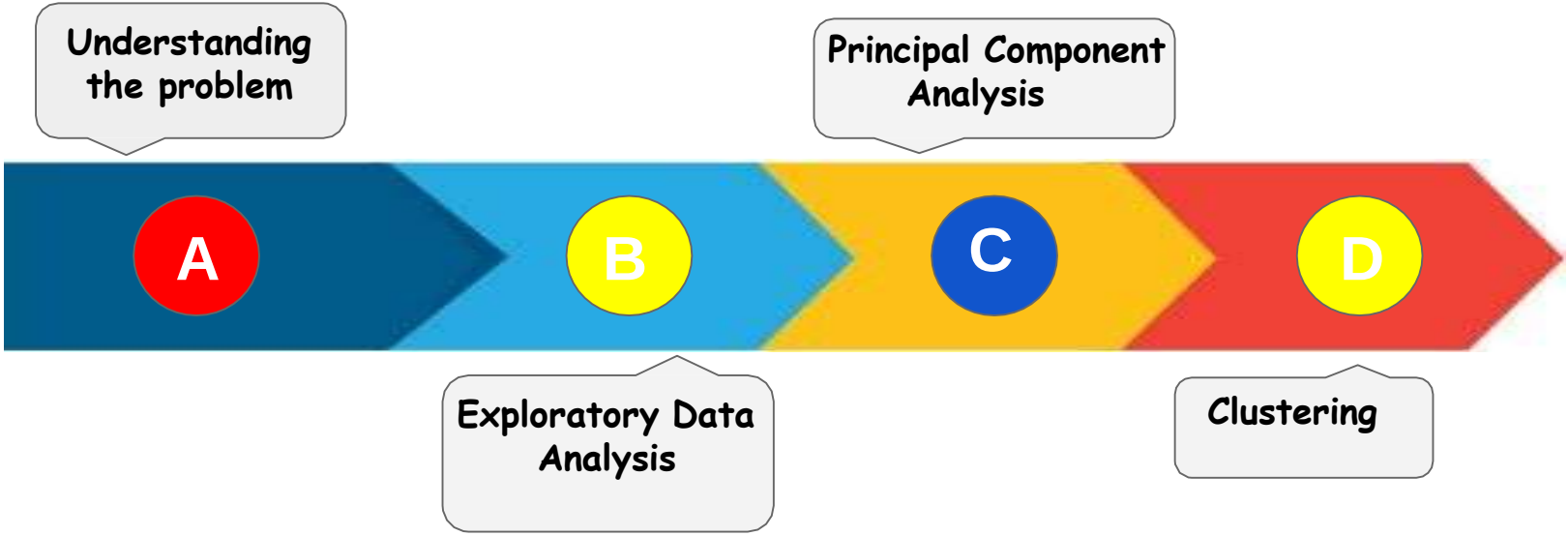
Data Analysis: Lower Back Pain Prediction Problem

