Jimma Institute of Technology

Faculty of Computing

Computer Science Program

Data mining Assignment-1

Title: - Data mining task on Given Ethio-telecom Sample csv file

Prepared by:

1. Chimdessa Tesfaye

January 2021

JIMMA, ETHIOPIA

Contents

Chapter One

Introduction

1.1. Introduction

Today, many organizations are doing analysis on their customers by using the historical data in their database. Most of the organizations do these surveys to improve their system and meet the needs of their customers.

Ethiotelecom is one of the organizations found in our country Ethiopia. At present, it is the only telecom company which gives services related to the telecommunication in our country. Because of this, the company has to work quite goodly to maintain the service deliverance and quality.

This report is done on the historical data gathered by the company through different times and circumstances. The data encompasses of the data of 1048576 customers.

The given data is mainly based on the *CALL_FEE* **and** *DATA_USAGE* service provided by the company.

The given data is divided into four columns named *CALL_START*,*CALL_END*,*DATA_USAGE* and *CALL_FEE* respectively.

Based on the given data we are going to cluster related users by the call fee and data usage. From this we can able to observe the rate of data usage vs. call fee

1.2. Statement of the Problem

The purpose of this report is to visualize and cluster related users based on the data usage and call fee.

1.3. Significance of the Study

There are four primary groups that may benefit from the study.

- The first group is the company itself
- The second group is the employees.
- The third group is the customers.
- The fourth group is us, the report writers. Because by doing this report we are able to do the analysis of large data sets, by using different data mining

concepts. Which help us to manage large companies, doing surveys on large data sets, which is the big task know a days.

1.4. Scope of the Report

This study was limited to only the data of sample 1048576 customers, only for one week between days 22-26-jun-19.

1.5. Methods of the Report

Source of data

Data for this study were given to us by our teacher as private data from the Ethiotelecom company, because it is not got by an individual else.

Statistical methods

Simple statistical techniques were used to tabulate and plot the results of this report. The primary data were analyzed using outcome of the report result.

Sample Selection

The number of customer's data sample taken is 1048576, and while clustering we use 50 samples(s=50).

Limitation of the study

The number of days takes as the sample is only four days. So, this have its own impact on the outcome of the report.

