# Web Mining (CSE3024)

## Lab Assignment 8

Name: **Kritika Mishra**

Registration Number: **16BCI0041**

Slot: L15+L16

Date: 3rd October 2018

Question:

**Implement a k-means algorithm with sklearn to partition observations in a dataset into a specific number of clusters in order to aid in analysis of the data.**

- **Use Sklearn Toolkit / Package to perform the process**
- **Import kmeans and PCA through the sklearn library**
- **Devise an elbow curve to select the optimal number of clusters (k)**
- **Generate and visualise a k-means clustering algorithms**

**Note : Dataset in CSV can be generated or downloaded from the internet. Please specify the source of the dataset in the documentation steps of this program.**

Dataset:

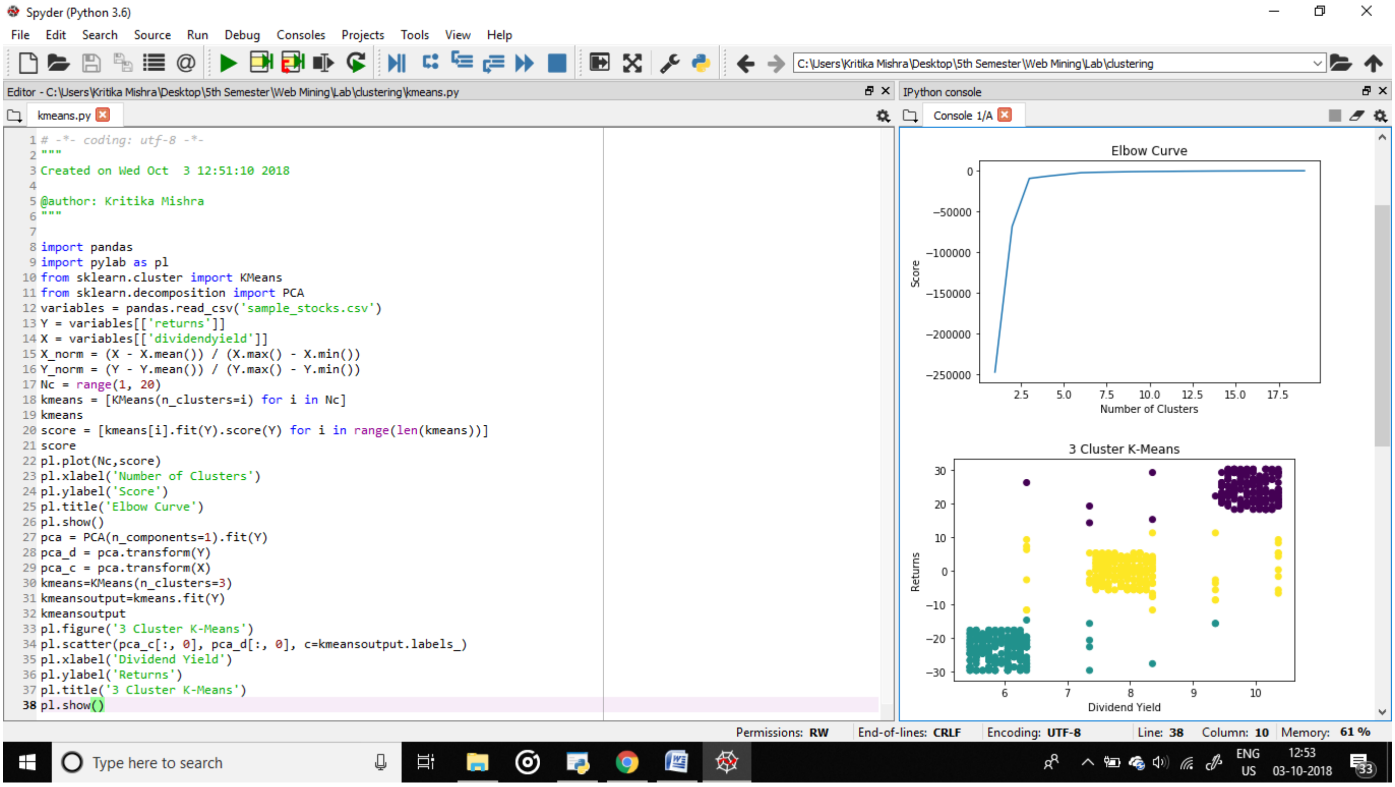
http://www.michaeljgrogan.com/datasets/ ->sample_stocks.csv

