Jan 2021

Module 4 Assignment: Clustering Data

This project examines a dataset that captures customers data from an international Mall (https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python)

MAIN GOAL : This module assignment **focuses on finding hidden patterns**, we will try **to cluster data** for a better view of customer segmentation. . Dataset is small and easy to work with, has no missing values and almost no outliers (will ignore it). We will use 3 different cluster methods : **K-Means, DBScan, and Hierarchical Clustering** and compare their results.

Dataset overview:

- **Customer ID**: Id of customer, this field will be dropped as it's not useful
- **Gender**: customer gender - female / male
- **Age**: age of customer, from 18 to 70 years
- **Annual Income**: income of customer, will be renamed to income only, values from 13 to 137
- **Spending Score**: Score assigned by the mall based on customer behavior and spending nature, values from 1 to 99

Data preparation

	Gender	Age	Income	Score
0	Male	19	15	39
1	Male	21	15	81
2	Female	20	16	6
3	Female	23	16	77
4	Female	31	17	40

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 4 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Gender  200 non-null    object
 1   Age     200 non-null    int64
 2   Income  200 non-null    int64
 3   Score   200 non-null    int64
dtypes: int64(3), object(1)
memory usage: 6.4+ KB
```

There are no null values and data are not skwed.

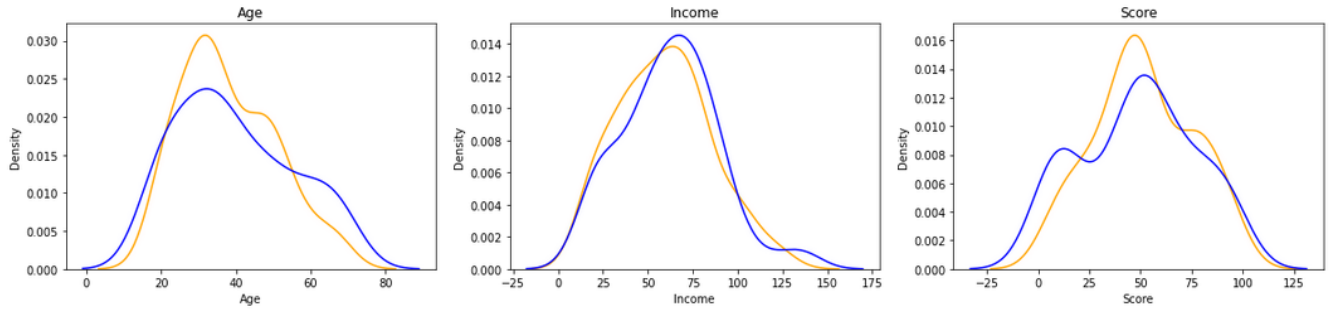
```
df.isnull().sum()
```

```
Gender    0
Age       0
Income    0
Score     0
dtype: int64
```

Exploratory data analysis

Distribution difference based on gender

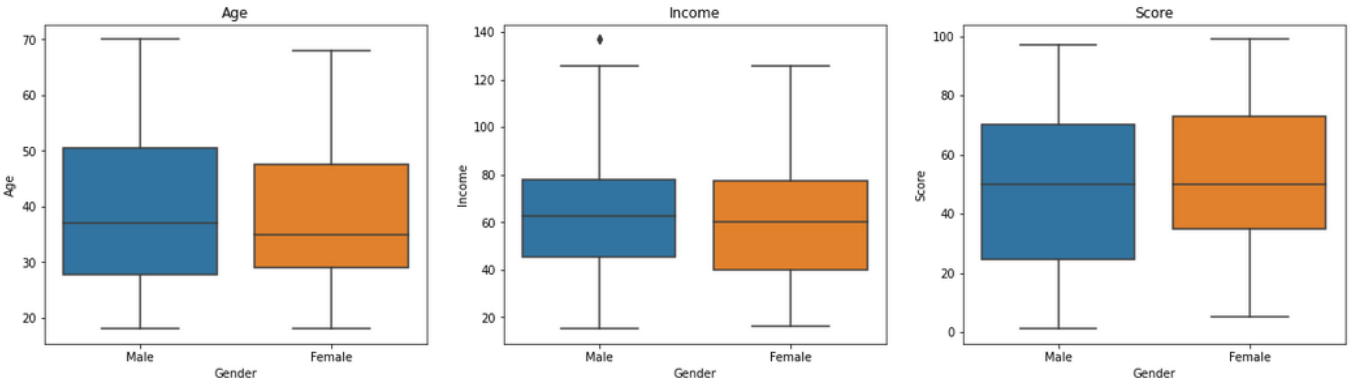
There is nothing significant, except slightly more females in age around 28 and slightly more females with score around 50.



