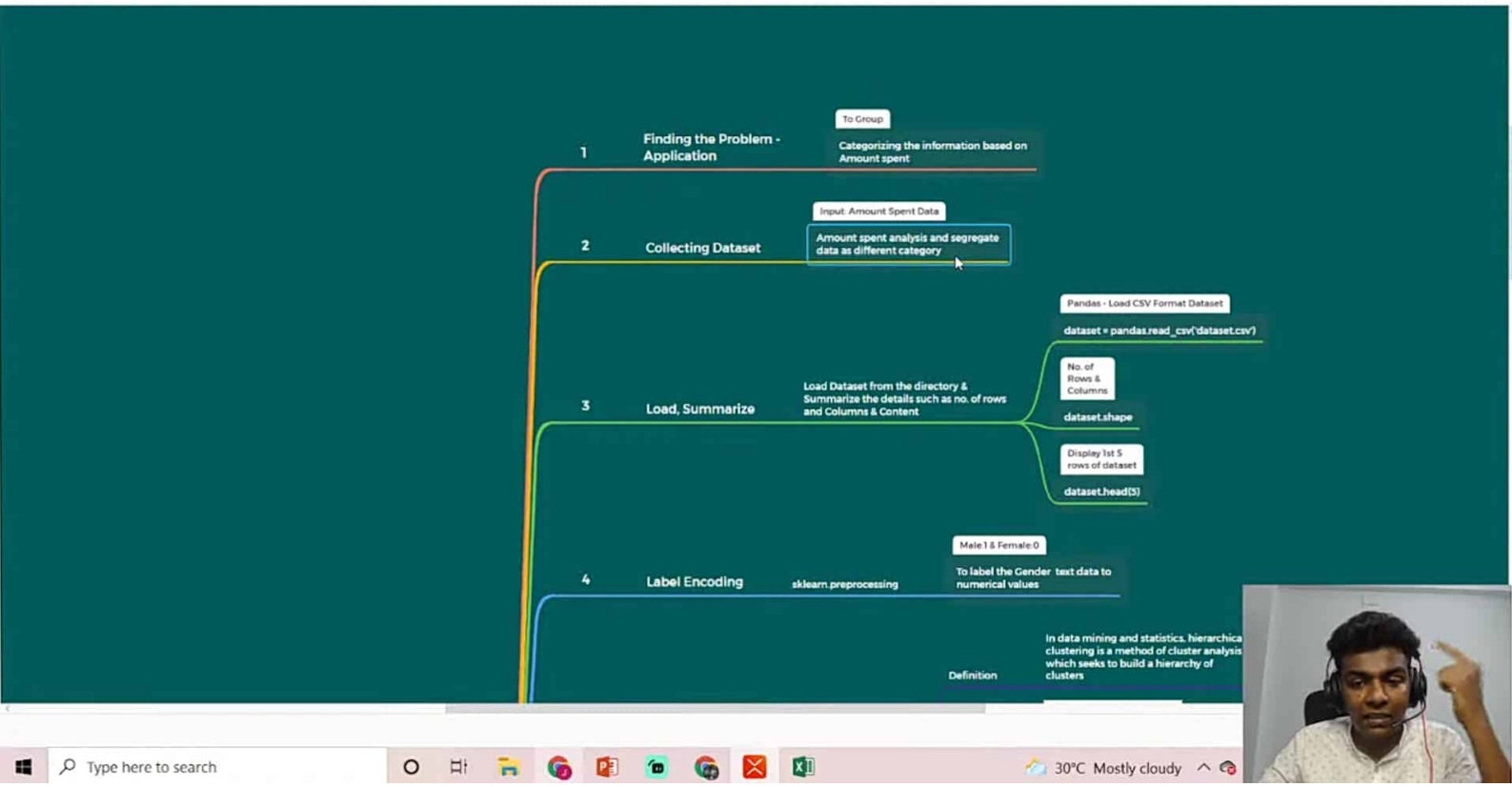


4 Label Encoding

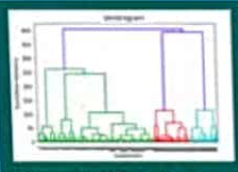
sklearn.preprocessing

Male:1 & Female:0

To label the Gender text data to numerical values

5 Algorithm

Hierarchical Clustering



Definition

In data mining and statistics, hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters

	A	B	C	D	E
B	16				
C	47	37			
D	72	57	40		
E	77	65	30	31	
F	79	66	35	23	10

Compute a distance matrix

EUCLIDEAN DISTANCE

The distance between two clusters has been computed based on the length of the straight line drawn from one cluster to another

Method

Need to determine from where distance is computed

Linkage Criteria

Two most similar parts of a cluster

single-linkage

Two least similar bits of a cluster

complete-linkage

Center of the clusters

mean or average-linkage

DEFAULT

ward

5 Algorithm

Hierarchical Clustering

Method

Linkage Criteria

complete-linkage

Center of the clusters

mean or average-linkage

DEFAULT

ward

This method works out which observations to group based on reducing the sum of squared distances of each observation from the average observation in a cluster

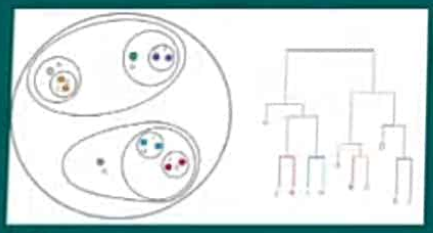
Types

Sequentially merging similar clusters

Agglomerative Hierarchical clustering

Initially grouping all the observations into one cluster, and then successively splitting these clusters

Divisive Hierarchical clustering



Steps

1 Identify the two clusters that are closes

2 Merge the two most similar clusters

Steps

1 Identify the two clusters that are closest

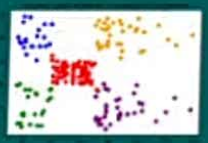
2 Merge the two most similar clusters

Hierarchical Data

Dendrogram

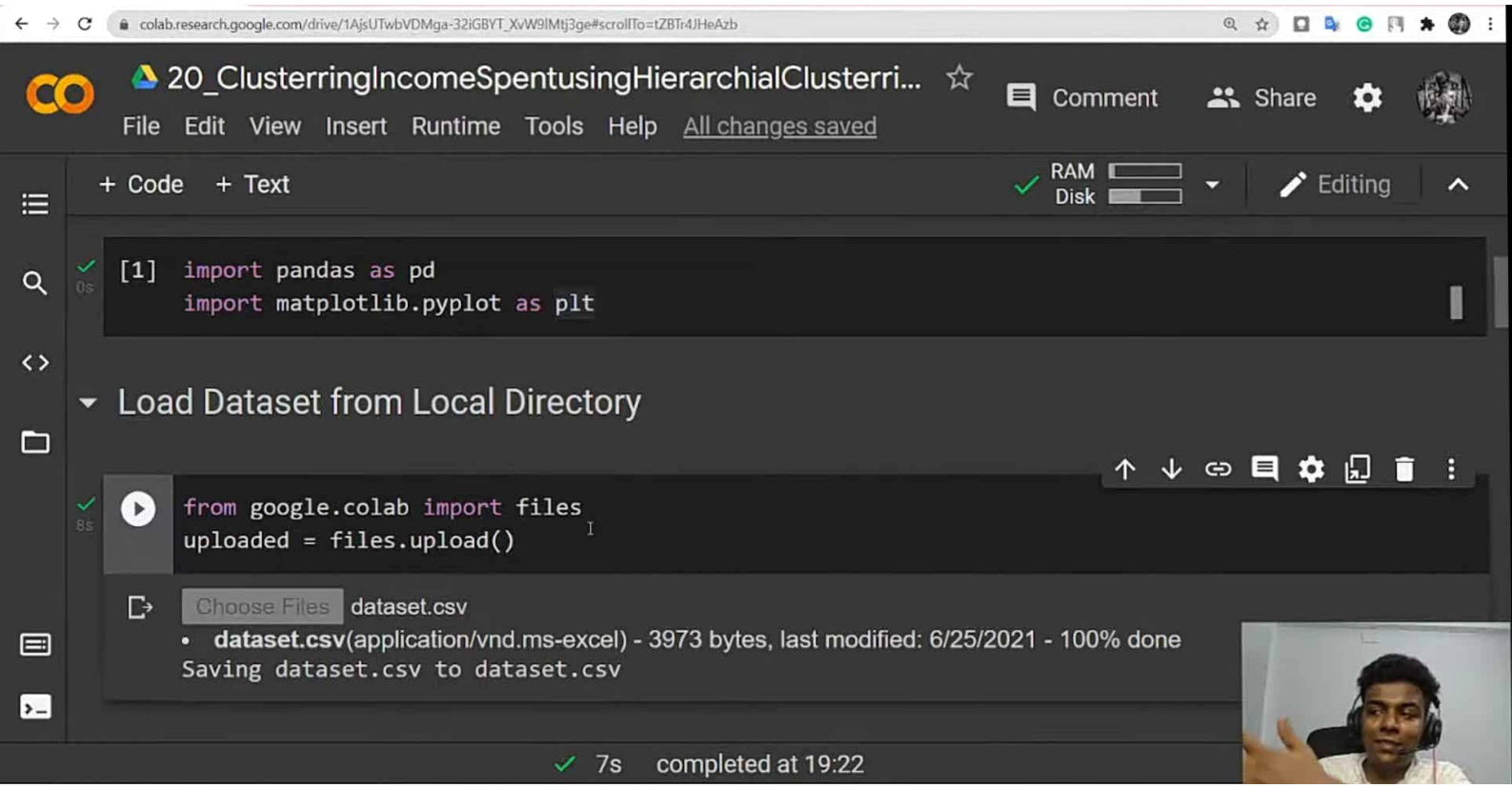
6 Data Visualization - O/P of Hierarchical Clustering

