Crime dataset clustering

Business Objective:- To cluster similar states on basis of their crime statistics and identify safest states

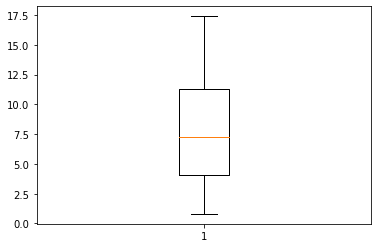
Constrainsts:- Reliable first primary source data is required

|  |  |  |  |
| --- | --- | --- | --- |
| Name of feature | Description | Type | Relevance |
| Unnamed:0 | states | Categorical,Nominal | All states in dataset |
| Murder | Murder | Quantitative,Nominal | Murder rates in that state |
| Assault | Assault | Quantitative,Nominal | Assault rates in that state |
| Rape | Rape | Quantitative,Nominal | Rape rates in that state |
| Urban pop | Urban population | Quantitative,Nominal | Urban population in that state |

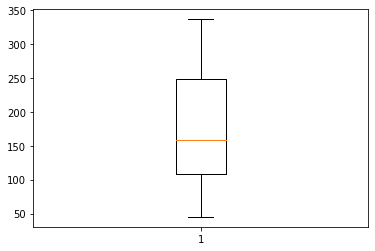
Data Cleaning/Feature Engineering

Boxplots to know if there are any outliers

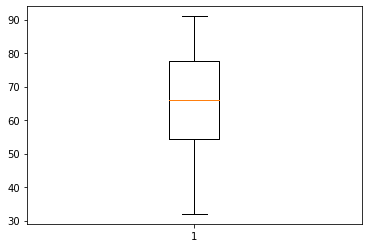
Murder column



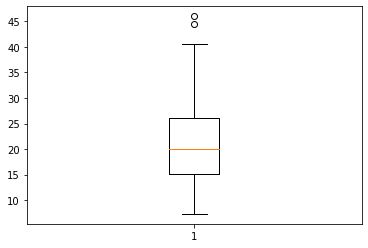
Assault column



Urban population



Rape

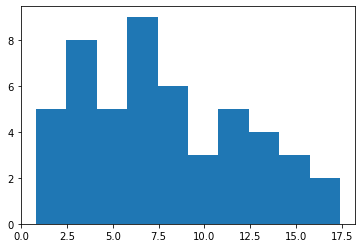


So as it can be seen two outliers for Rape column, since they are just few in number and size of dataset is small we will ignore them.