Differences in Age, Income and Score by gender

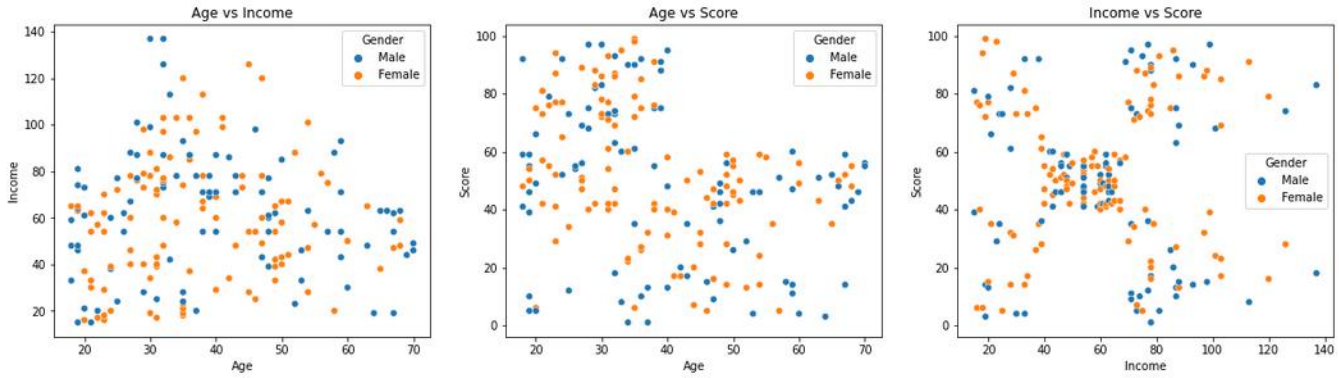
We will use Boxplot visualization to see also quartiles, distribution, median and outliers. You will find there is one outlier by Income/Male, but we will not handle it.

There is no significant difference or findings except that Female seems to have higher bound for lower score (first quartile)... This not unexpected, Females usually like to go to shopping malls more than Males.



Relation between variables

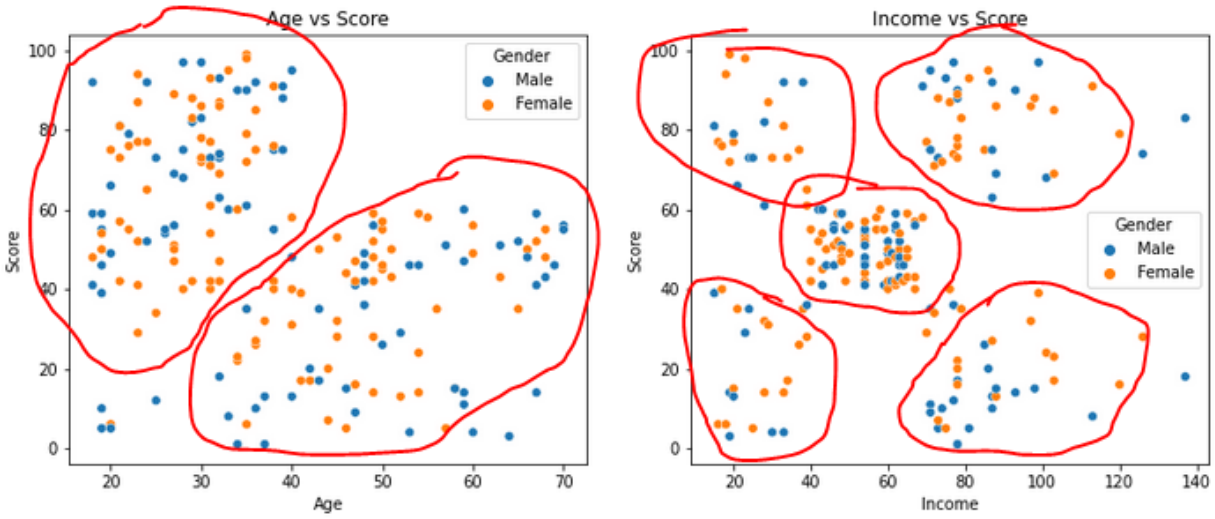
We want to check if there is significant relation between variables, (i.e. income increase with age or score decrease with age)



