

nrcm-hierarchical-clustering-2

August 28, 2023

1 R DURGAPRASAD GOUD

2 21XO5A6744

#CSE(DATASCIENCE)

3 project Title

analysis and prediction of “Mall_customers” of csv file of american mall market called as u requirment of dendrogram using using scipy graphic library with the help of double codes spicity.cluister.hierararchy

3.1 Problem statement:-

The American Finance Market Clients as per the rate of GDP of 2011 who has highest no_of growth in their business market.

As a data science engineer find out which hierachy cluster use maximum linkage in upcoming future #####TASK-1: ##### Import the library and dataset #####TASK-2: ##### Using the dendrogram to find theoptimal no_of clusters #####TASK-3: ##### Create a hierachy model and visualize the clusterwith help of matplotlib.lib library

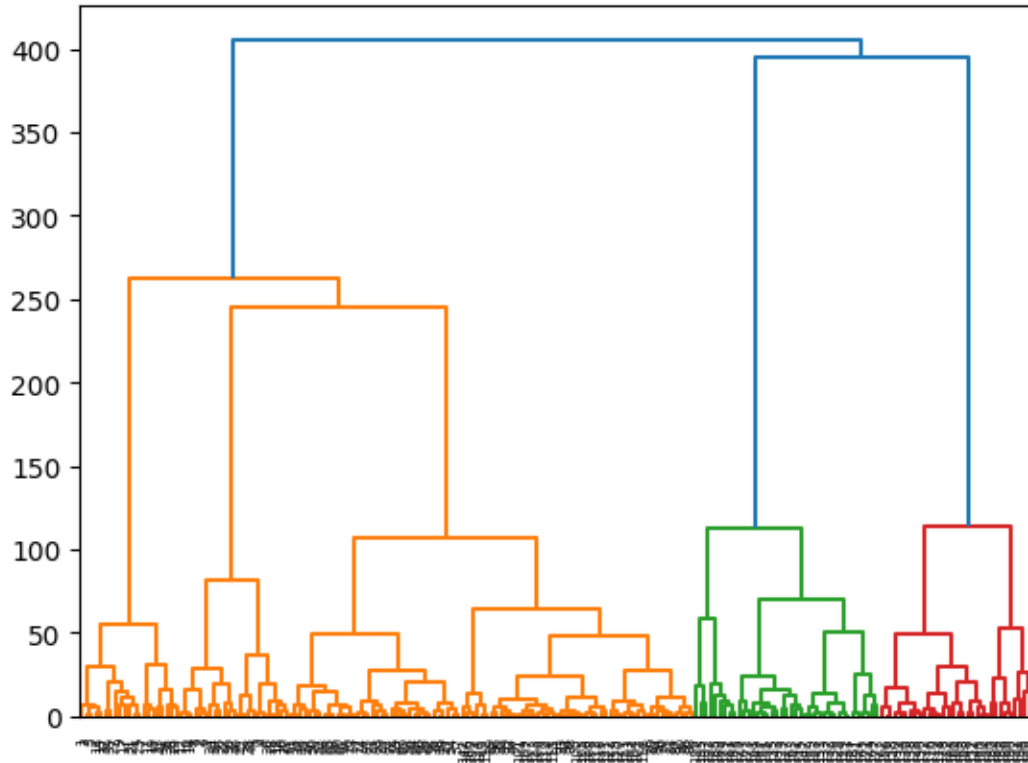
```
[7]: #Import the numpy, pandas , matplotlib, seaborn libery's
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
[8]: #Assign variable name "dataset" and the input variable as "X" indcludong select
      ↪all the row and index columns which you want [colum_index, Column_index].
dataset=pd.read_csv("Mall_Customers.csv")
dataset

X=dataset.iloc[:,[3,4]].values
```

