Coursework Project-Multivariate Analysis(MATH4068)

Preliminiary look at the dataset and Aim

The dataset file gap.csv that has been given to us contains the GDP per Capita, and the life expectancy for 142 different countries from the time period of 1952 to 2007.The data has been taken from gapminder.org.The aim of this report is to analyse the data given in the file by applying various multivariate anlaysis methods and taking inference from the results.

Exploratory Data Analysis

As usual before performing various operartions on the dataset like Principal Component Analysis and Multi-Dimensional analysis,it is important to perform Exploratory Data Analysis on the data to find some patterns and relationships between the dataset attributes.Since we are dealing with two different attributes GDP and Life expectancy we will split the dataset into two parts to make our plots more readable and easier to plot.

We will be trying to analyse the data and show how GDP and Life Expectancies have changed over the past 70 years.

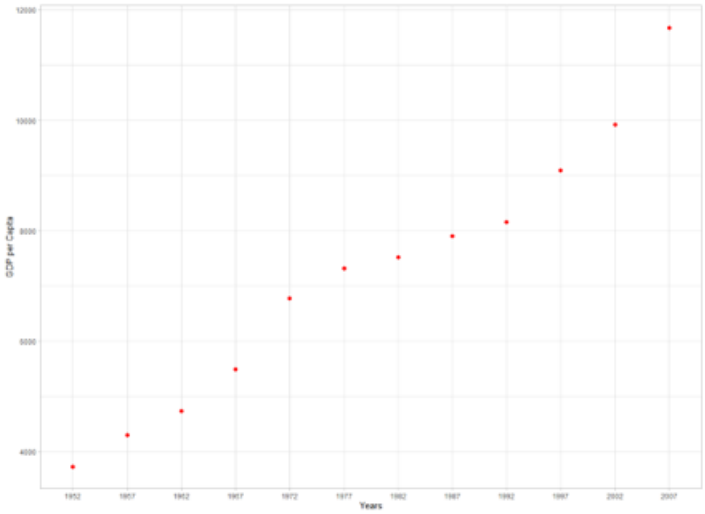

Figure 1: GDP CHANGES IN THE LAST 70 YEARS

As we can see from the plot, the GDP Per capita has had an increasing trend over time. This can be linked to the fact that factors like population of the countries have increased and there are more employment opportunities which means more money been generated over time.

As we can see in the plot below in the case of the life expectancy as well, there has been an increasing trend over time. This can be linked to the fact that a lot of technological advancements, cleaner environments, better medicinial treatments and hygiene have all reduced mortality rates increasing from less than 50 in the year 1952 to almost 70 in the year 2007.

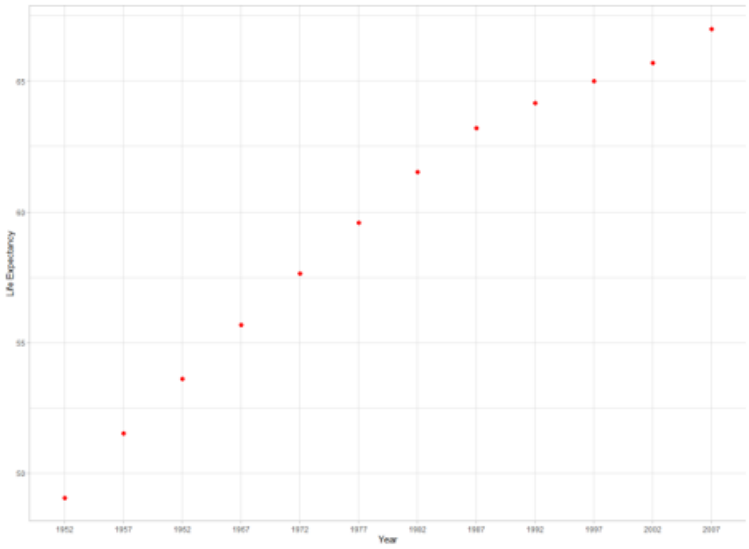

Figure 2: LIFE EXPECTANCY CHANGES IN THE LAST 70 YEARS

After taking a look at the following plots , it will be safe to infer that GDP and Life Expectancy are directly proportional to each other which would suggest that through increase in the economic growth of the country, it leads to the prolongation of longevity.

We can also take a look at if various GDP's and Life Expectancies are correlated to each other. The best way to do this would probably be by feature selection algorithms such as Principal Component Analysis but we can also do it using a corrplot which would tell us how intensely they are correlated to each other

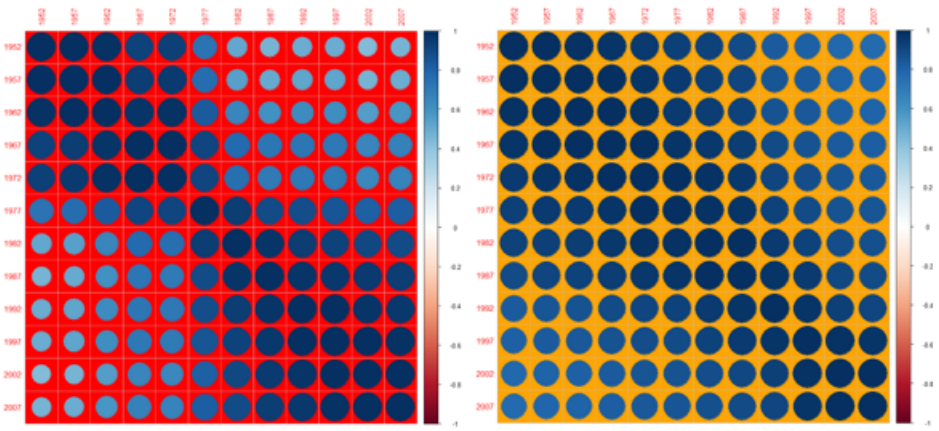

Figure 3: CORRELATION PLOT FOR GDP(RED)AND LIFE EXPECTANCY(ORANGE)

PERFORMING PRINCIPAL COMPONENT ANALYSIS

After performing exploratory data analysis we can perform principal component analysis on the data. Principal component analysis is important as a pre-processing technique as it allows us to only focus on those parts of the data that contain the most useful information through which we can reduce the dimesnions of our data and improve the model accuracy at a very low cost.

