Achraf Safsafi
DSC 550
Final Project: Customer Segmentation Analysis

Problem Statement:

Mall owners always seek to increase their profits by increasing their customers. Here comes the role of machine learning to achieve this goal, where customer data are used to reach a proper business strategy. In this project, I will use a mall customer dataset to perform a customer segmentation analysis using some basic unsupervised machine learning techniques . I want to divide a customer base into groups of individuals that are similar in specific ways relevant to marketing, such as spending habits in order to understand the customers so the mall administration can target the right ones and have sufficient visibility to design the best marketing strategy.

Data Set:

The dataset used in this project is published on the Kaggle website:

https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python

The dataset contains 200 observations of 5 variables, which are described below.

CustomerID: Unique ID assigned to the customer

Gender: Gender of the customer

Age: Age of the customer

Annual Income: Annual Income of the customer

Spending Score: Score assigned by the mall-based on customer behavior and spending nature.

Achraf Safsafi
DSC 550
Final Project: Customer Segmentation Analysis

Here are the first ten rows of the data set:

CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77
5	Female	31	17	40
6	Female	22	17	76
7	Female	35	18	6
8	Female	23	18	94
9	Male	64	19	3
10	Female	30	19	72

As we see, we have a dataset consist of 200 mall customers data . The data frame includes CustomerID, genre,age,annual income and spending score of each customer.

Next, we call describe () function on the dataset to see the descriptive statistics for each variable.

	Age	Annual Income (k$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000
mean	38.850000	60.560000	50.200000
std	13.969007	26.264721	25.823522
min	18.000000	15.000000	1.000000
25%	28.750000	41.500000	34.750000
50%	36.000000	61.500000	50.000000
75%	49.000000	78.000000	73.000000
max	70.000000	137.000000	99.000000

According to the results above, all data looks good. There are no evident outliers. All the value of the Spending score is between 1 and 100.

Achraf Safsafi
DSC 550
Final Project: Customer Segmentation Analysis

Checking for missing values:

0	Gender	200 non-null	object
1	Age	200 non-null	int64
2	Annual Income (k$)	200 non-null	int64
3	Spending Score (1-100)	200 non-null	int64

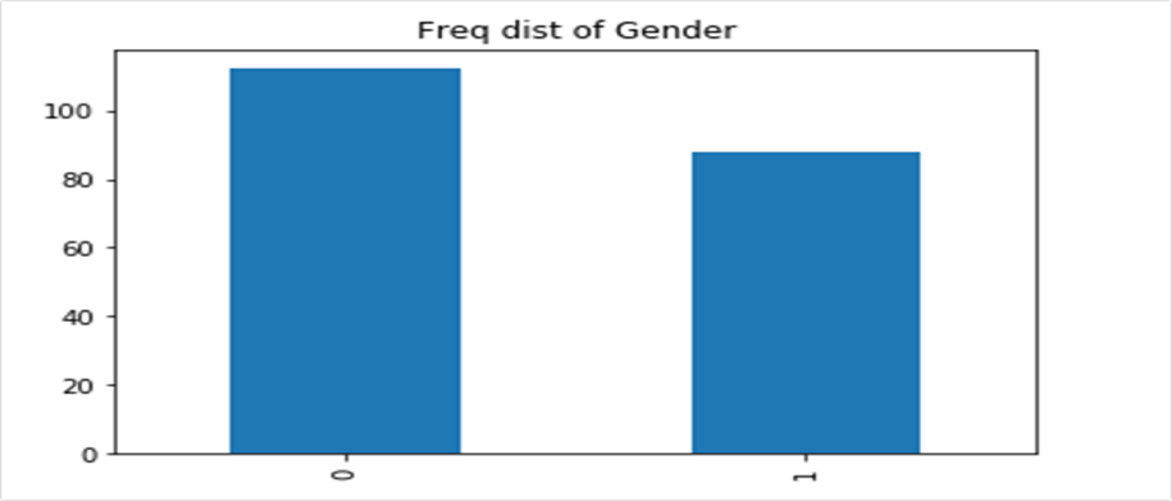
As we see there is no null values.

Feature scaling:

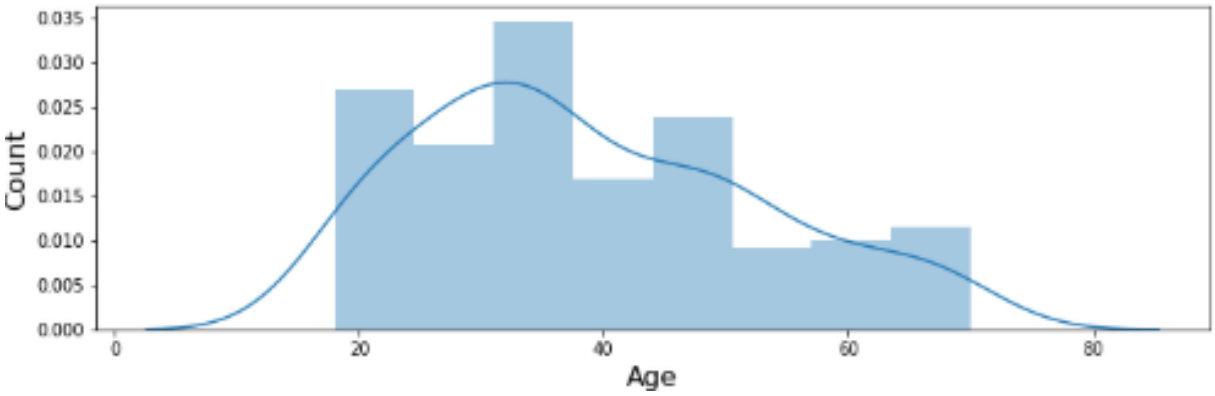
CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
1	1.128152	-1.424569	-1.738999	-0.434801
2	1.128152	-1.281035	-1.738999	1.195704
3	-0.886405	-1.352802	-1.700830	-1.715913
4	-0.886405	-1.137502	-1.700830	1.040418
5	-0.886405	-0.563369	-1.662660	-0.395980

