Graduate Rotational Internship Program : June 2021

The Sparks Foundation
Data Science & Business Analytics Tasks - 2

Predicting Optimum Clusters for Iris using Unsupervised ML

Owner: Reema Lad

Friday, June 18, 2021

Project Case

#2

Prediction using Unsupervised ML

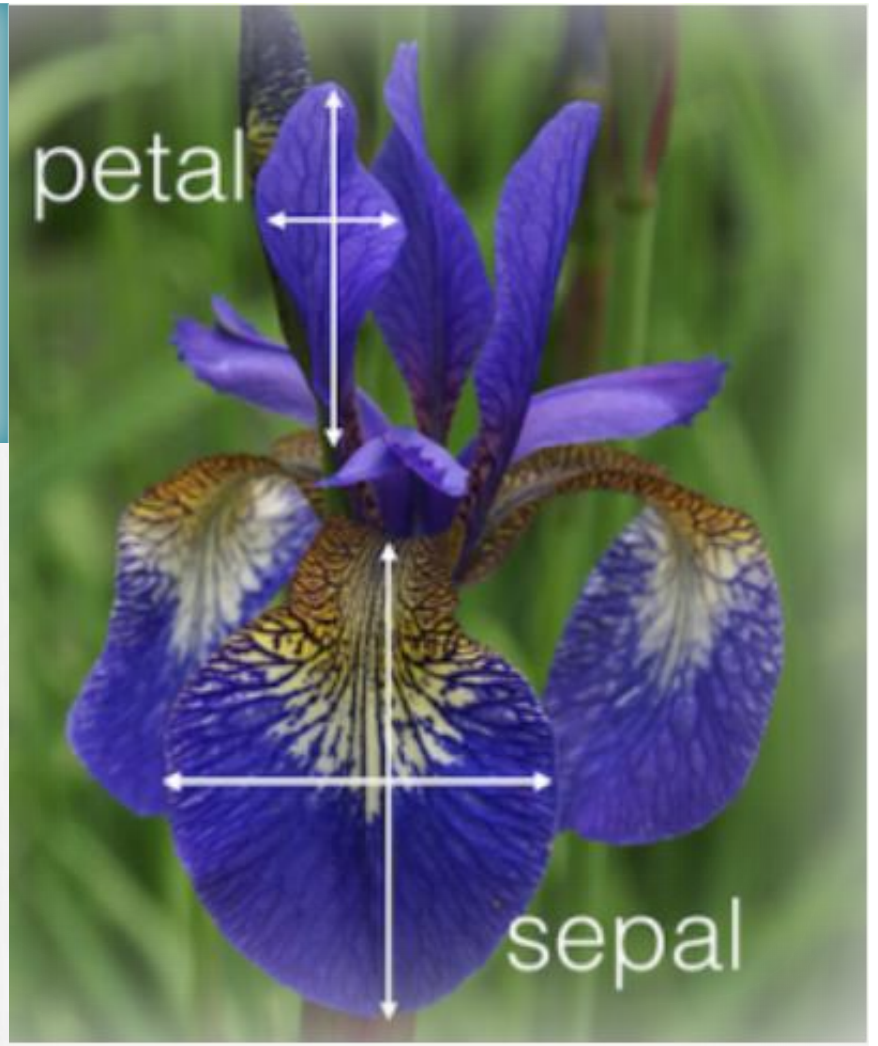
(Level – Beginner)

- From the given 'Iris' dataset, predict the optimum number of clusters and represent it visually.
- Use R or Python or perform this task

Iris Versicolor

Iris Setosa

Iris Virginica



Dataset Details

Problem Statement

From the given 'Iris' dataset, predict the optimum number of clusters and represent it visually

Data Dictionary

As Id is an unique reference number alloted to each observation, its insignificant for our study, hence dropped

Variable	Definition
Id	Hours Studied by student
SepalLengthCm	Length of sepals in centimeters
SepalWidthCm	Width of sepals in centimeters
PetalLengthCm	Length of petal in centimeters
PetalWidthCm	Width of petal in centimeters
Species	Species of Iris - Setosa, Virginica and Versicolor

Process

- Import Libraries
- Load Data
- Reading Raw Data
- Visualization, UniVariate - BiVariate Analysis, EDA
- Model Building
- Cluster Visualization

Data Insights

* **150 rows with 5 columns**

* **No missing data**

* **No Duplicate rows**

* **Target Variable : Species**

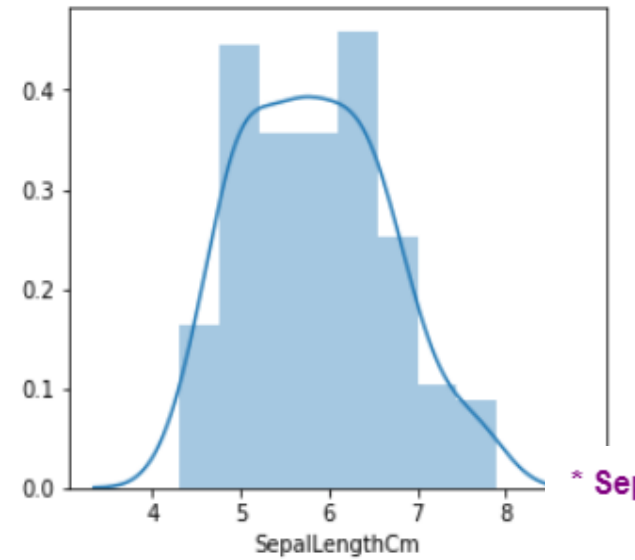
* **Predictor Variable : SepalLength, SepalWidth, PetalLength, PetalWidth**