PERFORMING PCA ON LOG(GDP)

Before we conduct PCA, it is important to convert the GDP into log(GDP) as the values vary over several orders of magnitude between countries.We will hence be conducting principal component analysis on log(GDP) and life expectancy data.

After performing the PCA we find that the first three components account for 99.3% of the cumulative proportion of the data. The Scree plot below shows a clearer view about the proportion of variances of all the 12 components.

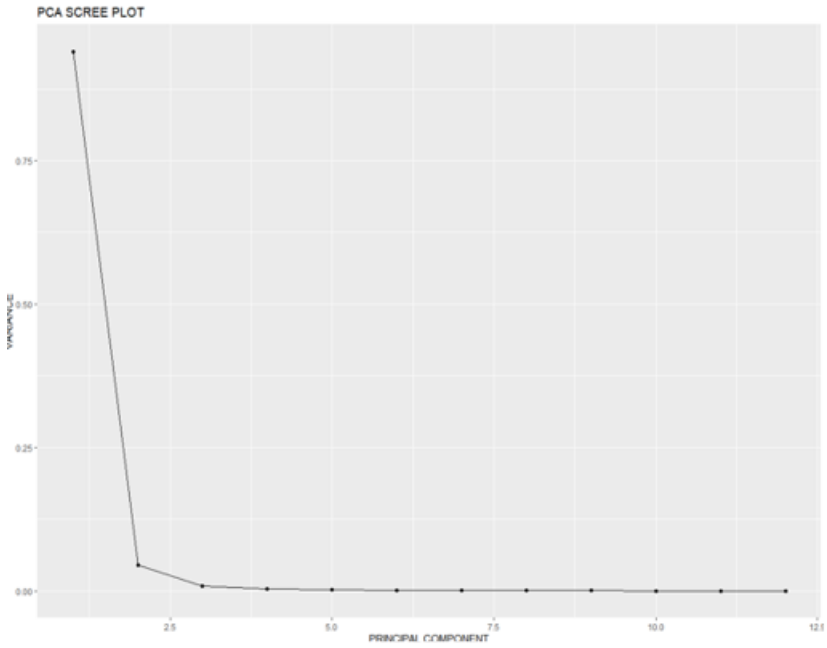

Figure 4: PRINCIPAL COMPONENT ANALYSIS-SCREE PLOT(LOG GDP)

Since we have decided to retain three components from the PCA(by elbow method), the following interpretations can be made about the components by just just checking the rotation attribute of the pca:

1. The PC1 component has all negative eigen values which would mean an inverse correlation with the log(GDP).
2. The PC2 component has negative eigen values from 1982 to 2007 which would suggest an inverse correlation while it has positive correlation from 1952 to 1977
3. The PC3 component has positive correlations in the years from 1952 to 1962 and from 1197 to 2007 whereas it is negatively correlated between the years 1967 to 1992.

PERFORMING PCA ON Life expectancy

After performing the PCA we find that the first three components account for 99.29% of the cumulative proportion of the data. The Scree plot below shows a clearer view about the proportion of variances of all the 12 components.

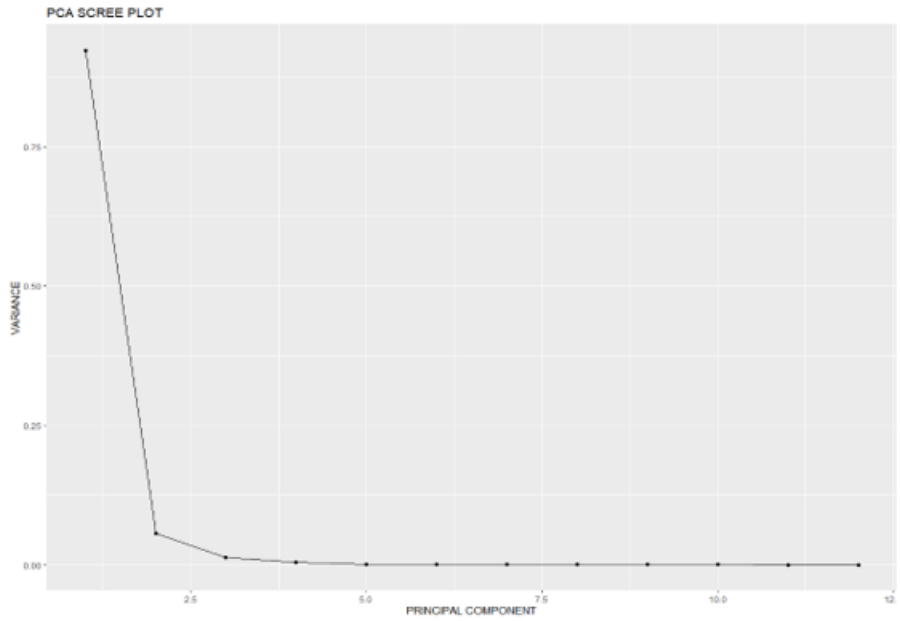

Figure 5: PRINCIPAL COMPONENT ANALYSIS-SCREE PLOT(Life Expectancy)

Since we have decided to retain three components from the PCA(by elbow method), the following interpretations can be made about the components by just just checking the rotation attribute of the pca: 1. The PC1 component has all positive eigen values which would mean an inverse correlation with the Life Expectancy. 2. The PC2 component has negative eigen values from 1987 to 2007 which would suggest an inverse correlation while it has positive correlation from 1952 to 1982 3. The PC3 component has negative correlations in the years from 1952 to 1967 and from 2002 to 2007 whereas it is positively correlated between the years 1972 to 1997.

SCATTER PLOTS OF COMBINATIONS OF PRINCIPAL COMPONENTS

From the GDP PCA plots shown below we can make the following observations:

All the European countries apart from a few like Bosnia and Herzegovina had a great GDP since data was recorded in 1952 and hence are clustered together. The African countries have negative eigen values and hence are inversely correlated with the GDP .There are extremities when studying the Middle East and the Asian countries. Kuwait acts as a bit of an outlier since it is located at the top-left part of the second principal component graph making it very rich when it comes to GDP. Massive economies now such as China and India do fairly poorly in 1952

From the Life expectancy plots below we can make the following observations:

All the European countries do very well in the life expectancy graphs since they all are positively correlated to life expectancy and Most African Countries are struggling to meet life expectancy standards while Asian countries are somehwere in the middle with Afghanistan being an outlier as it is the poorest country when it solely comes to Life Expectancy.

