###

FACULTY OF ENGINEERING AND NATURAL SCIENCES

COMPUTER ENGINEERING DEPARTMENT

2022-2023 FALL

PROJECT REPORT

STUDENT ID: ####

NAME LASTNAME: SEDA NUR KILIÇ

PROJECT TITLE: MALL CUSTOMERS SEGMENTATION

INTRODUCTION

This project aims to segment customers of a mall based on their shopping habits and demographics. The goal is to create customer segmentation and analysis for a retail store within a mall, further create an unsupervised machine learning model such as K-Means Clustering to provide insights to the marketing team.

MATERIALS AND METHODS

Materials:

1) Customer data, including demographic information (age, gender, income, etc.) and shopping habits (frequency of visits, average spend, etc.)
   Dataset : https://www.kaggle.com/datasets/shwetabh123/mall-customers
2) Software for data analysis and visualization, such as Python, Rapidminer
3) Libraries and packages for implementing the k-means algorithm

Methods:

1) Perform exploratory data analysis to identify patterns and relationships in the data.
2) Use the k-means algorithm to cluster the customers into distinct groups based on their characteristics.
3) Evaluate the performance of the clustering algorithm and fine-tune the model as needed.
4) Visualize the results of the clustering to understand the characteristics of each segment.
5) Use the segmented customer data to inform targeted marketing and promotional efforts.

K-means is a popular and widely-used clustering algorithm that groups similar data points together into clusters. The basic idea behind the algorithm is to define spherical clusters such that the points within a cluster are as close as possible to the centroid of the cluster, and as far as possible from the centroid of the other clusters.

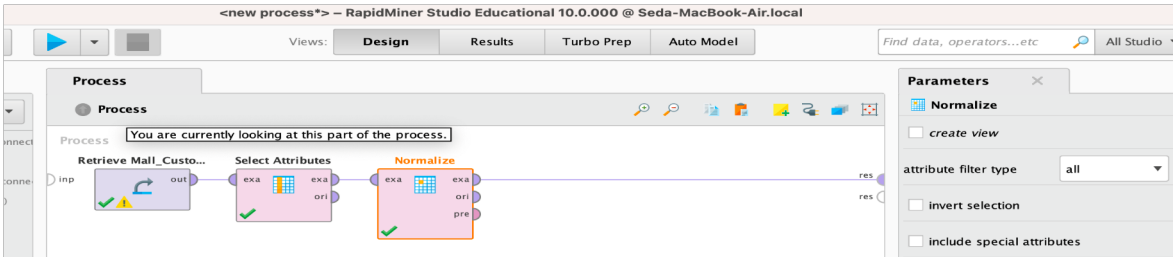
The algorithm works as follows:

1) Initialize K centroids, where K is the number of clusters you want to create. These centroids can be chosen randomly from the data points.
2) Assign each data point to the cluster whose centroid is closest to it.
3) Recalculate the centroid of each cluster by taking the mean of all data points assigned to that cluster.
4) Repeat steps 2 and 3 until the assignments of data points to clusters no longer change or a maximum number of iterations is reached.

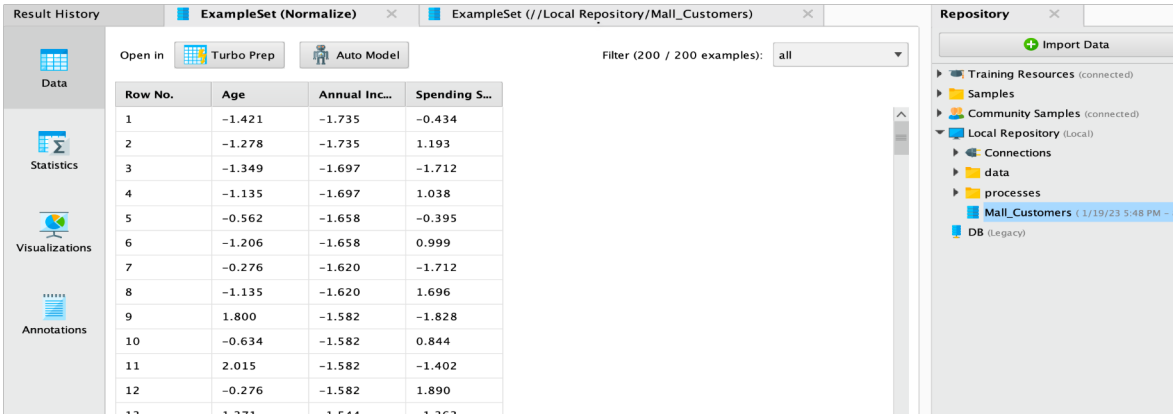
The main advantage of k-means is that it is simple and easy to understand, and it can be applied to a wide range of data types. However, it has some limitations as well. One of the main limitations is that it assumes that the clusters are spherical in shape and equally sized, which is not always the case in real-world data. Additionally, it is sensitive to the initial