* **Target Variable has 3 classes with equally distributed number of samples; so data is balanced.**

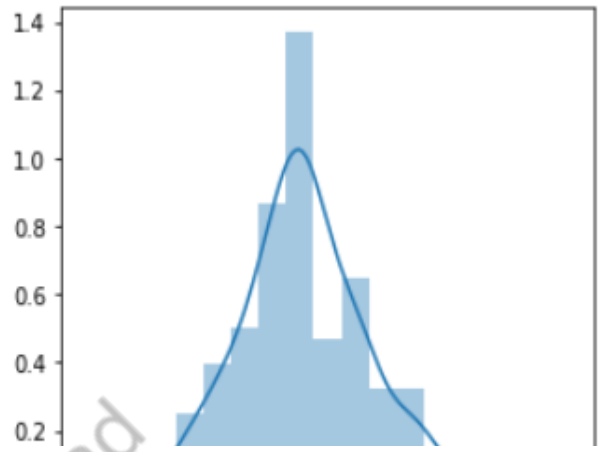
* **As observed visually basis value count, we see three classes; would try to establish the same using Clustering ML Models**

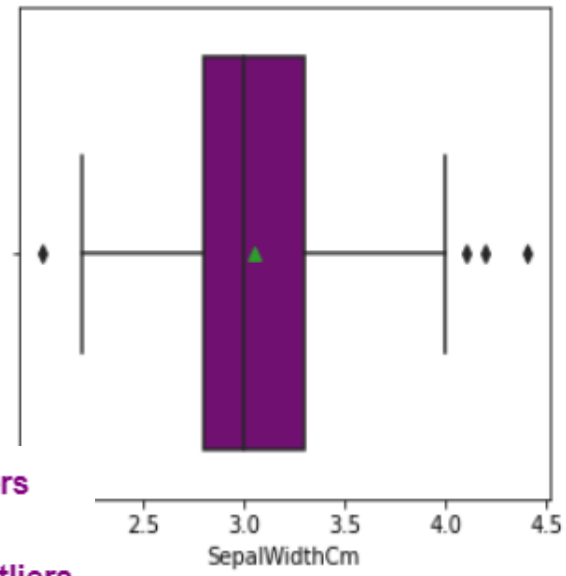
Understanding Data & EDA, Insights on Variables

Univariate Analysis of SepalLengthCm



Univariate Analysis of SepalWidthCm



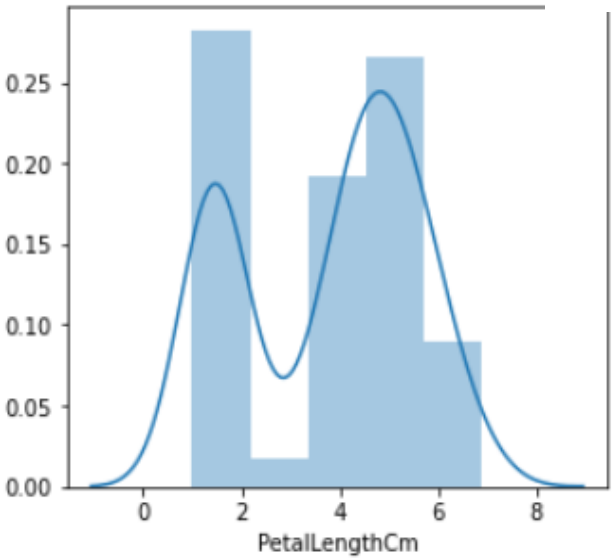


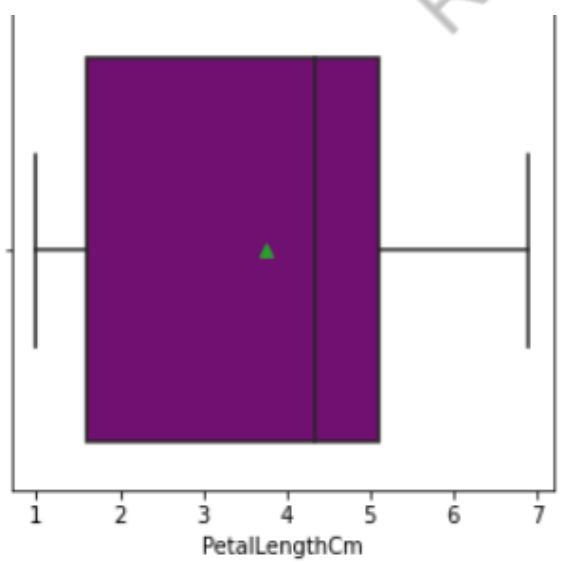
* Sepal Length and Sepal Width both are normally distributed; however later has certain outliers