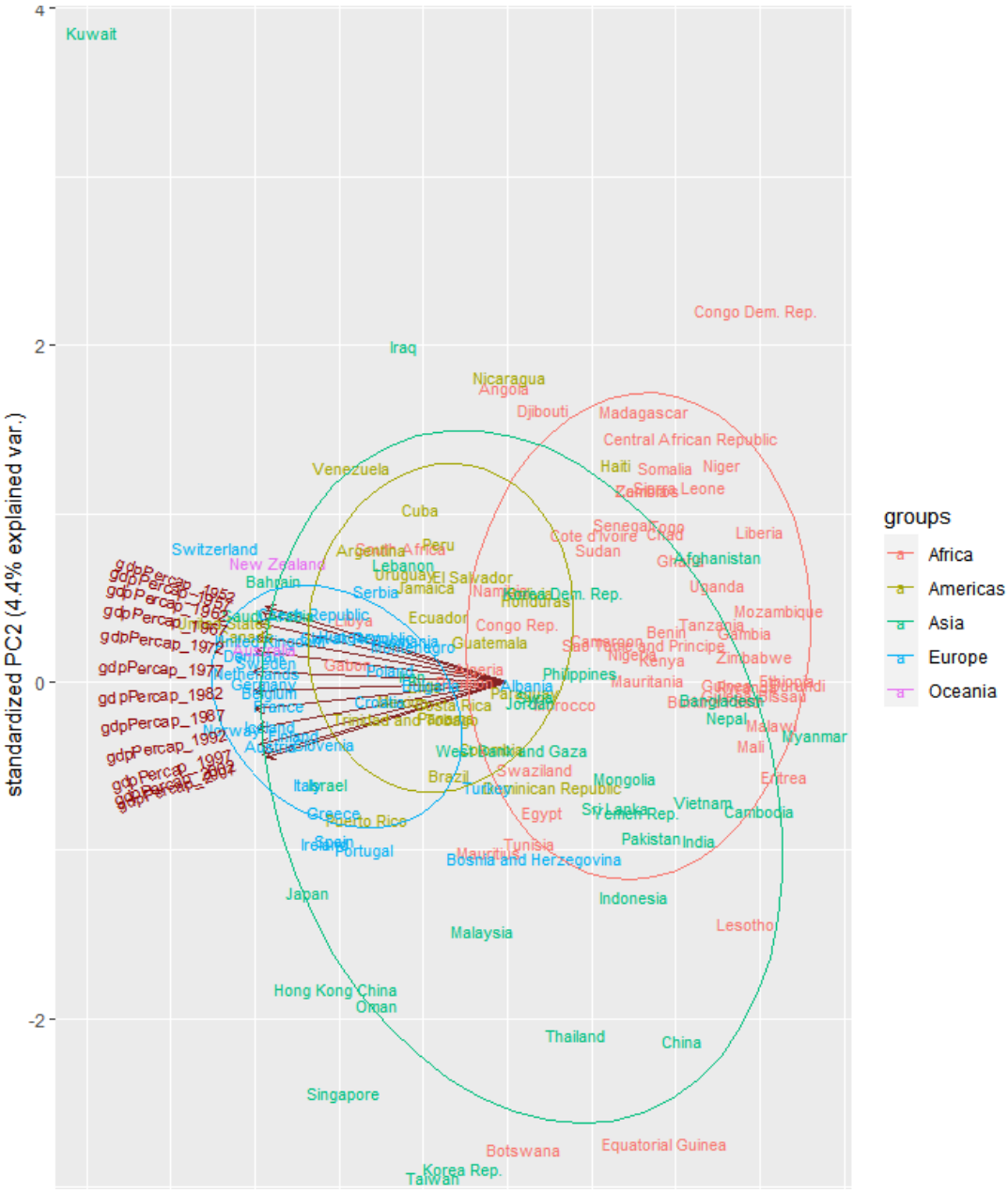

Figure 6: FIGURE SHOWING COUNTRIES ALONG WITH THEIR PCA EIGEN VALUES-GDP

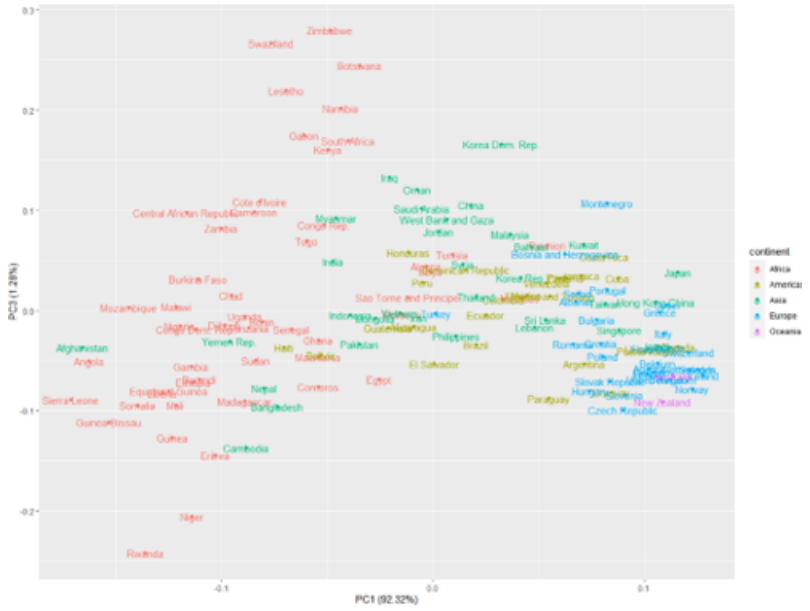

Figure 7: SCATTER PLOT BETWEEN PC1-PC3

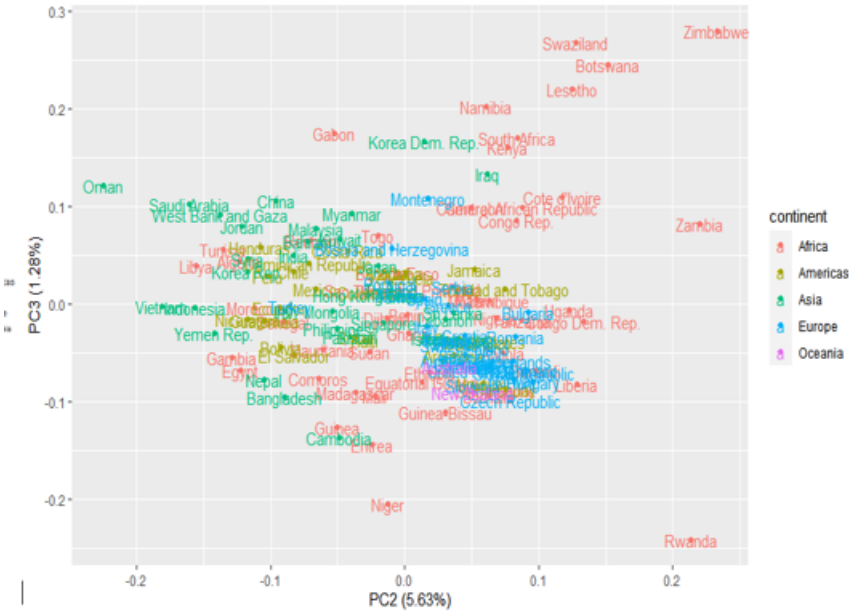

Figure 8: SCATTER PLOT BETWEEN PC2-PC3

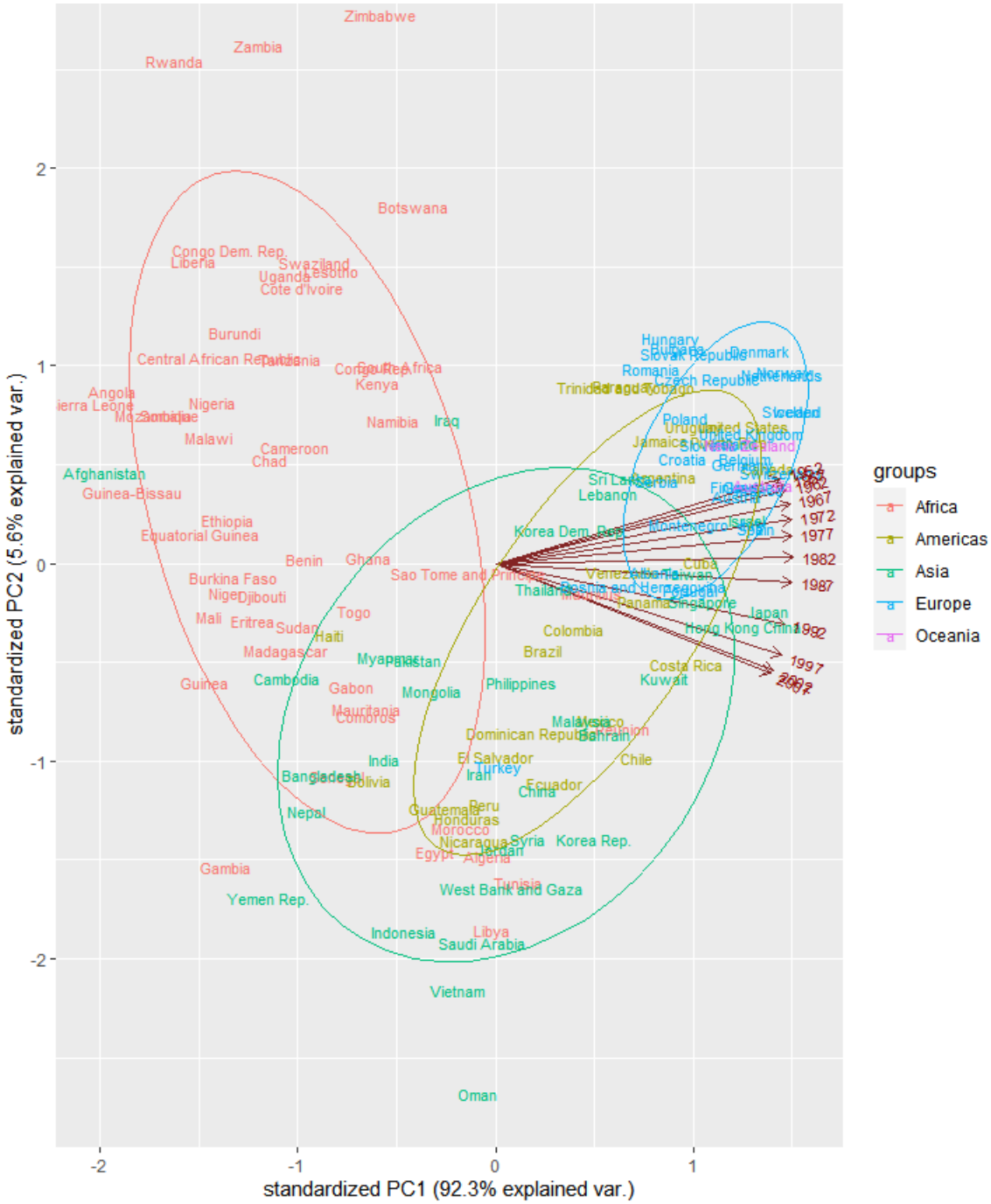

Figure 9: FIGURE SHOWING COUNTRIES ALONG WITH THEIR PCA EIGEN VALUES-LIFE EXPECTANCY

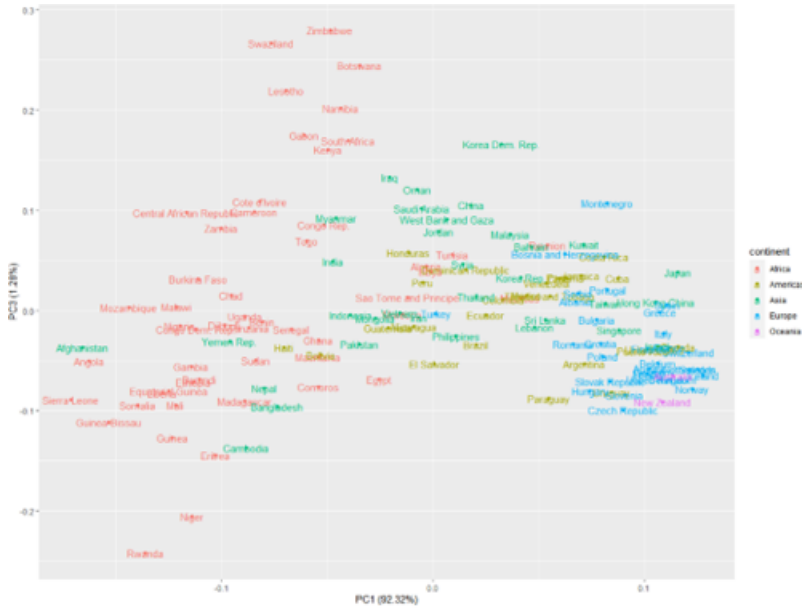

Figure 10: SCATTER PLOT BETWEEN PC1-PC3

