Life Expectancy

DCS 424, Advanced Data Analysis

Dhruv Chandulal Dobariya, Sidhant Thakur, Parth Babubhai Patel

Introduction:

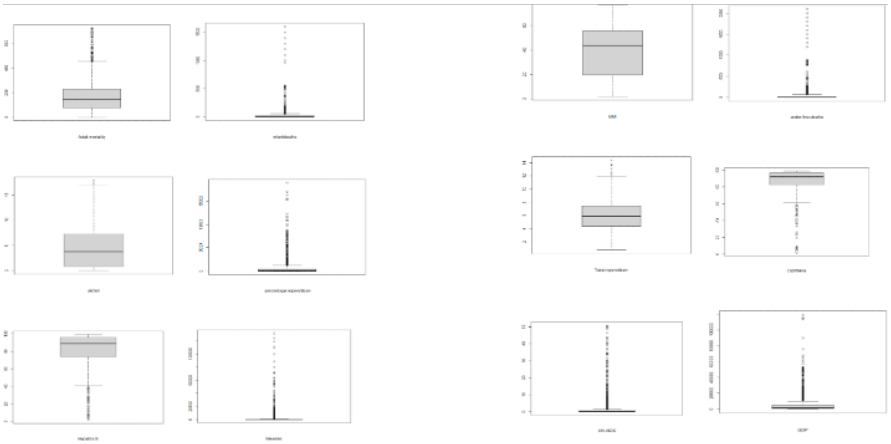
The dataset has 22 columns and 2938 rows. The dataset has 19 continuous variables and 3 categorical variables. The purpose of this research is to find the predicting factor that contributes to a lower life expectancy value. This will help a country decide which sectors should be targeted in order to effectively enhance its population's life expectancy

Exploratory Analysis of the Data:

The dataset included 2938 observations and 22 variables. Approximately 3.9 percent of the data is missing. As a result, we examined each variable by determining how much data was missing, determining whether the variable contained useful information, and addressing the missing data. We addressed missing data in the data set by substituting the mean for missing values.

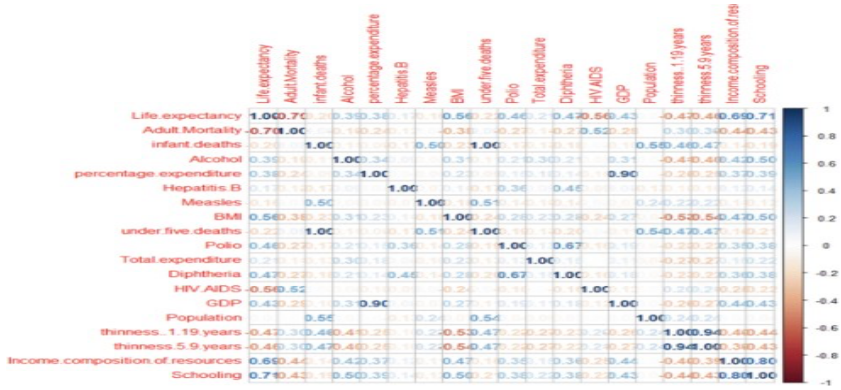
Checking Outliers and skewness:

From the boxplot, it is seen that our data is skewed and have outlier



To mitigate the skewness, we used log transformation and IQR method to remove outliers.

Heatmap:



As shown in the plot, there are variables that are correlated to each other which can cause issues with multicollinearity as the variables are predicted and affected by each other.