Code:

```
import pandas
import pylab as pl
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
variables = pandas.read_csv('sample_stocks.csv')
Y = variables[['returns']]
X = variables[['dividendyield']]
X_norm = (X - X.mean()) / (X.max() - X.min())
Y_norm = (Y - Y.mean()) / (Y.max() - Y.min())
Nc = range(1, 20)
kmeans = [KMeans(n_clusters=i) for i in Nc]
kmeans
score = [kmeans[i].fit(Y).score(Y) for i in range(len(kmeans))]
score
```

```
pl.plot(Nc,score)
pl.xlabel('Number of Clusters')
pl.ylabel('Score')
pl.title('Elbow Curve')
pl.show()
pca = PCA(n_components=1).fit(Y)
pca_d = pca.transform(Y)
pca_c = pca.transform(X)
kmeans=KMeans(n_clusters=3)
kmeansoutput=kmeans.fit(Y)
kmeansoutput
pl.figure('3 Cluster K-Means')
pl.scatter(pca_c[:, 0], pca_d[:, 0], c=kmeansoutput.labels_)
pl.xlabel('Dividend Yield')
pl.ylabel('Returns')
pl.title('3 Cluster K-Means')
pl.show()
```

## Output:

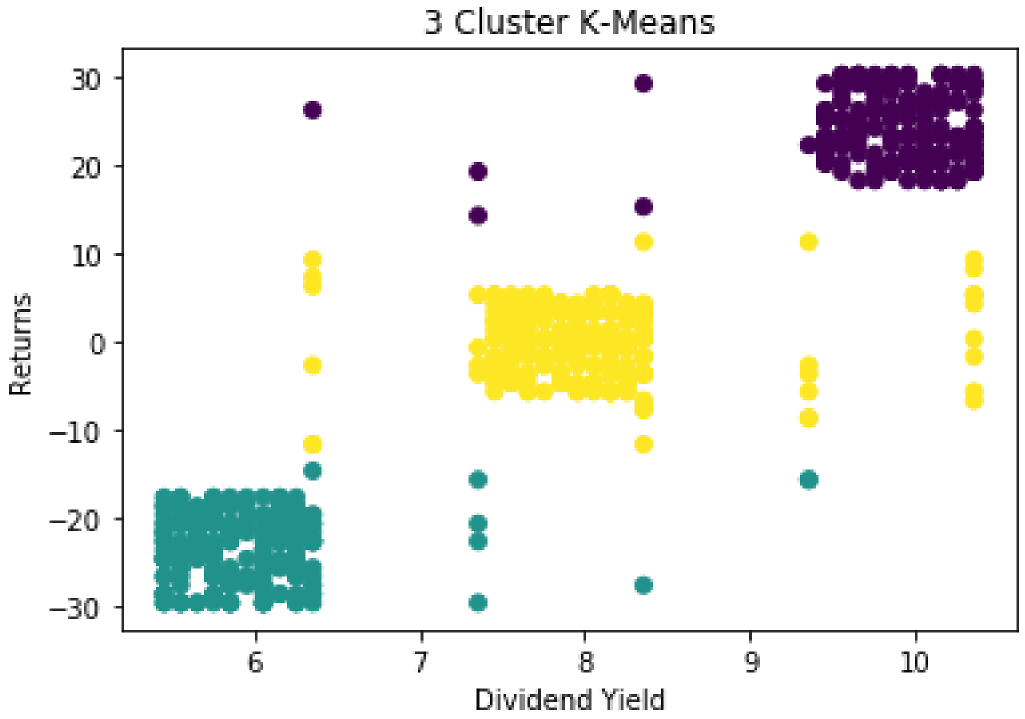


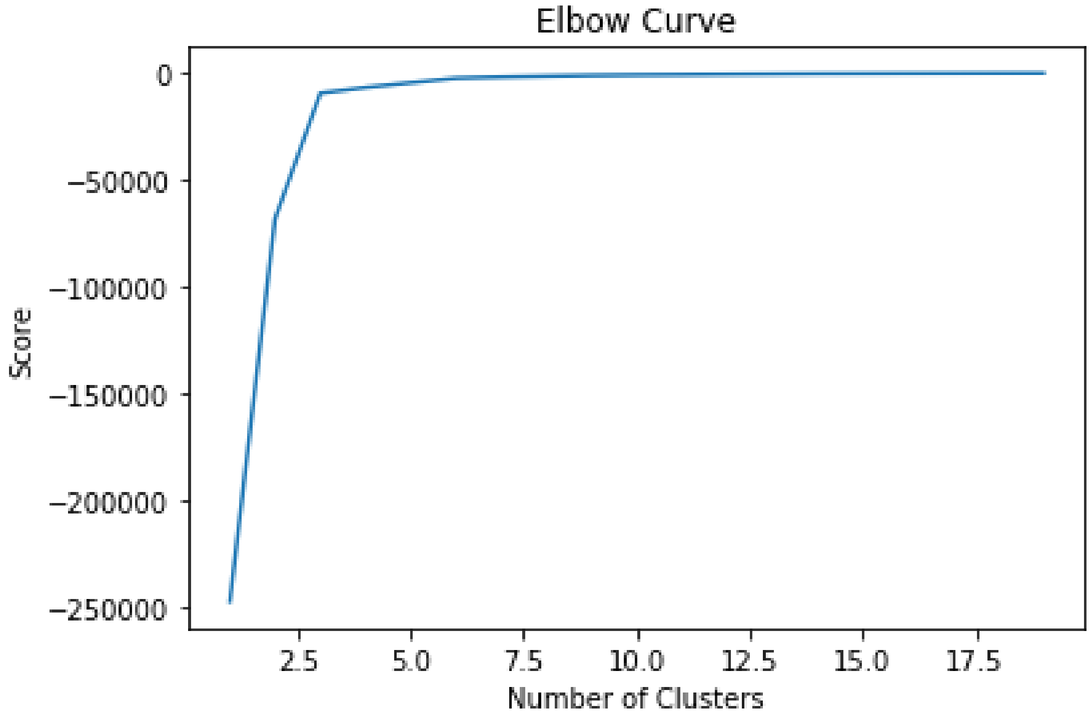
Spyder (Python 3.6)
File
Edit
Search
Source
Run
Debug
Consoles
Projects
Tools
View
Help
C:\Users\Kritika Mishra\Desktop\5th Semester\Web Mining\Lab\clustering
Editor - C:\Users\Kritika Mishra\Desktop\5th Semester\Web Mining\Lab\clustering\kmeans.py