Achraf Safsafi
DSC 550
Final Project: Customer Segmentation Analysis

Exploring the Data:



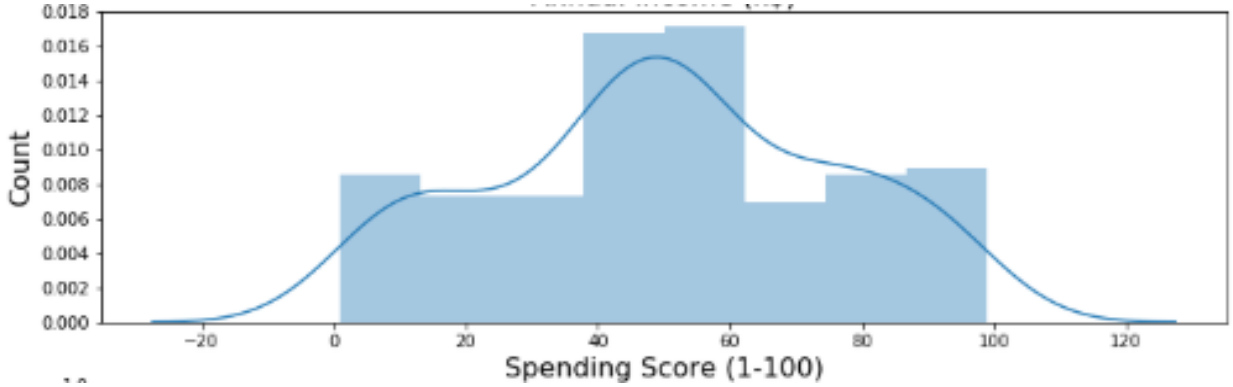
The plot shows that there are slightly more women than men in the dataset.



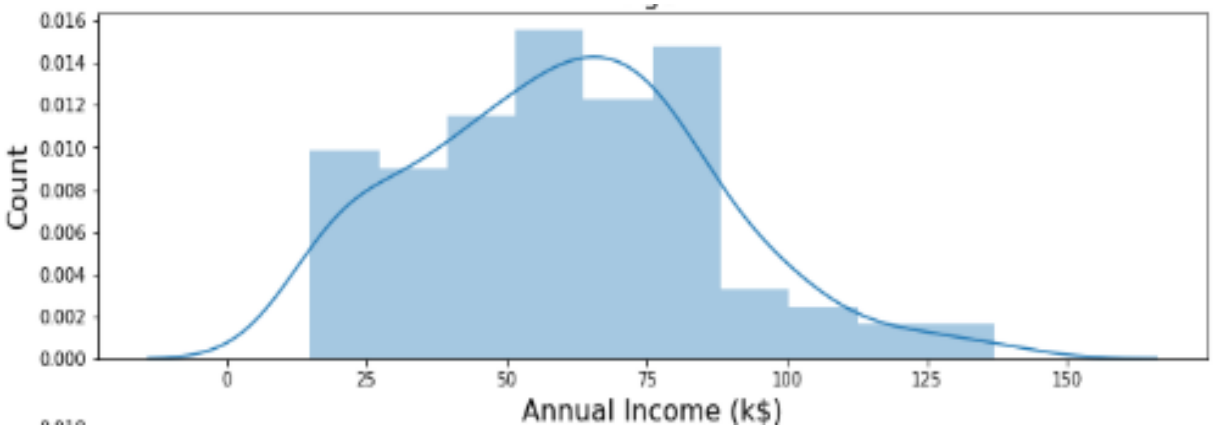
We can say that the Age is normally distributed. Its values are mostly between 35 and 40.

Also we observe that there are less old customers.