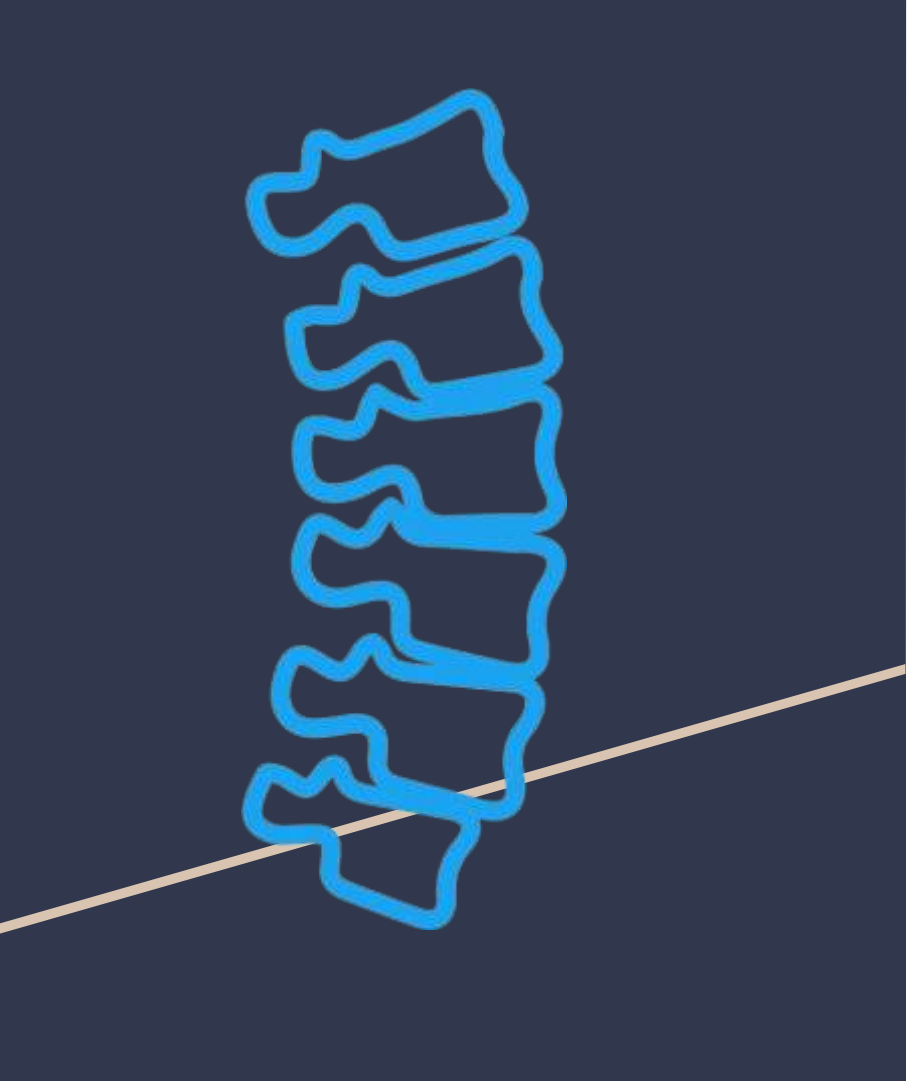
ROADMAP


Understanding the problem
A
Exploratory Data Analysis
B
Principal Component Analysis
C
Clustering
D

A Understanding the problem

This project aims to identify patients with lower back pain beforehand on the basis of spine misalignment and other factors.

We are provided with data on 12 potential causes and no dependent variable indicating towards unsupervised learning model.