Key Findings :

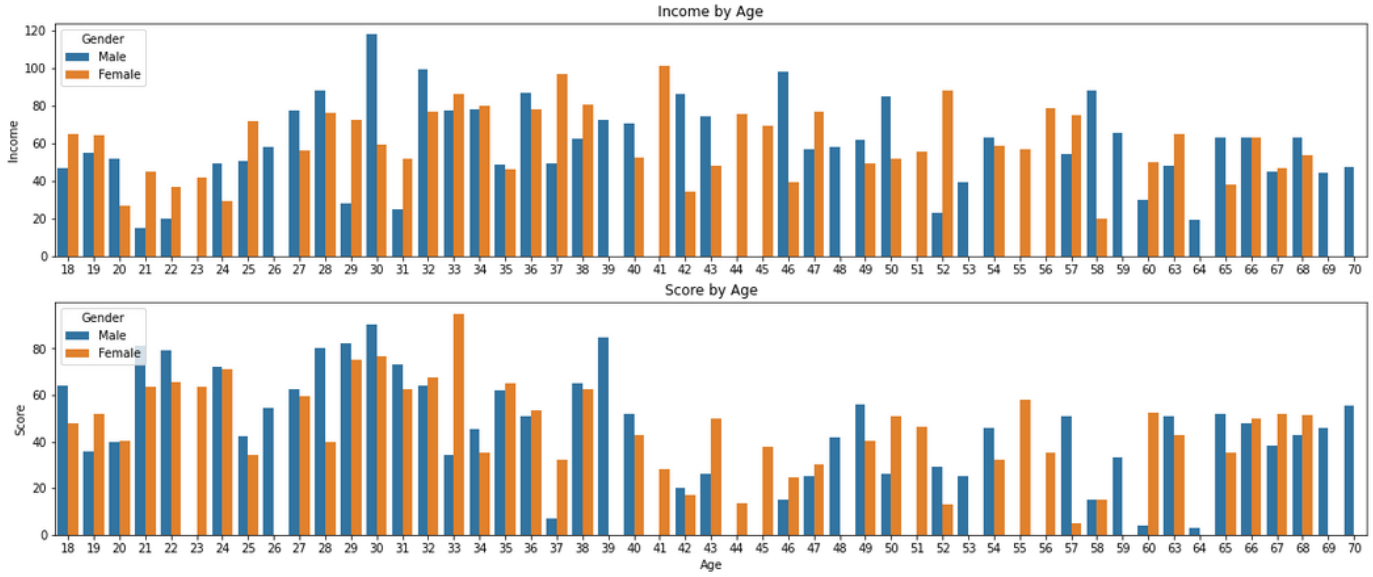
We can see there seems to be 2 groups of customers by age vs score (top left quarter vs bottom right quarter), where diagonal is delimiting them.

What is more important is chart Income vs Score where we can see 5 different groups of customers (corners & center).We've probably found ideal way to cluster our customers based on income and score.



Income & Score by Age

Last, check if there is significant difference (increasing/decreasing trend) when looking on Income or Score by Age.



Key Findings :

It’s interesting to note that 18 years people has almost same score as 60 years old.