7 Fitting Model to Clustering algorithm



8 Visualizing Clustered Result

CustomerID	Gender	Age	Annual Inc	Spending Score
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
11	Male	67	19	14
12	Female	35	19	99
13	Female	58	20	15
14	Female	24	20	77
15	Male	37	20	13
16	Male	22	20	79
17	Female	35	21	35
18	Male	20	21	66
19	Male	52	23	29
20	Female	35	23	98
21	Male	35	24	35
22	Male	25	24	73
23	Female	46	25	5
24	Male	31	25	73
25	Female	54	28	14
26	Male	29	28	82
27	Female	45	28	32
28	Male	35	28	61
29	Female	40	29	31


colab.research.google.com/drive/1AjsUTwbVDMga-32iGBYT_XvW9lMtj3ge#scrollTo=tZBTr4JHeAzb
20_ClusterringIncomeSpentusingHierarchialClusterri...
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
RAM
Disk
Editing
[1] import pandas as pd
import matplotlib.pyplot as plt
0s
Load Dataset from Local Directory
from google.colab import files
uploaded = files.upload()
8s
Choose Files dataset.csv
dataset.csv(application/vnd.ms-excel) - 3973 bytes, last modified: 6/25/2021 - 100% done
Saving dataset.csv to dataset.csv
7s completed at 19:22

colab.research.google.com/drive/1AjsUTwbVDMga-32iGBYT_XvW9lMtj3ge#scrollTo=q4vNcNRlQtjr

20_ClusterringIncomeSpentusingHierarchialClusterri...

Comment Share

File Edit View Insert Runtime Tools Help

+ Code + Text

RAM Disk

Editing

```
[3]  dataset = pd.read_csv('dataset.csv')
```

0s

Summarize Dataset

```
print(dataset.shape)
print(dataset.describe())
print(dataset.head(5))
```

Label Encoding

0s completed at 19:24

Waiting for colab.research.google.com...

Type here to search

30°C Mostly cloudy

colab.research.google.com/drive/1AjsUTwbVDMga-32iGBYT_XvW9lMtj3ge#scrollTo=wOuthXLIT0GI

20_ClusterringIncomeSpentusingHierarchialClusterri...