Achraf Safsafi
DSC 550
Final Project: Customer Segmentation Analysis



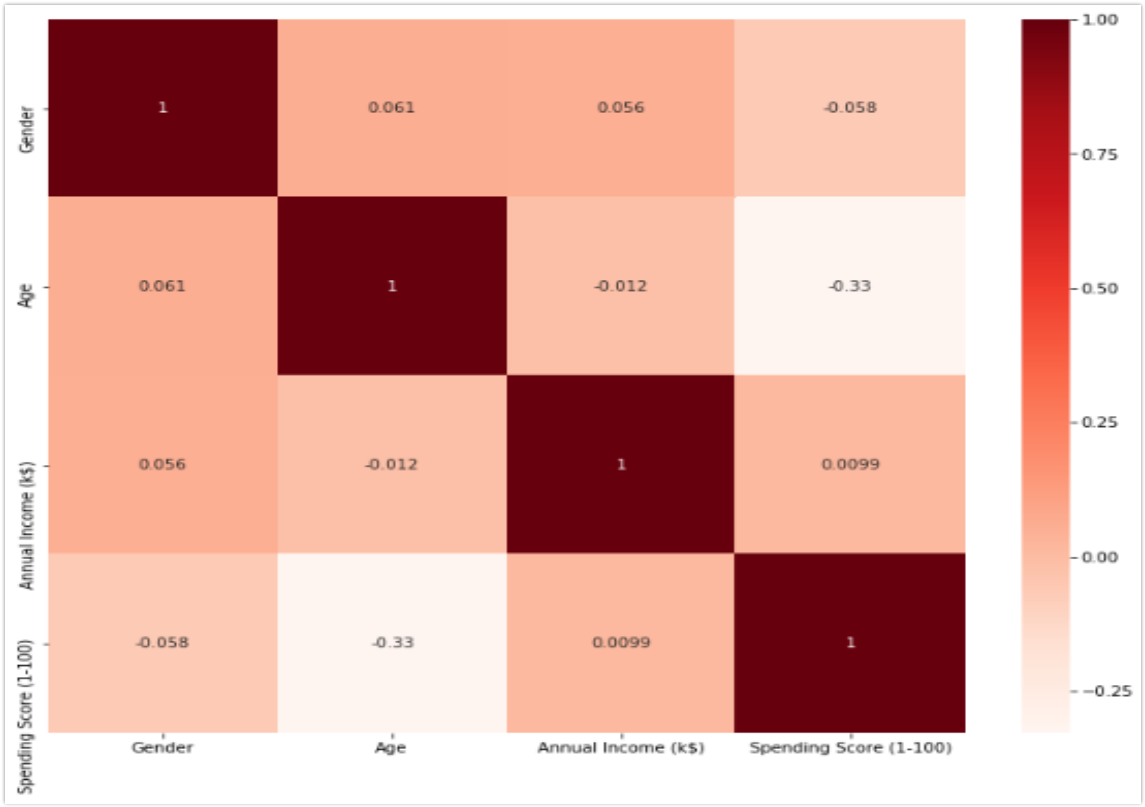
We can conclude that most of the customers have spending score between 40 and 60.



The distribution indicates that much of the incomes lies between 60,000 and 80,000 dollars.

Achraf Safsafi
DSC 550
Final Project: Customer Segmentation Analysis

Correlation:



Using Pearson Correlation, we conclude that the features Age and Spending Score has poor correlation. This means that the age cannot explain or predict the spending habits. Also, the results show that Age and Income has weakly negative correlation. This implies that also Age cannot explain Income. Then, Income and Spending Score has too slightly positive correlation. This shows that Income could explain about 0.99 % of Spending Score.

Achraf Safsafi
DSC 550
Final Project: Customer Segmentation Analysis

Clustering Analysis:

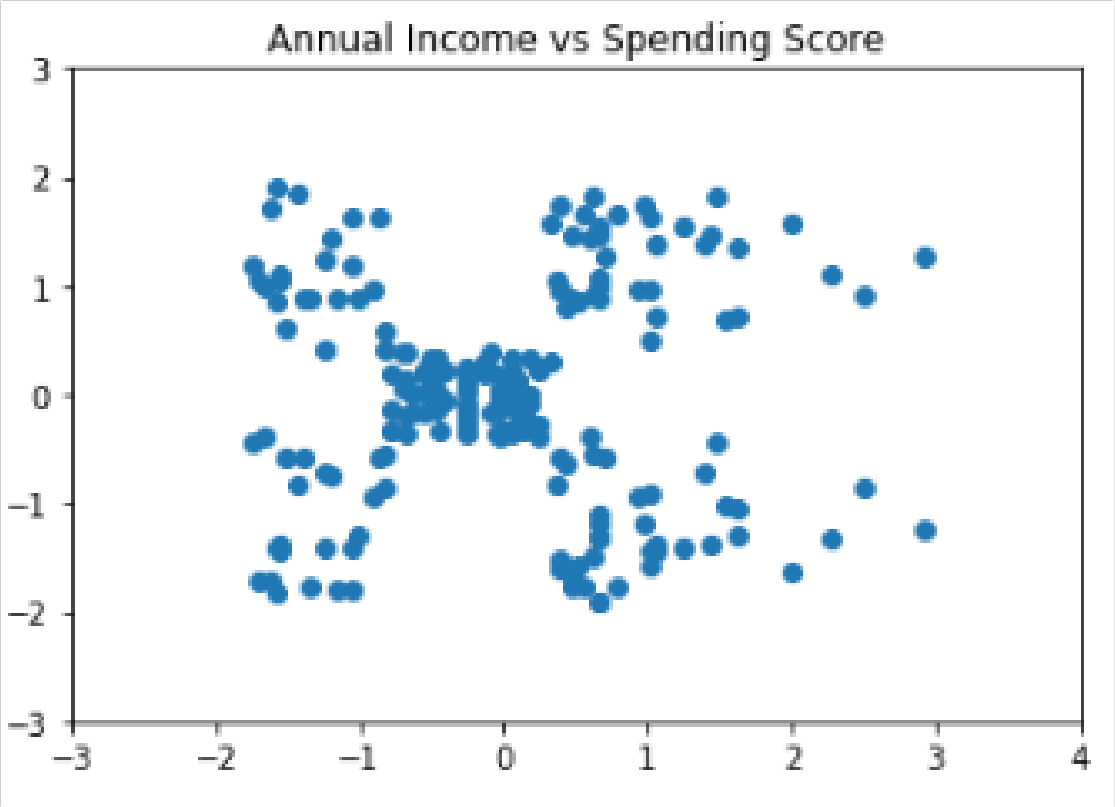
Let’s drop Age and Gender, and keep CustomerID as index, and we work only on Annul Income and Spending Score that the objective of the model.