There are multiple variables that have a high correlation but the variables that have the highest correlation value are percentage.expenditure and GDP, schooling and income.composition.of resources, adult mortality and life.expectancy

Checking Overfitting:

```
Call:
lm(formula = Life.expectancy ~ ., data = Train)

Residuals:
     Min       1Q   Median       3Q      Max
-15.0852  -2.1380  -0.1628   2.2479  16.3522

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     5.551e+01  8.239e-01  67.366  < 2e-16 ***
Adult.Mortality                -2.189e-02  1.122e-03 -19.516  < 2e-16 ***
infant.deaths                   1.044e-01  1.185e-02   8.816  < 2e-16 ***
Alcohol                         1.064e-01  3.449e-02   3.084  0.00208 **
percentage.expenditure          4.869e-05  1.405e-04   0.347  0.72896
Hepatitis.B                    -1.558e-02  5.447e-03  -2.860  0.00429 **
Measles                        -2.071e-05  9.911e-06  -2.089  0.03687 *
BMI                             3.785e-02  7.108e-03   5.324 1.17e-07 ***
under.five.deaths              -7.743e-02  8.782e-03  -8.816  < 2e-16 ***
Polio                           2.987e-02  6.468e-03   4.619 4.21e-06 ***
Total.expenditure               4.309e-03  5.164e-02   0.083  0.93351
Diphtheria                      3.882e-02  6.745e-03   5.755 1.05e-08 ***
HIV.AIDS                       -4.498e-01  2.492e-02 -18.047  < 2e-16 ***
GDP                             5.158e-05  2.251e-05   2.291  0.02208 *
Population                     -1.026e-09  2.033e-09  -0.505  0.61386
thinness..1.19.years           -9.582e-02  6.750e-02  -1.420  0.15592
thinness.5.9.years             -3.195e-03  6.665e-02  -0.048  0.96178
Income.composition.of.resources  5.562e+00  9.786e-01   5.683 1.60e-08 ***
Schooling                       7.078e-01  6.213e-02  11.391  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.114 on 1450 degrees of freedom
Multiple R-squared:  0.8139,    Adjusted R-squared:  0.8116
F-statistic: 352.4 on 18 and 1450 DF,  p-value: < 2.2e-16
```

```
> # Find the RMSE of the training set
> rmseOlsTrain = sqrt(mean(olsFit$residuals^2))
> rmseOlsTrain
[1] 4.087661
> # Predict on the test set
> olsPred = predict(olsFit, Test)
> # Compute the RMSE of the predictions on the test set
> rmseOlsTest = sqrt(mean((olsPred - Test$Life.expectancy)^2))
> rmseOlsTest
[1] 4.054692
```

There is a small difference between testing and training set so, there is no overfitting in the data.

Application of Analysis: The dependent variable is analyzed in three ways: linear discriminant analysis, principal component analysis, and multidimensional scaling and clustering analysis.Using both regularized regression and principal component analysis to calculate numeric performances on the independent variable, and multidimensional is to do exploratory analysis and to visualize the similarity/dissimilarity and cluster analysis to see how to cluster the independent variables so, they are close to each other on some set of variables.

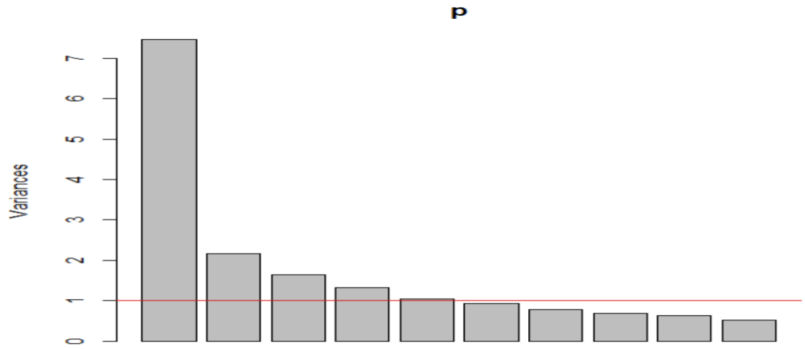
Principal Component Analysis:

In this section we will describe the process using Principal Component Analysis for variable selection.

```
> summary(ds)
 Life.expectancy Adult.Mortality    Alcohol        Hepatitis.B         BMI
 Min.   :45.30   Min.   :  1.0   Min.   : 0.010   Min.   :59.00   Min.   : 2.00
 1st Qu.:69.00   1st Qu.: 71.0   1st Qu.: 1.290   1st Qu.:80.94   1st Qu.:28.60
 Median :73.80   Median :126.5   Median : 4.603   Median :93.00   Median :51.25
 Mean   :72.91   Mean   :128.3   Mean   : 5.083   Mean   :89.62   Mean   :43.68
 3rd Qu.:77.00   3rd Qu.:179.0   3rd Qu.: 8.195   3rd Qu.:97.00   3rd Qu.:57.60
 Max.   :89.00   Max.   :441.0   Max.   :16.580   Max.   :99.00   Max.   :87.30
     Polio       Total.expenditure   Diphtheria   thinness..1.19.years thinness.5.9.years
 Min.   :52.00   Min.   : 0.650    Min.   :51.0   Min.   : 0.100       Min.   : 0.100
 1st Qu.:89.00   1st Qu.: 4.570    1st Qu.:89.0   1st Qu.: 1.400       1st Qu.: 1.400
 Median :95.00   Median : 5.938    Median :95.0   Median : 2.500       Median : 2.600
 Mean   :91.91   Mean   : 5.958    Mean   :91.8   Mean   : 3.633       Mean   : 3.653
 3rd Qu.:98.00   3rd Qu.: 7.410    3rd Qu.:98.0   3rd Qu.: 5.500       3rd Qu.: 5.400
 Max.   :99.00   Max.   :11.710    Max.   :99.0   Max.   :15.300       Max.   :15.500
 Income.composition.of.resources   Schooling       log_infant     log_percentageexpenditure
 Min.   :0.3080                  Min.   : 4.70   Min.   :0.0000   Min.   :0.000
 1st Qu.:0.6276                  1st Qu.:11.70   1st Qu.:0.0000   1st Qu.:2.119
 Median :0.7225                  Median :13.00   Median :0.8959   Median :4.955
 Mean   :0.7070                  Mean   :12.99   Mean   :1.3474   Mean   :4.321
 3rd Qu.:0.8030                  3rd Qu.:14.78   3rd Qu.:2.3979   3rd Qu.:6.404
 Max.   :0.9480                  Max.   :19.70   Max.   :6.1269   Max.   :9.877
  log_Measles     log_underfivedeaths  log_HIVAIDS         log_GDP       log_Population
 Min.   : 0.000   Min.   :0.000       Min.   :0.09531   Min.   : 2.536   Min.   : 7.858
 1st Qu.: 0.000   1st Qu.:0.000       1st Qu.:0.09531   1st Qu.: 7.055   1st Qu.:13.015
 Median : 1.792   Median :1.099       Median :0.09531   Median : 8.360   Median :15.309
 Mean   : 2.757   Mean   :1.469       Mean   :0.21578   Mean   : 8.065   Mean   :14.545
 3rd Qu.: 4.881   3rd Qu.:2.565       3rd Qu.:0.18232   3rd Qu.: 8.921   3rd Qu.:16.361
 Max.   :11.804   Max.   :6.326       Max.   :1.30833   Max.   :11.688   Max.   :19.332
```

From the summary of the dataset, we can see that there is a lot of variation in our dataset, so we need to scale it.

Scree Plot:



The appropriate number of factors to extract from the scree plot is four. Then we used varimax rotation to perform PCA. The number of components with eigenvalues greater than variance =1 shown on the scree plot suggests four components

Principal factor analysis (PFA)

```
> p = principal(ds,rotate = "varimax", nfactors=4)
> print(p$loadings, cutoff=.4, sort = T)

Loadings:
                               RC1    RC4    RC2    RC3
Life.expectancy                 0.768
Adult.Mortality                -0.656
Income.composition.of.resources 0.829
Schooling                       0.733
log_percentageexpenditure       0.596
log_HIVAIDS                    -0.603
log_GDP                         0.694
Alcohol                                0.586
Total.expenditure                      0.663
thinness..1.19.years                  -0.855
thinness.5.9.years                    -0.851
Hepatitis.B                                   0.892
Polio                                         0.889
Diphtheria                                    0.888
log_infant                     -0.450                 0.771
log_Measles                                           0.704
log_underfivedeaths            -0.468                 0.762
log_Population                                        0.609
BMI                                    0.449

                 RC1   RC4   RC2   RC3
SS loadings    4.430 3.068 2.889 2.203
Proportion Var 0.233 0.161 0.152 0.116
Cumulative Var 0.233 0.395 0.547 0.663
> |
```

