Data description:

I will be using the famous Iris dataset. It can be found in SAS help folder of the SAS library. The dataset is the measurement of 3 different species. It takes the measurements of Sepal Length, Sepal Width, Petal Length and Petal Width of three different species. The differences are used to make clusters to further affirm that these species are in fact different.

Columns

- Select all
- Species
- SepalLength
- SepalWidth
- PetalLength
- PetalWidth

| Property | Value |
|---|---|
| Label | Iris Species |
| Name | Species |
| Length | 10 |
| Type | Char |
| Format | |
| Informat | |

Total rows: 150 Total columns: 5 Rows 1-100

| | SepalLength | SepalWidth | PetalLength | PetalWidth |
|---|---|---|---|---|
| | 50 | 33 | 14 | 2 |
| | 46 | 34 | 14 | 3 |
| | 46 | 36 | 10 | 2 |
| | 51 | 33 | 17 | 5 |
| | 55 | 35 | 13 | 2 |
| | 48 | 31 | 16 | 2 |
| | 52 | 34 | 14 | 2 |
| | 49 | 36 | 14 | 1 |
| | 44 | 32 | 13 | 2 |
| 0 | 50 | 35 | 16 | 6 |
| 1 | 44 | 30 | 13 | 2 |
| 2 | 47 | 32 | 16 | 2 |
| 3 | 48 | 30 | 14 | 3 |
| 4 | 51 | 38 | 16 | 2 |
| 5 | 48 | 34 | 19 | 2 |
| 6 | 50 | 30 | 16 | 2 |
| 7 | 50 | 32 | 12 | 2 |
| 8 | 43 | 30 | 11 | 1 |
| 9 | 58 | 40 | 12 | 2 |

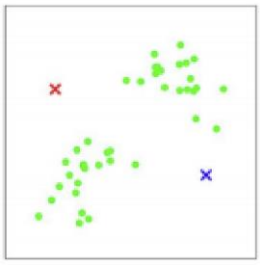
**Objective of analysis**:

The Objective of analysis is use K-means to find the correct cluster count with distinct mean values. We already know how many species there are. K-means will affirm our knowledge.

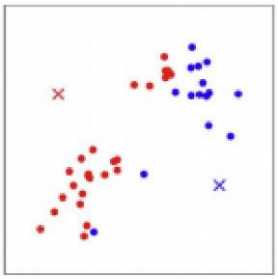
**Introduction to K-means clustering**:

K-means clustering assumes that different groups of people, things, attributes, and objects can be separated using an iterative algorithm that separates groups by their mean.

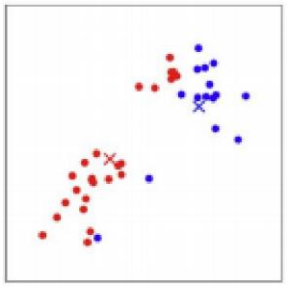
The algorithm finds an 1,2,3, or 4 etc initial random coordinates on the graph.



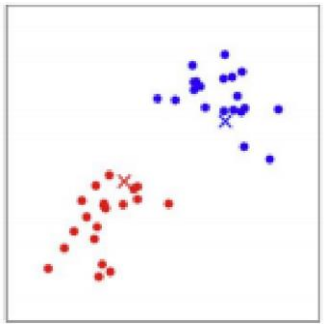
It calculates the nearest cluster to those random points.



Then it the initial random point iterates toward the center of that the cluster using the average of the vectors. See below how it iterates toward the center of the cluster.



Then the algorithm re-finds the closest points on the graph after moving to the center. Thus a new cluster is made below.



The algorithm does this until the variation is minimum. For example: For two two distinct groups, one cluster will have a lot of variation. 2 Clusters will have the right amount of variation. 3 clusters will have even less variation, but the rate of variation decrease will go down from 2 groups. That's why 2 groups is sufficient for two distinct groups.

**How you implement the methods**:

Standardizing of variables should be done beforehand with mean 0 standard deviation 1.It can be done using: method=std. I believe Standardizing makes variables and models more comparable or relative. Values should be unique. Really influential outliers can be dropped because they might make the algorithm believe clusters should be added.

Sometimes principle components are used to make the clusters more visually interpretable.

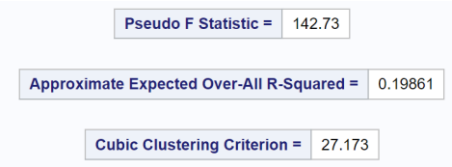
One of the ways to see if algorithms is working well is to find the cluster count that maximizes Cubic clustering criterion(CCC) and Psudo-F stat.

**Complete interpretations of the results and conclusions**:

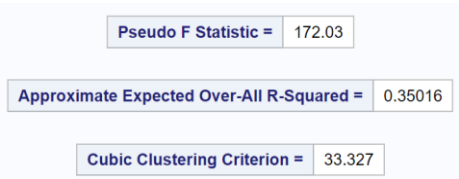
One cluster gave me F stat of 0 and Cubic cluster criterion of 0. Two cluster gave me F stat of 142.7, CCC was 27.17. Three clusters: F stat was 172.03 while CCC was 33.32. Four clusters output was for F stat was 170.65 while CCC was 31.79. Notice that fall after three clusters. This means three clusters is sufficient.