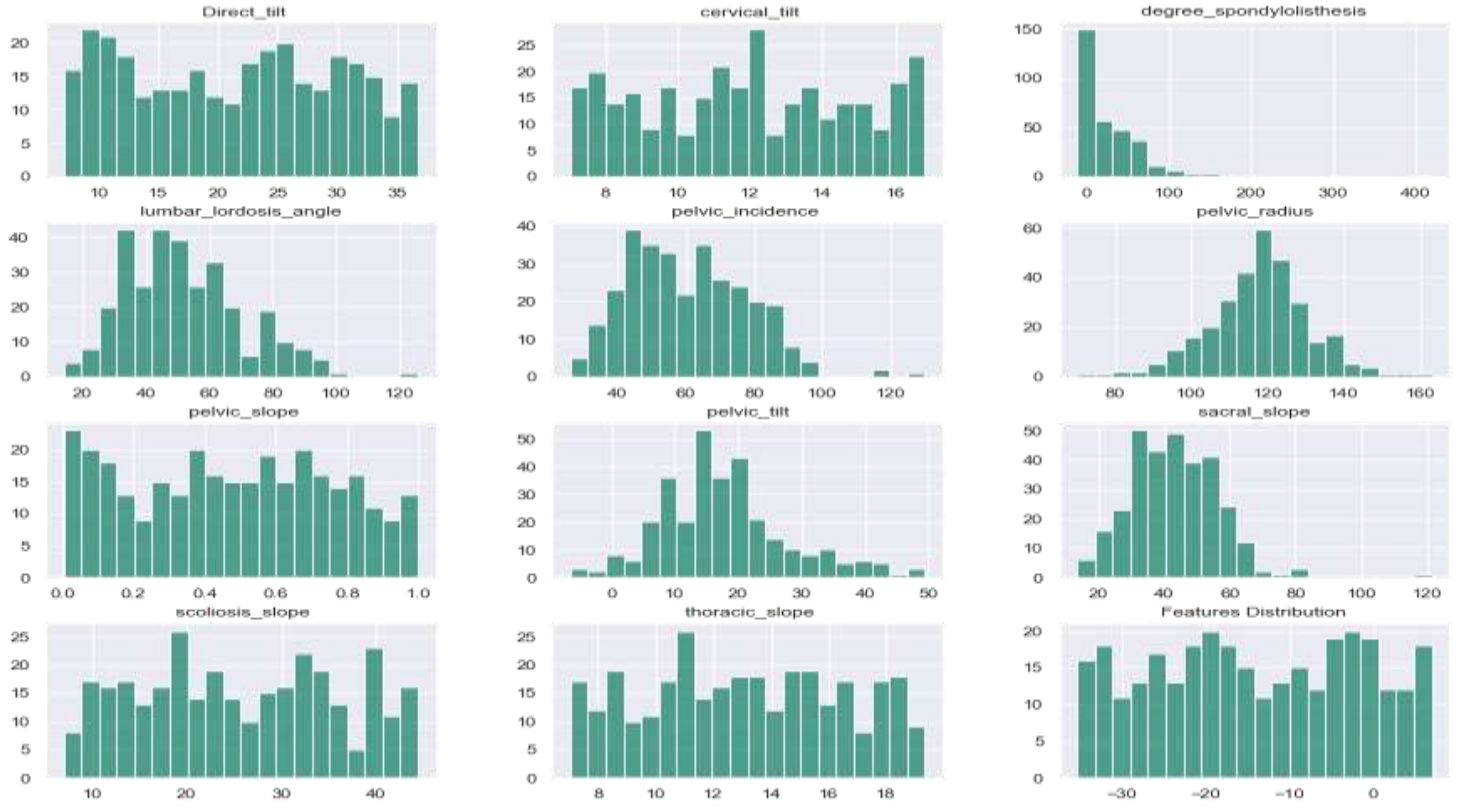
VARIABLES GIVEN IN THE PROBLEM

- Pelvic_incidences
- Pelvi_tilt
- Lumbar_lordosis_angle
- Sacral_slope
- Pelvic_radius
- Degree_spondylolisthesis
- Pelvic_slope
- Direct_tilt
- Thoracic_slope
- Cervical_tilt
- Sacrum_angle
- Scoliosis_slope