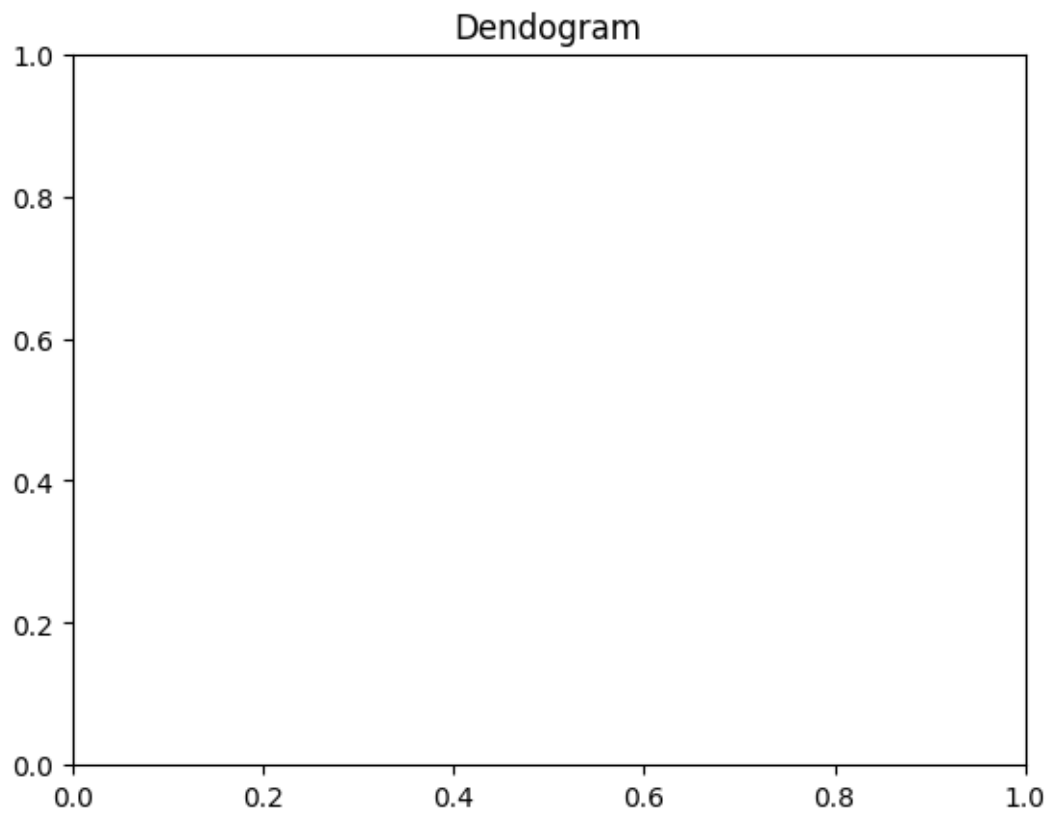
```
[9]: #import scipy cluster using attribute "scipy.cluster.hierarchy" as sch alias
from scipy.cluster import hierarchy as sch
```

```
[10]: #Using the dendrogram to find the optimal number of clusters
# Assign a variable as dendrogram and declares the "sch.dendrogram(sch.
      ↳linkage(X, method = 'ward'))"
dendrogram = sch.dendrogram(sch.linkage(X,method = 'ward'))
```