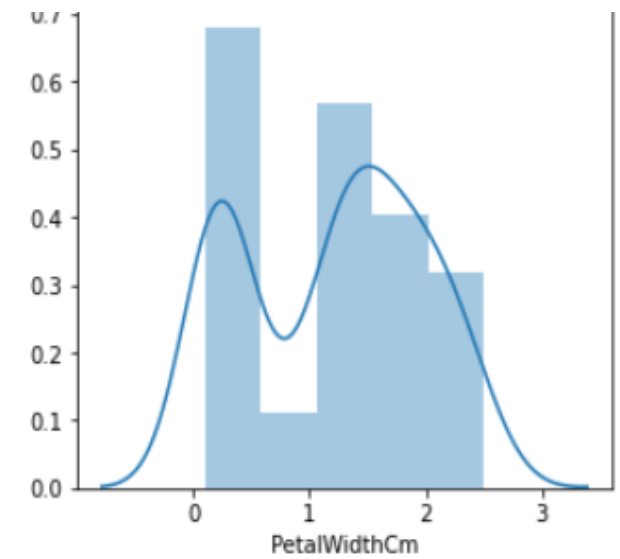
* Petal Length and Petal Width both do not follow normal distribution however there are no outliers

* To understand further, we need to analyze certain more points

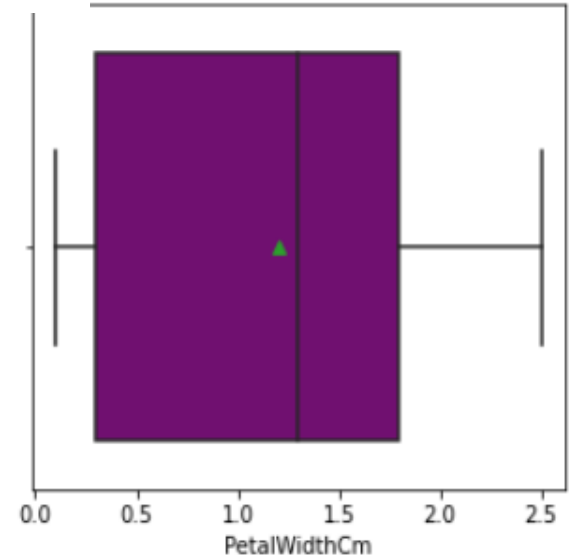