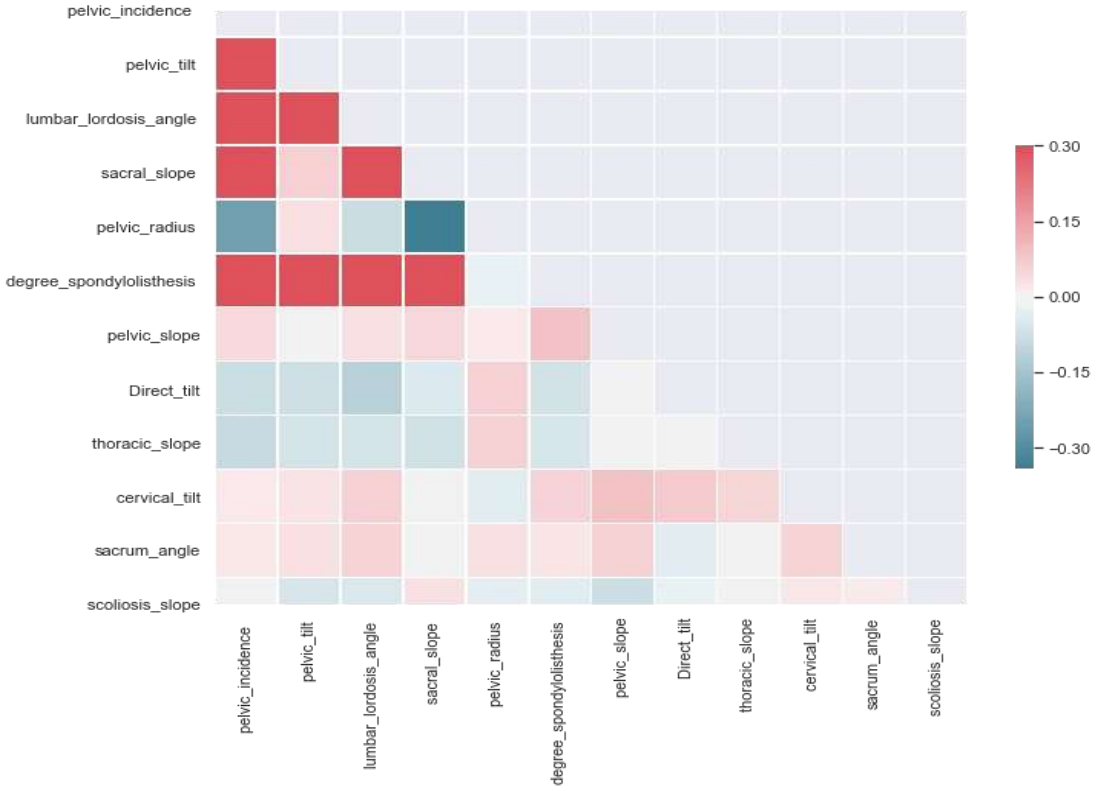
B

Exploratory Data Analysis

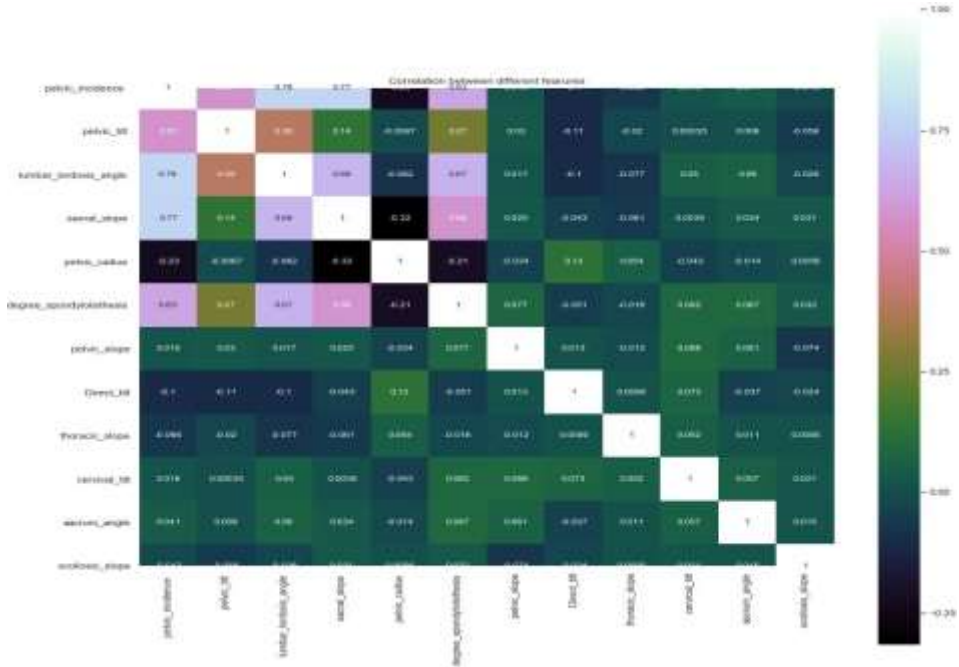
Distribution Of Features



Correlation Matrix

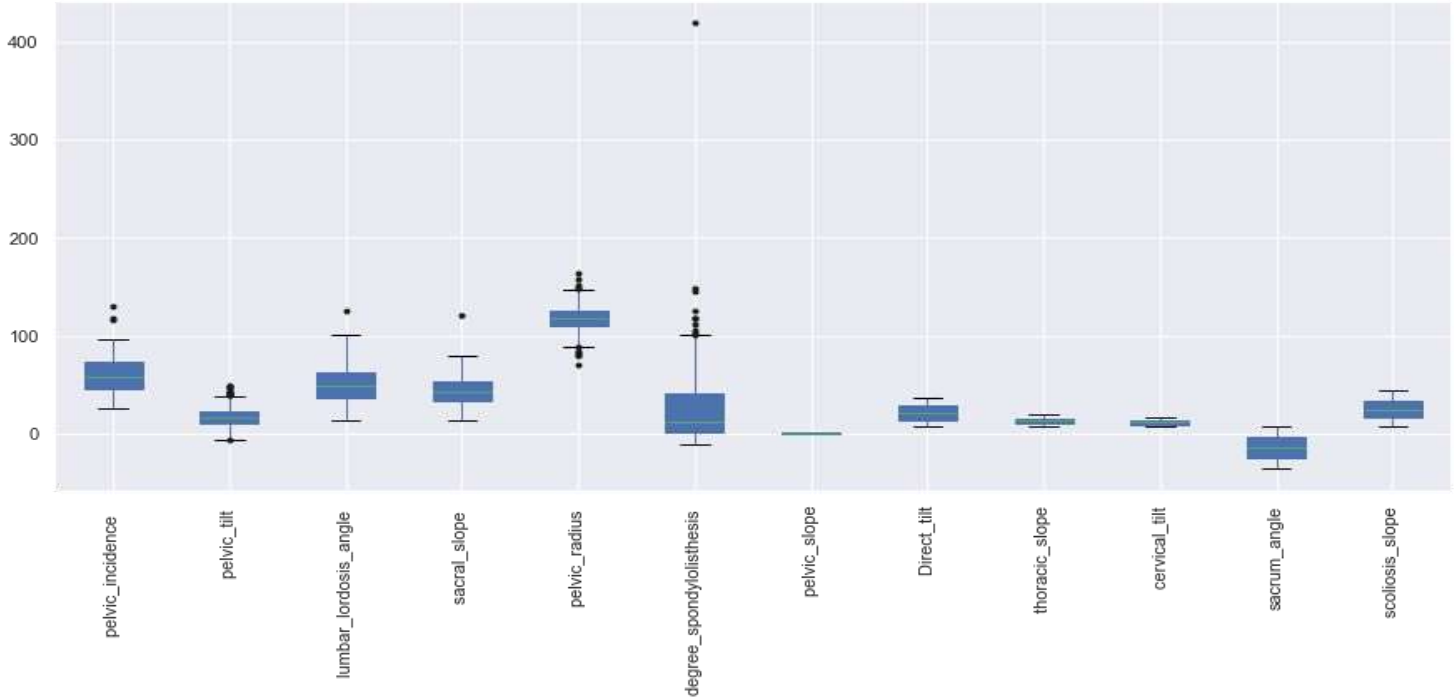


Correlation of different features/variables



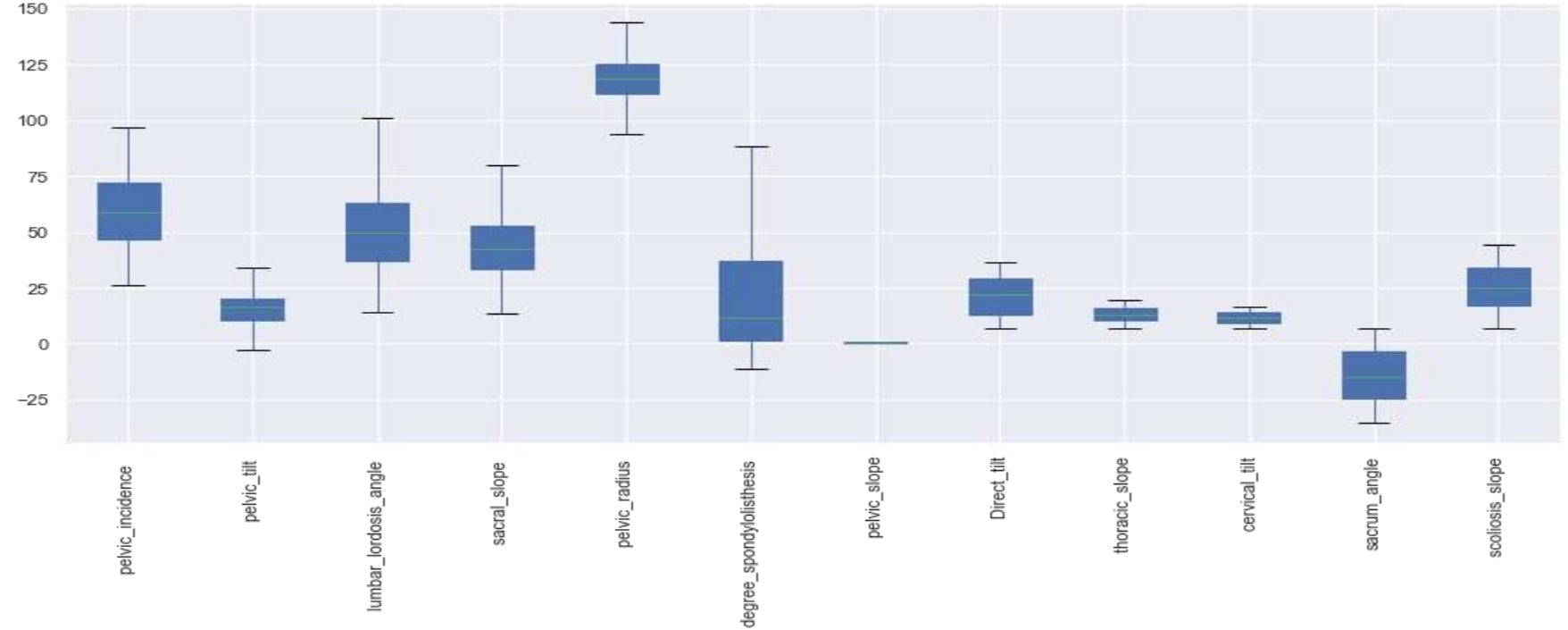
* First six variables have high correlation amongst them

Boxplot with outliers



*Values greater than 75 percentile and less than 25 percentile are taken as outliers.
***Pelvic tilt, pelvic radius and degree spondylolisthesis** have a lot of outliers.

