**Question 0: Salary Prediction**

**Preliminary Visualization**

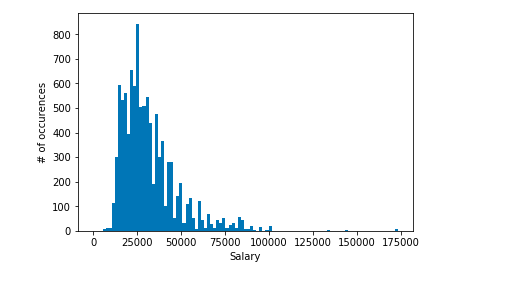
salary = originalsalary

plt.hist(originalsalary['SalaryNormalized'], bins=[n\*1750 **for** n **in** range(100)], histtype='stepfilled')

plt.xlabel("Salary")

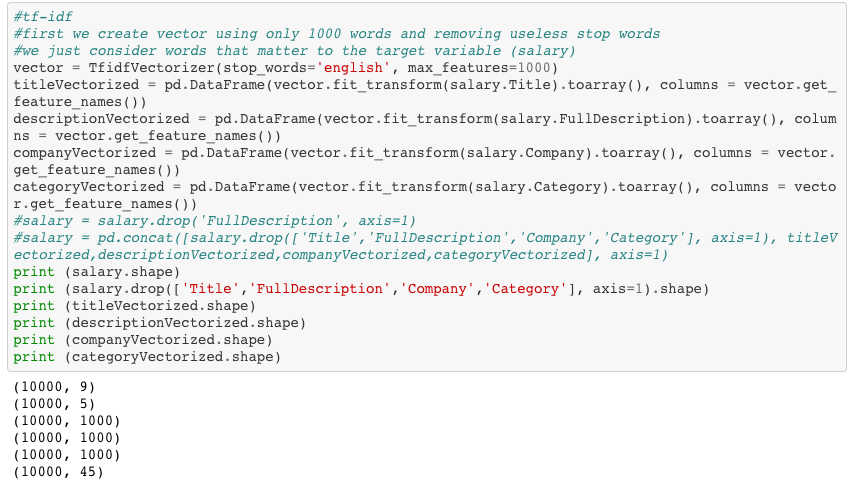
plt.ylabel("# of occurences")

plt.show()



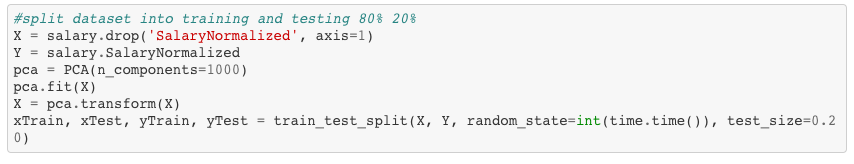
**Data Preprocessing**

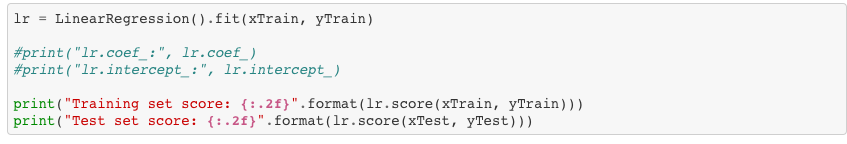
****

****

****

**Training and Testing Models**

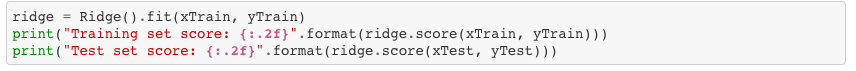
**  
Linear Regression**

****

Training set score: 0.73

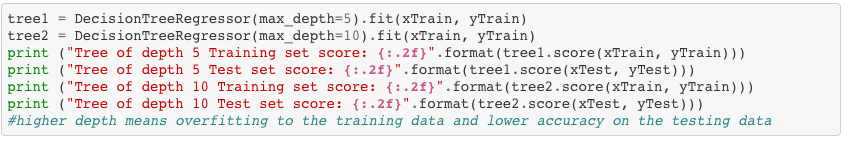
Test set score: 0.60

**Ridge Regression**

****

Training set score: 0.79

Test set score: 0.60 **Decision Trees**

****

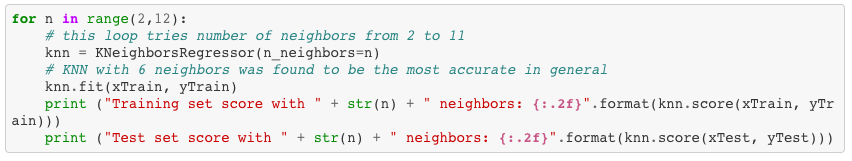
Tree of depth 5 Training set score: 0.50

Tree of depth 5 Test set score:0.30

Tree of depth 10 Training set score:0.77

Tree of depth 10 Test set score:0.55

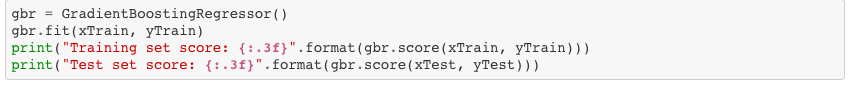
**K-Nearest Neighbors**

****

Training set score with 6 neighbors: 0.39

Test set score with 6 neighbors: 0.18

**Gradient Boosted Regression**[**¶**](https://render.githubusercontent.com/view/ipynb?commit=ca63444e3db91b833abd909b24255e7d089c2700&enc_url=68747470733a2f2f7261772e67697468756275736572636f6e74656e742e636f6d2f726f68697472616a6373652f535441543531372f636136333434346533646239316238333361626439303962323432353565376430383963323730302f5175657374696f6e305f50726f322e6970796e62&nwo=rohitrajcse%2FSTAT517&path=Question0_Pro2.ipynb&repository_id=149576428&repository_type=Repository#Gradient-Boosted-Regression)