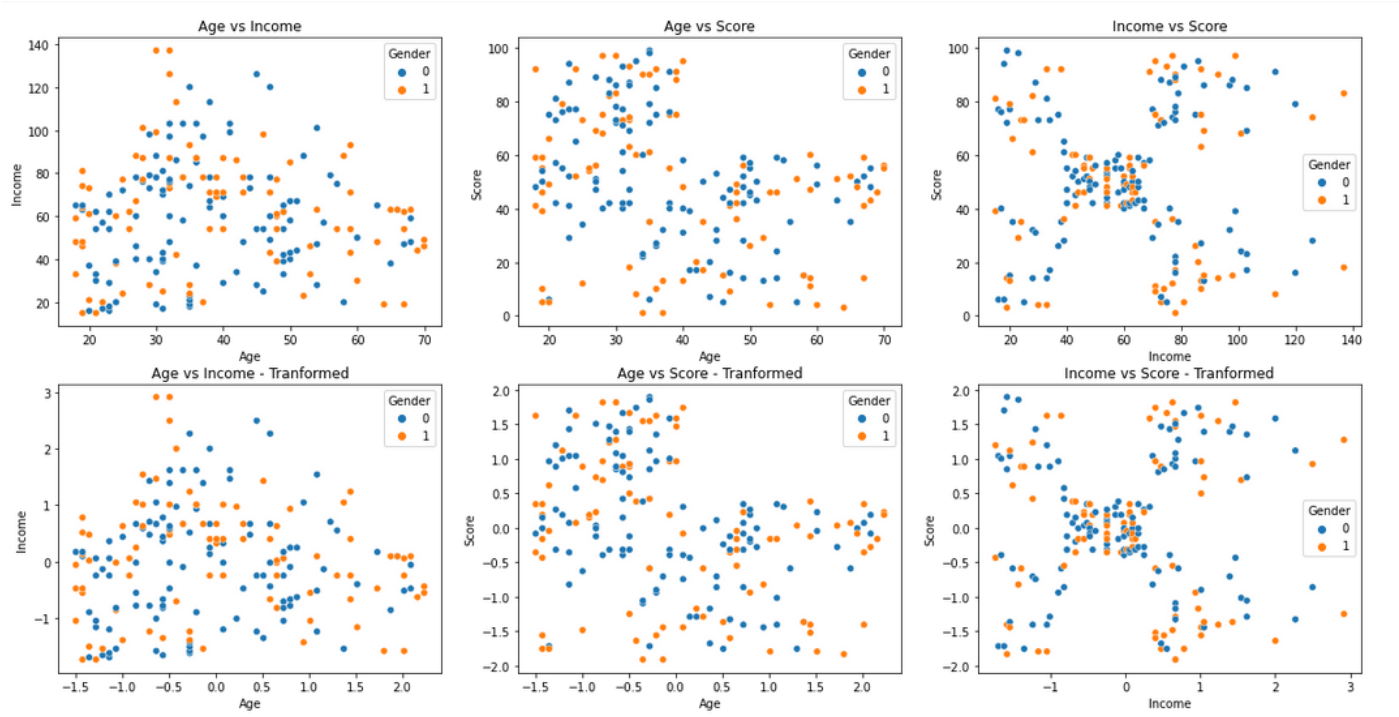
You may notice that income seems to be highest for age group 25-50 comparing to others and similarly, score is higher for group of people in age 20-40 comparing to others.

Data preparation

Gender column will be **one-hot encoded** into 0/1 values, new data frame will be created as well having age, income and score normalized. This is done just to compare if there will be any difference in elbow method, however all columns except age has pretty much same values so we should not see anything extreme differences here (values in range 0-100).

Transformed vs Original

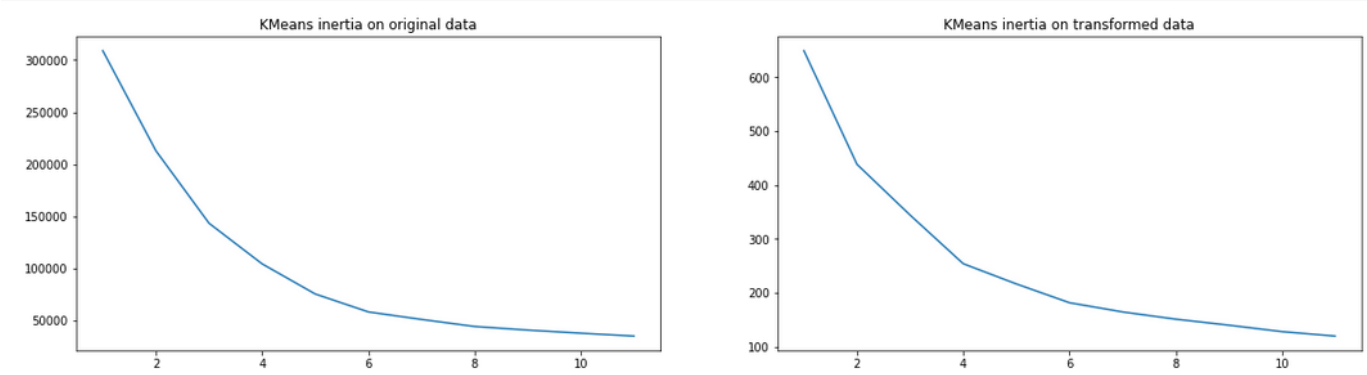
We have found possible clustering on chart income vs score and age vs score. Just quickly check how it looks like on normalized and raw data. No difference to report



Clustering using KMeans