CustomerID	Annual Income (k$)	Spending Score (1-100)
1	-1.738999	-0.434801
2	-1.738999	1.195704
3	-1.700830	-1.715913
4	-1.700830	1.040418
5	-1.662660	-0.395980
...	...	...
196	2.268791	1.118061
197	2.497807	-0.861839
198	2.497807	0.923953
199	2.917671	-1.250054

200	2.917671	1.273347

Visualizing the data:

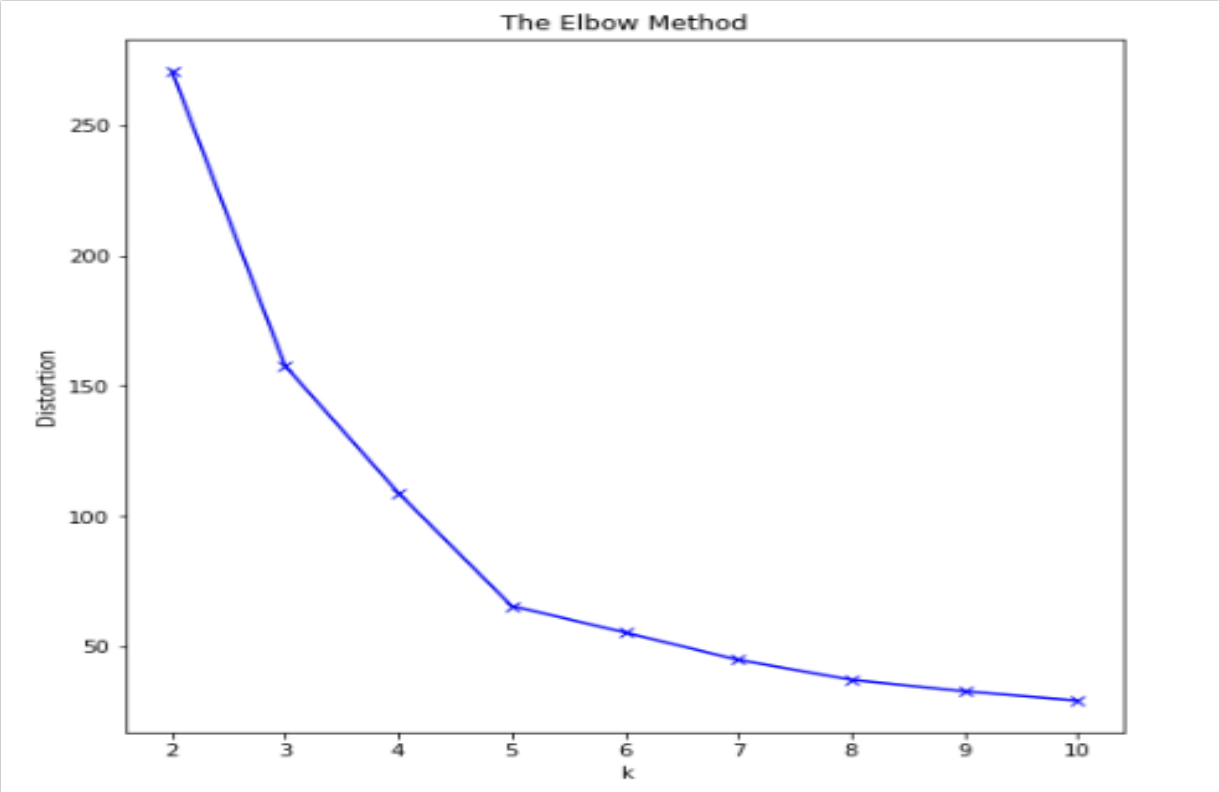
This scatter plot is for the data before clustering.