Chapter Two

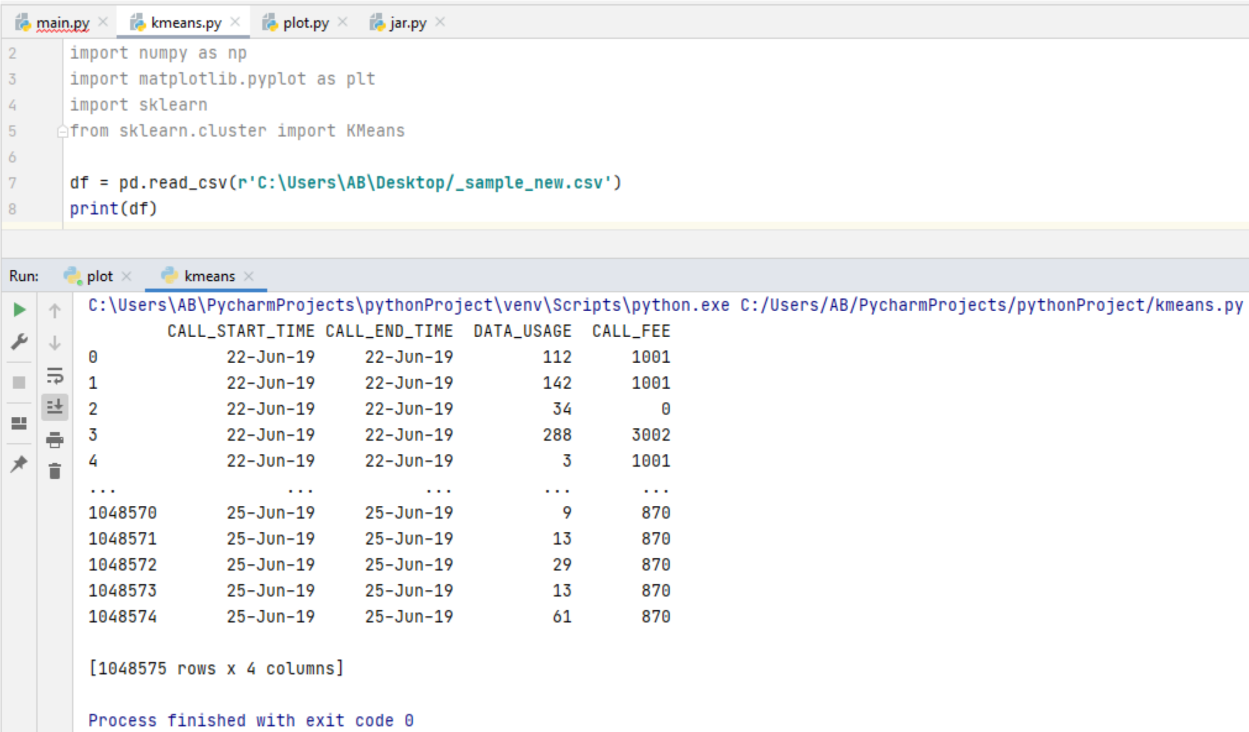
2. Implementing Mining Task

2.1. Introduction

We use data mining concepts in the way to gain the result we intended from the samples. We implement the data mining concepts on the python. By using the data mining concepts we can able to read the .csv files and display them. After that we select the columns that are more relevant to the outcome we intended for and omit the columns that are not relevant for the outcome based on the DM concepts. Then we use the data mining tasks like clustering to group the overall customers into groups, to simplify the outcome generation. The whole steps we follow in generating our outcome are listed below.

Importing Data

- ***Viewing the data by importing the whole .csv file from our folder***

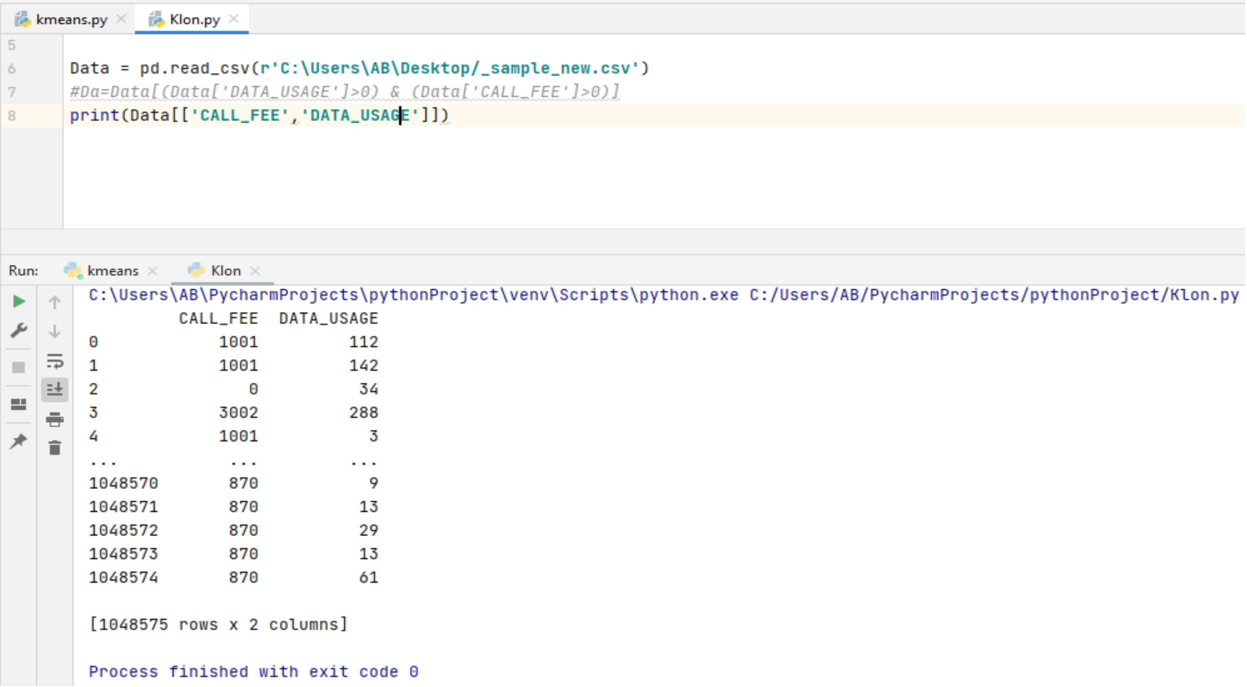


```
import numpy as np
import matplotlib.pyplot as plt
import sklearn
from sklearn.cluster import KMeans

df = pd.read_csv(r'C:\Users\AB\Desktop/_sample_new.csv')
print(df)
```

```
C:\Users\AB\PycharmProjects\pythonProject\venv\Scripts\python.exe C:/Users/AB/PycharmProjects/pythonProject/kmeans.py
          CALL_START_TIME CALL_END_TIME  DATA_USAGE  CALL_FEE
0               22-Jun-19     22-Jun-19         112      1001
1               22-Jun-19     22-Jun-19         142      1001
2               22-Jun-19     22-Jun-19          34         0
3               22-Jun-19     22-Jun-19         288      3002
4               22-Jun-19     22-Jun-19           3      1001
...                   ...           ...         ...       ...
1048570         25-Jun-19     25-Jun-19           9       870
1048571         25-Jun-19     25-Jun-19          13       870
1048572         25-Jun-19     25-Jun-19          29       870
1048573         25-Jun-19     25-Jun-19          13       870
1048574         25-Jun-19     25-Jun-19          61       870

[1048575 rows x 4 columns]

Process finished with exit code 0
```