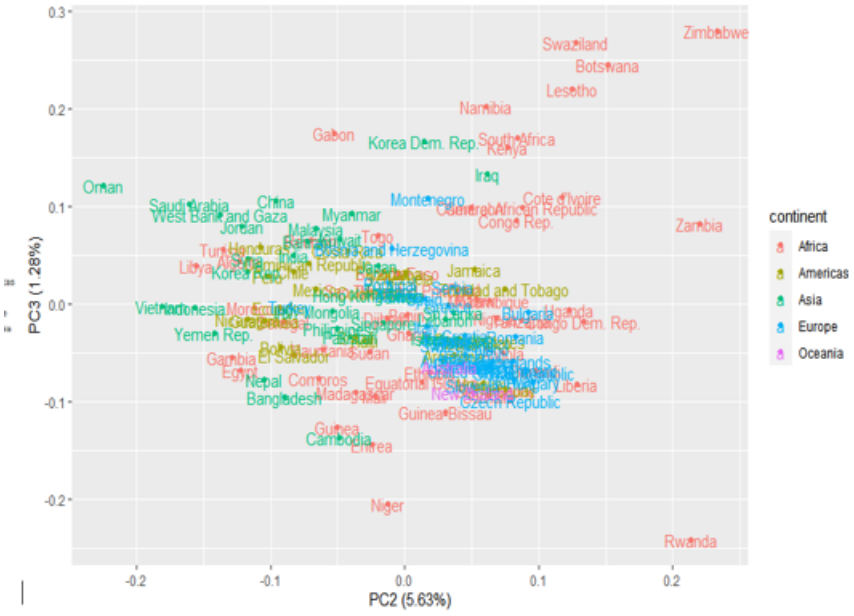

Figure 11: SCATTER PLOT BETWEEN PC2-PC3

Performing Multidimensional Scaling

Multidimensional Scaling is a means of visualizing the level of similarity in the dataset by creating a map displaying the relative positions of a number of objects giving only a table of euclidean distances between them.

We will perform MDS on the combined dataset of log(gdp) and life expectancy values.The mapped data with MDS is seen below