kmeans.py
IPython console
Console 1/A

```
# -*- coding: utf-8 -*-
"""
Created on Wed Oct  3 12:51:10 2018

@author: Kritika Mishra
"""

import pandas
import pylab as pl
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
variables = pandas.read_csv('sample_stocks.csv')
Y = variables[['returns']]
X = variables[['dividendyield']]
X_norm = (X - X.mean()) / (X.max() - X.min())
Y_norm = (Y - Y.mean()) / (Y.max() - Y.min())
Nc = range(1, 20)
kmeans = [KMeans(n_clusters=i) for i in Nc]
kmeans
score = [kmeans[i].fit(Y).score(Y) for i in range(len(kmeans))]
score
pl.plot(Nc,score)
pl.xlabel('Number of Clusters')
pl.ylabel('Score')
pl.title('Elbow Curve')
pl.show()
pca = PCA(n_components=1).fit(Y)
pca_d = pca.transform(Y)
pca_c = pca.transform(X)
kmeans=KMeans(n_clusters=3)
kmeansoutput=kmeans.fit(Y)
kmeansoutput
pl.figure('3 Cluster K-Means')
pl.scatter(pca_c[:, 0], pca_d[:, 0], c=kmeansoutput.labels_)
pl.xlabel('Dividend Yield')
pl.ylabel('Returns')
pl.title('3 Cluster K-Means')
pl.show()
```


3 Cluster K-Means
Returns
30
20
10
0
−10
−20
−30
6
7
8
9
10
Dividend Yield


Elbow Curve
Score
0
−50000
−100000
−150000
−200000
−250000
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Number of Clusters