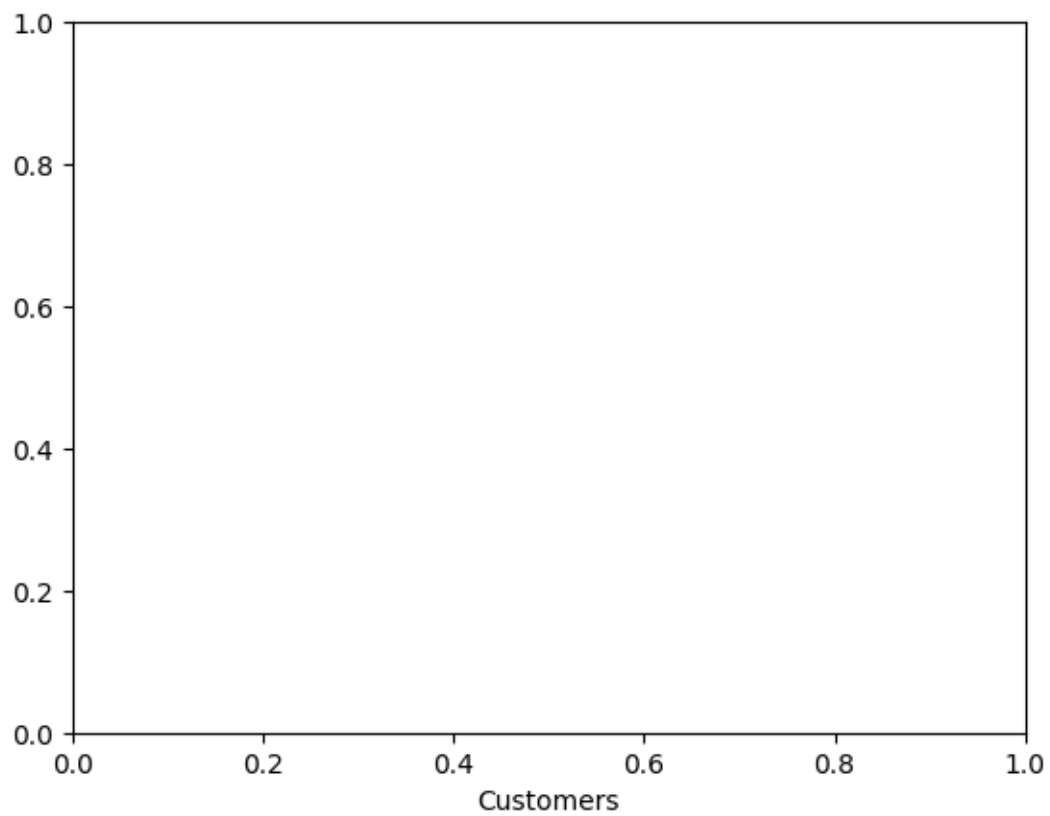
```
[11]: #Assign the title as "Dendograms"
plt.title("Dendogram")
```

```
[11]: Text(0.5, 1.0, 'Dendogram')
```



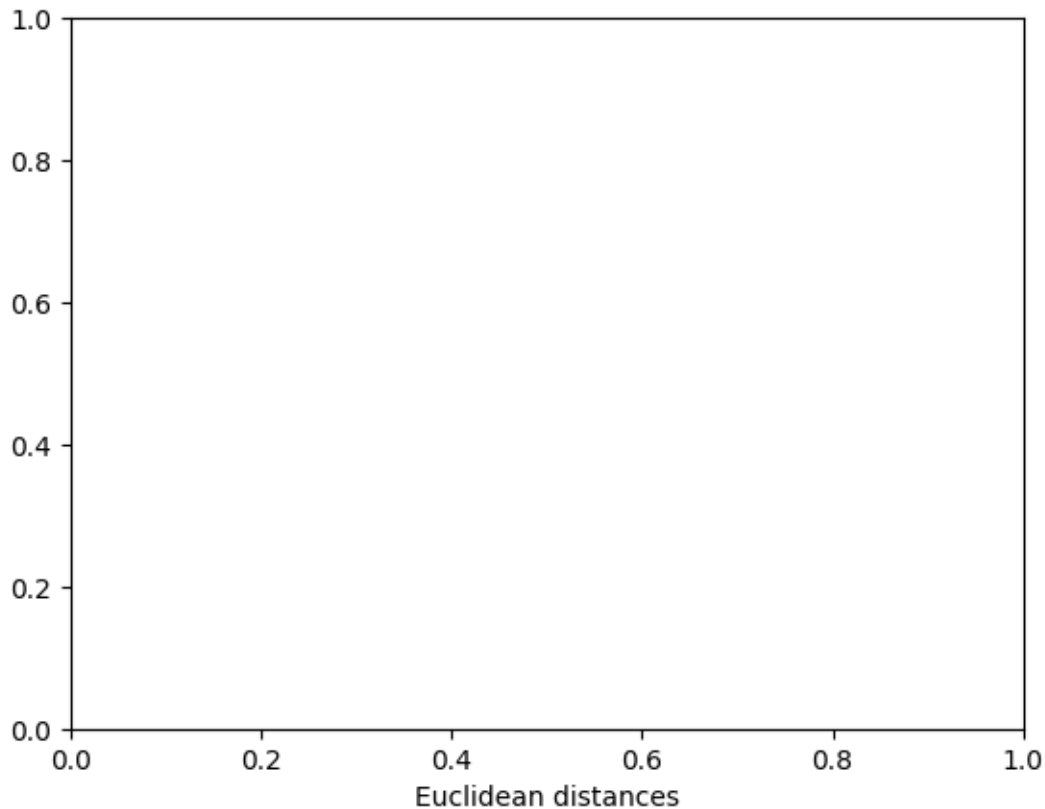
```
[12]: #Label X axis as "Customers"  
plt.xlabel("Customers")
```

```
[12]: Text(0.5, 0, 'Customers')
```



```
[13]: #Label Y axis as 'Euclidean distances'  
plt.xlabel("Euclidean distances")
```

```
[13]: Text(0.5, 0, 'Euclidean distances')
```



```
[14]: # from "sklearn.cluster" attribute import "AgglomerativeClustering" default
      ↪ argument.
```

```
from sklearn.cluster import AgglomerativeClustering
```

```
[15]: #Create a cluster for five or nth cluster which you want.
```

```
hc = AgglomerativeClustering(n_clusters = 5, affinity = 'euclidean', linkage =
      ↪ 'ward')
```

```
y_hc = hc.fit_predict(X)
```

```
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_agglomerative.py:983:
FutureWarning: Attribute `affinity` was deprecated in version 1.2 and will be
removed in 1.4. Use `metric` instead
```

```
warnings.warn(
```