Univariate A





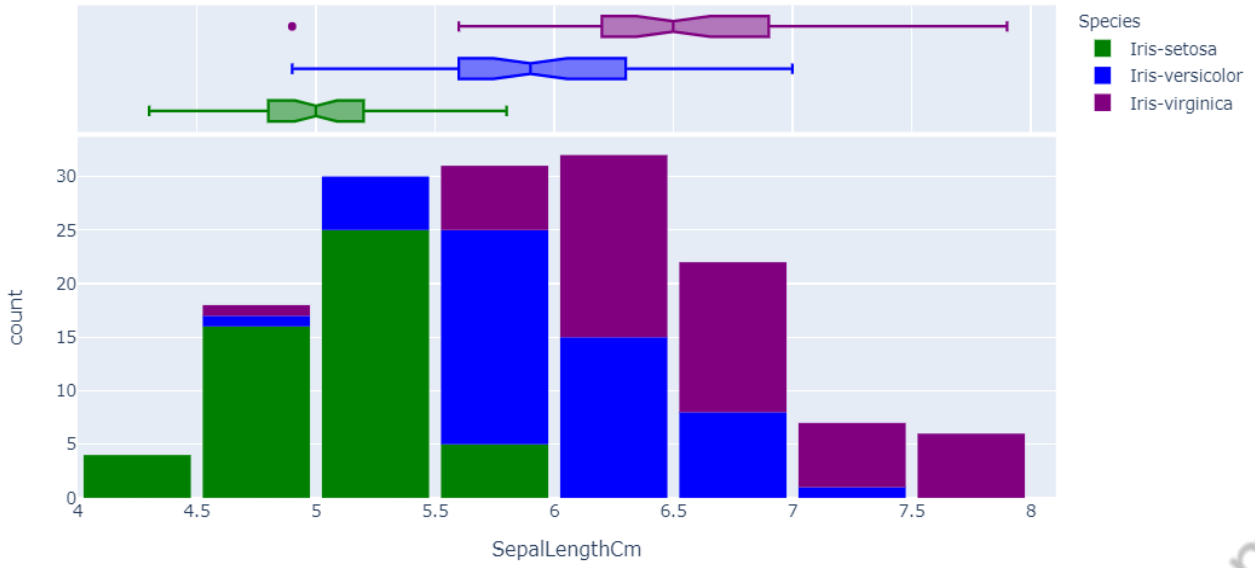


VidthCm

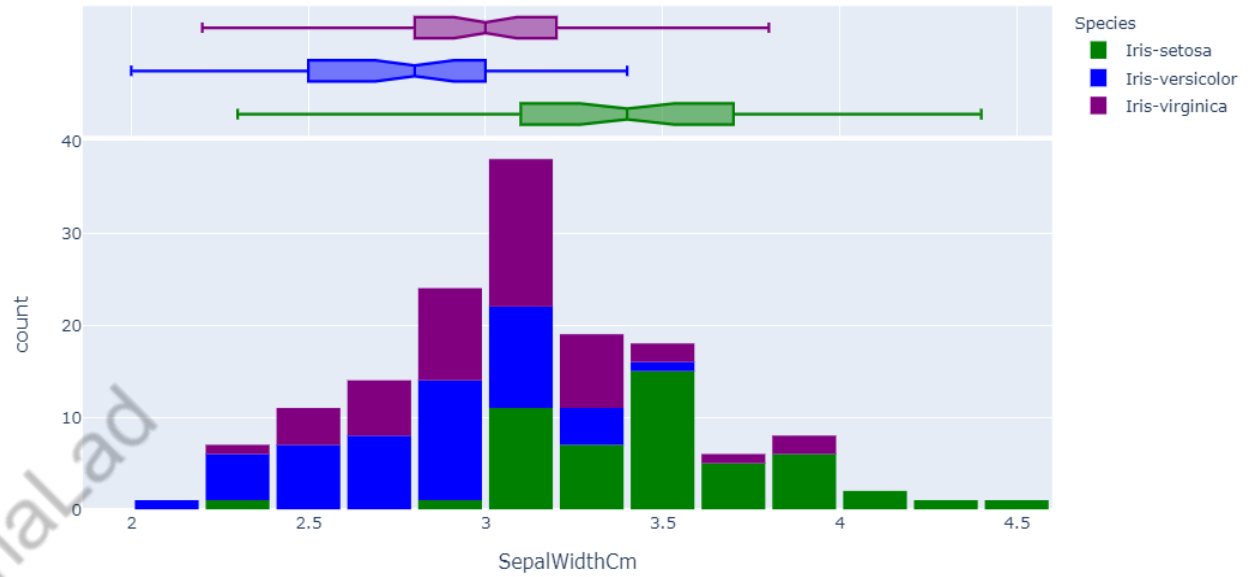


Understanding Data & EDA, Insights on Variables

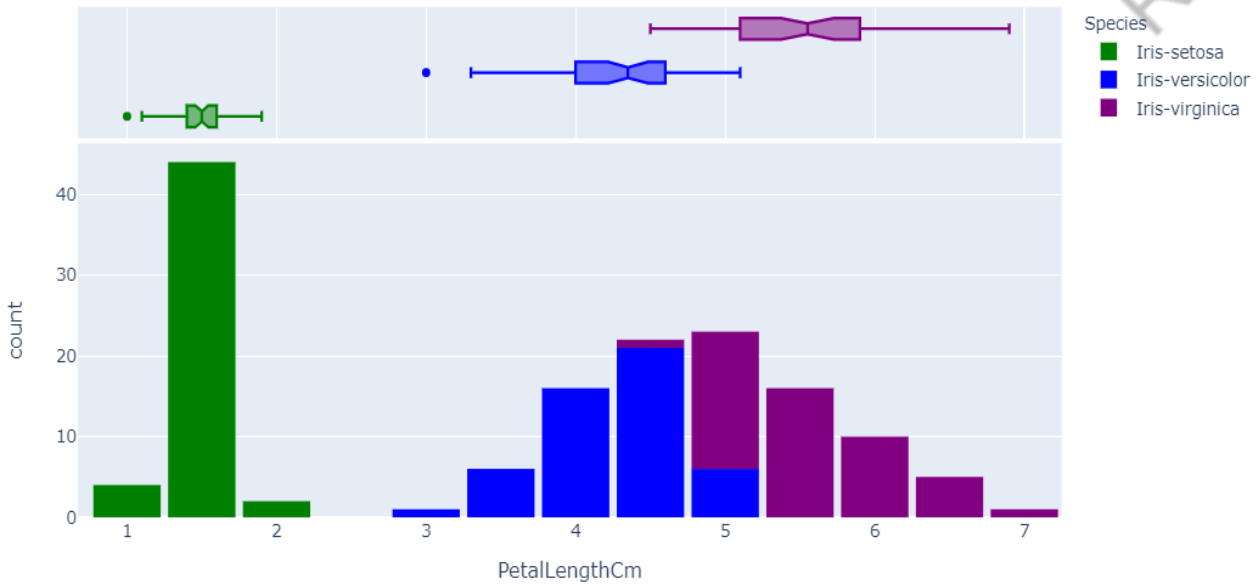
Study of Sepal Length - Species Wise