placement of centroids, which can lead to different results depending on the starting point. In addition to these, it also requires to specify the number of clusters in advance which may be difficult sometimes, and it is also sensitive to the scale of the data and the presence of outliers.

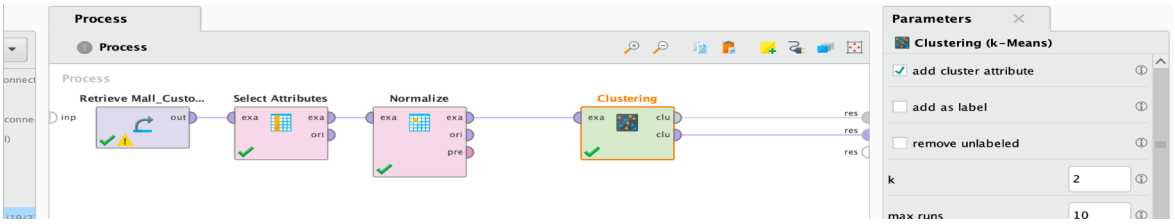
TOOL



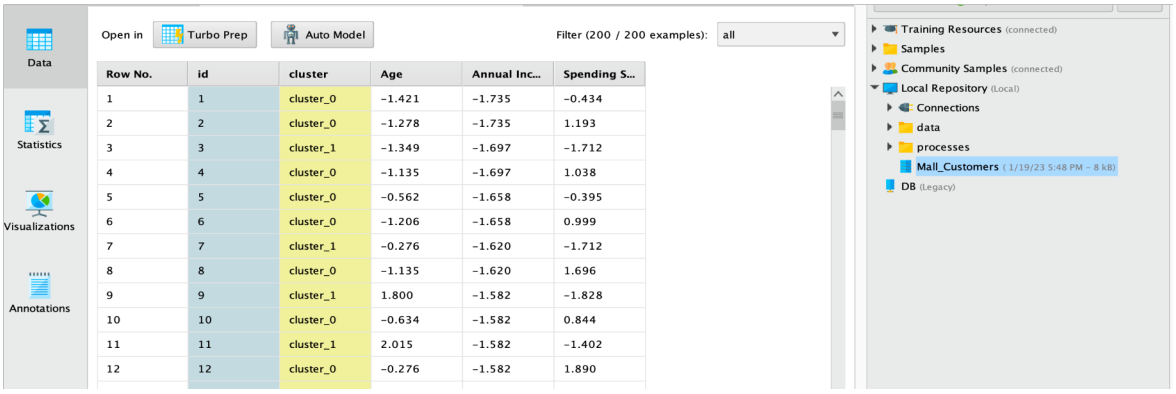
There are several different ways to normalize. We have used the z-transform in this example.



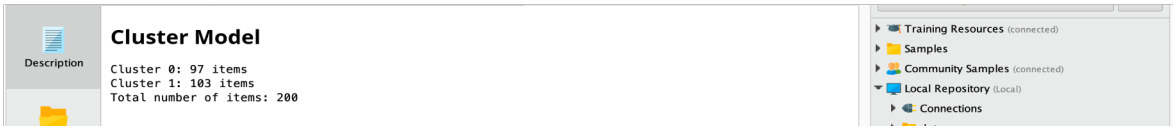
Each feature of the customer you don't see the absolute values anymore but just it just the standardized score. For example this first customer needs one minus one point four standard deviations so this is a really young customer.