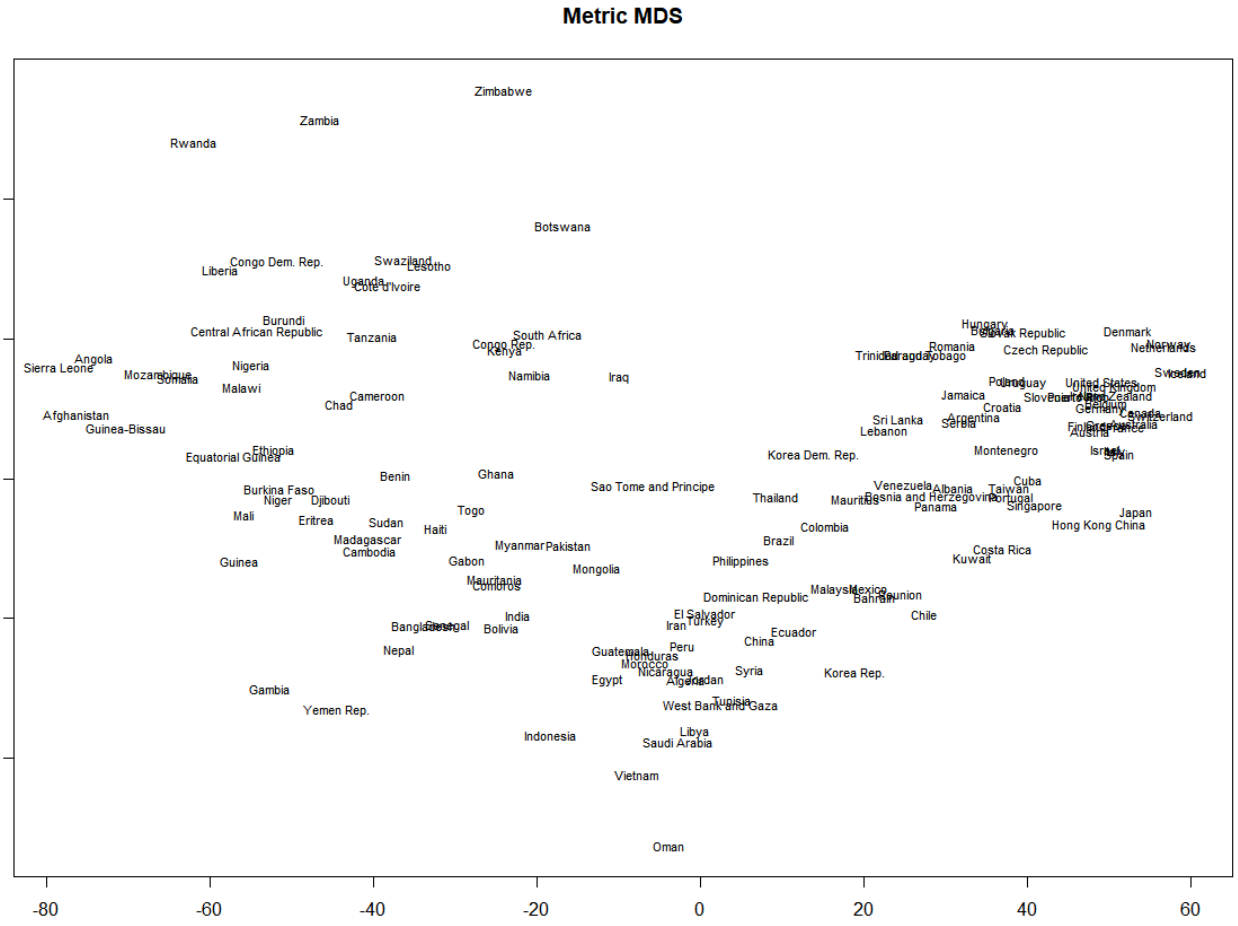

Figure 12: Mapped MDS

The following plot and the one that was produced when we were plotting the PC1-2 life expectancy are very identical. This is because we have used all data-points and given the distance measure between the points as euclidean which has lead to identical results. Therefore if the data input to these algorithms is different,they will produce different results because technically they both do different things.