****

Training set score: 0.73

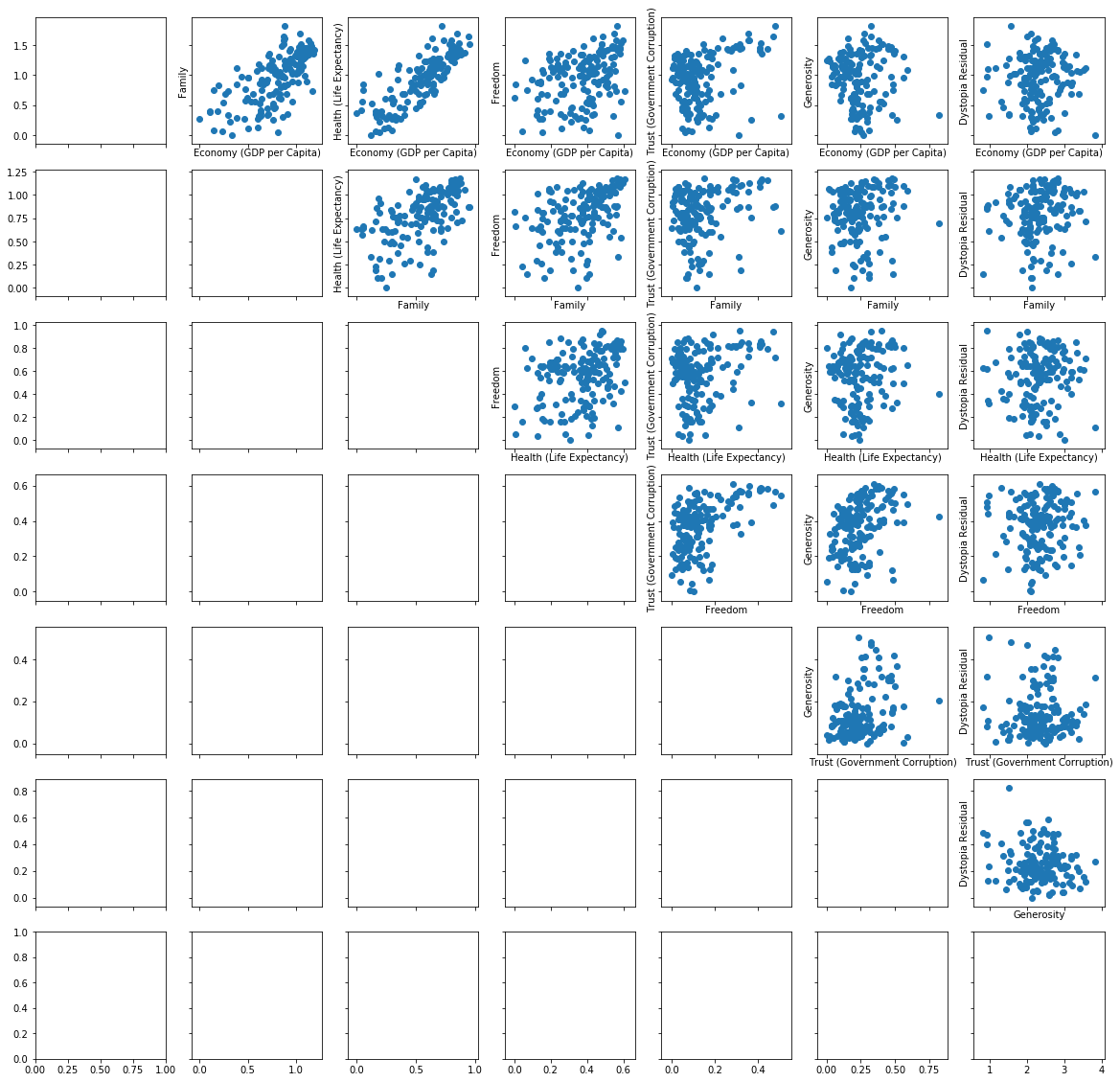
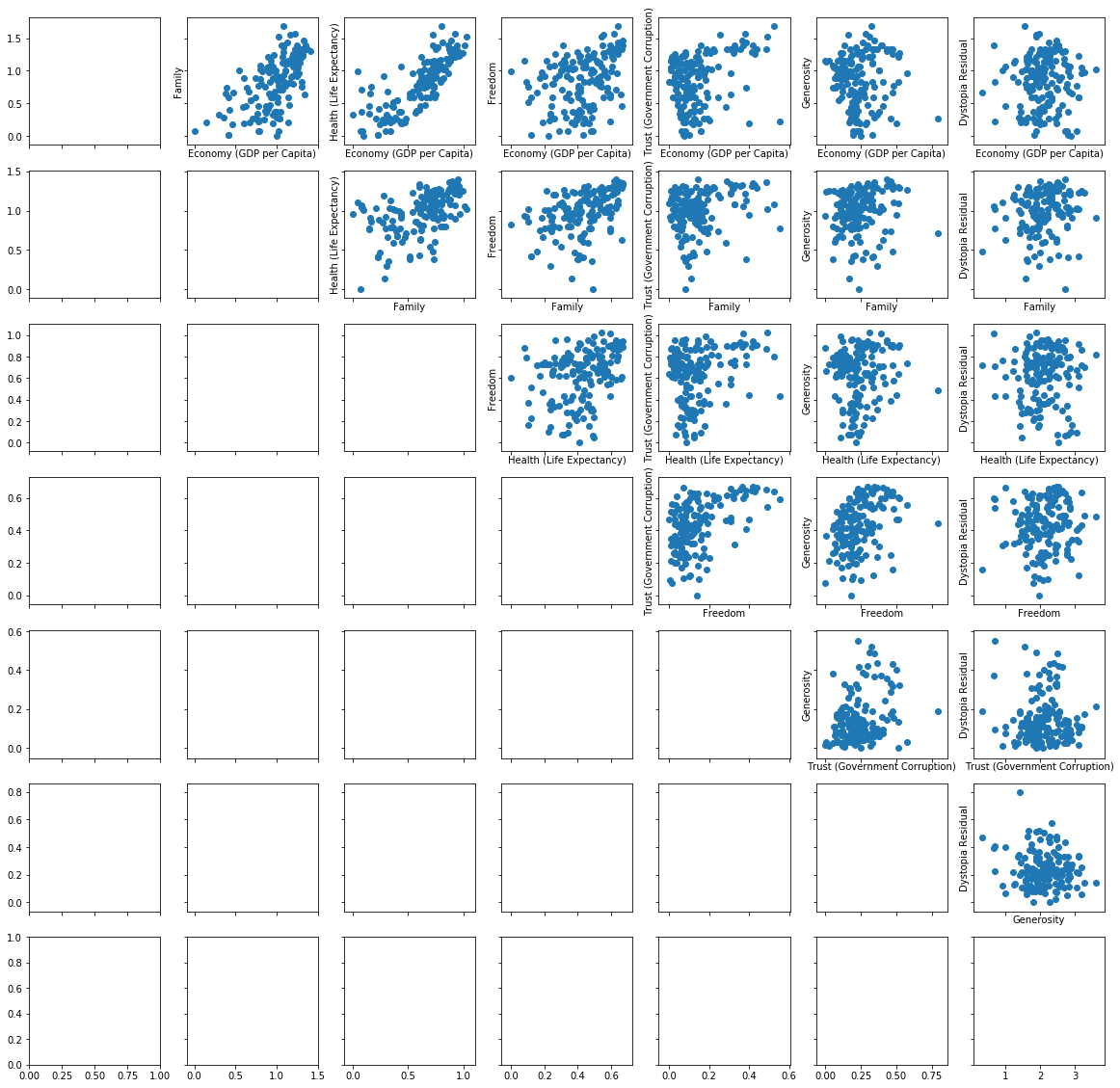
Test set score: 0.50