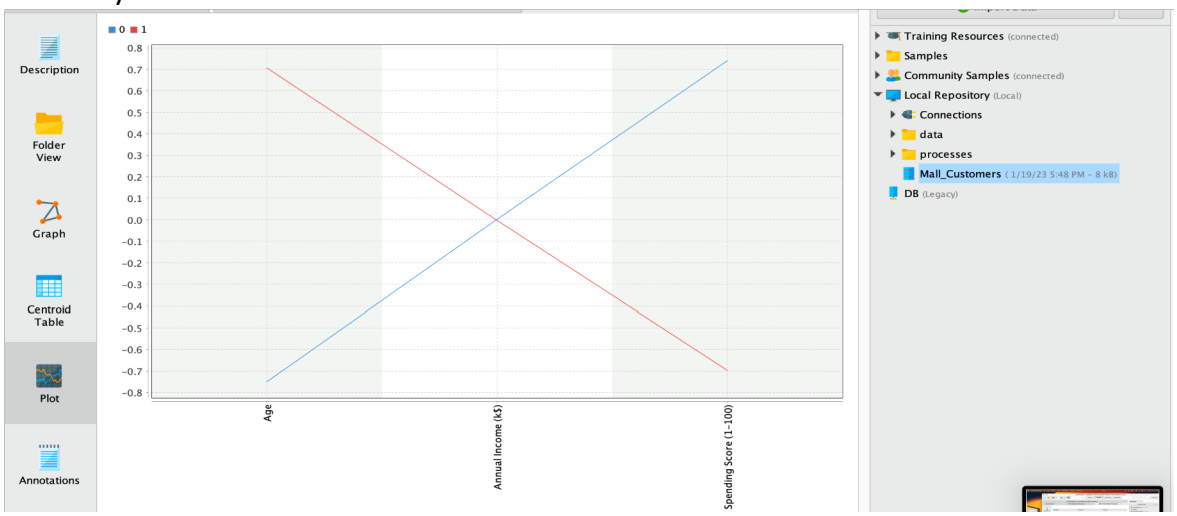
Put k equal to 2 because that is the smallest value that I could have if I'm trying to partition a dataset.

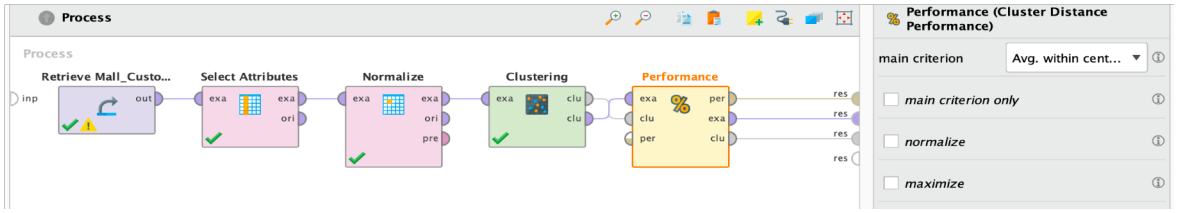


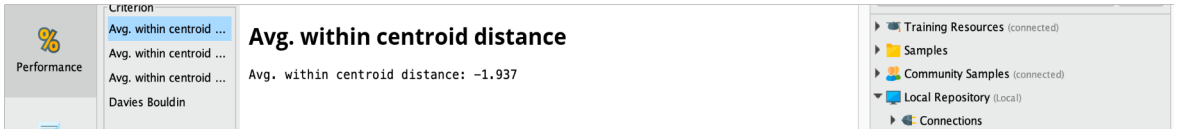
So from one of those ports you can see you have the three properties for each customer and for each customer you also know which cluster they were assigned to so for example customer 3 was assigned to cluster 1.



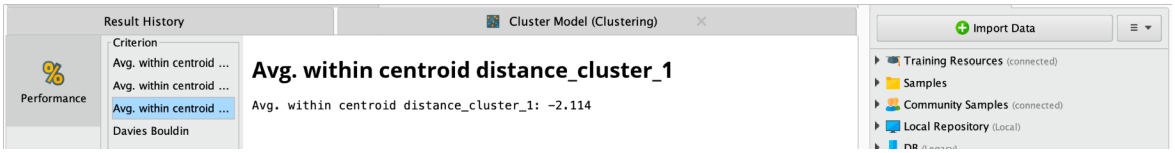
Summary data about the clusters.