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk Editing

[4]

2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

Label Encoding

```
from sklearn import preprocessing
label_encoder = preprocessing.LabelEncoder()
dataset['Gender'] = label_encoder.fit_transform(dataset['Gender'])
dataset.head()
```

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score
0	1	1	19	15	39

1s completed at 19:25

colab.research.google.com/drive/1AjsUTwbVDMga-32iGBYT_XvW9lMtj3ge#scrollTo=2PuOMjABSCXw

20_ClusterringIncomeSpentusingHierarchialClusterri...

Comment Share

File Edit View Insert Runtime Tools Help All changes saved

+ Code + Text

RAM Disk

Editing

4	5	0	31	17	40

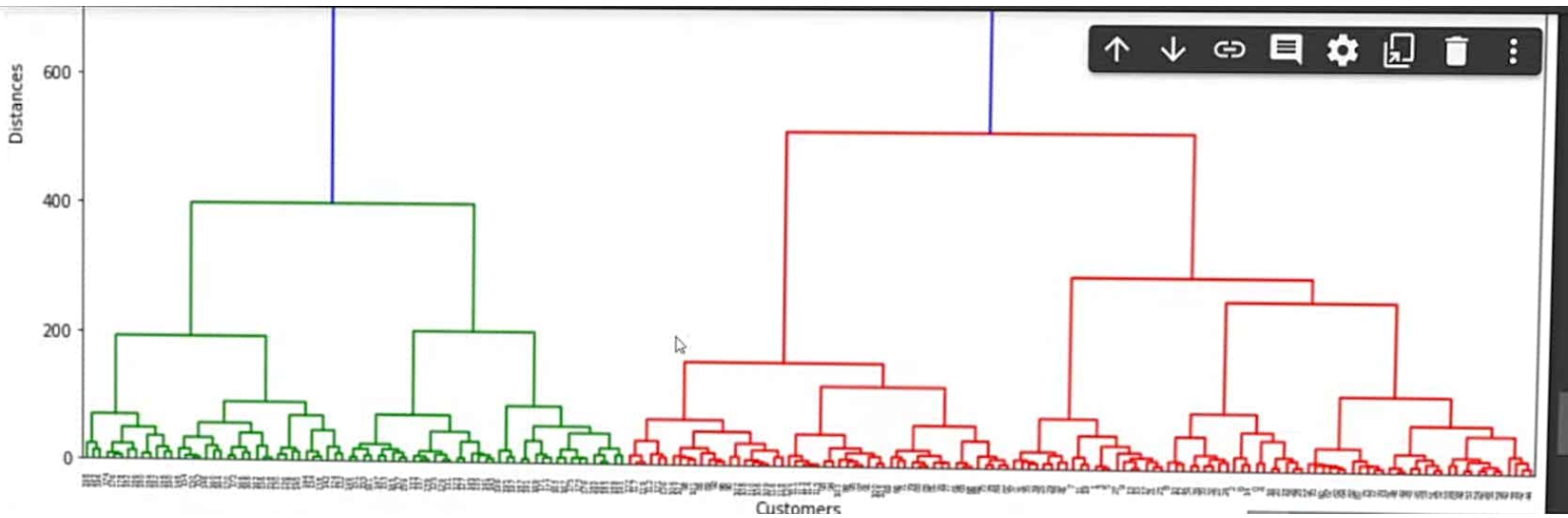
Dendrogram Data visualization

```
import scipy.cluster.hierarchy as clus

plt.figure(1, figsize = (16 ,8))
dendrogram = clus.dendrogram(clus.linkage(dataset, method  = "ward"))

plt.title('Dendrogram Tree Graph')
plt.xlabel('Customers')
plt.ylabel('Distances')
plt.show()
```

1s completed at 19:25


Distances
600
400
200
0
Customers

20_ClusterringIncomeSpentusing

colab.research.google.com/drive/1AjsUTwbVDMga-32iGBYT_XvW9lMtj3ge#scrollTo=0K7mijvHUW6Z

20_ClusterringIncomeSpentusingHierarchialClusterri...