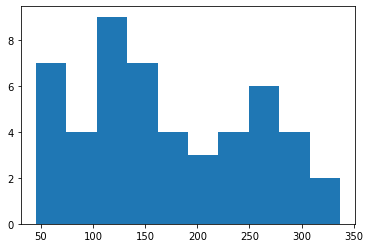
EDA

We will draw histograms to know distribution of data

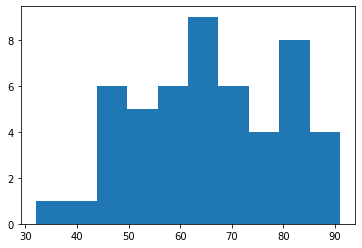
Murder feature



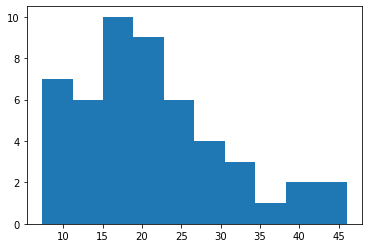
Assault feature



Urban Population



Rape feature



Mean,Media,Mode of dataset

(1) mean :-

Murder 7.788

Assault 170.760

UrbanPop 65.540

Rape 21.232

(2) median:-

Murder 7.25

Assault 159.00

UrbanPop 66.00

Rape 20.10

(3)Mode:-

Murder Assault UrbanPop Rape

2.1 120.0 66.0 14.9

(4) skewness

Murder 0.393956

Assault 0.234410

UrbanPop -0.226009

Rape 0.801200

(5) Kurtosis

Murder -0.827488

Assault -1.053848

UrbanPop -0.738360

Rape 0.353964

Dataset was normalized after this

Murder Assault UrbanPop Rape

count 50.000000 50.000000 50.000000 50.000000

mean 0.420964 0.430685 0.568475 0.360000

std 0.262380 0.285403 0.245335 0.242025

min 0.000000 0.000000 0.000000 0.000000

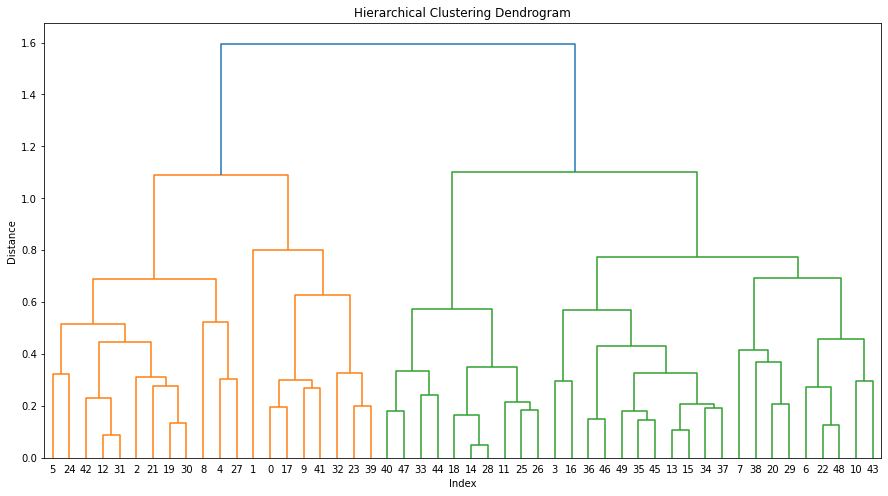
25% 0.197289 0.219178 0.381356 0.200904

50% 0.388554 0.390411 0.576271 0.330749

75% 0.629518 0.698630 0.775424 0.487726

max 1.000000 1.000000 1.000000 1.000000

After this Clustering dendrogram was viewed to make sense of clustering



After this Agglomerative Clustering was performed

And new labels were stored in a Pandas series as follows

array([0, 0, 3, 1, 3, 3, 1, 1, 3, 0, 1, 2, 3, 1, 2, 1, 1, 0, 2, 3, 1, 3,

1, 0, 3, 2, 2, 3, 2, 1, 3, 3, 0, 2, 1, 1, 1, 1, 1, 0, 2, 0, 3, 1,

2, 1, 1, 2, 1, 1], dtype=int64)

After this cluster feature was added to original dataframe

Murder Assault UrbanPop Rape clust

0 0.746988 0.654110 0.440678 0.359173 0

1 0.554217 0.746575 0.271186 0.961240 0

2 0.439759 0.852740 0.813559 0.612403 3

3 0.481928 0.496575 0.305085 0.315245 1

4 0.493976 0.791096 1.000000 0.860465 3

After this States feature was added to original dataframe

clust States Murder Assault UrbanPop Rape

0 0 Alabama 0.746988 0.654110 0.440678 0.359173

1 0 Alaska 0.554217 0.746575 0.271186 0.961240

2 3 Arizona 0.439759 0.852740 0.813559 0.612403

3 1 Arkansas 0.481928 0.496575 0.305085 0.315245

4 3 California 0.493976 0.791096 1.000000 0.860465

After this states were grouped by clusters to see if clustering has happened properly

Murder Assault UrbanPop Rape

clust

0 0.800452 0.711473 0.364407 0.445413

1 0.296084 0.298973 0.657627 0.279070

2 0.143373 0.115411 0.293220 0.111886

3 0.607430 0.725742 0.785311 0.644703

At the end this clustered dataframe was stored as a new dataset