Performing Hypothesis Testing

Hypothesis Testing is a form of statistical inference that uses data from a sample to draw conclusions about a parameter of a population probability distribution.First we need to make a tentative assumption about the data and we call this the null hypothesis.

We need to conduct a multivariate hypothesis test to test whether there was a statistically significant difference between the mean log(GDP) and life expectancy of Asian and European countries in the year

2007 and then do the same for the year 1952.We will use the Hotelling's two sample T2-test to conduct the hypothesis test

```
> colMeans(gap_gdppercap_2007)
gdpPercap_2007
      9.333722
> colMeans(gap_lifeexp_2007)
lifeExp_2007
    74.02378
>
> HotellingsT2(gap_gdppercap_2007, gap_lifeexp_2007)

        Hotelling's two sample T2-test

data:  gap_gdppercap_2007 and gap_lifeexp_2007
T.2 = 5234.3, df1 = 1, df2 = 124, p-value < 2.2e-16
alternative hypothesis: true location difference is not equal to c(0)
```

Figure 13: Result of the hypothesis Testing for 2007

Low P-values signify that there is infact a significant difference between the life expectancy and gdp in 2007 of Asian and European Countries so we will fail to reject the null hypothesis.The result is also the same for the year 1952

```
> colMeans(gap_gdppercap_1952)
gdpPercap_1952
      7.876614
> colMeans(gap_lifeexp_1952)
lifeExp_1952
    54.93063
>
> HotellingsT2(gap_gdppercap_1952, gap_lifeexp_1952)

        Hotelling's two sample T2-test

data:  gap_gdppercap_1952 and gap_lifeexp_1952
T.2 = 944.67, df1 = 1, df2 = 124, p-value < 2.2e-16
alternative hypothesis: true location difference is not equal to c(0)
```

Figure 14: Result of the hypothesis Testing for 2007

Performing Linear Discriminant Analysis

Linear Discriminant Analysis is a supervised machine learning method that is used to create machine learning models.It focuses on finding a feature subspace that maximizes the seperability between the groups.We need to Use linear discriminant analysis to train a classifier to predict the continent of each country using the log(GDP) and life expectancy from 1952-2007 and predict its accuracy

First we will divide the data into training and testing dataset with the ratio of 70:30

```
ind <- sample(2, nrow(gap),
              replace = TRUE,
              prob = c(0.7, 0.3))
gdp.train <- gap[ind==1,]
gdp.test<- gap[ind==2,]
```

Figure 15: TRAINING AND TESTING

We will use the lda function to build a classifier that will predict the continent on the basis of the gdp all over the years.

	Africa	Americas	Asia	Europe	Oceania
Africa	16	1	1	0	0
Americas	1	4	0	0	0
Asia	1	3	10	0	0
Europe	0	1	1	11	1
Oceania	0	1	0	0	0

Figure 16: CONFUSION MATRIX

The predictive accuracy comes out to be 78.846%.As we can see the model has done fairly well apart from the fact that it misclassified 11 terms.

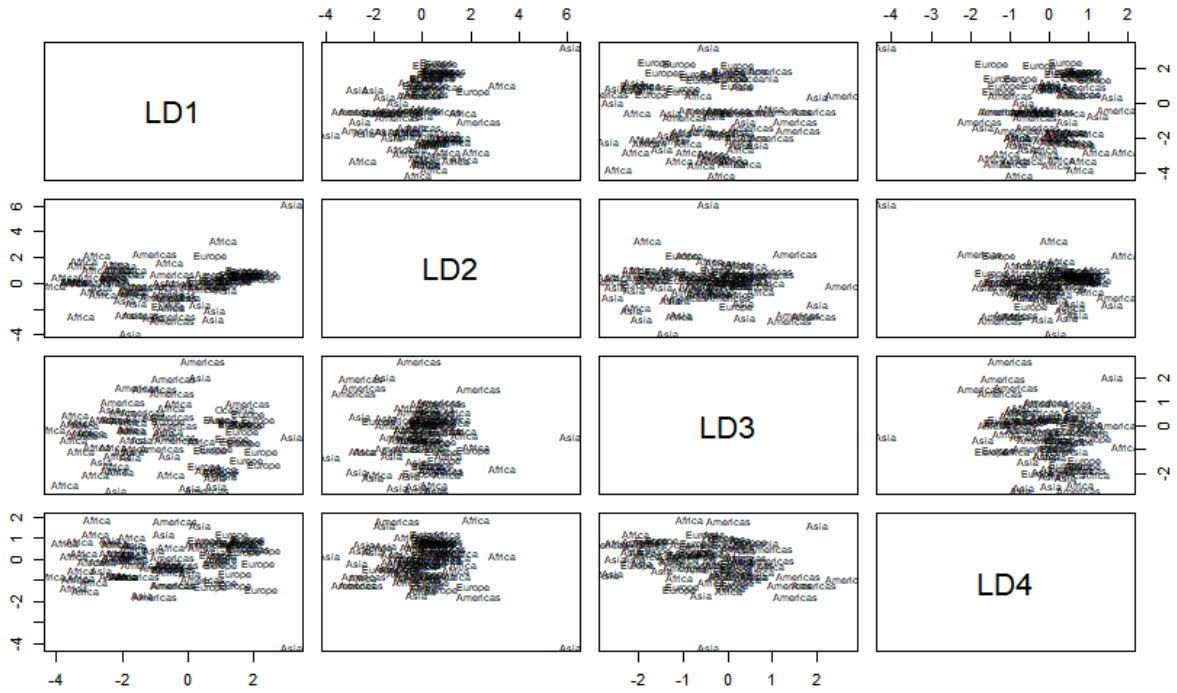

Figure 17: CLASSIFIER EFFECTIVENESS

Performing Clustering

Clustering Analysis or Clustering is the task of grouping a set of objects in such a way that the objects clustered together are more similar to the ones that are not. We are asked to perform k-means clustering on the data and give the plot of the final clusters.