Loading Results

The first rotated component (RC1) accounts for 23 percent of the variance, with RC4 accounting for 16 percent, RC2 accounting for 15 percent, RC3 accounting for 11 percent.
Above figure shows the rotated components with each variable's loadings. We did this because it makes division clear and easy to be interpreted with the cut off 0.4.
The components are sorted according to their eigenvalues, with the accumulated variance captured in each component.

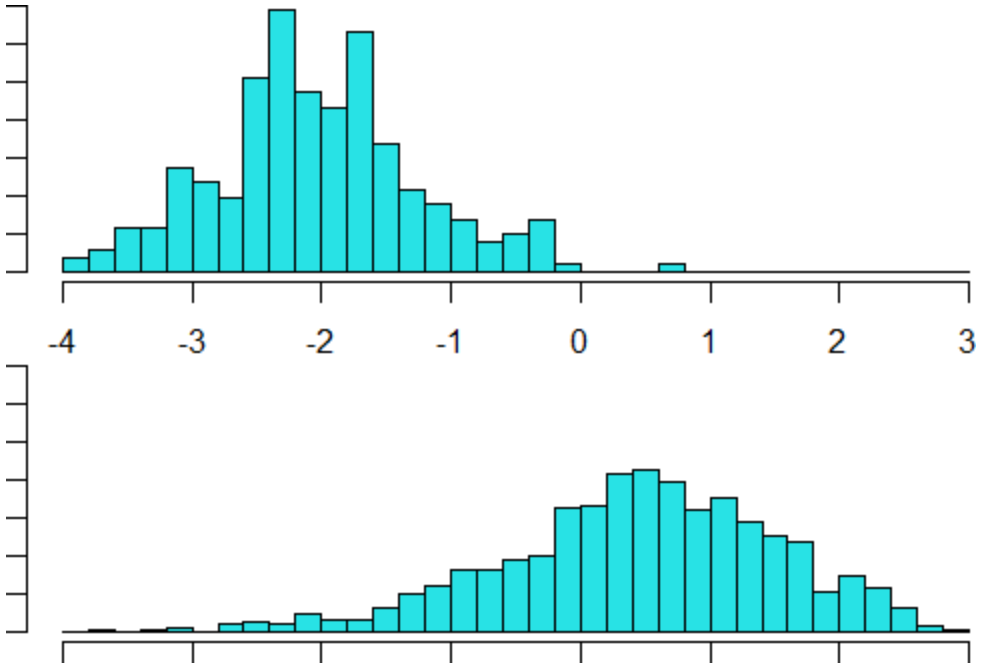
The first component (RC1) comprises four positive contributions from Life. Expectancy, income composition of resources, schooling, percentage expenditure, GDP and two negative contributions of adult mortality and HIV AIDS. This component can be interpreted as a country which has high life expectancy.

The second component (RC4) consists of three positive contributions of Alcohol, total expenditure, BMI, and two negative contributions: thinness 1.19 years, thinness 5.9 years. This could represent a country group where individuals consume more alcohol.
The third component (RC2) is composed of all positive contributions of Hepatitis, polio, Diphtheria. This might refer to a collection of countries whose populations are plagued by these three diseases.

The last component (RC3) consists of all positive contributions of infants, measles, under five deaths and population. This could point to a country where deaths from these diseases are more common in the population.

Finally, It avoids multicollinearity while capturing a significant amount of variance (66 percent).As a result, PCA extracts factors or hidden latent variables from this data set.