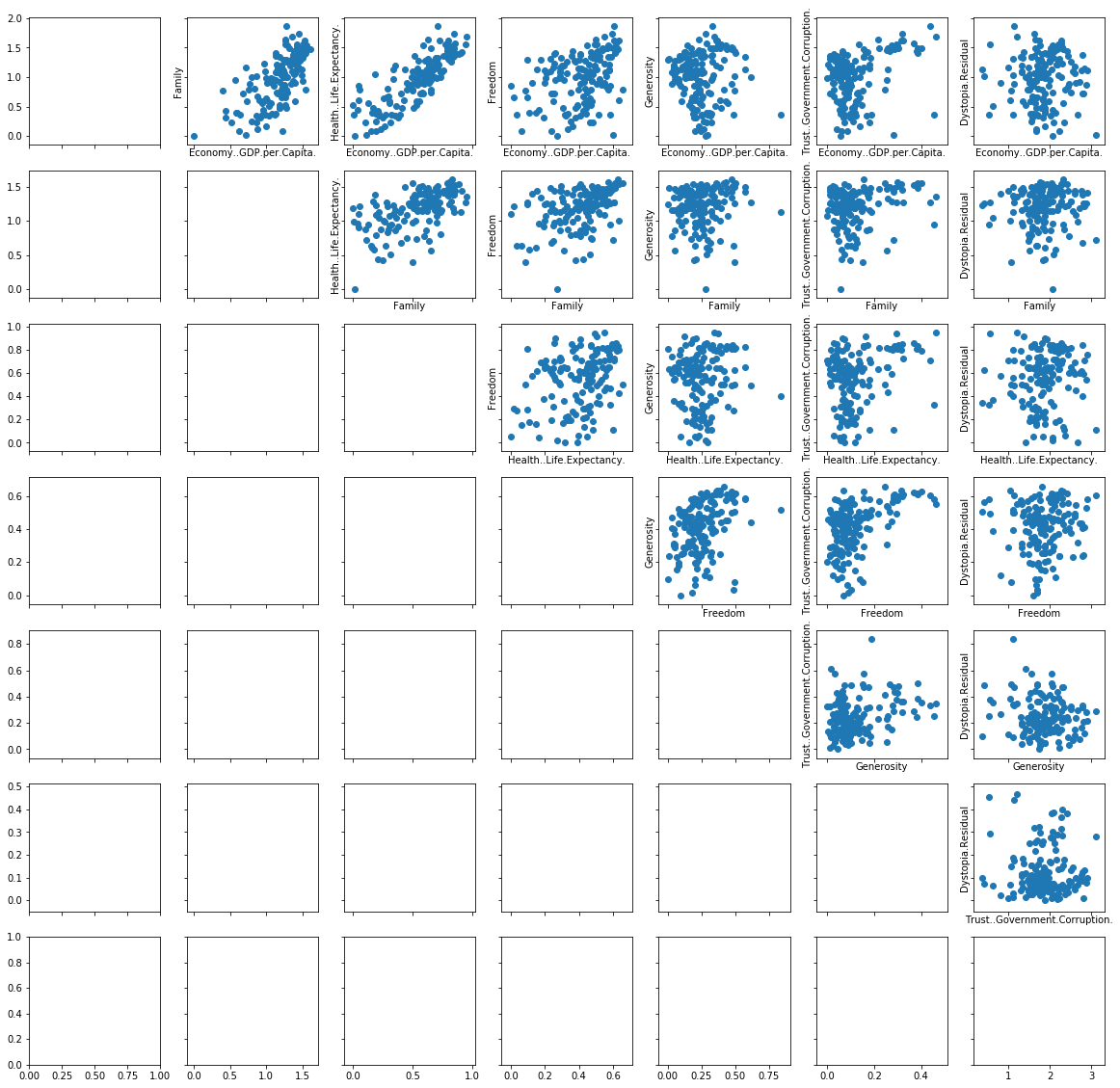
**Results and Conclusion:**

Datastet was imported and data preprocessing has been done by getting rid of all the useless columns, and replaced the null values with mode which is most frequent occurrence in the dataset, and then implemented tf-idf(term frequency-inverse document frequency) where I first created vector using only 1000 words and removing useless stop words because we just consider words that matter to the target variable (salary) then preprocess the categorical features and then splitted my data in 80% of training and 20% test set and applied different models in which It seems the linear and ridge regression the most accurate and best among all in training and test score than other models.

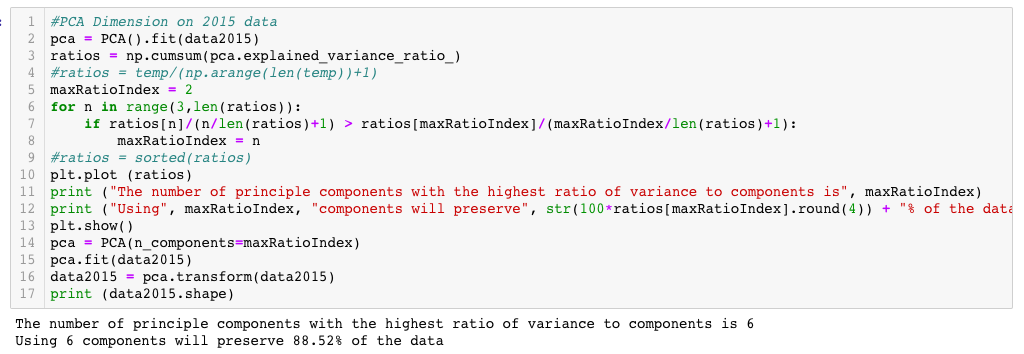
**Question 1: Happiness**

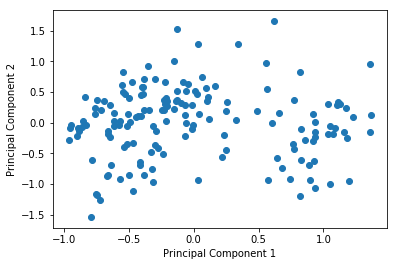
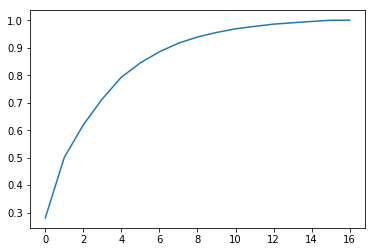
**Preliminary Visualization**

****

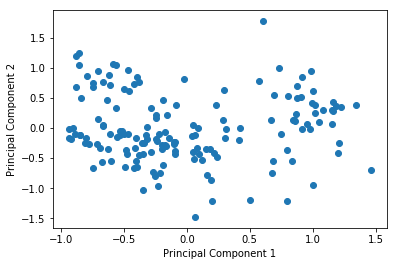
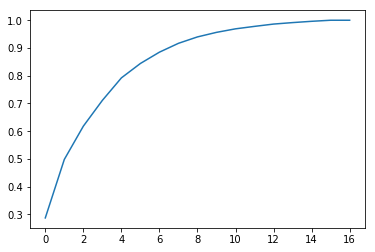
****