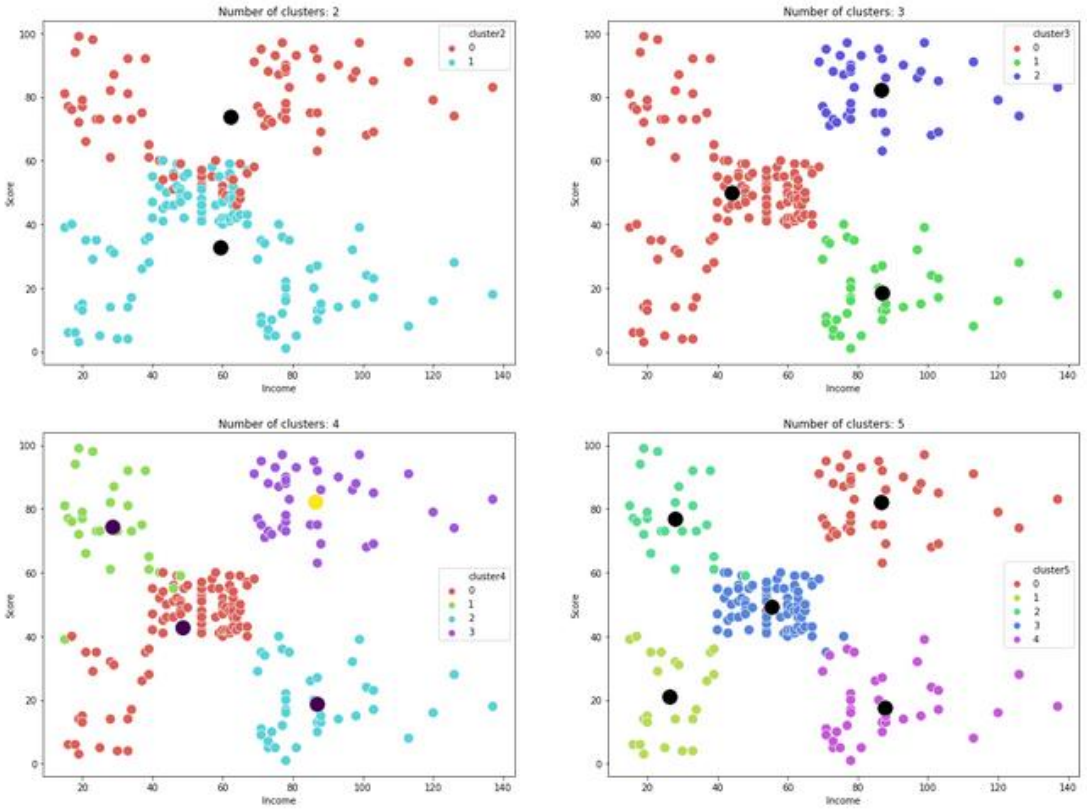
First, we will try to find ideal number of clusters for KMeans using elbow method, where we draw inertia for number of clusters (usually) in range 1-12 and try to find ones with highest gap angle. Then we will use these values to draw clusters and decide which one is most suitable for us.

We draw inertia of KMeans on raw as well as normalized data, just to see if it makes any difference.



Elbow results

When looking on inertia for original data, 3 and 5 seems to be our candidates for number of clusters. When looking on inertia in transformed data, 2 and 4 seems to be best. We simply check how clustering looks like when using 2, 3, 4 and 5 clusters.



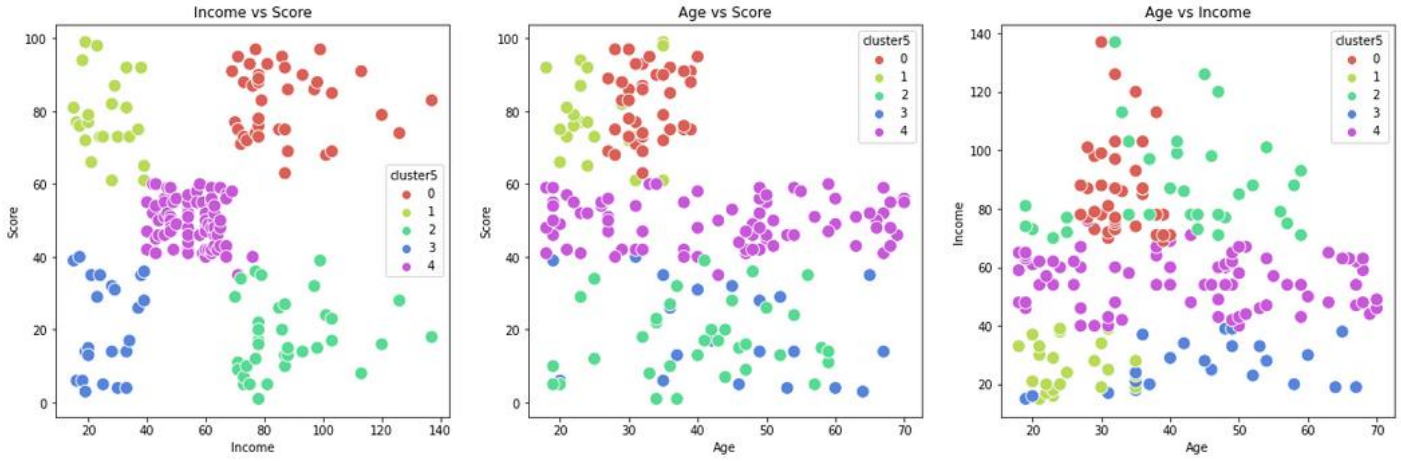
Key Findings:

Based on business needs, we should choose one that we can describe bests and will be useful for our business.