One cluster:

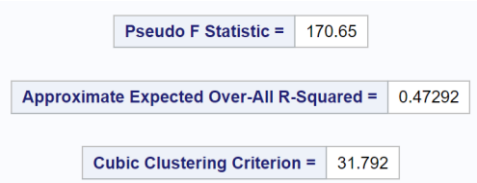

Two clusters



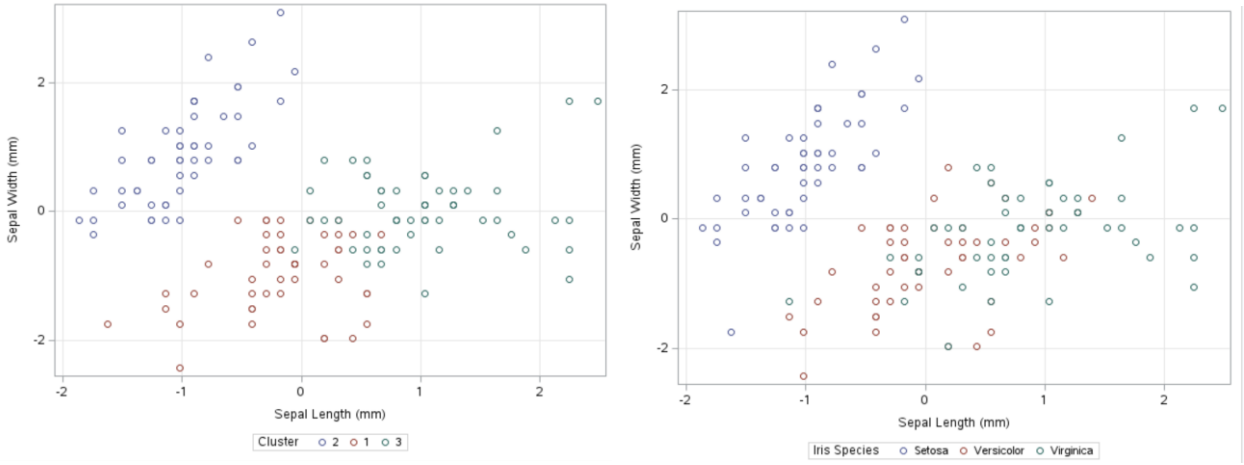

Three clusters gave me:




Four clusters gave me:



This is the scatter plot of the three clusters. Right scatter plot is correct cluster.

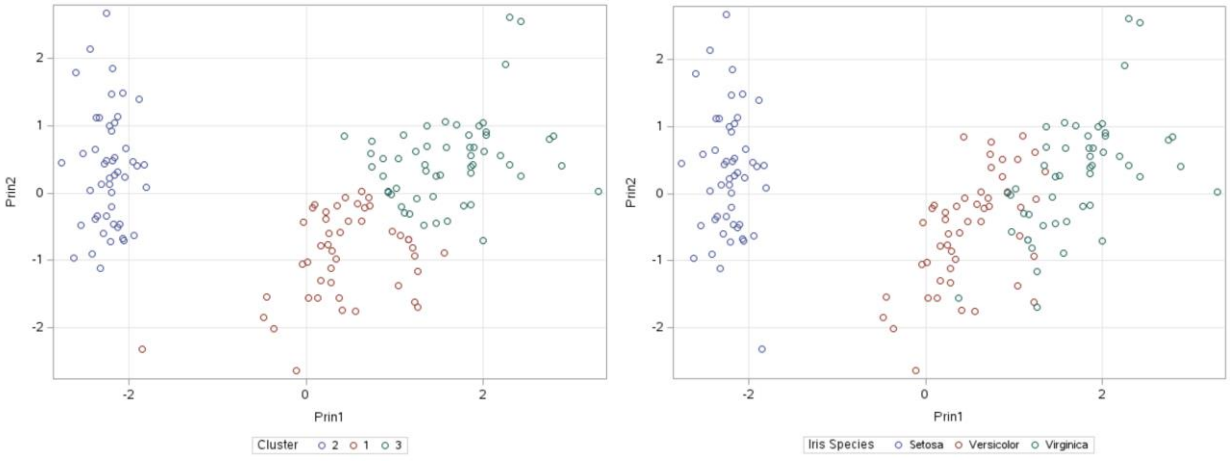



I am going to use PCA to make the three clusters more visually interpretable for two dimensions. After running PCA code I found that two PCAs can be used to explain 95% of the variation in the K- means problem.

**Eigenvalues of the Correlation Matrix**

| | Eigenvalue | Difference | Proportion | Cumulative |
|---|---|---|---|---|
| 1 | 2.91849782 | 2.00446735 | 0.7296 | 0.7296 |
| 2 | 0.91403047 | 0.76727360 | 0.2285 | 0.9581 |
| 3 | 0.14675688 | 0.12604204 | 0.0367 | 0.9948 |
| 4 | 0.02071484 | | 0.0052 | 1.0000 |

Scatter plot after using two principle components. The right scatter plot is the correct species.



**Appendix**:

K-means

```
ods noproctitle;

proc stdize data=SASHELP.IRIS out=Work._std_ method=std;
    var SepalLength SepalWidth PetalLength PetalWidth;
run;

proc fastclus data=Work._std_ maxclusters=3 out=work.Fastclus_scores0001;
    var SepalLength SepalWidth PetalLength PetalWidth;
run;

proc delete data=Work._std_;
run;
```

PCA code for K-means:

```
ods noproctitle;
ods graphics / imagemap=on;

proc princomp data=WORK.FASTCLUS_SCORES0001 plots(only)=(scree)
        out=work.Princomp_scores;
    var SepalLength SepalWidth PetalLength PetalWidth;
run;
```

Code for scatter plot of two Principle components.

```
ods graphics / reset width=6.4in height=4.8in imagemap;

proc sgplot data=WORK.PRINCOMP_SCORES;
    scatter x=Prin1 y=Prin2 / group=CLUSTER;
    xaxis grid;
    yaxis grid;
run;

ods graphics / reset;
```