**PCA Dimension Reduction on data2015, data2016 and data2017**

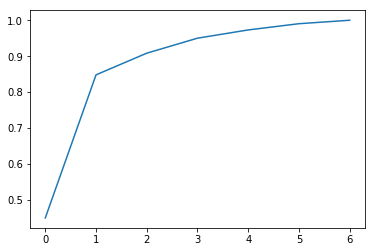
****

****

**Fig: PCA Dimension reduction on Data2015**

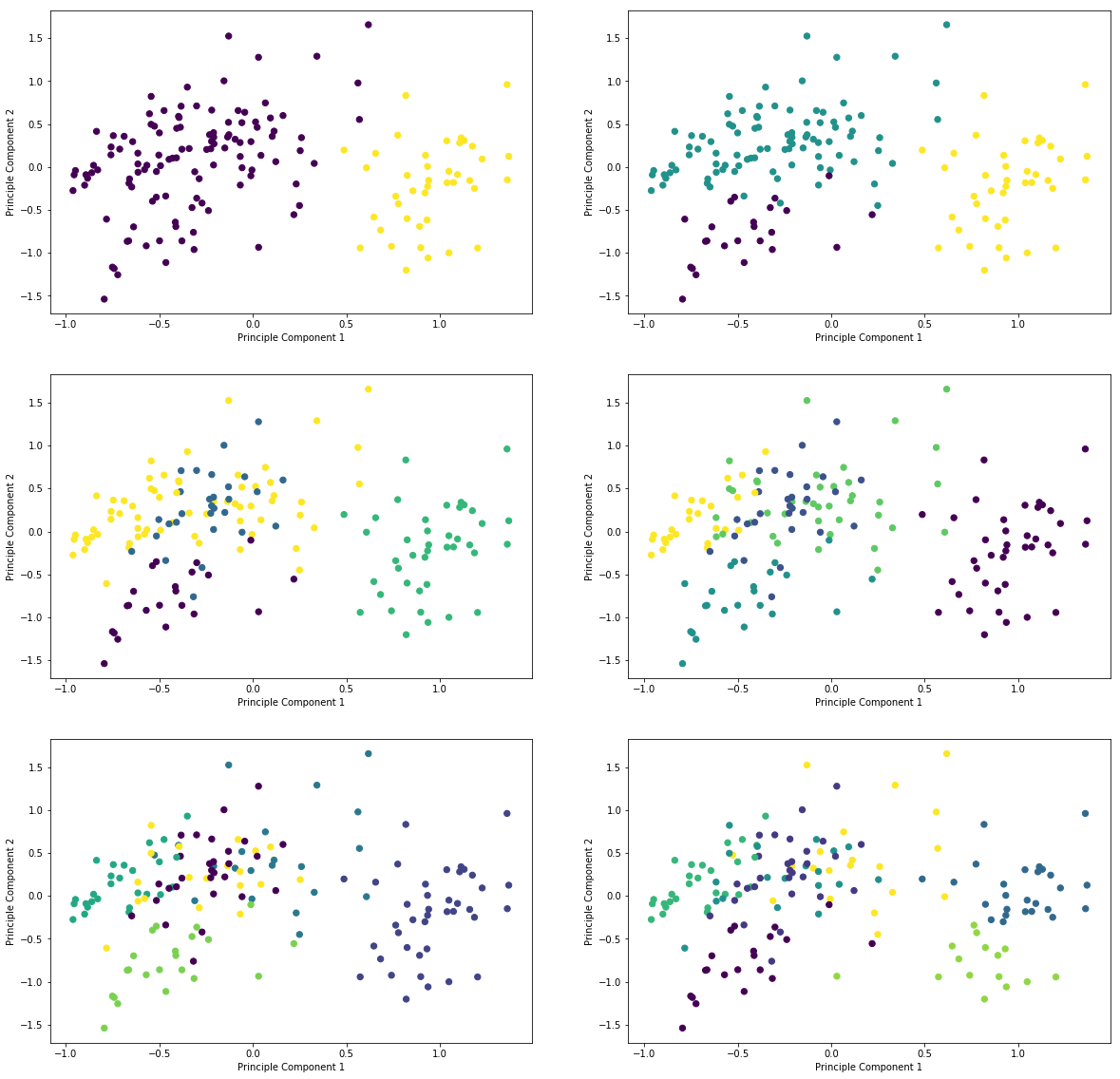


**Fig: PCA Dimension reduction on Data2016**

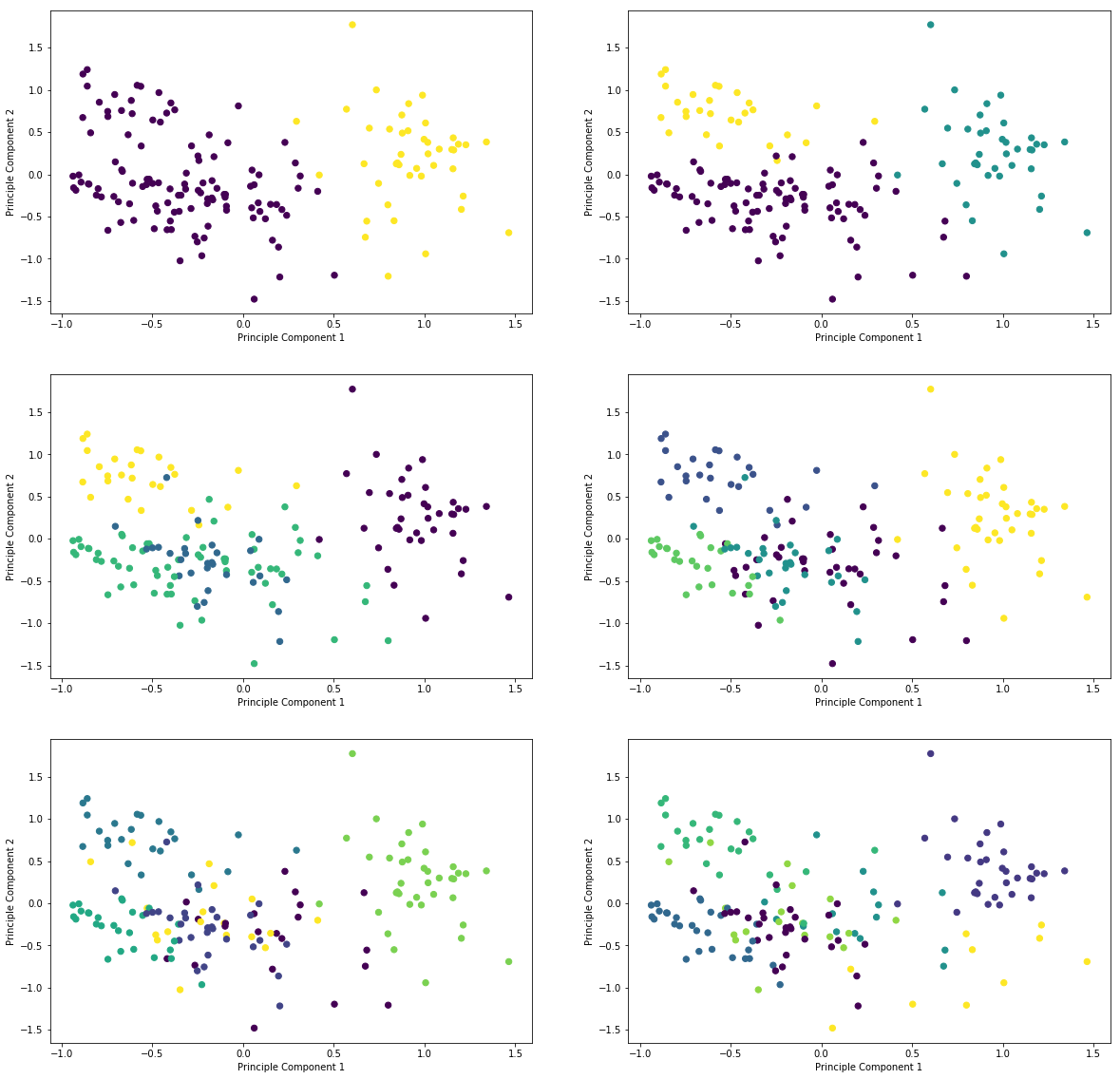
****

**Fig: PCA Dimension reduction on Data2017**

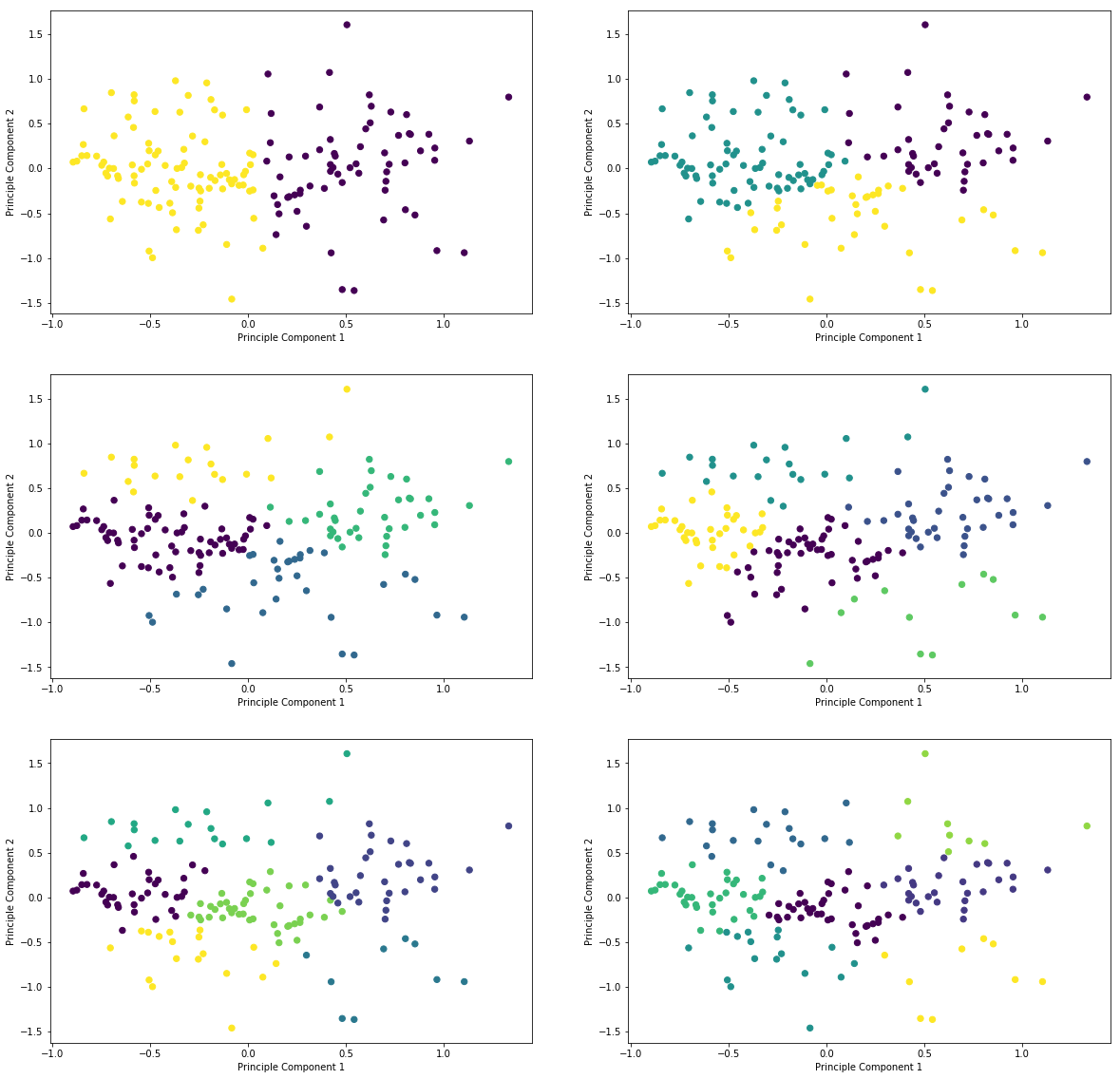
**KMeans**

****

**Fig: Kmeans on Data2015**

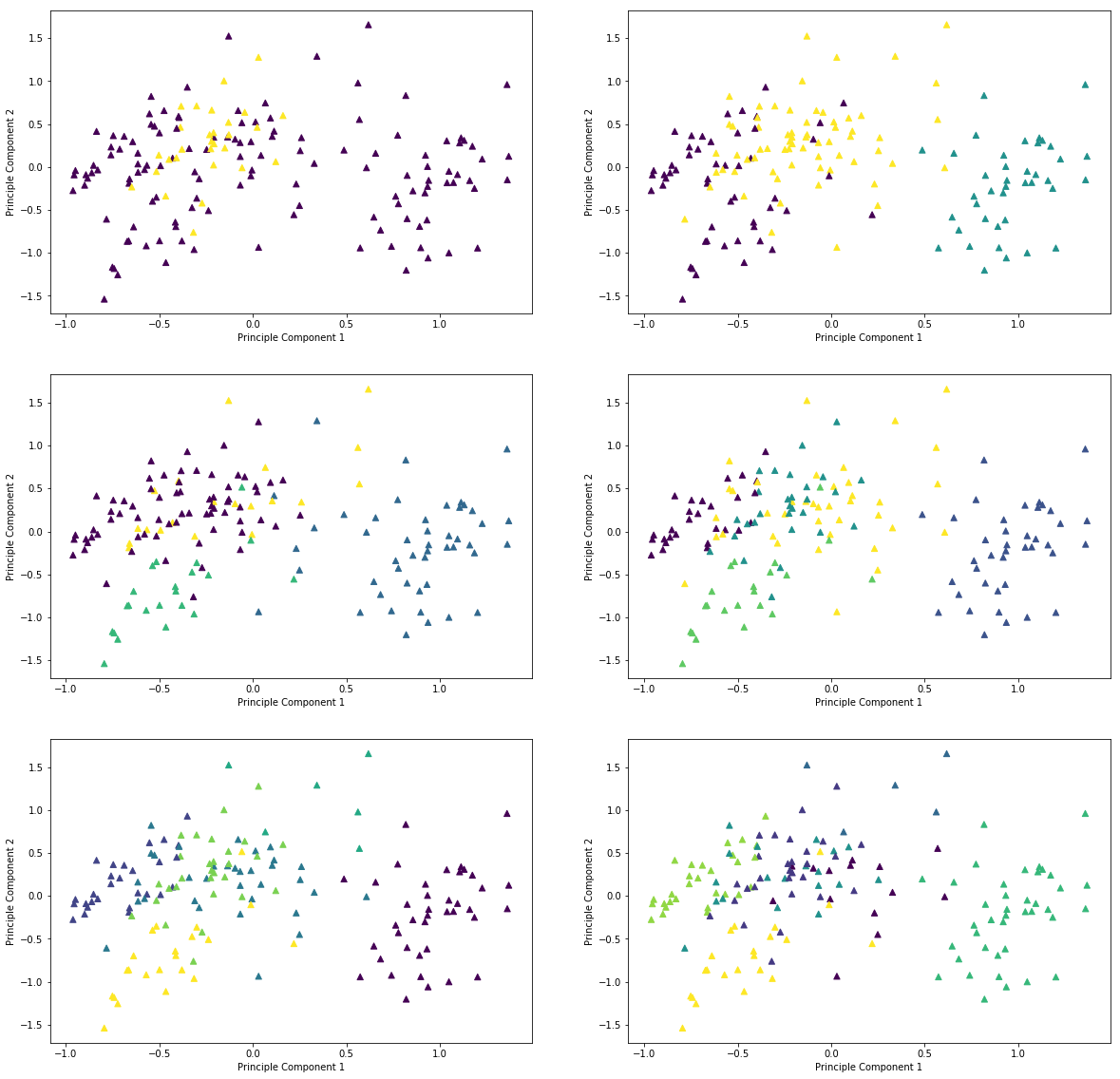
****

**Fig: Kmeans on Data2016**

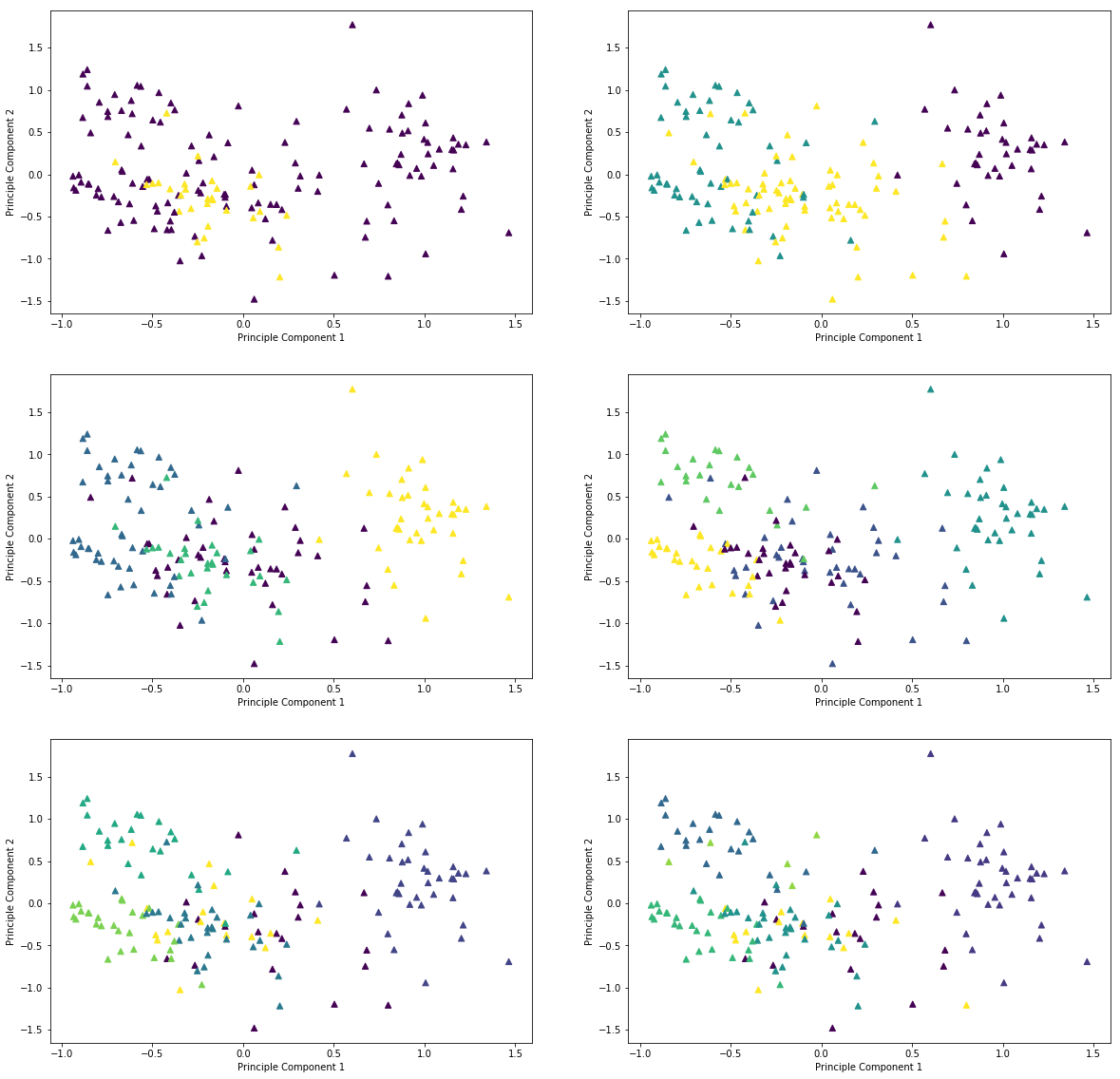
****

**Fig: Kmeans on Data2017**

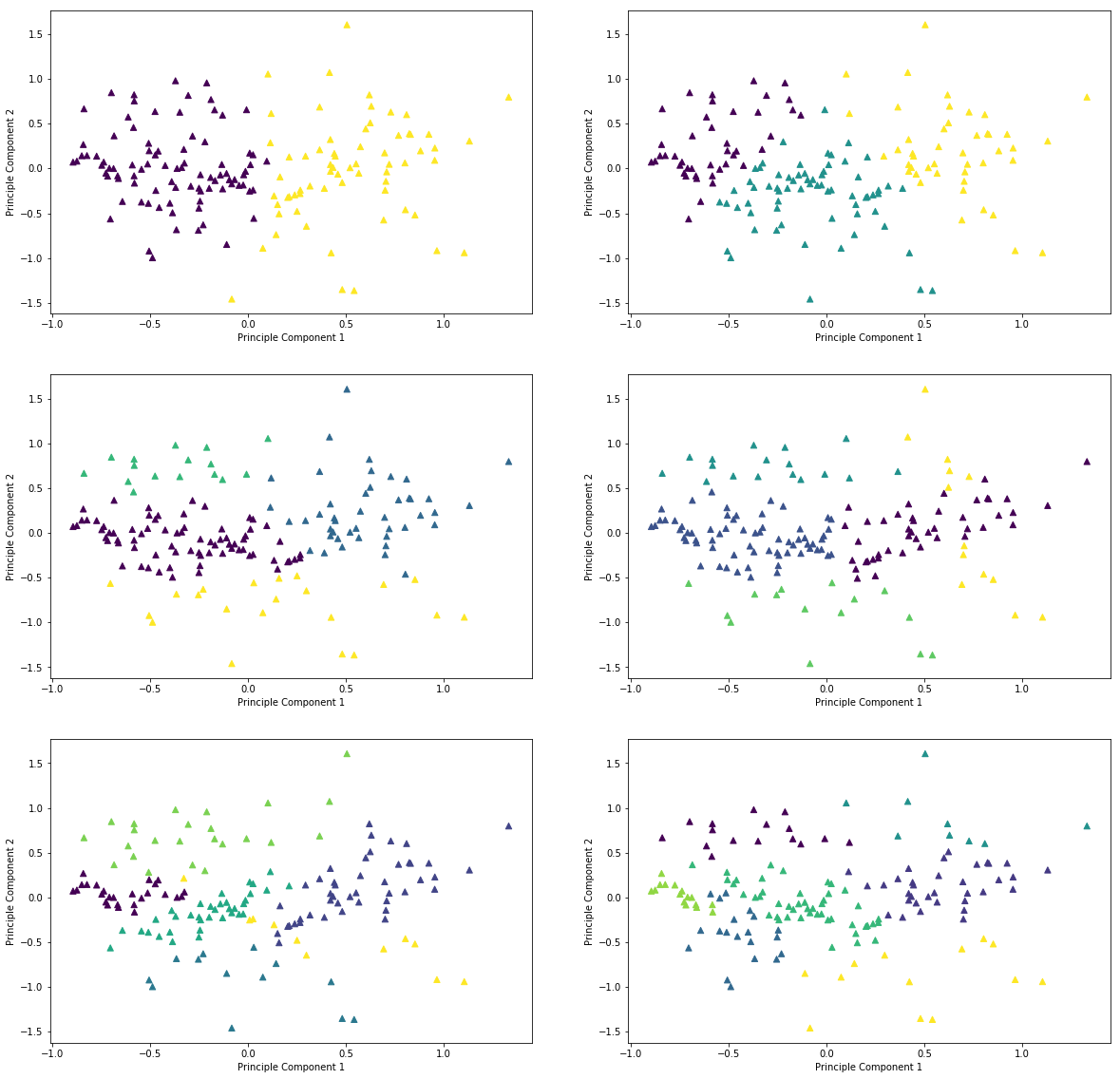
**GMM**

****

**Fig: GMM on Data2015**

****

**Fig: GMM on Data2016**

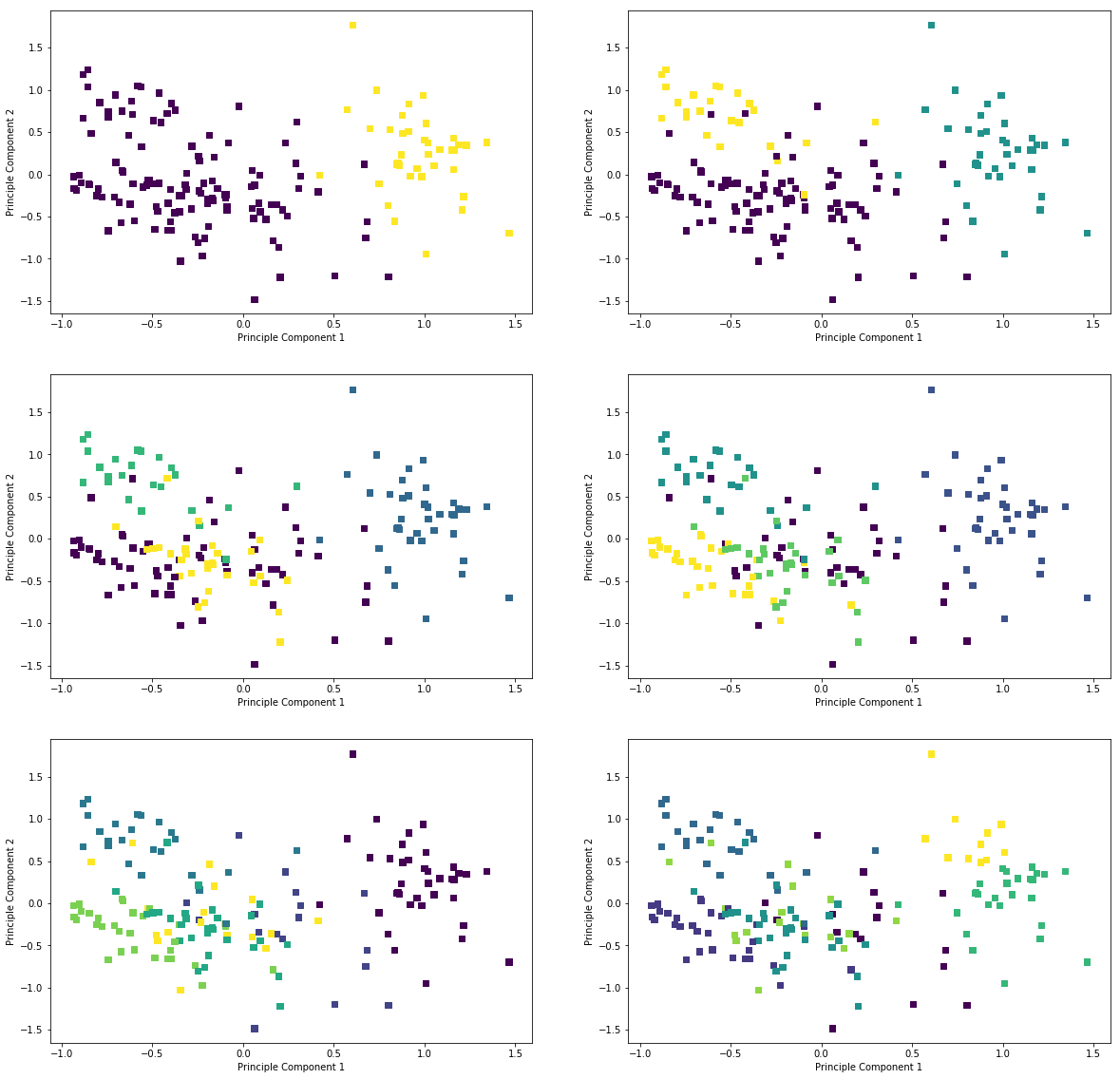
****

**Fig: GMM on Data2017**

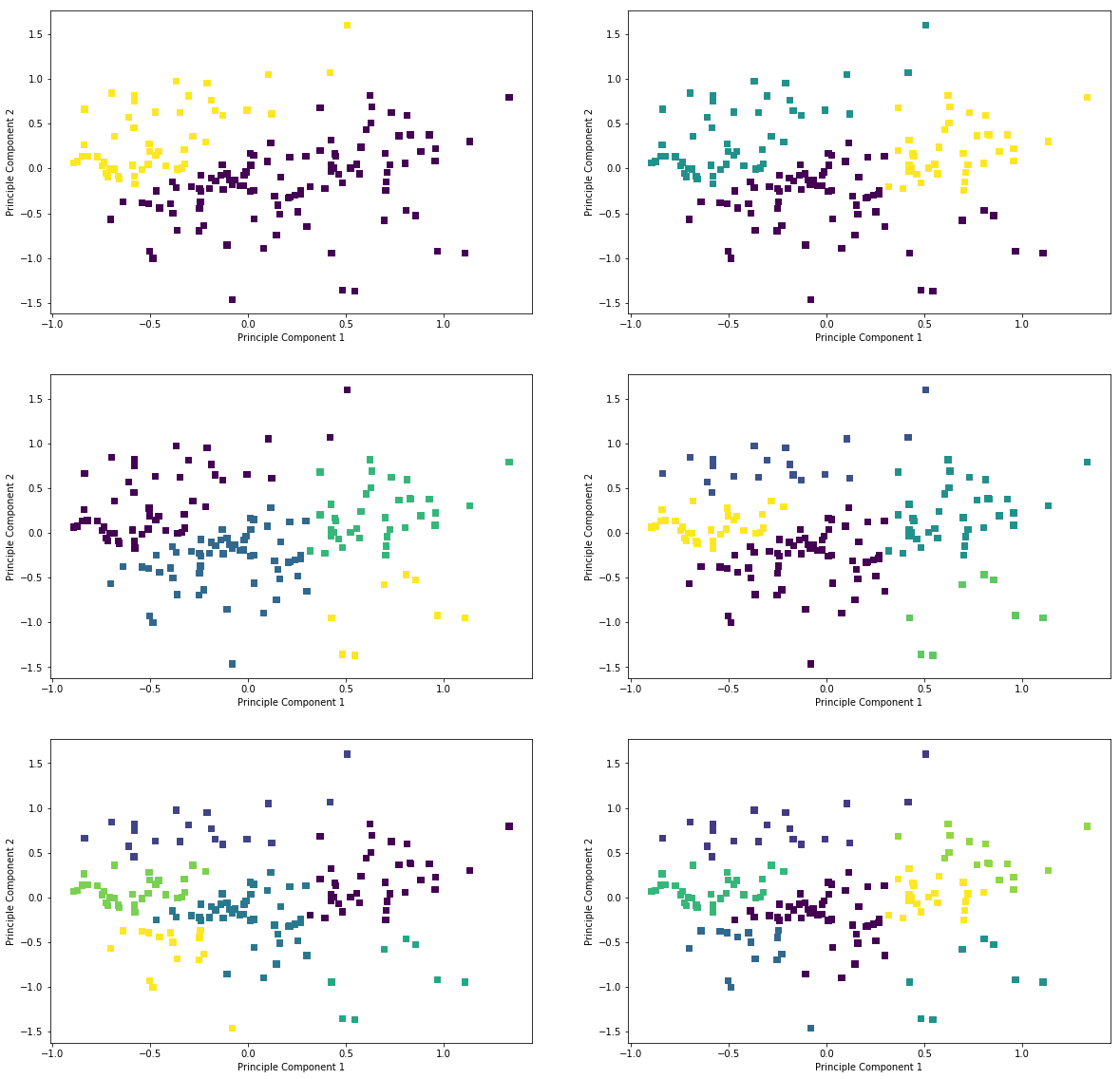
## AgglomerativeClusutering

## 

**Fig: AgglomerativeClustering on Data2015**

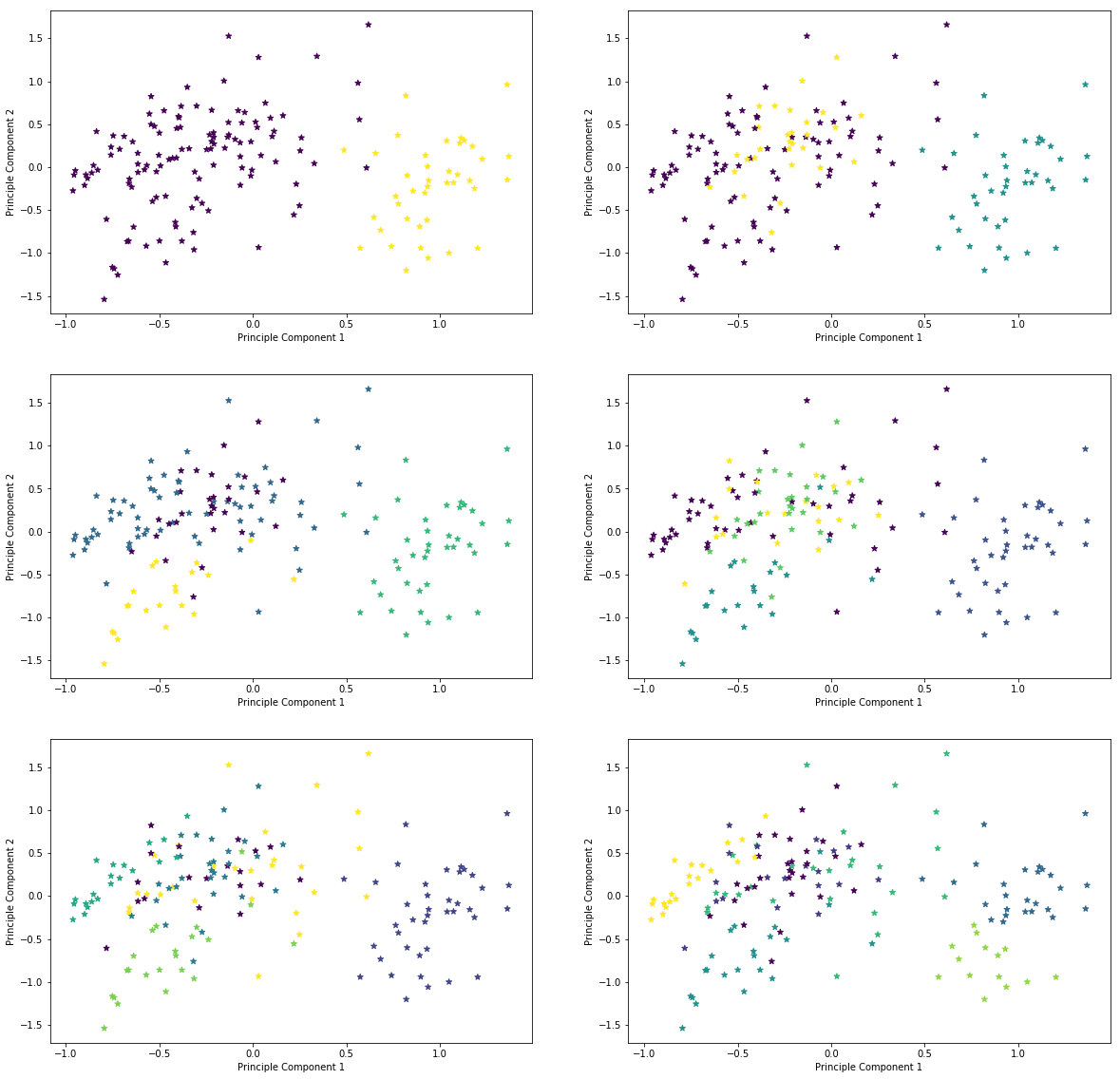
****

**Fig: AgglomerativeClustering on Data2016**

****

**Fig: AgglomerativeClustering on Data2017**

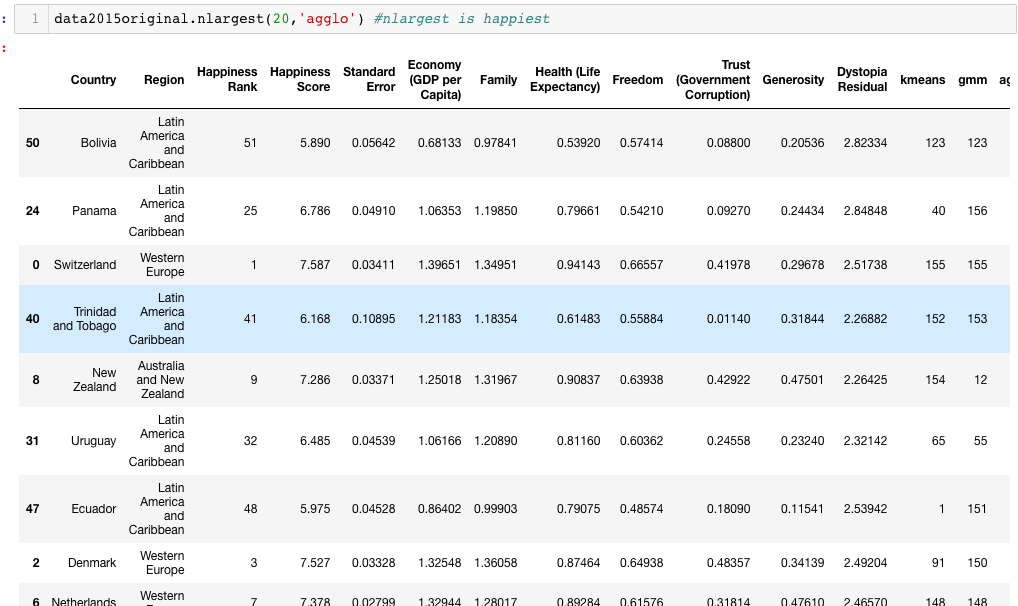
**Spectral Clustering**

****

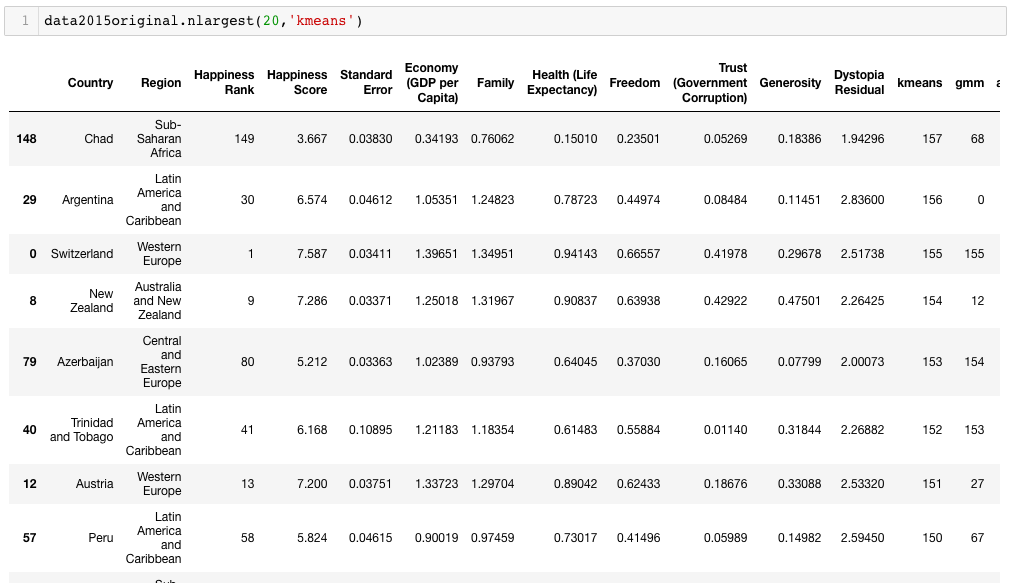