Achraf Safsafi
DSC 550
Final Project: Customer Segmentation Analysis

K means Algorithm:

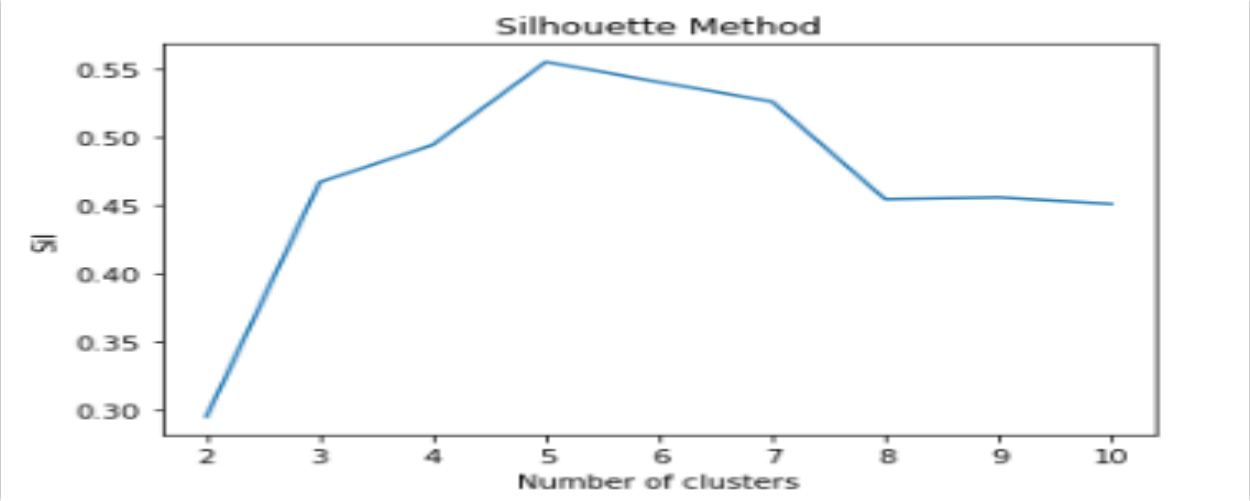
Elbow Method to Find the Optimal k:



Achraf Safsafi
DSC 550
Final Project: Customer Segmentation Analysis

Before we start implanting the algorithm, we need to determine the value of K, the number of clusters. So according to the graph above, we can say that the value of K could be 5.

To confirm the result, let's perform Silhouette Analysis for Selecting k.



2:0.29512063001659344

3:0.46658474419000145

4:0.4939069237513199

5:0.5546571631111091

6:0.5398800926790663

7:0.5256026931619203

8:0.4541279523637649

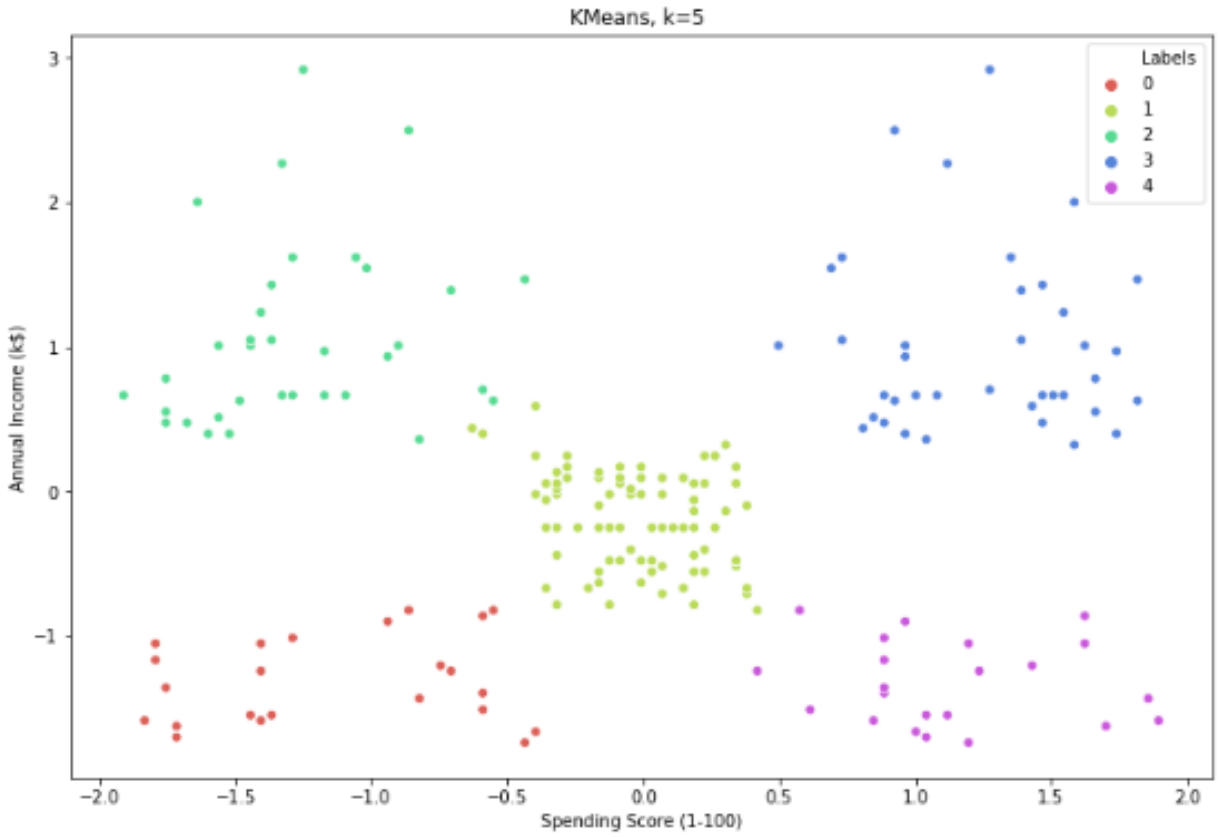
9:0.455477460588866

10:0.4507937437744966

The Silhouette method also shows that the optimal value of K is 5.