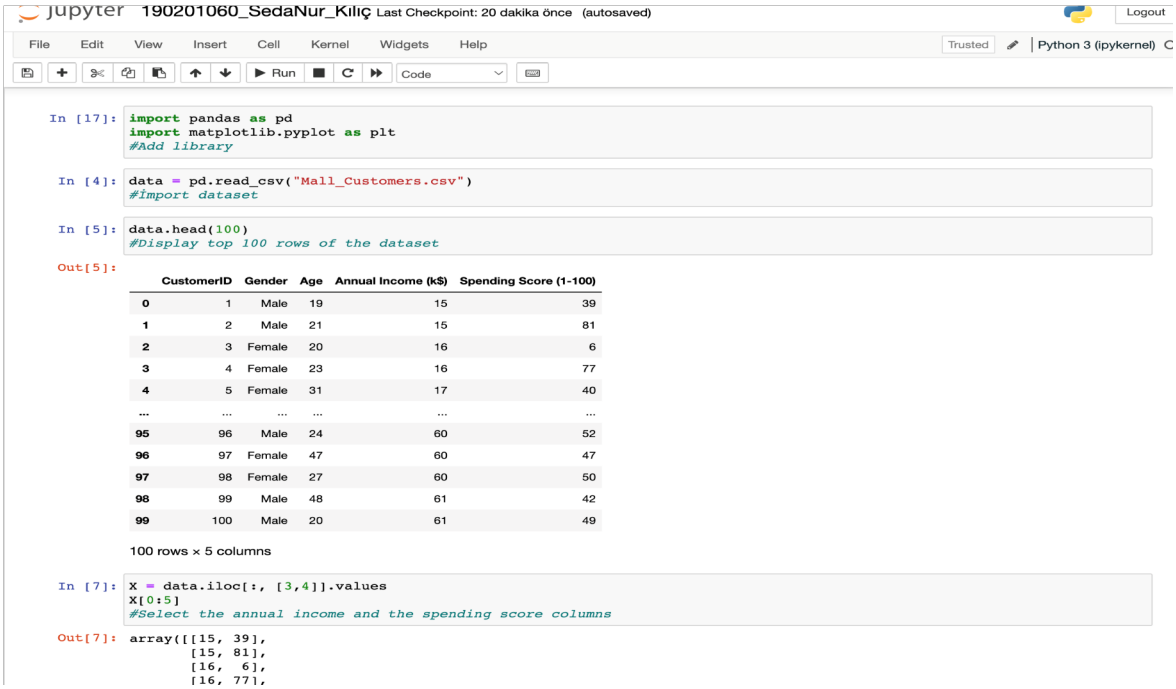


The first number shows average within centroid distance.

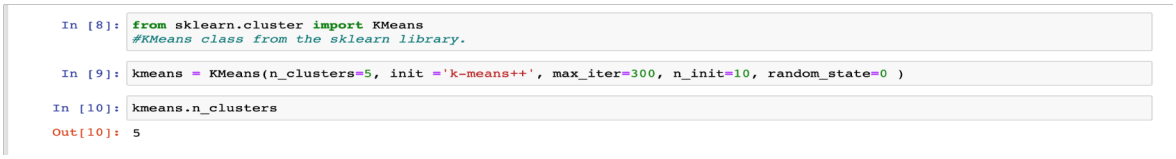


You can see that the average distance for the second cluster is minus 2.1 the average distance within the first cluster is minus 1.7. So this first average distance is simply the weighted average of those two alright.

CODES



Elbow method is one of the used to find out the optimal number of clusters. In this method, the sum of distances of observations from their cluster centroids, called Within Cluster Sum of Squares (WCSS). This is computed as; Yi is centroid for observation Xi. The below part show this method.:



We are going to use the fit predict method that returns for each observation which cluster it belongs to. The cluster to which client belongs and it will return this cluster numbers into a single vector that is called y_kmeans.

```
In [11]: y_kmeans = kmeans.fit_predict(X)

In [12]: type(y_kmeans)
Out[12]: numpy.ndarray

In [14]: y_kmeans
Out[14]: array([4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3,
                4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 1,
                4, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
                1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
                1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
                1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 0, 2, 1, 2, 0, 2, 0, 2,
                1, 2, 0, 2, 0, 2, 0, 2, 0, 2, 1, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2,
                0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2,
                0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2,
                0, 2], dtype=int32)

In [15]: data['cluster'] = y_kmeans
         data.head()
```

Out[15]:

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)	cluster
0	1	Male	19	15	39	4
1	2	Male	21	15	81	3
2	3	Female	20	16	6	4
3	4	Female	23	16	77	3
4	5	Female	31	17	40	4