**Fig: Spectral Clustering on Data2015**

Since we already clustered with numClusters equal to the number of countries, we simply rank the countries by which cluster they are in.

**Agglomerative seriation Ranking chart:**



**KMeans Seriation Ranking table:**

****

1) Norway tops the global happiness rankings for all three years 2015, 2016, and 2017.

Answer: Norway tops in the original dataset of happiness ranking but Bolivia, serbia and vietnam tops the ranking according to seriation analysis of data2015, data2016, data2017 respectively. My clustering have found Chad, serbia and south sudan tops the Kmeans seriation in data2015, data2016 and data2017 respectively.

2) All top ten countries rank highly on all the main features found to support happiness.

Answer: It seems top ten countries indeed having the high rankings and scores in some features but low in some features like Kmeans seriation score for instance Kmeans of panama and Ecuador is very low in numbers of 40 and 1 respectively.

3) Happiness is both social and personal.

Answer: Yes Happiness founds to be both social and personal as all factors involved which influence happiness by ecomic and government.

4) Unemployment causes a major fall in happiness, and even for those in work the quality of work can cause major variations in happiness

Answer: This is true indeed unemployment caused a major fall in happiness also other factors mattered and happiness get influenced by those like economy and health.

5) China are no happier than most countries, though richer and longer longevity Answer: China did not appeared anywhere in the happiness ranking which seems as no more the happier country as I found Serbia being the most happiest according to both seriation analysis.

6) Much of Africa is struggling

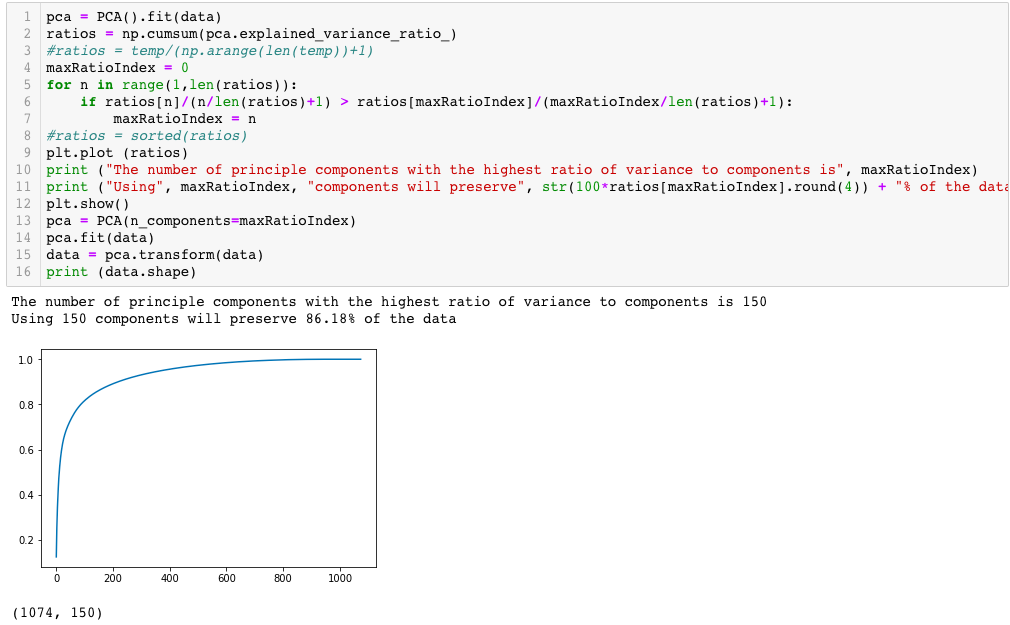
Answer: Sub-saharan Africa and central african republic seems to be the most saddest in all analysis of years

7) Happiness has fallen in America

Answer: It seems America did not topped but is not lowest as well It pretty in middle and average in happiness regions of all years.