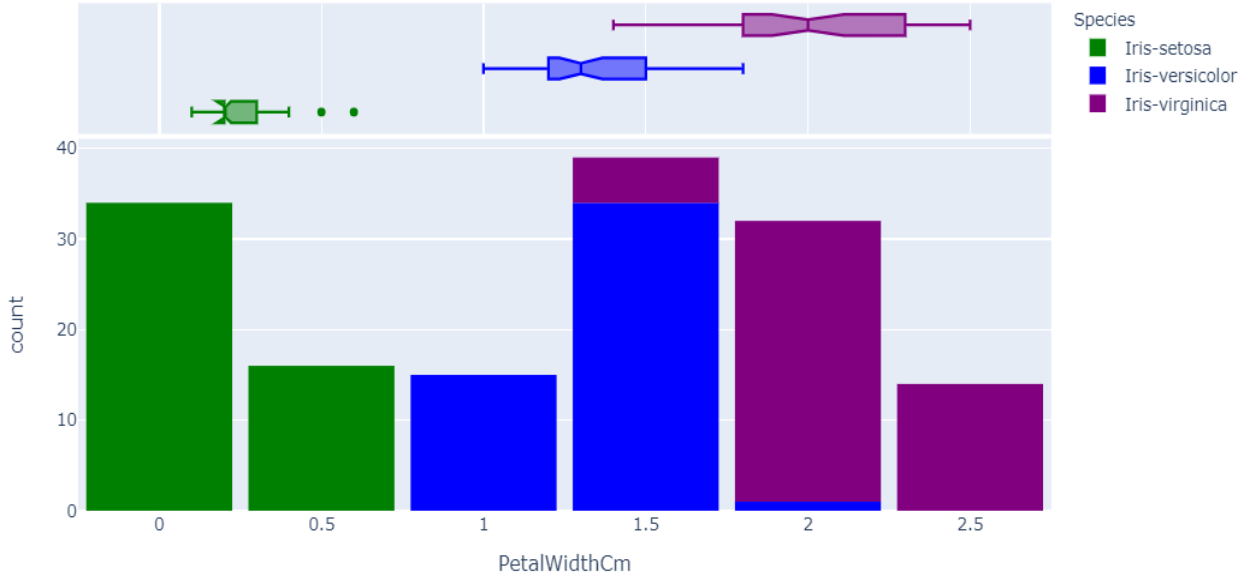
Study of Sepal Width - Species Wise



Study of Petal Length - Species Wise



Study of Petal Witdth - Species Wise



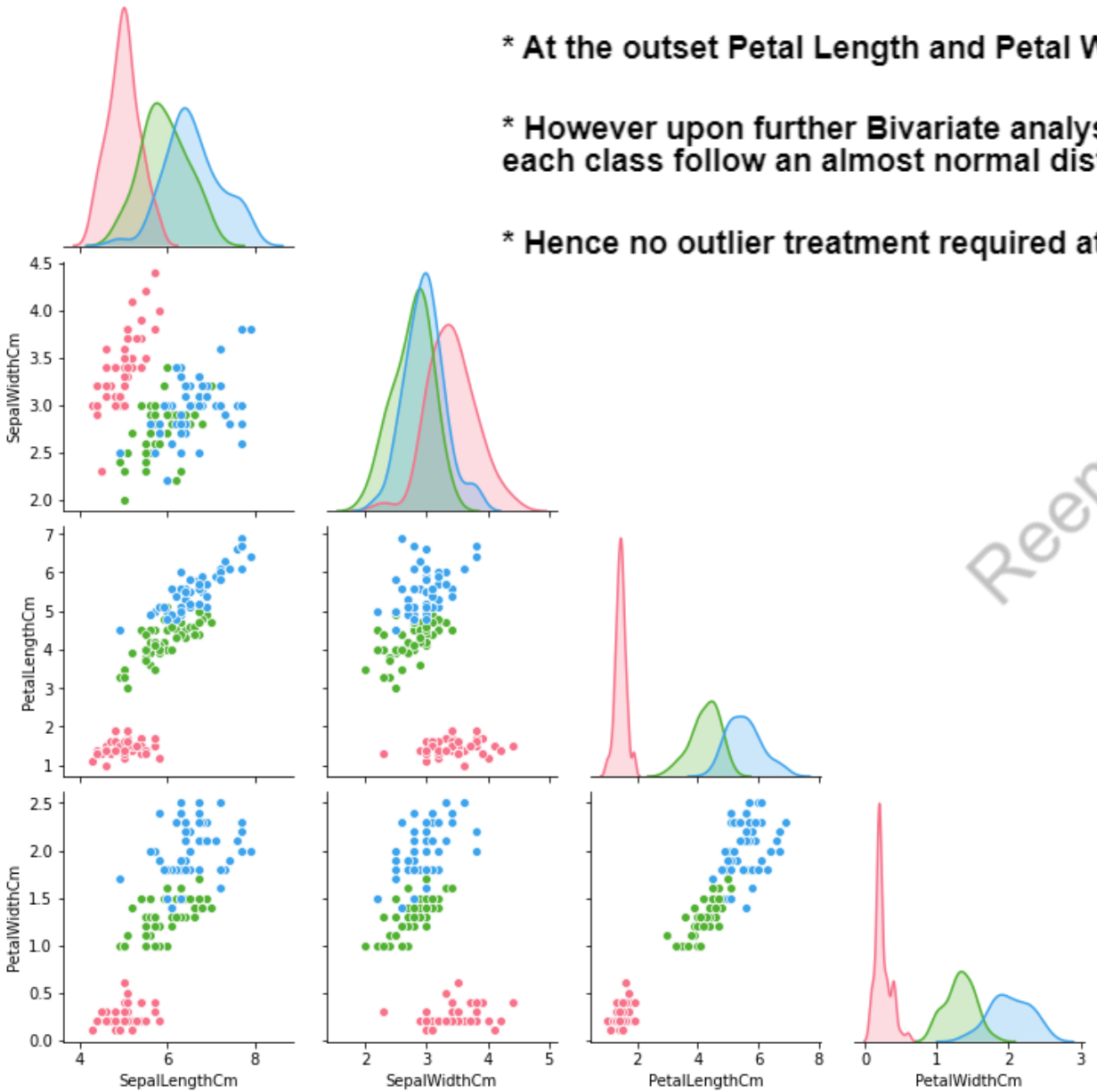
Understanding Data & EDA, Insights on Variables