Comment Share

File Edit View Insert Runtime Tools Help All changes saved

+ Code + Text

RAM Disk

Editing

[6]

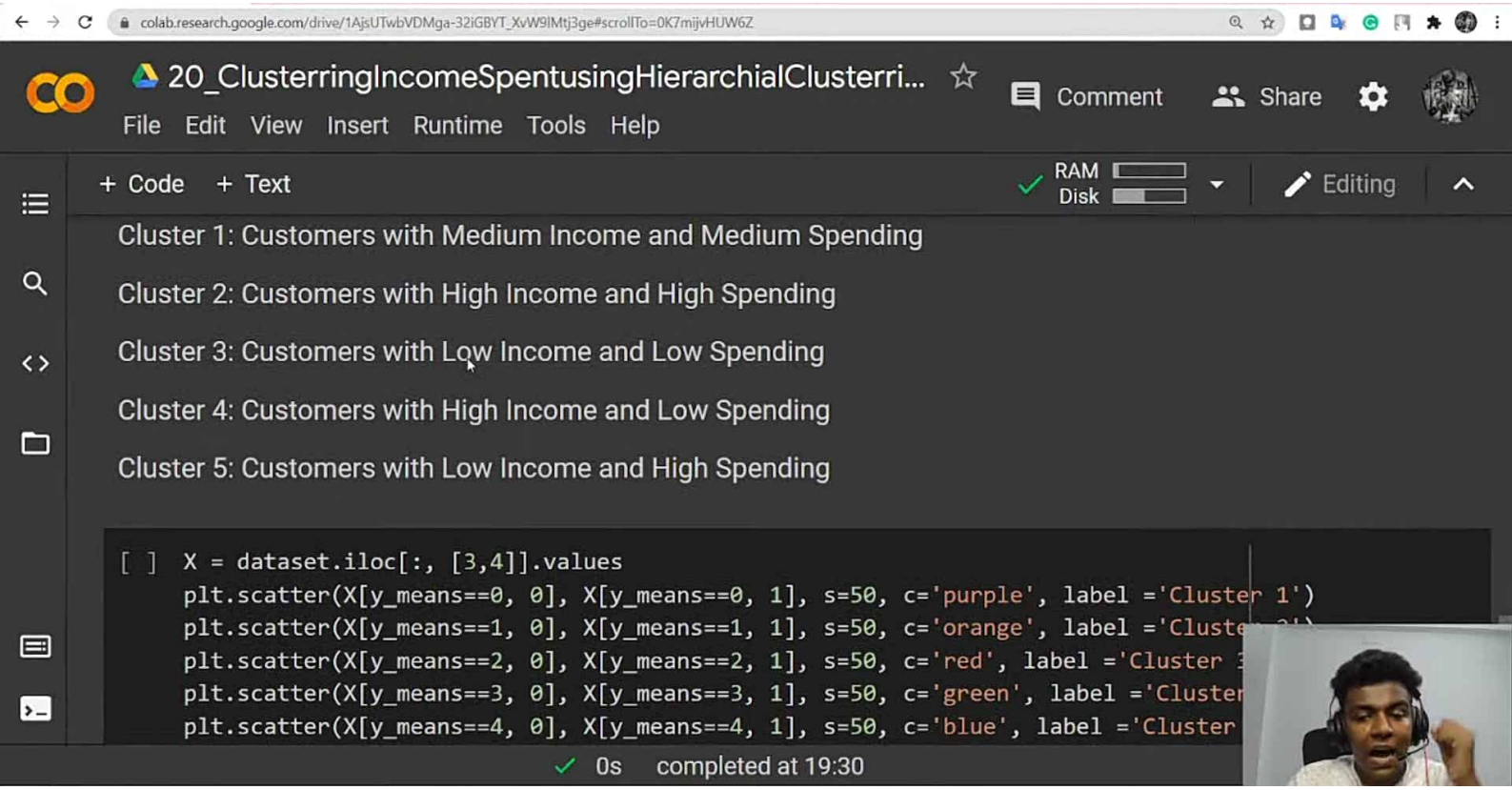
4s

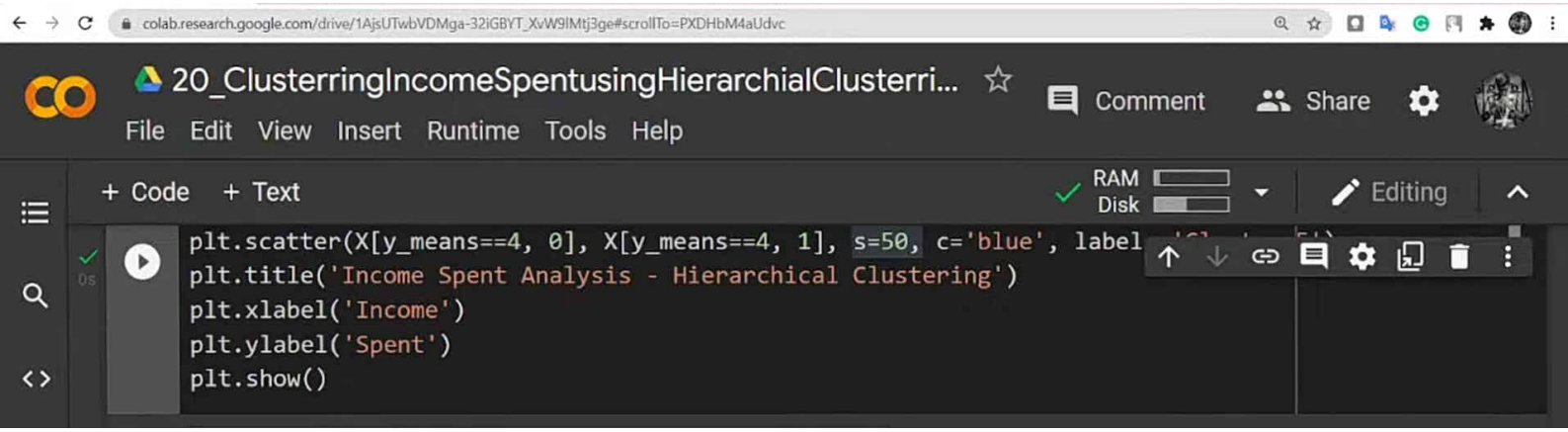
Fitting the Hierarchial clustering to the dataset with n=5

```
from sklearn.cluster import AgglomerativeClustering
model = AgglomerativeClustering(n_clusters = 5, affinity = 'euclidean', linkage ='average')
y_means = model.fit_predict(dataset)
y_means
```

Visualizing the number of clusters n=5

4s completed at 19:28


colab.research.google.com/drive/1AjsUTwbVDMga-32iGBYT_XvW9lMtj3ge#scrollTo=0K7mijvHUW6Z
20_ClusterringIncomeSpentusingHierarchialClusterri...
Comment
Share
File Edit View Insert Runtime Tools Help
+ Code + Text
RAM
Disk
Editing
Cluster 1: Customers with Medium Income and Medium Spending
Cluster 2: Customers with High Income and High Spending
Cluster 3: Customers with Low Income and Low Spending
Cluster 4: Customers with High Income and Low Spending