To get started with the Knn clustering we would need to scale our data first so as to prevent poorer convergence.After we are done scaling the data, we will use the elbow method to determine the number of clusters that we need to build.

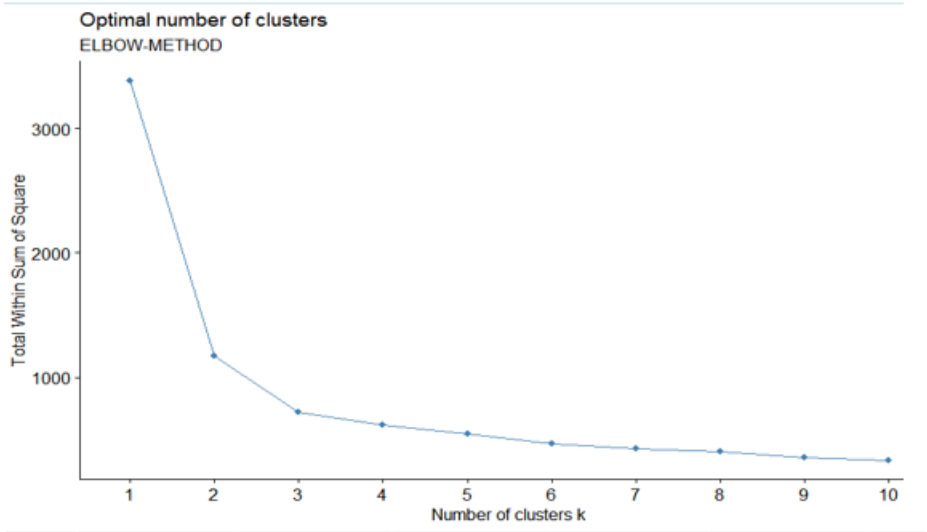

Figure 18: DETERMINING NUMBER OF CLUSTERS(ELBOW METHOD)

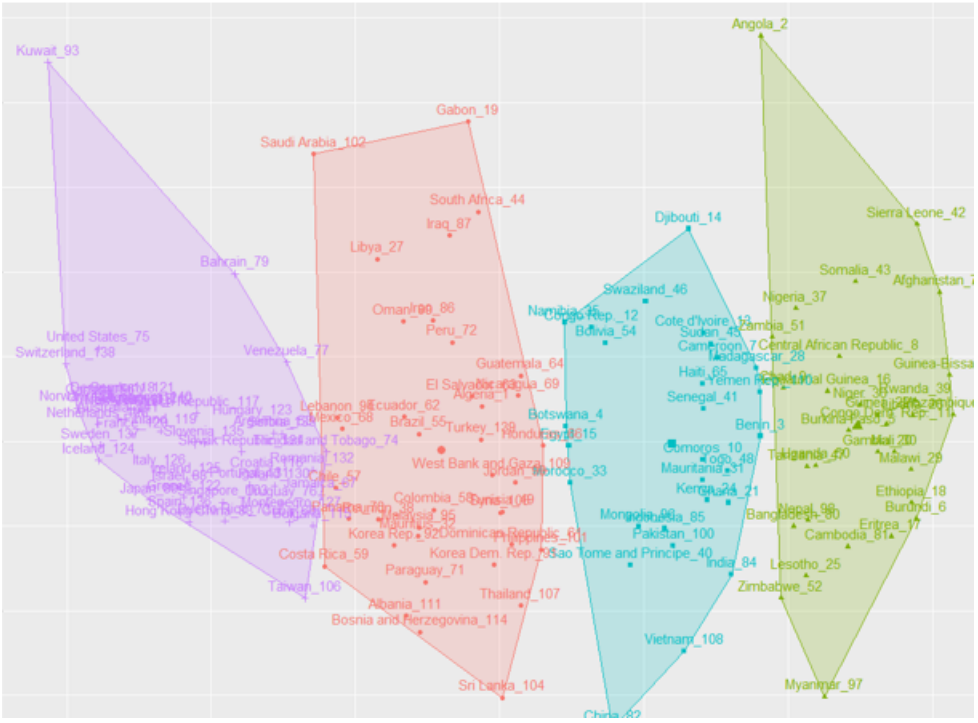

Figure 19: CLUSTERING USING KNN

As we can see from the elbow method plot above,for k=4 the slope changes the slowest and remains less changing as compared to other k's, so we will choose to make 4 clusters.After knn clustering, we are supposed to perform agglomerative hierarchical clustering.To get started with this approach it is important to show the bottom up approach of the dendogram.Each observation starts as its own cluster and merges with other observations.