- **Then we select the columns that are mandatory for our mining task.**

  We are going to cluster the customers based on the *Data_usage* and *Call_fee.*

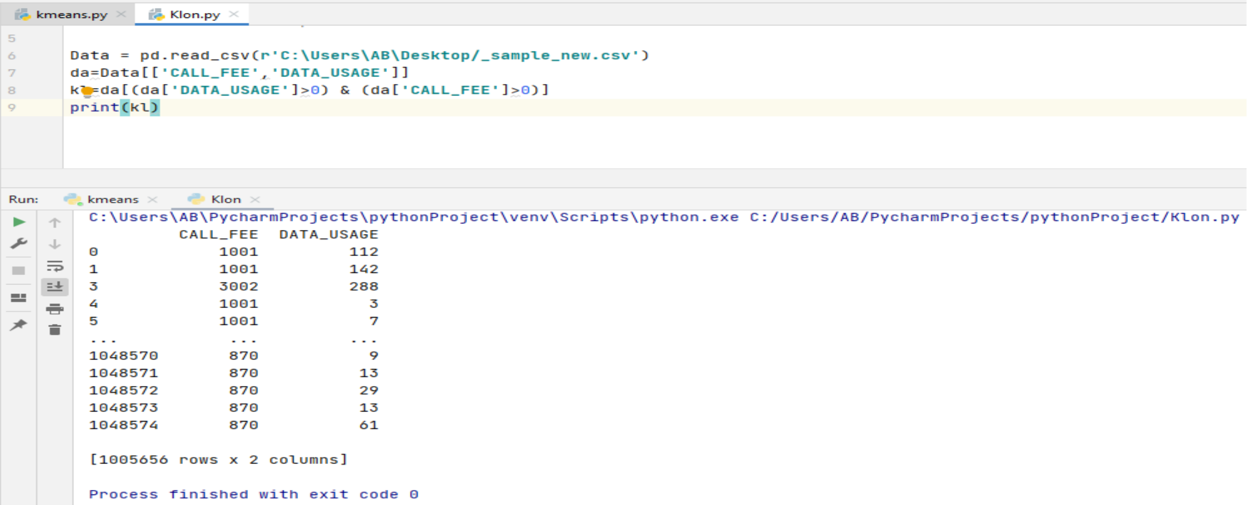


```
Data = pd.read_csv(r'C:\Users\AB\Desktop/_sample_new.csv')
#Da=Data[(Data['DATA_USAGE']>0) & (Data['CALL_FEE']>0)]
print(Data[['CALL_FEE','DATA_USAGE']])
```

```
C:\Users\AB\PycharmProjects\pythonProject\venv\Scripts\python.exe C:/Users/AB/PycharmProjects/pythonProject/Klon.py
         CALL_FEE  DATA_USAGE
0            1001         112
1            1001         142
2               0          34
3            3002         288
4            1001           3
...           ...         ...
1048570       870           9
1048571       870          13
1048572       870          29
1048573       870          13
1048574       870          61

[1048575 rows x 2 columns]

Process finished with exit code 0
```