I think 2 is not enough clusters and just divide customers into 2 groups - score under 50 and score over 50 so does not look well.

Using **5 clusters**, we are getting 5 different groups of customers that separates well from each other and we could run different campaigns on each customer group.

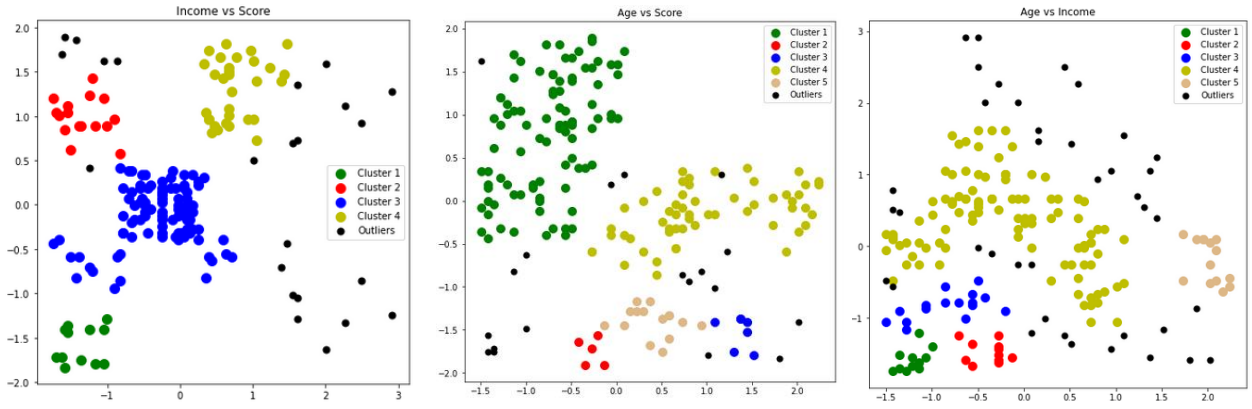
Final Clustering with 5 Clusters



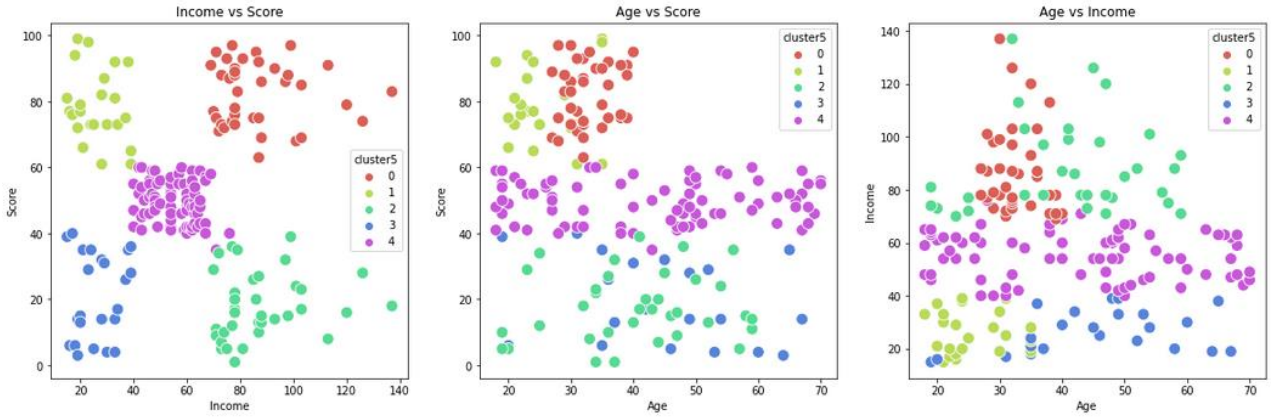
Customer segmentation with DBCSAN

Let’s see how DBScan cluster our Data. We had to tune the Hyperparameters to get good results.