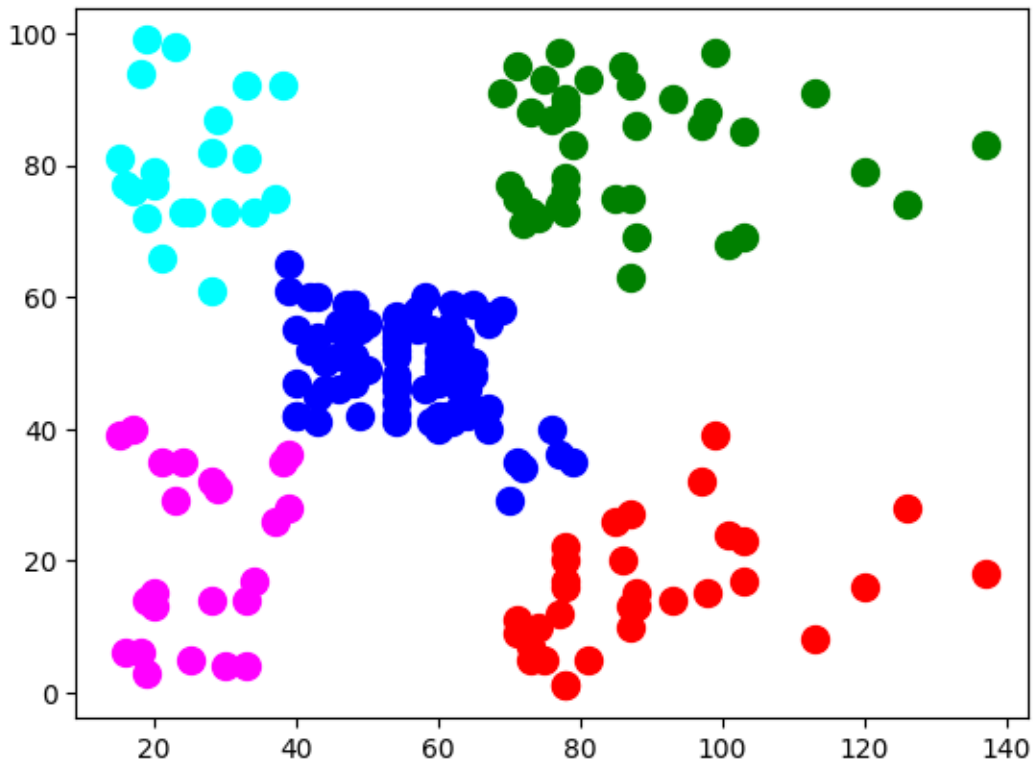
```
[16]: #Plot the scatter plot for scatter visualization.
```

```
plt.scatter(X[y_hc == 0,0], X[y_hc == 0,1], s = 100, c = 'red', label = 'Cluster
      ↪ 1')
```

```
plt.scatter(X[y_hc == 1,0], X[y_hc == 1,1 ], s = 100, c = 'blue', label =
      ↪ 'Cluster 2')
```