Achraf Safsafi
DSC 550
Final Project: Customer Segmentation Analysis

Applying k-means algorithm:



The K means clustering analysis shows that there are clearly five clusters of customers. The data point in the bottom right, the purples ones, belong to the customers with low annual income but high spending. The data point in the bottom left, the red ones, belong to the customers with low annual income and low spending. The data point in the top left belong to the customers with high annual income but low spending. The data point in the bottom right, the purples ones, belong to the customers with low annual income but high spending. And The data point in the

middle, the green ones, belong to the customers with average annual income and average spending.

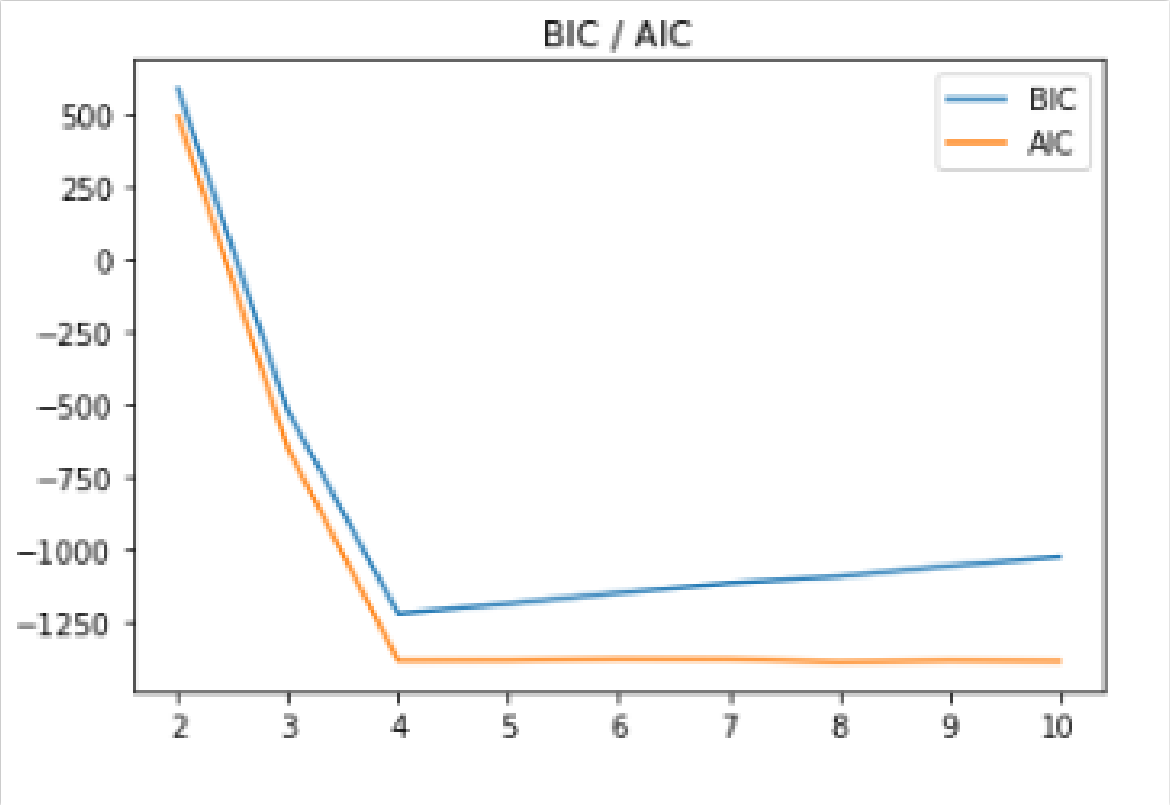
The data point in the top right, the blue ones, belong to the customers with high annual income

and high spending. And this our target. This type of the customers we look for. Besides, we can also focus on the customers that are in the middle as most of customers belong to this class.

