Ashpak Sheikh

KMEANS CLUSTERING

BATCH : DSWDMOD 020421

Problem Statement:- Perform clustering for the crime data and identify the number of clusters formed and draw inferences

About Data:- we have given data containing crime rates in various places around .

Analysis With Python:-

```
import pandas as pd

import numpy as np

crime_data=pd.read_csv("D:/DataScience/Class/assignment working/h_clustering/crime_data.csv")
```

Checking EDA

```
crime_data.describe()
```

```
In [110]: crime_data.describe()
Out[110]:
         Murder     Assault   UrbanPop       Rape
count  50.00000   50.000000  50.000000  50.000000
mean    7.78800  170.760000  65.540000  21.232000
std     4.35551   83.337661  14.474763   9.366385
min     0.80000   45.000000  32.000000   7.300000
25%     4.07500  109.000000  54.500000  15.075000
50%     7.25000  159.000000  66.000000  20.100000
75%    11.25000  249.000000  77.750000  26.175000
max    17.40000  337.000000  91.000000  46.000000
```

```
crime_data.columns.values

#removing categorical column

crime=crime_data.drop("Unnamed: 0",axis=1)

crime=crime.astype("int")

crime.isna().sum()
```

```
In [114]: crime.isna().sum()
Out[114]:
Murder      0
Assault     0
UrbanPop    0
Rape        0
dtype: int64
```

Checking outliers

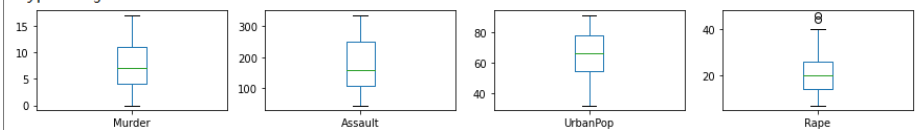
Ashpak Sheikh

KMEANS CLUSTERING

BATCH : DSWDMOD 020421

```
crime.plot(kind="box",subplots=True,layout=(4,4),figsize=(15,8))
```

```
In [117]: crime.plot(kind="box",subplots=True,layout=(4,4),figsize=(15,8))
Out[117]:
Murder          AxesSubplot(0.125,0.71587;0.168478x0.16413)
Assault     AxesSubplot(0.327174,0.71587;0.168478x0.16413)
UrbanPop    AxesSubplot(0.529348,0.71587;0.168478x0.16413)
Rape        AxesSubplot(0.731522,0.71587;0.168478x0.16413)
dtype: object
```



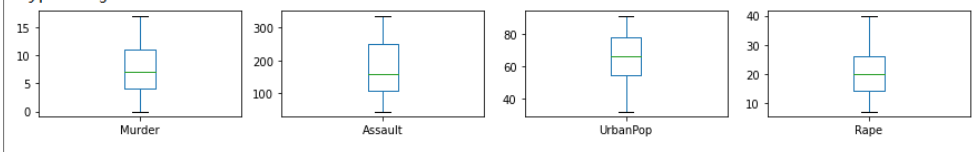
Outlier treatment

```
from scipy.stats.mstats import winsorize

crime["Rape"]=winsorize(crime["Rape"],limits=(0.01,0.04))

crime.plot(kind="box",subplots=True,layout=(4,4),figsize=(15,8))
```

```
In [123]: crime.plot(kind="box",subplots=True,layout=(4,4),figsize=(15,8))
Out[123]:
Murder          AxesSubplot(0.125,0.71587;0.168478x0.16413)
Assault     AxesSubplot(0.327174,0.71587;0.168478x0.16413)
UrbanPop    AxesSubplot(0.529348,0.71587;0.168478x0.16413)
Rape        AxesSubplot(0.731522,0.71587;0.168478x0.16413)
dtype: object
```



Zero variance

```
crime.var()
```

```
In [124]: crime.var()
Out[124]:
Murder        19.473061
Assault     6945.165714
UrbanPop     209.518776
Rape          79.301633
dtype: float64
```

Normalizing data

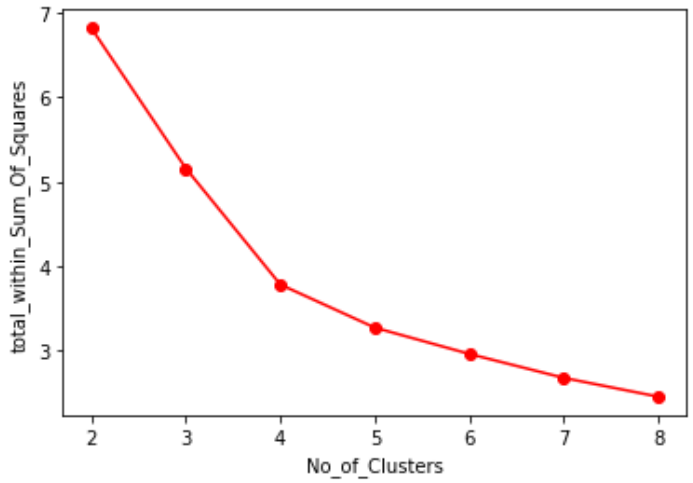
```
def norm(x):
    z=(x-x.min())/(x.max()-x.min())
```

```
    return z


crime_norm=norm(crime)
```

Plotting scree plot

```
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
twss=[]
x=range(2,9)
for i in x:
    kmeans=KMeans(n_clusters=i)
    kmeans.fit(crime_norm)
    twss.append(kmeans.inertia_)
plt.plot(x,twss,"ro-");plt.xlabel("No_of_Clusters");plt.ylabel("total_within_Sum_Of_Squares")
```



Making clusters

```
kmeans=KMeans(n_clusters=4)
kmeans.fit(crime_norm)
crime_data["Clusters"]=kmeans.labels_
crime_data.columns.values
crime_data=crime_data[['Clusters','Unnamed: 0', 'Murder', 'Assault', 'UrbanPop', 'Rape']]
```

Ashpak Sheikh
KMEANS CLUSTERING
BATCH : DSWDMOD 020421

crime_data.groupby(by="Clusters").mean()

```
In [139]: crime_data.groupby(by="Clusters").mean()
Out[139]:
              Murder     Assault   UrbanPop       Rape
Clusters
0           3.600000   78.538462  52.076923  12.176923
1          13.937500  243.625000  53.750000  21.412500
2          10.815385  257.384615  76.000000  33.192308
3           5.656250  138.875000  73.875000  18.781250
```

Summary and insights from data:

- From Kmeans its clear that we can divide crime severity into 4 types
- There are areas with low crime rate and few areas with higher crime rates
- We can see first cluster is with lower crime rates compared to fourth cluster
- Where Assaults are maximum their Rape rate is also maximum
- By lowering assaults ,rate of rapes can be bring down