Boxplot after removing Outliers



* Note : The outliers have been replaced by the median.

C

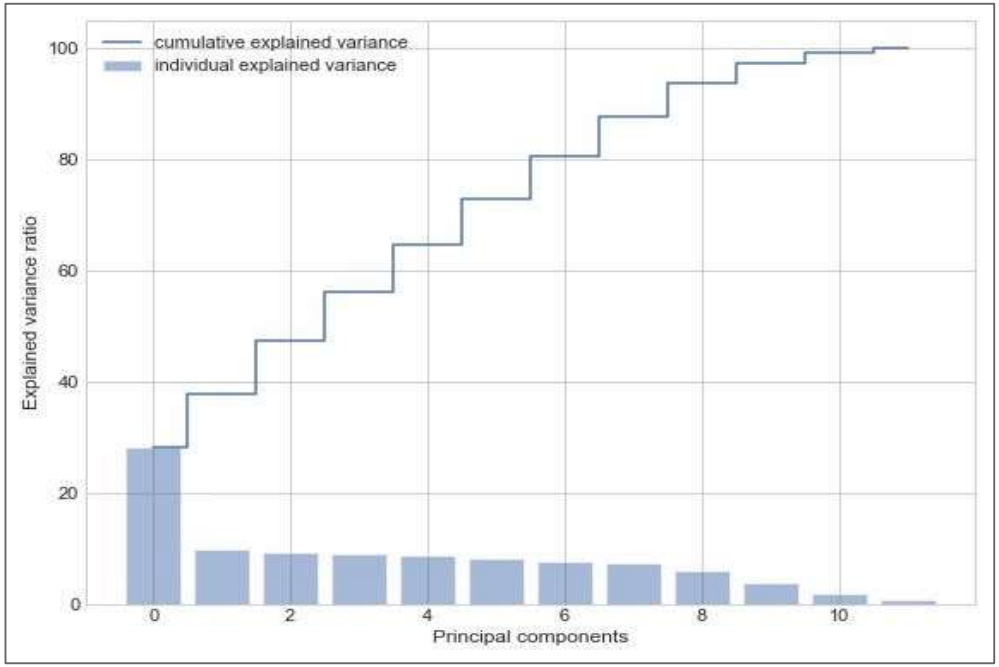
Principal Component Analysis

Eigenvalues

- Eigenvectors that have the least eigenvalues hold very less information and can be eliminated altogether.
- Eigenvalues in descending order:

3.3832351935916263
1.177002380565475
1.1247875308974582
1.0731624629895875
1.0383153061016597
0.9744020922082253
0.9105411146785651
0.8686858544435843
0.7239578891507754
0.4353419429426945
0.2271490458176812
0.10225413806897596

Individual and Cumulative Explained Variance




```
explained_variance=pca.explained_variance_ratio_
```

```
explained_variance.sum()
```

