

✓ MALL CUSTOMERS CLUSTERING



Introduction In Previous Lesson, we did Exploratory Data Analysis(EDA) on Bank data.

In This Part we will do Data Clustering, Clustering is Unsupervised Machine Learning Technique. This means That No testing of our ML Models, Unlike Classification/Regression where models are tested. We will be looking at Mall Customers Data.

What is Clustering? - Clustering is an unsupervised machine learning method of identifying and grouping similar data. Data is placed into Groups based on Similarities An Example - A bank might cluster all members who make deposits of 10000 on average in one group and Cluster those who deposit 500 on average in another group. This will help the bank understand ts customers better an give them different offers based on the Groups.

Clustering is also known as Segmentation.

Clustering is a Unsupervised Learning where we ONLY Train the Model, No Testing of the Model

In this example we see how a supermarket can Group its Customers, this will allow the bank target specific group with different Easter Offers.

This is a Customer Segmentation Problem.

Step 1: Read Data

Below we read the data, it has columns such Gender, Age, Customer Salary/Income, Spending Score Out of 100. (0 - Bad and 100 - Good)

```
# Regression and Classification are Supervised Learning - There is model testing
# Clustering - is Unsupervised Learning - No model testing.
# Clustering Gruping data to different groups called clusters
# Each of These clusters has similarities, something in common.
# Applications: Customer Segmentation, Data Segmentation
import pandas
data = pandas.read_csv("https://coding.co.ke/datasets/Mall_Customers.csv")
data.head()
# Clustering Numerical data - Kmeans
```

	CustomerID	Gender	Age	Annual Income	Spending Score
0	1	Male	19	15.0	39.0
1	2	Male	21	15.0	81.0
2	3	Female	20	16.0	6.0
3	4	Female	23	16.0	77.0

Step 2 : Split Data

Below we split data into the variable we are interested in our clustering, Here we want to cluster Mall Customer by Age, Income and Spending Score. Age will help us know the demographics, Income will help understand the Customer shopping capabilities and Spending Score will assist in knowing their previous shopping habits. Hence we Only Take from Row 2 to 4: that is 2:5 (Recall minus one).

Please note that Y is missing, Only X is available, hence Unsupervised Learning

```
# drop all empties, if present
data.dropna(inplace=True)
array = data.values
X = array[:, 2:5] # Age, Income, Score
# No Variable Y
```

Step 3: Fit Data to Model

Below we Fit X to Kmeans Model What is Kmeans - K-means is a data clustering approach for unsupervised machine learning that can separate unlabeled data into a predetermined number of disjoint groups of with similarities.

Please Note we specify 10 as our Number of Clusters.

You can Specify any Number of Clusters, The more Clusters the Better the Groups will Contain Good Selection. Random State helps in Randomizing the data During Grouping.

At this Point Kmeans Creates all the Clusters.

To import KMEANS we need to Scikit Learn, Open <https://scikit-learn.org/stable/> Check that Clustering is one of the Items on this website.

Sklearn is a Python Machine Learning Library.

```
from sklearn.cluster import KMeans
model = KMeans(n_clusters=10, random_state=42)
model.fit(X)

/usr/local/lib/python3.9/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning:
  warnings.warn(
  KMeans
  KMeans(n_clusters=10, random_state=42)
```

Step 4: Find Clusters

Next, We Find the Clusters that K means Created - 10 Clusters. K Means shows Each Clusters and Its Means/Average For Age, Income, Spending Score. The Average Per Cluster Helps you know that the records in That Group in close to the means shown.

```
# Lets see the 10 Clusters Kmeans created.
means = model.cluster_centers_
clusters = pandas.DataFrame(means, columns = ['Age', 'Income', 'Score'])
clusters
```

	Age	Income	Score
0	61.500000	51.230769	50.230769
1	25.103448	56.068966	50.758621
2	43.750000	106.500000	19.875000
3	25.272727	25.727273	79.363636
4	48.750000	24.583333	9.583333
5	32.900000	108.900000	84.200000
6	46.500000	61.625000	46.000000
7	38.230769	30.384615	35.076923
8	32.740741	78.037037	81.592593
9	39.190476	79.190476	13.190476

Step 5: Explain Clusters

Above we see 10 Clusters, We can Extract Any 3 Clusters that are Notable.

Group/Cluster 2: 43.750000 106.500000 19.875000

We can see the Group has members who in their 40s, Earning Very well Upto 106K Per Month, But We Notice that there Shopping Score is Low, Meaning they do not Shop a lot. The Supermarket needs to Give them best Experience and Motivate them improve their Shopping Since they have Good Earnings. Also this might be the same for Group 9.

Group/Cluster 5: 32.900000 108.900000 84.200000

This is an Interesting Cluster, We see it has young people in early 30s Earning upto 108K a month and Shopping Experience is Great at 84/100, The supermarket should award, give offers to this Group in order to maintain them , as they seem to be bringing more profit.

Group/Cluster 4: 48.750000 24.583333 9.583333

This Group shows that the members are in their 40s, These might be Parents, who earn 24k a month and their shopping Score is as low as 9/100. The supermarkets might come up with cheap packages of Food Stuffs that they can target this Group, which will motivate them to Buy.

Group/Cluster 3: 25.272727 25.727273 79.363636

Lets Look at this Group, They are Young at 25 Years, Earning arund 25k a month, But they seem to have a Good shopping Score. They might be those who completed College and Started working, They are buying Stuff to start off, The supermarket must encourage them by doing good Cheap Packages, Bonus Points, That they can redeem after some period of shopping, Also sell items that target the youth.

What do you see in Group 8?


Step 6: Get Specific Cluster Data

In this Step, All Records were Put into either Group 0 to Group 9. SO, we do model.labels_ to Get each Record and the Group No it was allocated to. In the Next Step we show all records in Cluster 5.

Please observe that they are within the means below For Age , Income, Score respectively

32.900000 108.900000 84.200000

```
# Let see who is in group 2
data['no'] = model.labels_ # get every record group number
cluster5 = data[data['no'] == 5]
cluster5
```



	CustomerID	Gender	Age	Annual Income	Spending Score	no
179	180	Male	35	93.0	90.0	5
181	182	Female	32	97.0	86.0	5
183	184	Female	29	98.0	88.0	5
185	186	Male	30	99.0	97.0	5
189	190	Female	36	103.0	85.0	5
191	192	Female	32	103.0	69.0	5
193	194	Female	38	113.0	91.0	5
195	196	Female	35	120.0	79.0	5
197	198	Male	32	126.0	74.0	5
199	200	Male	30	137.0	83.0	5

Above can be Done to other Clusters too.

Step 7: Put Specific Cluster Data in CSV

Below we get Cluster 5 Members in their own CSV, This will help us get the members of this Group in a Single CSV. This CSV can now be Forwarded to Marketing Team to Make more recomendations based on this Group.

The CSV is Saved in Your Current Folder.

```
data[data['no'] == 5].to_csv('Cluster5.csv')
```