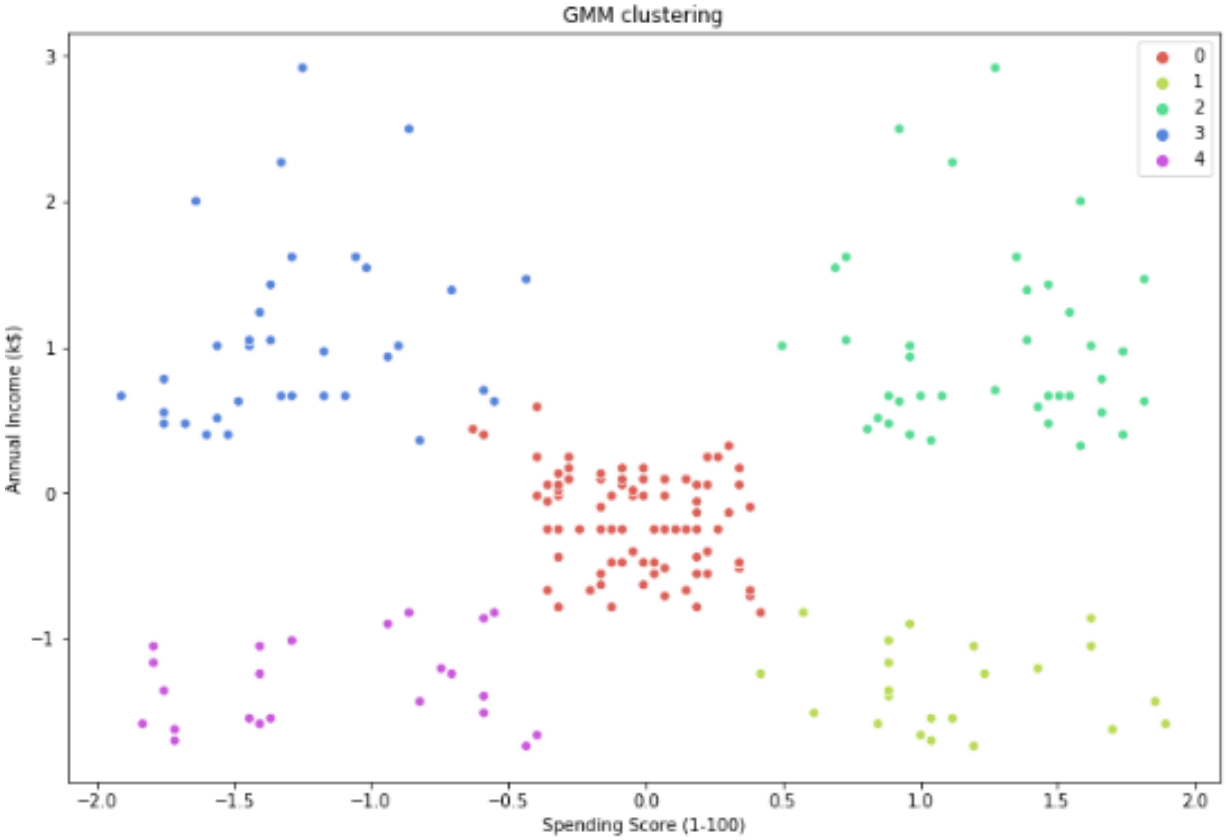
GMM Clustering:

Let’s now build a Gaussian Mixture Model (GMM) . But first let’s find the Optimal Numbers of Clusters using BIC/AIC Plot.