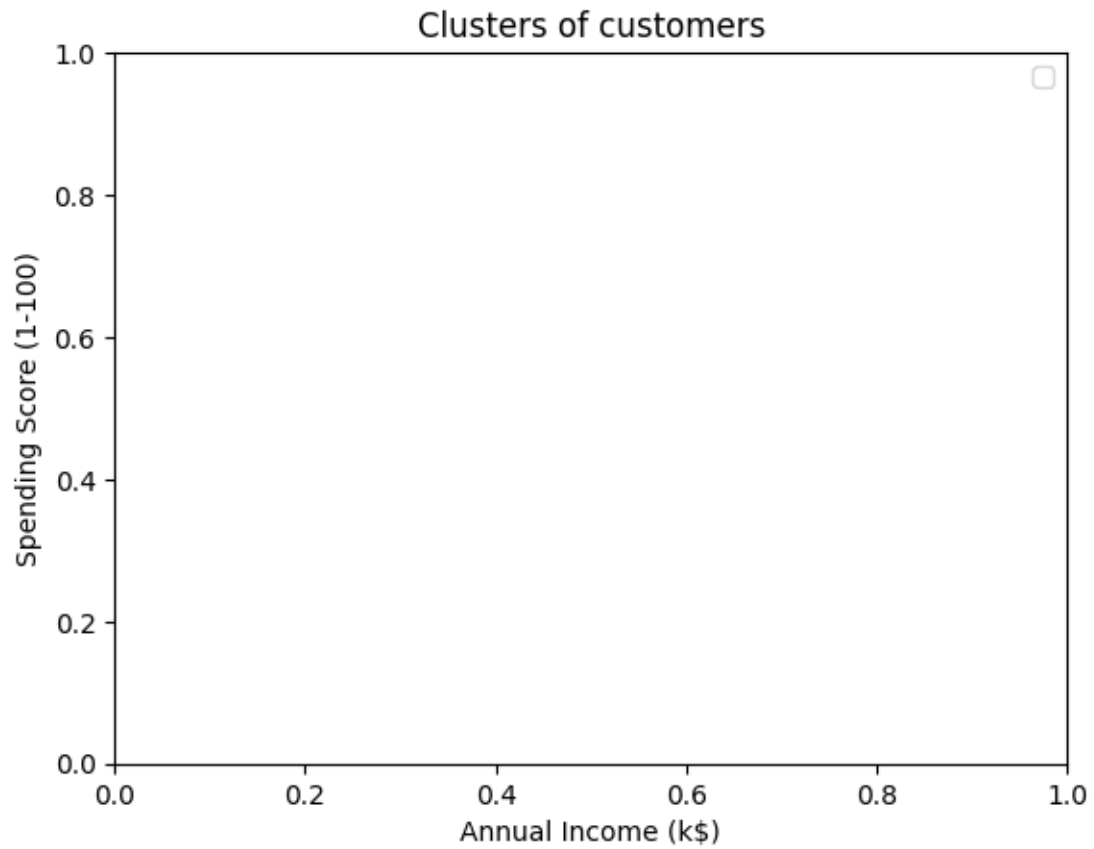
```
plt.scatter(X[y_hc == 2,0 ], X[y_hc == 2,1], s = 100, c = 'green', label =_
↳'Cluster 3')
plt.scatter(X[y_hc == 3,0], X[y_hc == 3,1], s = 100, c = 'cyan', label =_
↳'Cluster 4')
plt.scatter(X[y_hc == 4,0 ], X[y_hc == 4,1], s = 100, c = 'magenta', label =_
↳'Cluster 5')
```

[16]: <matplotlib.collections.PathCollection at 0x7b55cc111d80>



```
[17]: plt.title('Clusters of customers')
plt.xlabel('Annual Income (k$)')
plt.ylabel('Spending Score (1-100)')
plt.legend()
plt.show()
```

WARNING:matplotlib.legend:No artists with labels found to put in legend. Note that artists whose label start with an underscore are ignored when legend() is called with no argument.



4 Conclusion

according to the model building as a engineer my prediction is cluster no-3 has the highest no_of linkage