Insights - EDA & Data Pre-processing

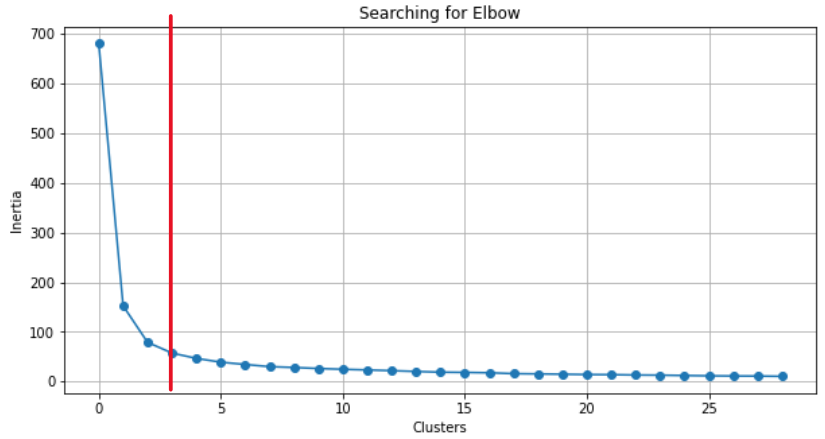
* At the outset Petal Length and Petal Width reflected outliers

* However upon further Bivariate analysis its identified that when seen as per the existing classes, each feature of each class follow an almost normal distribution

* Hence no outlier treatment required at this point of time



Model – K-Means Clustering




```
------------------------------------------------------------
 Original Classes & Values
------------------------------------------------------------
 Iris-virginica     50
Iris-setosa         50
Iris-versicolor     50
Name: Species, dtype: int64

------------------------------------------------------------
 Classes & Values Classified by KMeans Clustering
------------------------------------------------------------
 Iris-Versicolor    62
Iris-Sentosa        50
Iris-Virginica      38
Name: KMCLabelName, dtyp
 Total Values :  150
```

Optimal n_clusters determined to be 3 - as beyond that there is no significance change in inertia