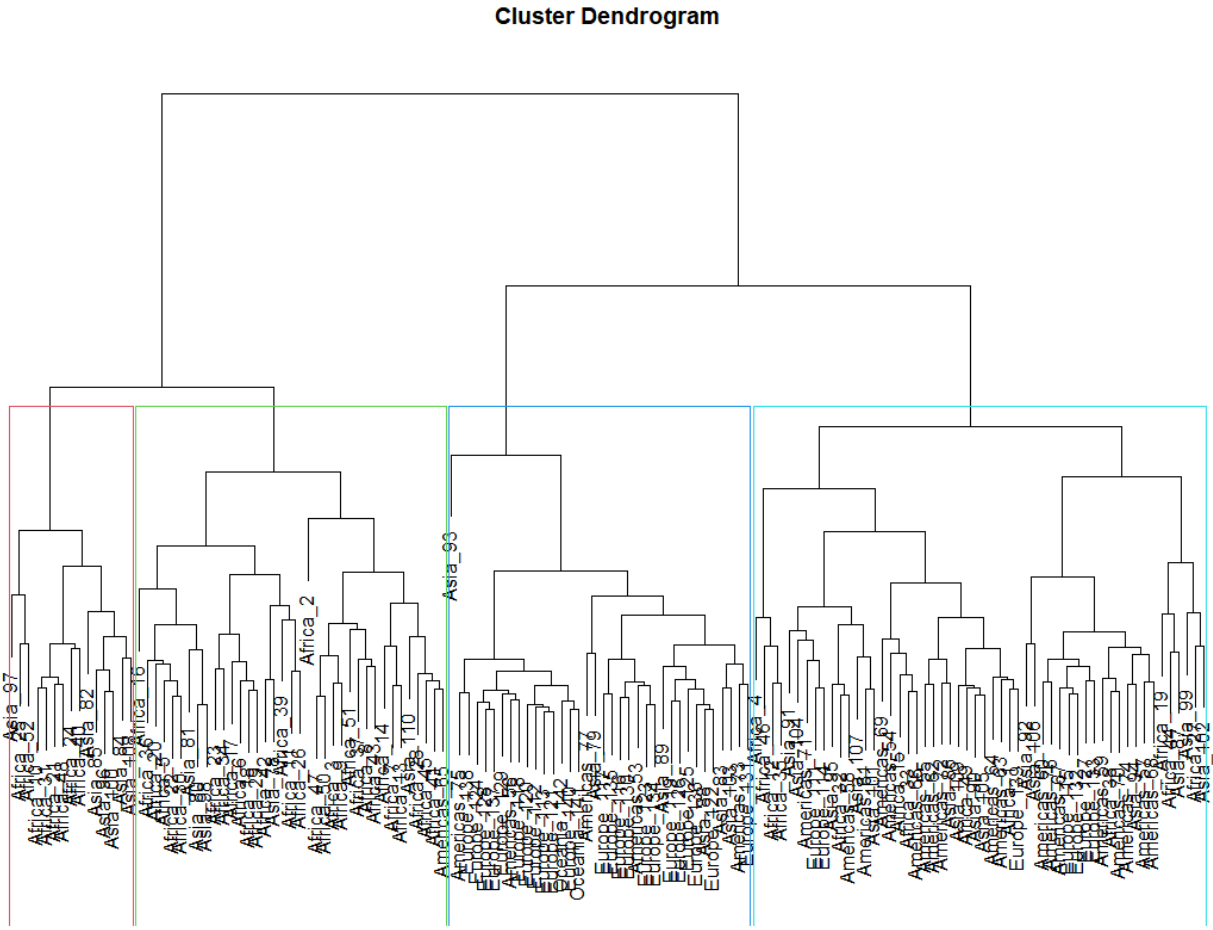

Figure 20: CLUSTER DENDOGRAM

After trying various methods such as average, single and complete for clustering, the clusters obtained from complete and average were the most promising and had minimal overlapping involved. The average hierarchical clustering could be considered to be the best out of all of the clustering algorithm. Kuwait being an outlier has not been clustered and it also uses the least amount of clusters.

Figure 21: HIERARCHICAL-CLUSTERING AVERAGE(RIGHT) AND COMPLETE(LEFT)

After going through all of the clusters above, it would be safer to infer that apart from a few outliers most of the countries have been clustered in a way that the whole cluster is made up of countries present in the same continent.The left most cluster has majority of the European,North American and Oceanian countries that had high GDP and life expectancy.

The African and the low gdp ridden asian countries are present in the right most cluster whereas majority of the asian and middle eastern countries are present in the middle.

K-NN and Hierarchical Clustering are based on the same ground idea but they work in an opposite way albeit having some similarities including the fact that the number of clusters to be made can be set in the same way.Hence both of them have 4 clusters

PERFORMING PRINCIPAL COMPONENT REGRESSION

principal component regression (PCR) is a regression analysis technique that is based on principal component analysis (PCA).In PCR, instead of regressing the dependent variable on the explanatory variables directly, the principal components of the explanatory variables are used as regressors.

We will use PCR to predict the life expectancy in 2007 from the GDP values

After fitting the model using the pcr function of the pls library and enabling cross validation, we need to find how many components we will need to retain to ensure best performance

After going through the plot we can infer that we need 4 components and the RMSE value is 12.43

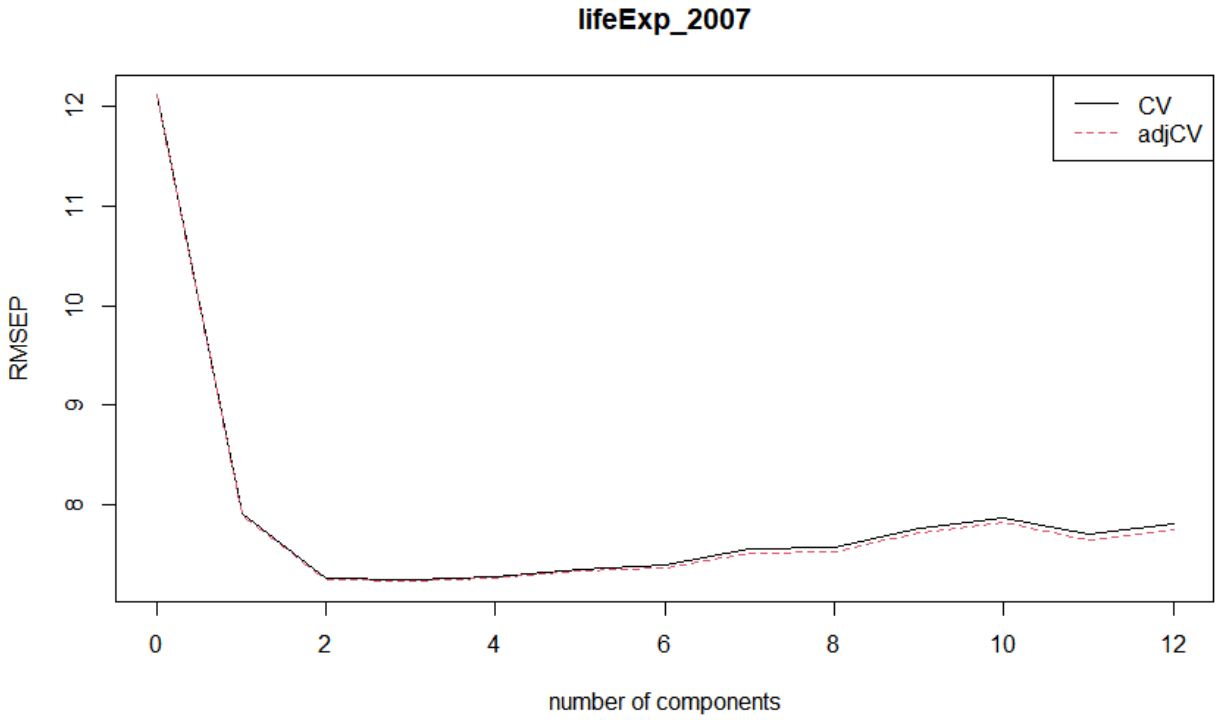

Figure 22: RMSEP VS COMPONENTS