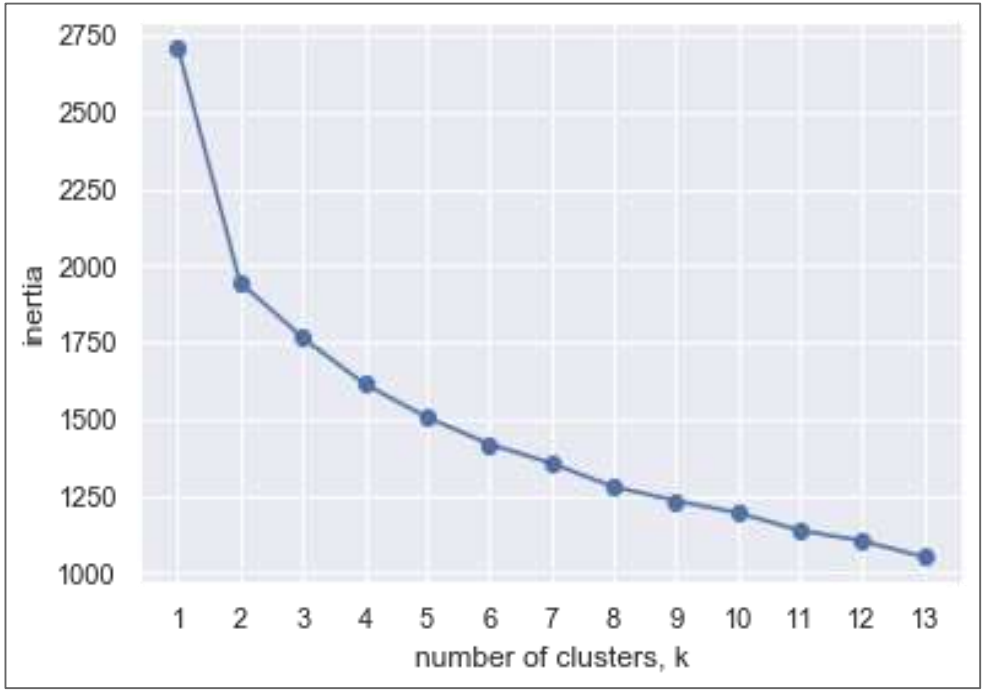
```
0.8006612912359714
```

*80% of the data is explained by 6 components.

D

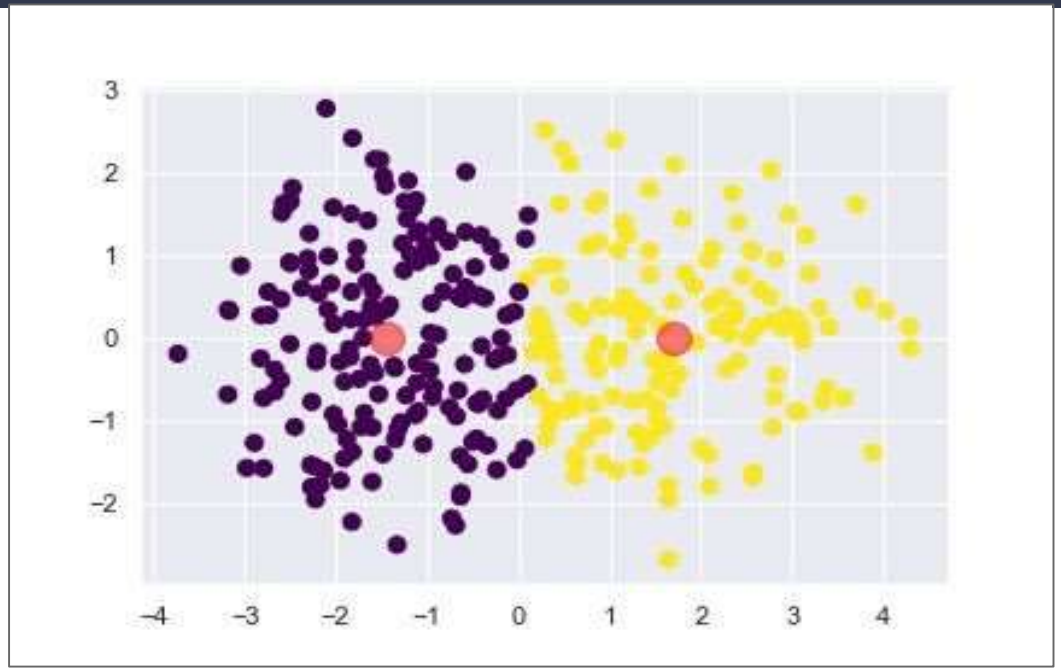
Clustering

Elbow method for K Means Clustering



*The curve bends sharply at k=2 and then declines at a constant rate.

Visualization of K Means Clusters with 2 centroids



*Intuitively, one represents abnormal cases and the other normal cases

CONCLUSION:

After performing Exploratory Data Analysis,Principal Component Analysis and Clustering the observations are as follows:

- The optimal number of clusters are 2 each of which intuitively represent normal and abnormal cases.
- It is impossible to define which cluster represents which case without the label information.
- Since the actual data is not available, we cannot judge the accuracy of the model.

*The dataset is believed to be incomplete and a concrete interpretation is hence unfeasible.

THANK YOU