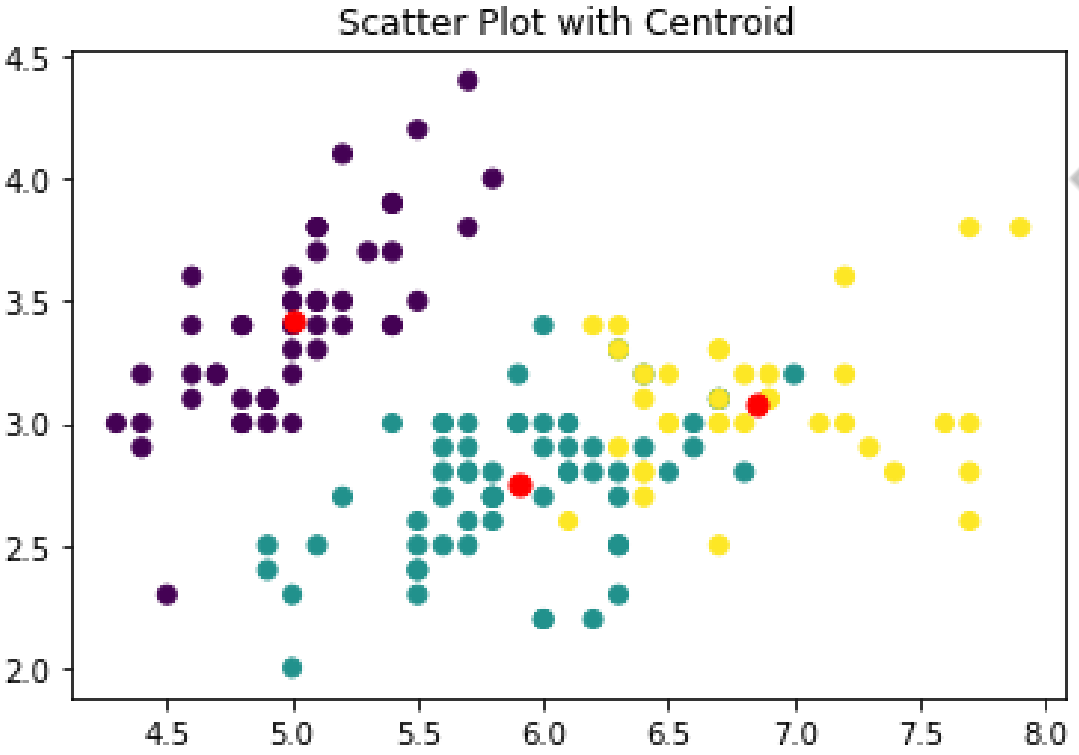
Model Insights

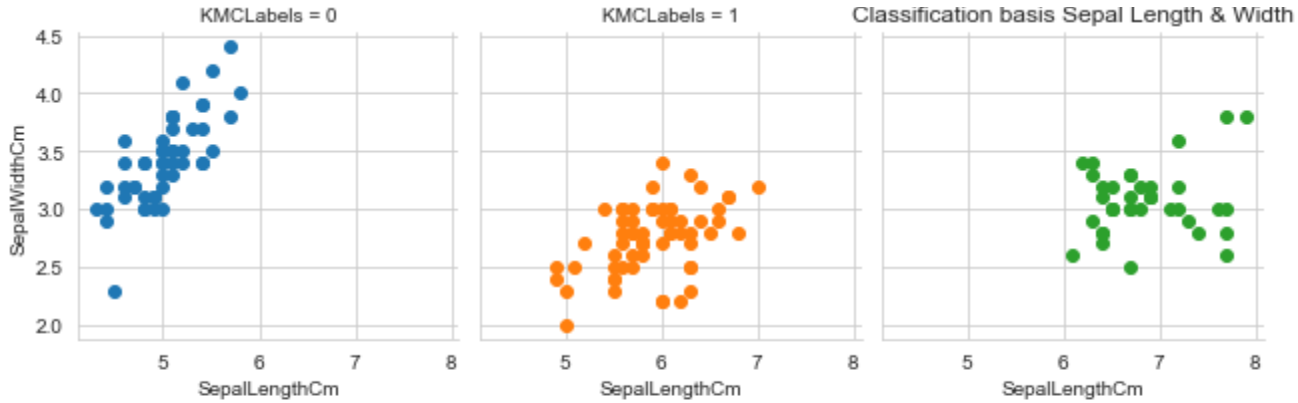
* With the Elbow Cut method identified optimal value of K = 3

* Hence three clusters are formed which is similar to our original data clusters

* However slight error in identifying the correct class for Versicolor & Virginica. Sentosa class is perfectly clustered

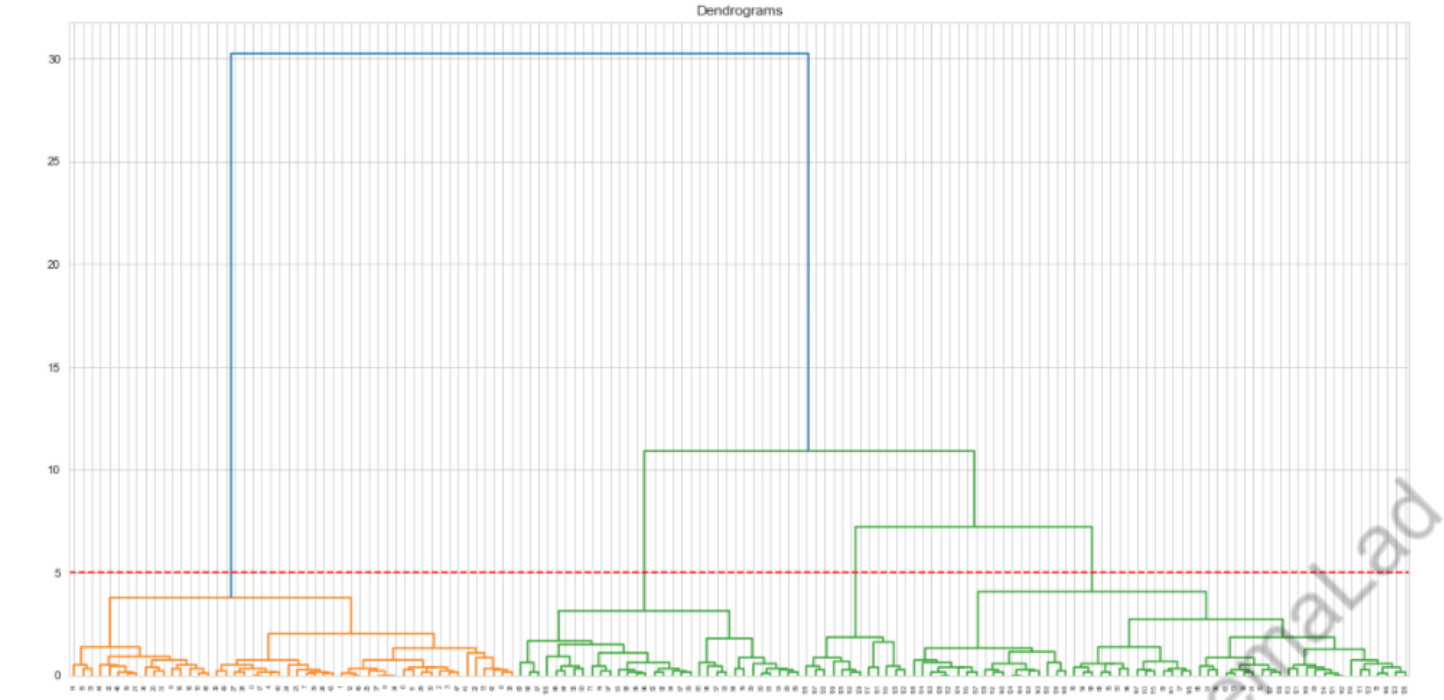
* Clusters are almost well seperated, for little overlap in Versicolor and Virginica