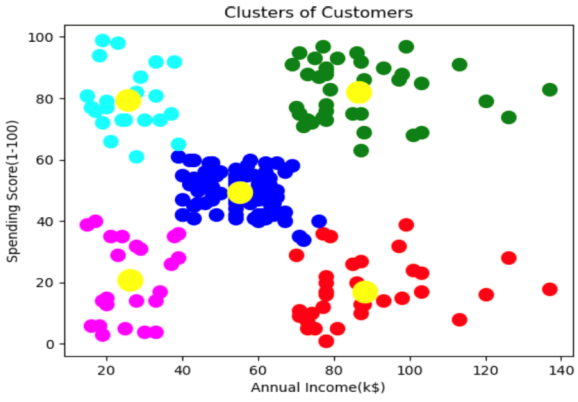
The operations performed in the comment lines are also explained.

RESULTS

In last step, we plot according to their annual income and spending scores.

```
In [18]: plt.scatter(X[y_kmeans==0, 0], X[y_kmeans==0, 1], s=100, c='red', label ='Cluster 1')
         plt.scatter(X[y_kmeans==1, 0], X[y_kmeans==1, 1], s=100, c='blue', label ='Cluster 2')
         plt.scatter(X[y_kmeans==2, 0], X[y_kmeans==2, 1], s=100, c='green', label ='Cluster 3')
         plt.scatter(X[y_kmeans==3, 0], X[y_kmeans==3, 1], s=100, c='cyan', label ='Cluster 4')
         plt.scatter(X[y_kmeans==4, 0], X[y_kmeans==4, 1], s=100, c='magenta', label ='Cluster 5')
         #Visualising the clusters

         plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1], s=300, c='yellow', label = 'Centroids')
         plt.title('Clusters of Customers')
         plt.xlabel('Annual Income(k$)')
         plt.ylabel('Spending Score(1-100)')
         plt.show()
         #Plot the centroid. This time we're going to use the cluster centres
         #attribute that returns here the coordinates of the centroid.
```



DISCUSSION AND CONCLUSIONS

The results of this mall customer segmentation project can be used to better understand the characteristics of different groups of customers and tailor marketing and promotional efforts to each group. For example, if a segment of customers is found to have a high income and a preference for luxury brands, targeted promotions for high-end products and services may be more effective for this group.

The k-means algorithm was used to cluster customers based on their demographics and shopping habits. The performance of the algorithm was evaluated and fine-tuned as

needed. The results of the clustering were then visualized to understand the characteristics of each segment.

It is important to note that while the k-means algorithm is a commonly-used clustering method, it has its limitations. For example, it assumes that clusters are spherical and equally sized, which may not be the case in all data sets. Additionally, the algorithm is sensitive to the initial placement of centroids and the scale of the data, which can affect the results. Therefore, it is important to consider these limitations when interpreting the results of the analysis.

In conclusion, this mall customer segmentation project used data mining techniques to identify distinct groups of customers with similar characteristics. The results of the analysis can be used to inform targeted marketing and promotional efforts and improve the overall performance of the mall. However, it is important to consider the limitations of the k-means algorithm and the specific characteristics of the data when interpreting the results.

REFERENCES

1. "Customer segmentation in retail industry: a case study of a shopping center" by J. Kukkonen and K. Kukkonen (2015). This study used cluster analysis to segment customers of a shopping center in Finland based on their shopping behavior and demographic characteristics.
2. "Customer segmentation using RFM analysis: a case study of a department store" by S. C. Hsu and C. C. Chiu (2006). This study used RFM (recency, frequency, monetary) analysis to segment customers of a department store in Taiwan and found that the segments had different purchasing behaviors.
3. "A Study on Customer Segmentation for Shopping Mall" by B. H. Kim and J. W. Lee (2011). This study used cluster analysis to segment customers of a shopping mall in South Korea based on their demographics, shopping behaviors, and brand preferences.
4. "Customer Segmentation Based on Shopping Behaviour in a Supermarket" by P. G. Teixeira, M. J. Ramos, and P. R. S. Castro (2015). This study used cluster analysis to segment customers of a supermarket based on their shopping behaviors and found that the segments had different purchasing patterns.
5. "Consumer Segmentation in Retail Industry: A Study of Shopping Center Customers" by H. P. Tan, C. K. K. Soon, and K. K. Lai (2009). This study used cluster analysis to segment customers of a shopping center in Malaysia based on their demographics, shopping behaviors, and brand preferences.