Achraf Safsafi
DSC 550
Final Project: Customer Segmentation Analysis



Achraf Safsafi
DSC 550
Final Project: Customer Segmentation Analysis

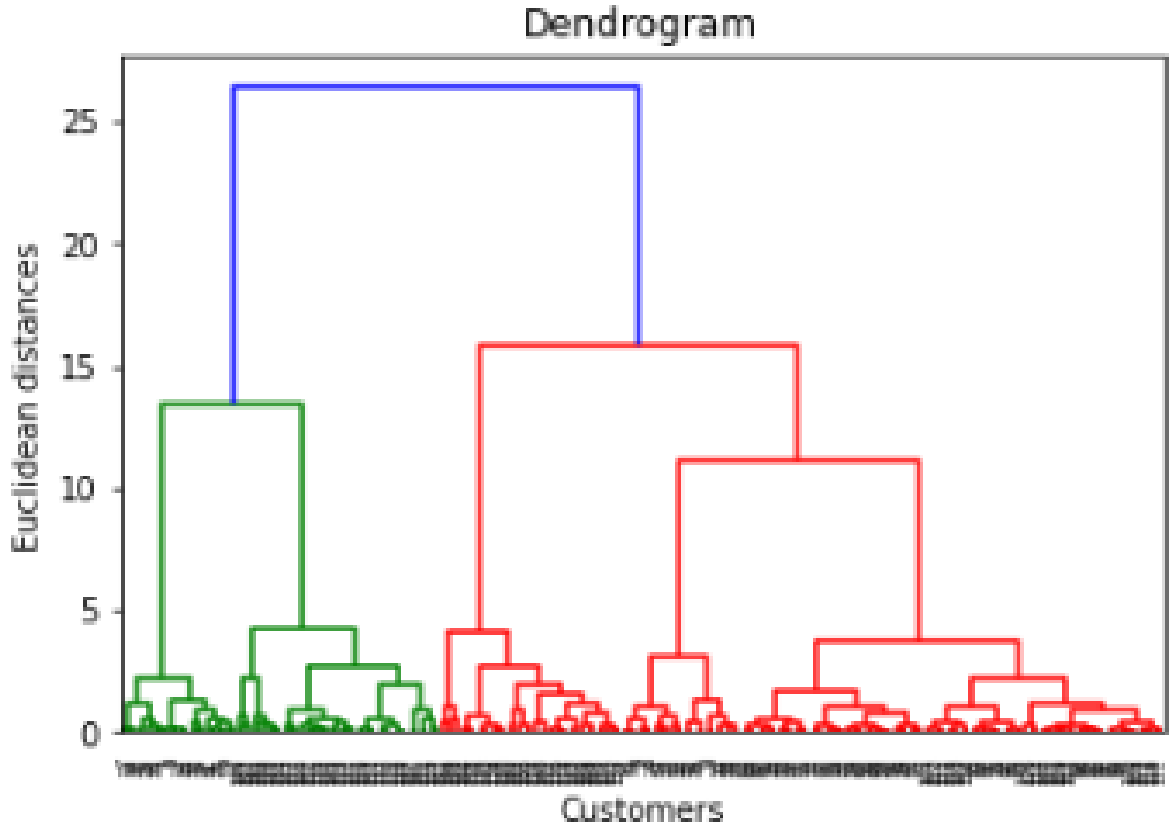


The GMM clustering analysis also shows that there five clusters of customers.

Hierarchical Clustering:

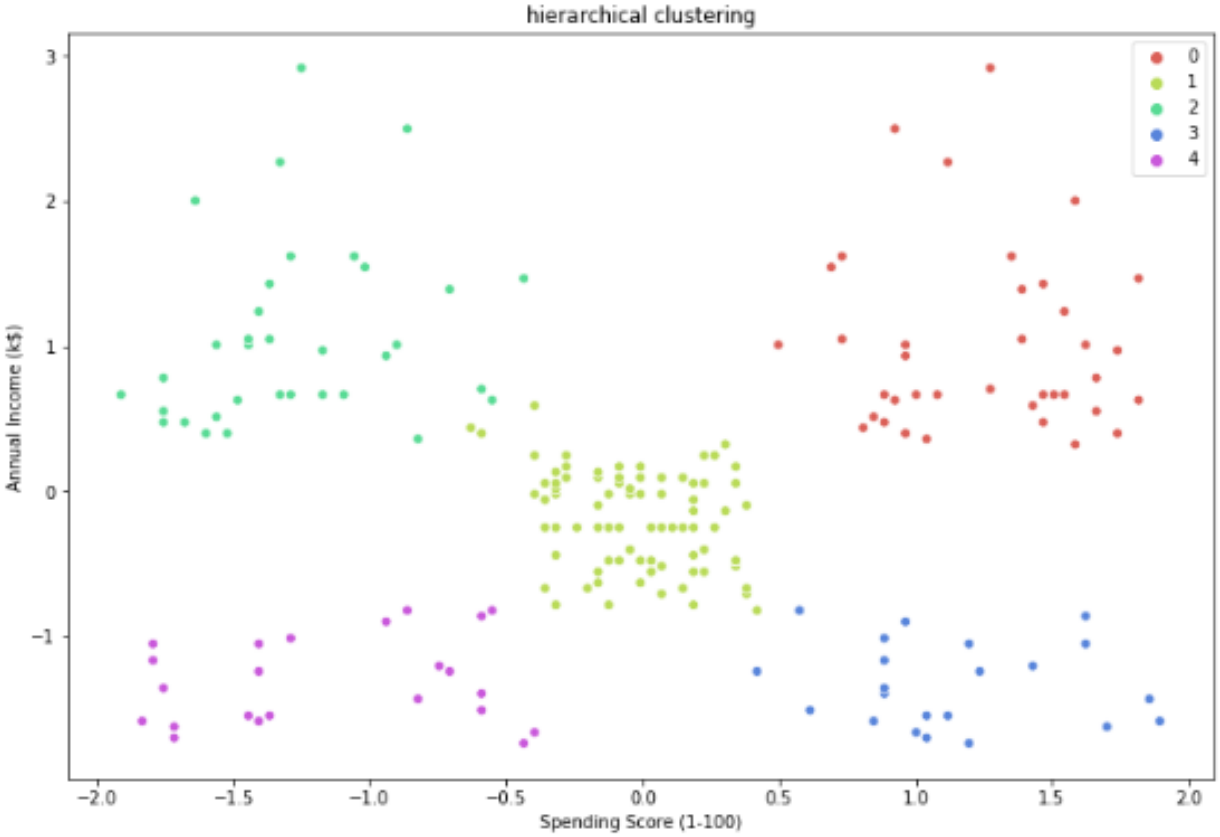
Achraf Safsafi
DSC 550
Final Project: Customer Segmentation Analysis

First, we will Find the optimal numbers of clusters before performing the hierarchical clustering.



In the above diagrams, we observe that the combination of 5 lines are not joined on the Y-axis from 5 to 13.So the optimal number of clusters will be also 5.

Achraf Safsafi
DSC 550
Final Project: Customer Segmentation Analysis



Applying Hierarchical Clustering Algorithm led to the same results as we got in K means clustering and GMM clustering.

Achraf Safsafi
DSC 550
Final Project: Customer Segmentation Analysis

Conclusion:

All three clustering methods led to similar results. We got 5 different clusters grouped by income and Spending Score. Among these clusters, there are two groups of customers that the mall administration must consider while designing their marketing strategy. They can target the consumers with high annual income and high spending and the customers with average annual income and average spending as well.