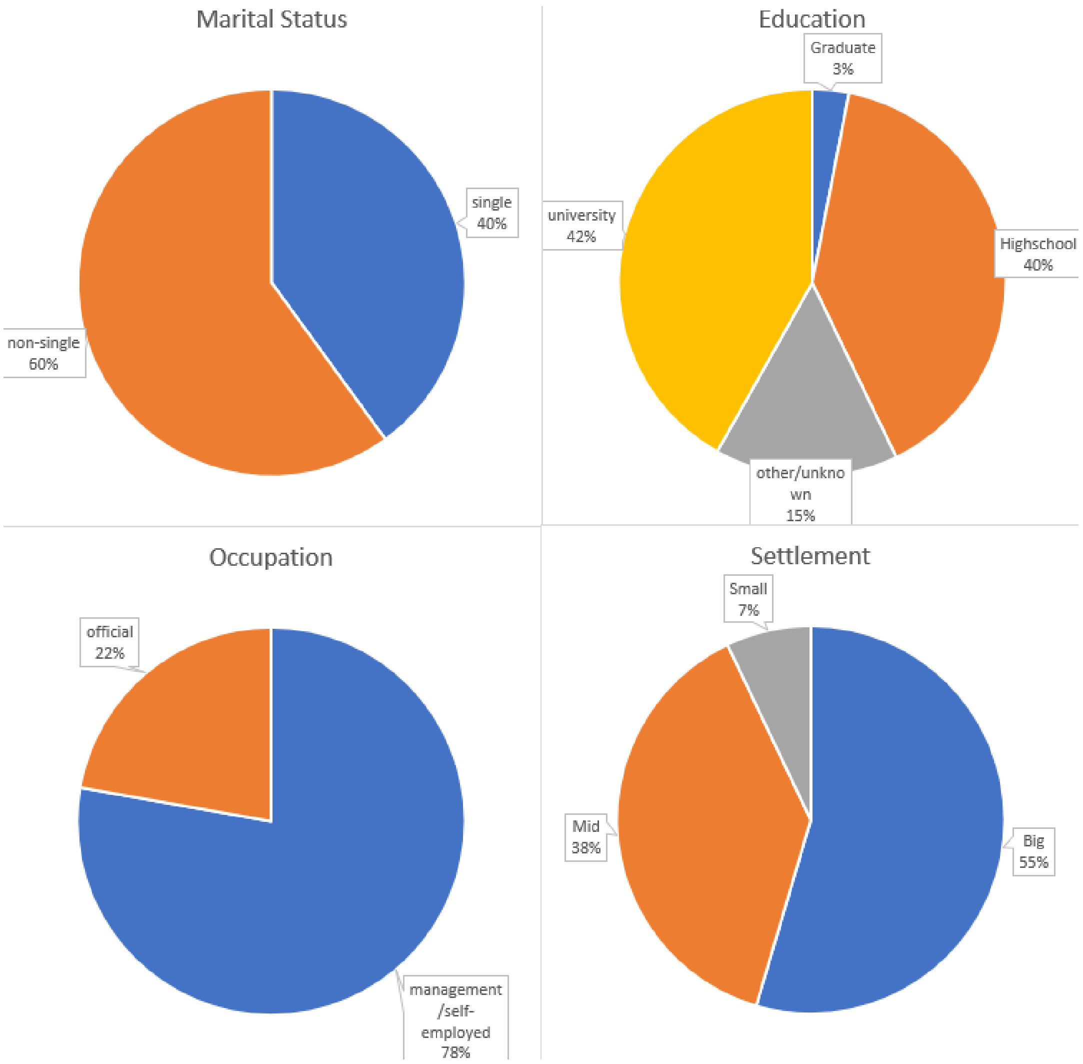
Average Age: 34 / Average Income: 114869



As you can see Cluster 2, their last status of education is in high school or university. And they almost have official jobs and live in a small city.

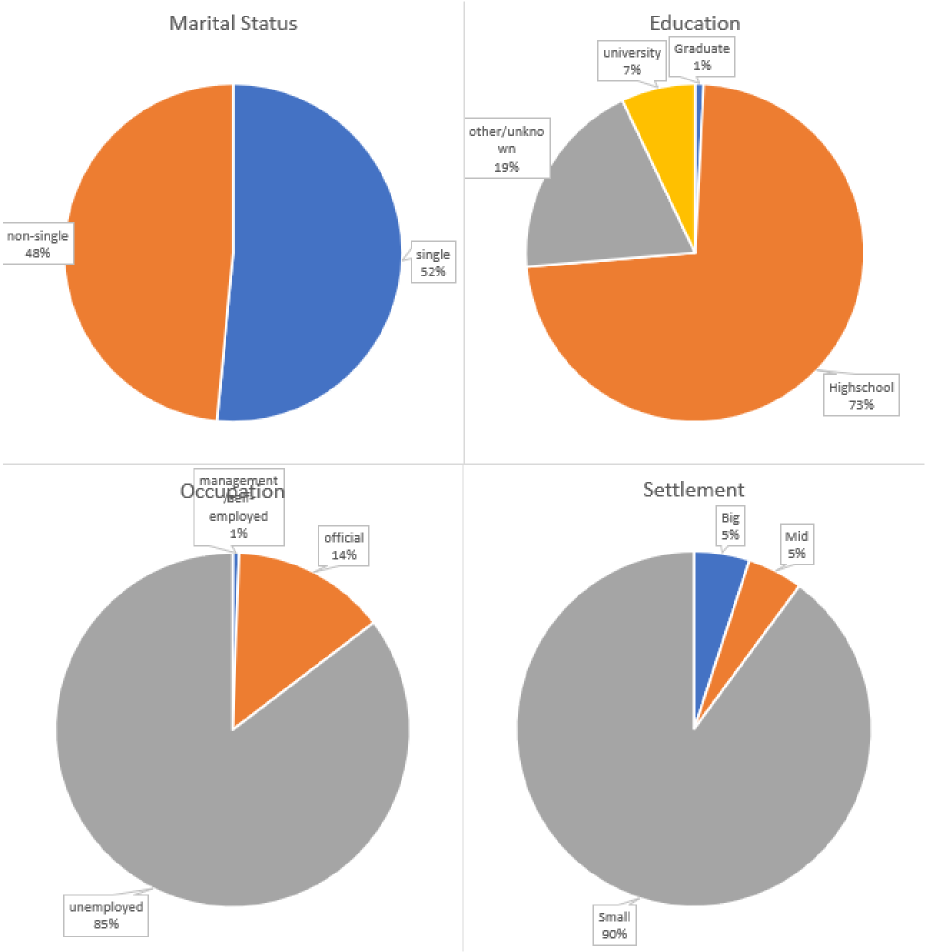
- Cluster 3

Average Age: 44/ Average Income: 22514



- Cluster 4

Average Age: 32 / Average Income: 79519



Total Summary

	Age(average)	Income(avg)	Most Frequent Marital Status	Most Frequent Education	Most Frequent Occupation	Most Frequent Settlement
Cluster 1	40	154,327	almost even	High School	Official	mid-sized
Cluster 2	34	114,869	almost even	High School	Official	small
Cluster 3	44	225,124	almost even	University	management self-employed	big
Cluster 4	42	79,519	almost even	High School	unemployed	small

As you can see in the total summary section it is hard to divide the clusters by average age, due to their similar values. However, when you see the income columns, the differences are quite conspicuous. Also, by combining Education, Occupation, and Settlement columns it is much easier to segment the customers. By clustering the customers the supermarket can develop a marketing strategy for each cluster’s customers.