Data preprocessing

- **Removing null values from columns Call_fee and Data_usage. b/c they results in incorrect outcomes while mining the knowledge from the data.**



```
Data = pd.read_csv(r'C:\Users\AB\Desktop/_sample_new.csv')
da=Data[['CALL_FEE','DATA_USAGE']]
kl=da[(da['DATA_USAGE']>0) & (da['CALL_FEE']>0)]
print(kl)
```

```
C:\Users\AB\PycharmProjects\pythonProject\venv\Scripts\python.exe C:/Users/AB/PycharmProjects/pythonProject/Klon.py
         CALL_FEE  DATA_USAGE
0            1001         112
1            1001         142
3            3002         288
4            1001           3
5            1001           7
...           ...         ...
1048570       870           9
1048571       870          13
1048572       870          29
1048573       870          13
1048574       870          61

[1005656 rows x 2 columns]

Process finished with exit code 0
```

There are 42,918 null values for the columns Data usage and call fee. So we remove the null values from these columns.

Then we apply the clustering algorithm on these two columns.

- ***Clustering is a data mining (machine learning) technique that finds similarities between data according to the characteristics found in the data & groups similar data objects into one cluster***
- ***Given a set of points, with a notion of distance between points, group the points into some number of clusters, so that members of a cluster are in some sense as close to each other as possible.***
- ***While data points in the same cluster are similar, those in separate clusters are dissimilar to one another.***

We use the k-means clustering algorithm

- ***Each cluster is represented by the center of the cluster***

Algorithm:

•Select K cluster points as initial centroids (the initial centroids are selected randomly)

–Given k, the k-means algorithm is implemented as follows:

•Repeat

–Partition objects into k nonempty subsets

–Recompute the centroids of each K clusters of the current partition

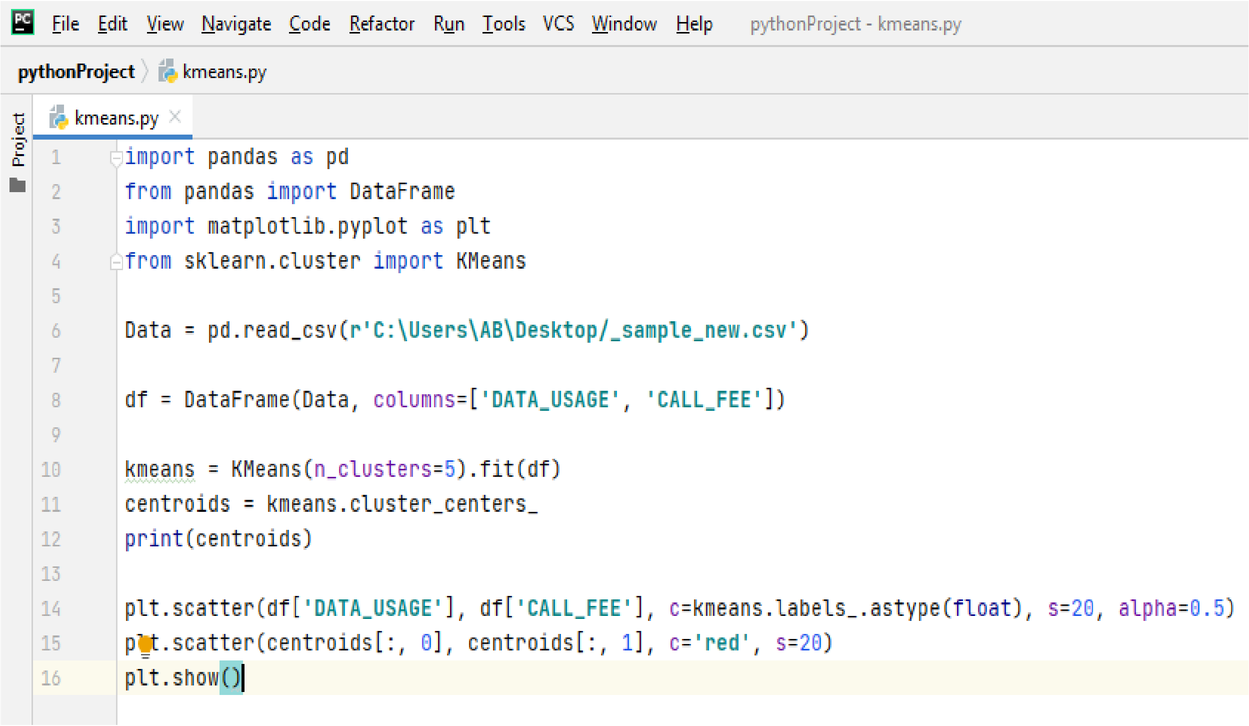
(The centroid is the center, i.e., mean point, of the cluster)