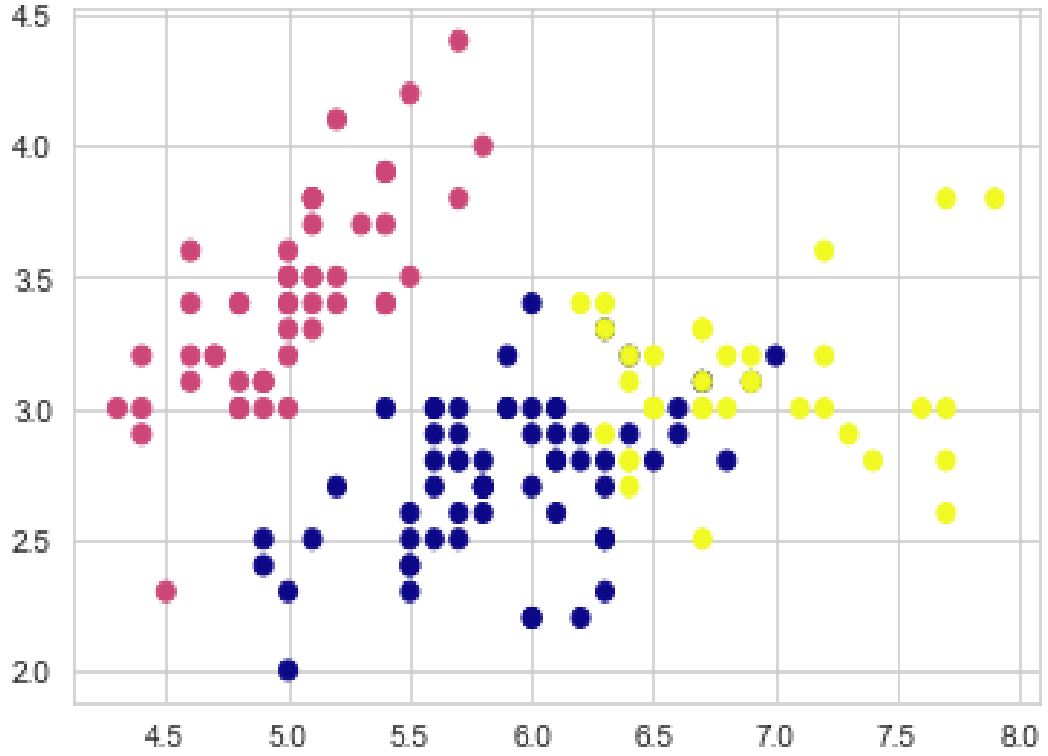


Cluster 0: Iris-Setosa
Cluster 1: Iris-Versicolor
Cluster 2: Iris-Virginica

Model – Hierarchical Clustering



Optimal n_clusters determined to be 3 - as below that there is no significant change in cluster distance



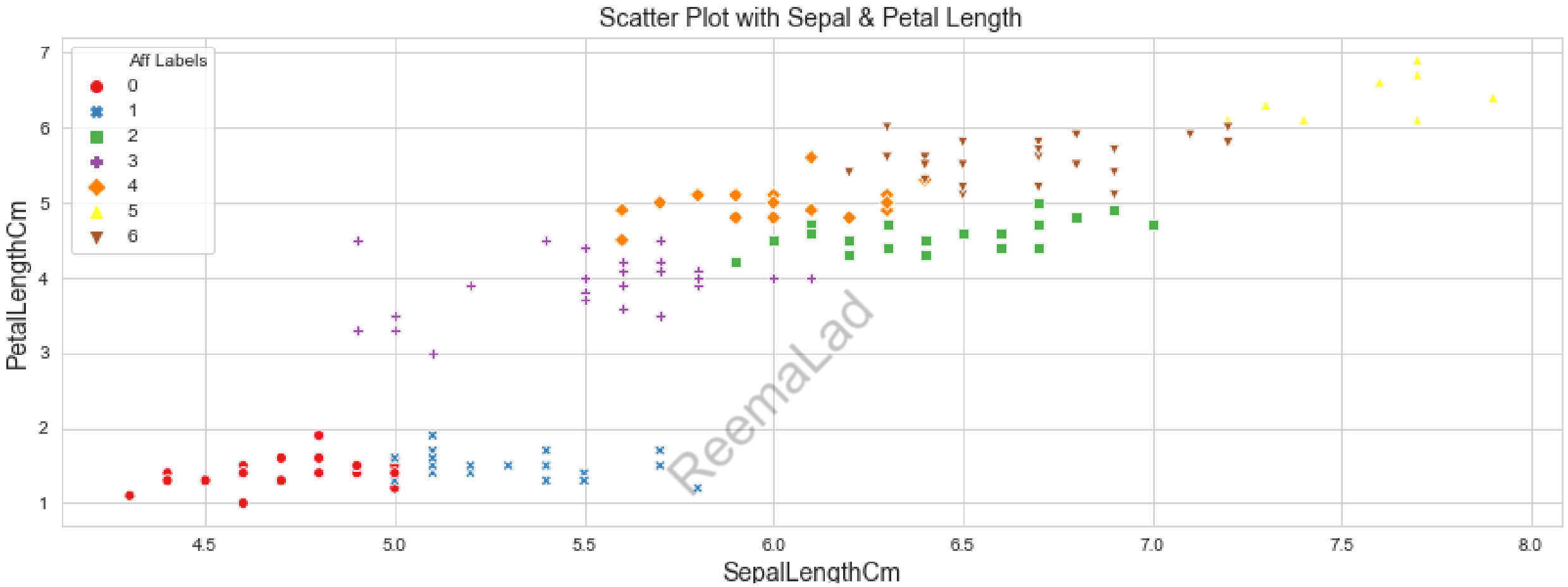
Model Insights

* Optimal number of clusters basis dendogram is 3

* Here too three classes are well classified with slight errors in Sentosa and Virginica.

* In this method Versicolor is identified accurately

Model – Affinity Propagation



Model Insights

* Optimal number of clusters as concluded by the model are 6

* Model is able to classify the classes with certain more depth and better boundries, hence distingishuing and demarking in to completely seperate clusters without any overlap; which inturn creates more minuscule but clearly demarked clusters.

Thank You