We've tried range of eps but selected 0.09 which gives adequate number of clusters. By the way, default value is 0.5 but here we have points with higher density. As for the min_samples, it seems logical to use approximately 10 or more as totally there are 200 customers in the dataset, however it leads to huge amount of points considered as outliers and inadequate result. We decreased it slightly and decided to use the value of 5.



Compared with K-Means Clustering:



Conclusion Using K-Mean and DBScan Clustering :

With K-Mean we have selected 5 clusters, meaning 5 customer groups. Let's try to describe them for marketing team:

- **Poor and not-spender** - customers with low income and low spending score (cluster #4)
- **Poor and spender** - customers with low income, but spending a lot (cluster #1)
- **Neutral** - customers with mid income and mid spending score (cluster #0)
- **Rich and not-spender** - customers with high income and low spending score (cluster #2)
- **Rich and spender** - customers with high income and high spending score (cluster #3)

With DBScan :

DBScan successfully **found 5 clusters** that show the spending score of customers depending on their age or annual income. Having the results of two algorithms it looks like K-means performs better for this need than DBSCAN in this task, it shows many outliers that should rather be interpreted as actual customers, but this is how the algorithm works.

NEXT STEPS or FUTURE IMPROVEMENTS

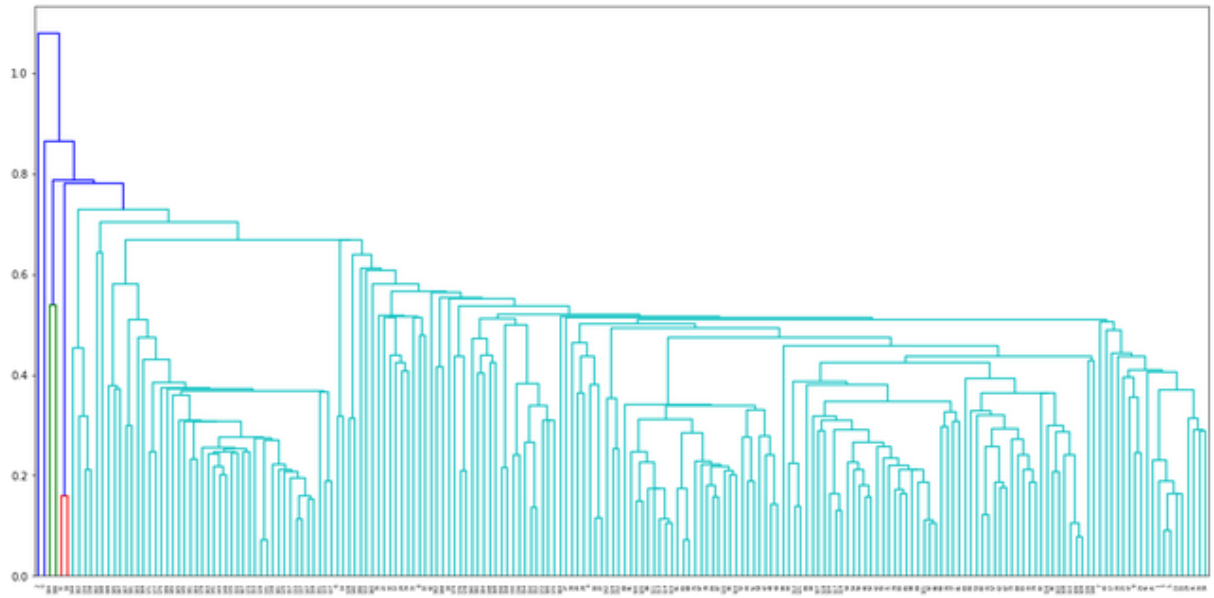
With this small data sat we reached a good clustering performance with K-Means Model and Hierarchical Clustering. We had a good customer segmentation and a clear data clustering.

We should, of course, need more observations to make, hopefully, data set density more uniform to give a chance to DBScan Model that usually performs better than K-Means in this environment.

We should also wait the results from a Marketing Campaign based on our suggestions from the clusters analysis.