–Assign each object to the cluster with the nearest seed point

•Until the centroid don’t change

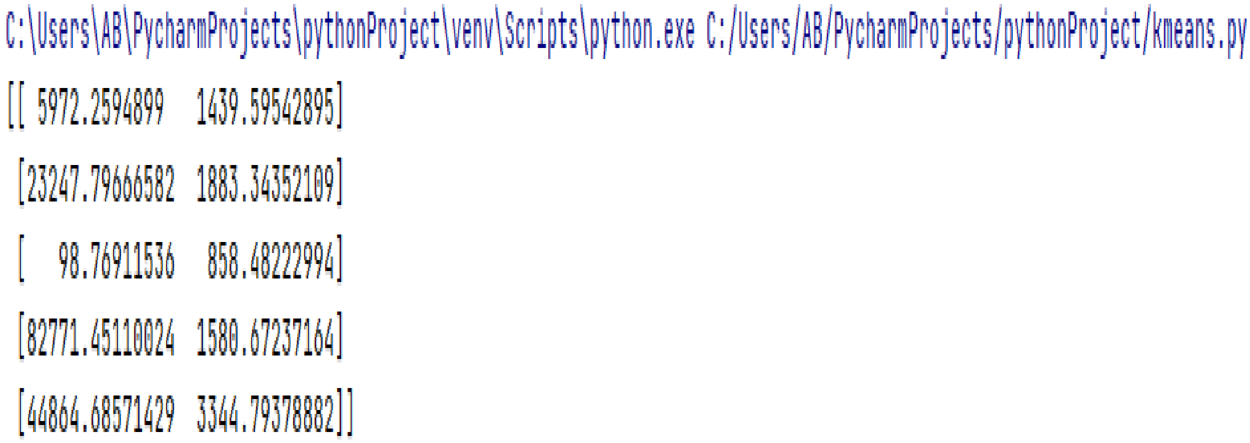
How the algorithm work on our data set is presented below.

- We select 5 clusters (*n_clusters=5)*
- *X-axis =’DATA_USAGE’*
- *Y-axis=’CALL_FEE’*
- *s=50*