**Question 2: Mitochondria**

## PCA Dimension Reduction

****

## KMeans

## 

## GMM AgglomerativeClusutering

## Spectral Clustering

## 

## 

## Fig: KMeans Fig: GMM

## 

## Fig: agglomerative clustering. Fig: Spectral clustering

## 

## This shows that although the clustering algorithms sometimes found more clusters than haplogroups, they are finding similar enough clusters to be considered accurate.

## Question 3: Data Mining the Bible

## Preliminary Visualization

## Word cloud is installed in python for this representation of 66 different most frequent words are reprsented which appeared in both the testaments

## 

## Fig: 66 word cloud

## PCA Dimension Reduction

## 

## KMeans Clustering

## 

## DBSCAN Clustering

## 

## Agglomerative Clustering

## 

## Spectral Clustering

## 

## Cluster Visualization

## 

## 

## Results and Discussions:

## A. What is the optimal number of clusters of these 66 Books? Find these clusters and describe them. Are you surprised at your finding? Why/Why not? Graph and color your clusters (probably on the first two PC's). On the graph, show your clusters in colors, the Testaments in plot symbols, and the Sections in sizes.

## *Answer A: Kmeans with optimal number of clusters 3 found to be more reliable in testing. yes I was surprised by findings as I thought that there must be 66 different clusters but It isn't.*

## B. How would Association Analyses help to reveal characteristic word clusters? Produce word clouds for the top 10 words clusters with the top 100 most frquent words. Describe these word clusters, and what they are telling you about the Bible. How do these top 10 words clouds represent the 2 Testaments and the 7 Sections?

## *Answer B: It will help by comparing the most frequently words occured together and from which clusters they belong as for example from the word cloud which is generated having top 100 words from each section and each testaments which consists of most frequently word used in old and new testaments like son,david, king etc.*

## C. How would Seriation Analyses help to reveal the structure of these 66 Books?

## *Seriation will put the books in order based on patters in the data. I don't think the books would be put in the same order as in the bible, because this order is fairly arbitrary. It's more likely that the seriation analyses will put the books in order of complexity, or writing style.*