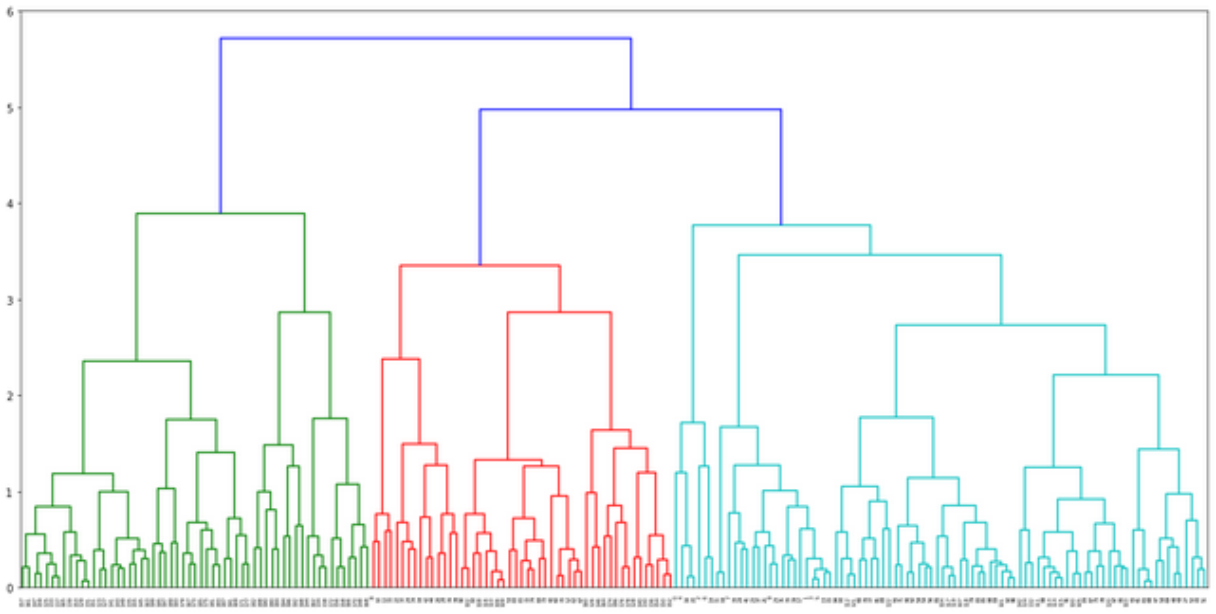
Hierarchical Clustering

Dendrogram with single Linkage method:



The Elbow curve suggests proceeding with 4 Clusters.

Dendogram with Complete Linkage method:



We will opt for 4 as cluster.

With the following customers for each cluster:

Out[40]:

```
3    20
2    39
1    51
0    90
Name: Cluster_Id, dtype: int64
```

And re-grouping with mean values we find :

	Age	Annual Income (k$)	Spending Score (1-100)
Cluster_Id			
0	32.466667	45.588889	54.488889
1	56.372549	53.960784	31.588235
2	32.692308	86.538462	82.128205
3	34.900000	94.100000	16.100000

Cluster 0 are those people whose

- Avg Age: 32
- Avg Annual Income (k$) : 45.5k
- Avg Spending Score (1-100): 54

We can label them **Medium Spender**

Cluster 1 are those people whose

- Avg Age: 56
- Avg Annual Income (k$) : 54 k
- Avg Spending Score (1-100): 31

We can label them **Low Spender**

Cluster 2 are those people whose

- Avg Age: 32
- Avg Annual Income (k$): 86 k
- Avg Spending Score (1-100): 82

We can label them **Extra Spender**

Cluster 3 are those people whose

- Avg Age: 35
- Avg Annual Income (k$) : 94 k
- Avg Spending Score (1-100): 16

We can label them **Very Low Spender**

Conclusion with HC:

Suggestions for Marketing Team:

- Target Cluster 1 with more offers
- Reward Cluster 2 people for being loyal customer.
- Improve the services to attract Cluster 3
- Target Cluster 0 with better employees' support

Final Conclusion

K-Means clustering performs better than **DBScan** providing a better and clearer **Cluster Separation.** K-Means can intercept more customer within Cluster while for DBScan they are outliers. The reason is the non-uniform data set density.

Hierarchical Clustering is an excellent way to visualize and it gives an excellent insight regarding the meaning of clusters. You can find all suggestions for the Marketing Team inside this work following each model.