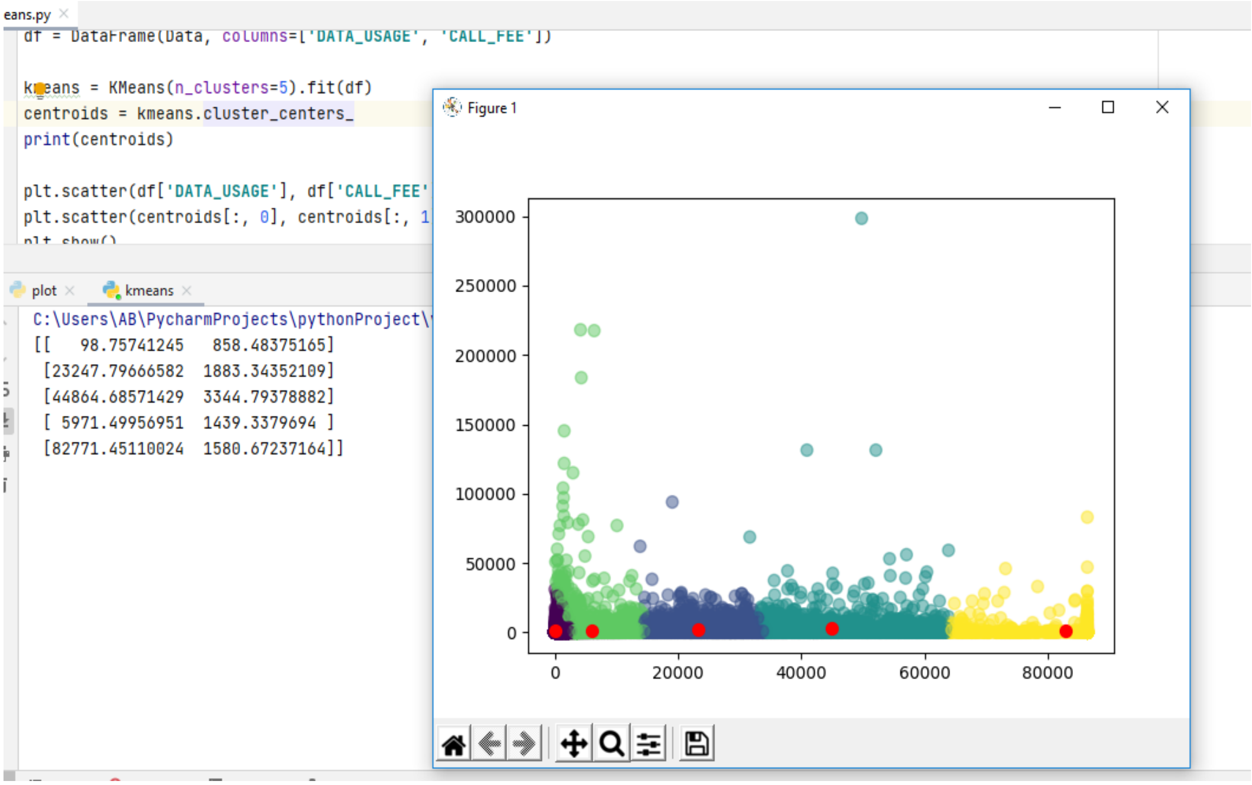
```python
import pandas as pd
from pandas import DataFrame
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans

Data = pd.read_csv(r'C:\Users\AB\Desktop/_sample_new.csv')

df = DataFrame(Data, columns=['DATA_USAGE', 'CALL_FEE'])

kmeans = KMeans(n_clusters=5).fit(df)
centroids = kmeans.cluster_centers_
print(centroids)

plt.scatter(df['DATA_USAGE'], df['CALL_FEE'], c=kmeans.labels_.astype(float), s=20, alpha=0.5)
plt.scatter(centroids[:, 0], centroids[:, 1], c='red', s=20)
plt.show()
```

- The output looks like...The five clusters range



```
C:\Users\AB\PycharmProjects\pythonProject\venv\Scripts\python.exe C:/Users/AB/PycharmProjects/pythonProject/kmeans.py
[[ 5972.2594899  1439.59542895]
 [23247.79666582  1883.34352109]
 [   98.76911536   858.48222994]
 [82771.45110024  1580.67237164]
 [44864.68571429  3344.79378882]]
```