Cluster 5: Customers with Low Income and High Spending
[ ] X = dataset.iloc[:, [3,4]].values
plt.scatter(X[y_means==0, 0], X[y_means==0, 1], s=50, c='purple', label ='Cluster 1')
plt.scatter(X[y_means==1, 0], X[y_means==1, 1], s=50, c='orange', label ='Cluste
plt.scatter(X[y_means==2, 0], X[y_means==2, 1], s=50, c='red', label ='Cluster
plt.scatter(X[y_means==3, 0], X[y_means==3, 1], s=50, c='green', label ='Cluster
plt.scatter(X[y_means==4, 0], X[y_means==4, 1], s=50, c='blue', label ='Cluster
0s completed at 19:30


colab.research.google.com/drive/1AjsUTwbVDMga-32iGBYT_XvW9lMtj3ge#scrollTo=PXDHbM4aUdvc
20_ClusterringIncomeSpentusingHierarchialClusterri...
Comment
Share
File Edit View Insert Runtime Tools Help
+ Code + Text
RAM
Disk
Editing
plt.scatter(X[y_means==4, 0], X[y_means==4, 1], s=50, c='blue', label
plt.title('Income Spent Analysis - Hierarchical Clustering')
plt.xlabel('Income')
plt.ylabel('Spent')
plt.show()