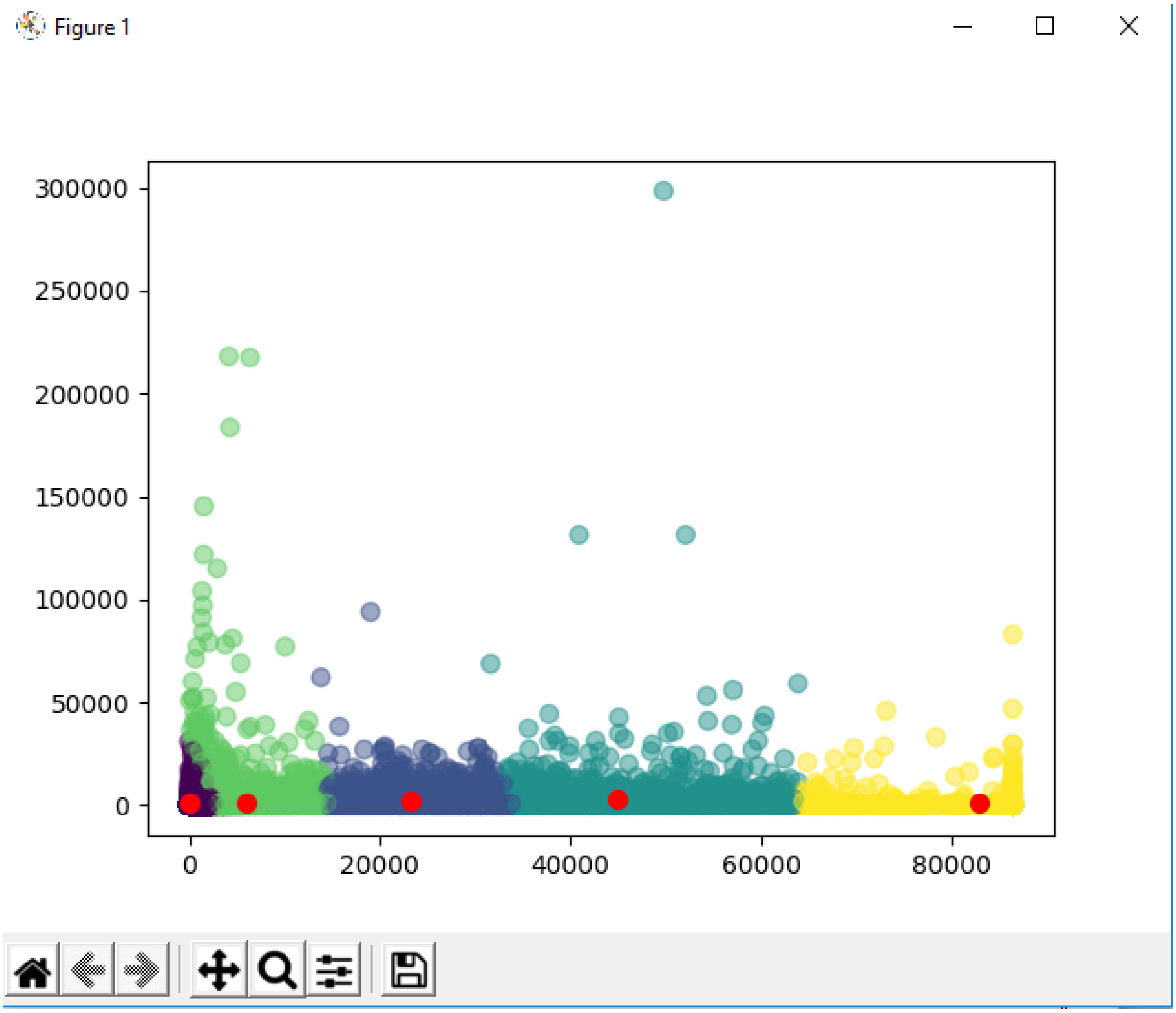
- Diagrammatically the clustered output looks like…



```
df = DataFrame(Data, columns=['DATA_USAGE', 'CALL_FEE'])

kmeans = KMeans(n_clusters=5).fit(df)
centroids = kmeans.cluster_centers_
print(centroids)

plt.scatter(df['DATA_USAGE'], df['CALL_FEE'
plt.scatter(centroids[:, 0], centroids[:, 1
plt.show()
```

```
C:\Users\AB\PycharmProjects\pythonProject\
[[   98.75741245   858.48375165]
 [23247.79666582  1883.34352109]
 [44864.68571429  3344.79378882]
 [ 5971.49956951  1439.3379694 ]
 [82771.45110024  1580.67237164]]
```

- For clear visualization..

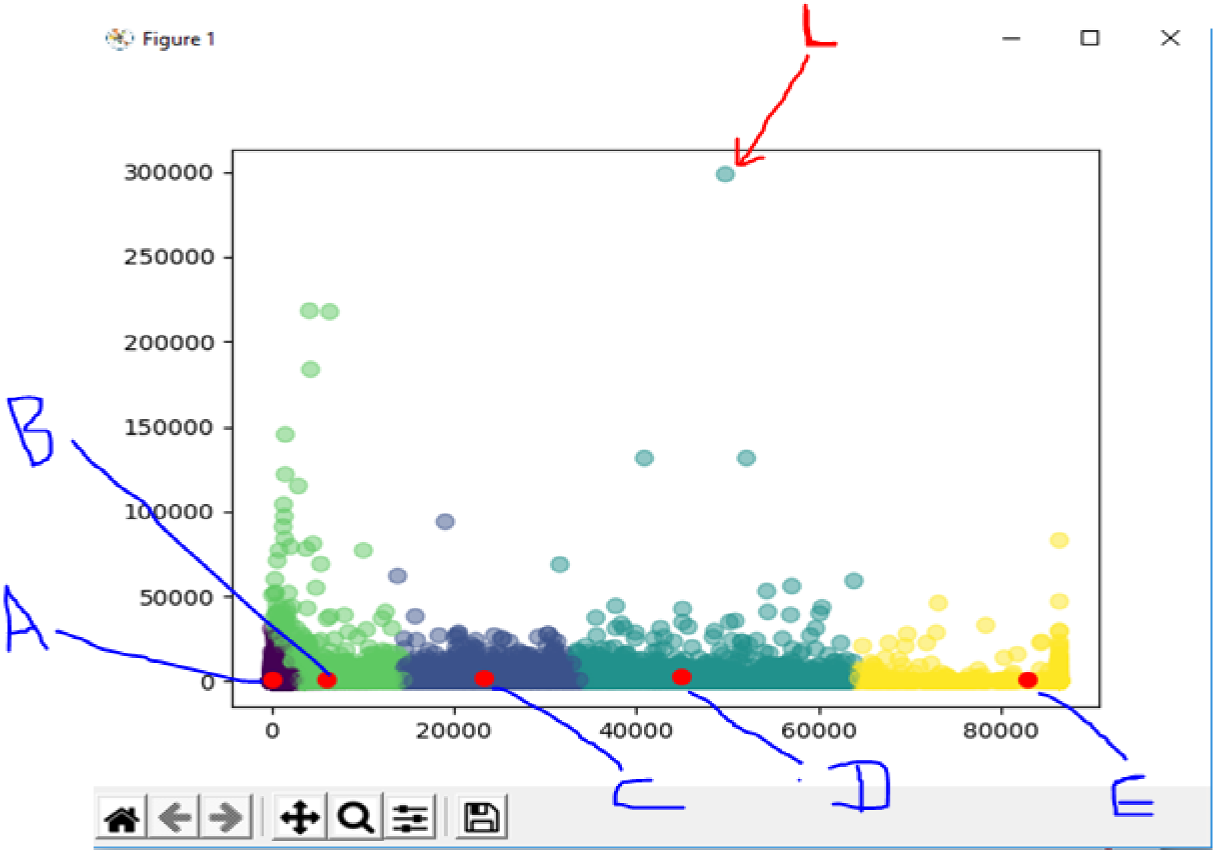


Chapter Three

Conclusions

On the basis of the findings, several conclusions can be made:-

- The observations in each cluster are similar to one another with respect call fee and data usage, and that the two clusters are quite distinct from each other.
- Users with cluster 'A' have similar or most related call fee and data usage.
- Users with cluster 'B' have similar or most related call fee and data usage.
- Users with cluster 'C' have similar or most related call fee and data usage.
- Users with cluster 'D' have similar or most related call fee and data usage.
- Users with cluster 'E' have similar or most related call fee and data usage.



- For example the user from the samples (denoted by L) has the highest Call fee near to 300,000, which means have high call duration(High call fee is result of high call duration). Which have high difference from the other users exists. Because of that it's not clustered around other users with related and most similar call fee and data usage. And other so many conclusions made from this.

Problems and challenges

While performing this data mining task using clustering we face several problems and they are stated below:-

- The data is very huge; it needs proper understanding of it to mine all the data given to as.
- We use the algorithm for first time, because of that it takes as much time to understand it.
- Due to the occasions we exist in, the time we have is very limited. Because of that we can't able to have much time to do the mining task.

We face the problems by:

- Try to understand the given data by using different techniques, like separating the columns, sorting them and understand the behavior of them and etc.
- We try to read online materials to understand the algorithm we use(kmeans algorithm for clustering)
- Try to share other assignments and projects to each other in a group to manage our time.