Linear Discriminant Analysis:




```
> plot(Trainlda.values$x[, 1], Trainlda.values$x[, 1], pch=16)
> # Compute a confusion matrix
> table(Train$Status, Trainlda.values$class)

             Developed Developing
  Developed        199         36
  Developing        61       1173
> source("C:/Users/dobar/Desktop/DePaul_assignments/3rd quarter/DSC 424/Confusion.R")
> confusion(Trainlda.values$class, Train$Status)
                 Accuracy  Prior Frequency.Developed Prior Frequency.Developing
                    0.934                      0.177                       0.823
```

```
Call:
lda(Status ~ ., data = Train)

Prior probabilities of groups:
 Developed Developing
  0.172226   0.827774

Group means:
           Life.expectancy Adult.Mortality  Alcohol Hepatitis.B      BMI    Polio
Developed         78.85850        82.72332 9.516245    85.06949 52.31265 93.44664
Developing        67.21464       180.87844 3.566146    80.04613 35.22464 79.97044
           Total.expenditure Diphtheria thinness..1.19.years thinness.5.9.years
Developed           7.494905    92.59289             1.362451           1.333202
Developing          5.628509    80.69454             5.595456           5.587743
           Income.composition.of.resources Schooling log_infant log_percentageexpenditure
Developed                        0.8323659  15.50217  0.4743287                  5.805417
Developing                       0.5823061  11.24946  2.0536602                  3.532393
           log_Measles log_underfivedeaths log_HIVAIDS  log_GDP log_Population
Developed     2.810326           0.5458789  0.09531018 9.112529       13.99980
Developing    3.555996           2.2489436  0.58562725 7.400335       14.55123
```

```
Coefficients of linear discriminants:
                                        LD1
Life.expectancy                -4.026606e-02
Adult.Mortality                 8.524779e-04
Alcohol                        -1.955806e-01
Hepatitis.B                    -7.553944e-05
BMI                             2.295199e-04
Polio                          -4.619332e-04
Total.expenditure              -1.035328e-01
Diphtheria                      4.427386e-03
thinness..1.19.years           -7.050625e-03
thinness.5.9.years              2.915852e-02
Income.composition.of.resources -8.402702e-01
Schooling                      -3.899956e-02
log_infant                      1.159300e+00
log_percentageexpenditure      -4.975337e-02
log_Measles                    -1.072171e-01
log_underfivedeaths            -1.012247e+00
log_HIVAIDS                    -8.643057e-02
log_GDP                        -1.575576e-02
log_Population                  1.915650e-02
>
```

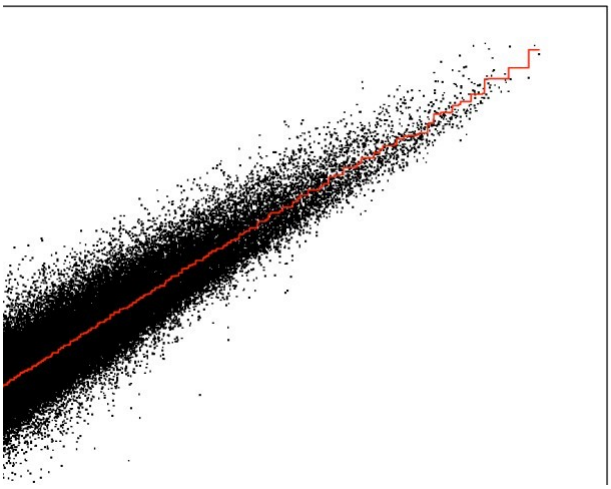
We tried applying LDA in a binary variable which is STATUS. Based on the confusion matrix we can say the true values on the rows and the predicted values on the columns and gain some valuable insight. Based on the above plots we can say that separation between Developing and Developed is good but with somewhat overlapping.

Multidimensional Scaling and Clustering Analysis

The final component of our project explains how we used clustering. Clustering is an unsupervised machine learning approach for discovering intriguing patterns. The same initial dataset was used with some extra data purification to prepare it for the clustering method, such as scaling, excision of outlier and conducting log transformations.

We have tried different dimension by non-matric MDS for clustering on distance similarity matrix of the scaled data and by examining the scree plot of stress and dimension we have select the dimension which best fit for data representation of original data.

Dimension	1	2	3	4	5
Stress	25.80%	17.31%	12.44%	9.61%	7.68%

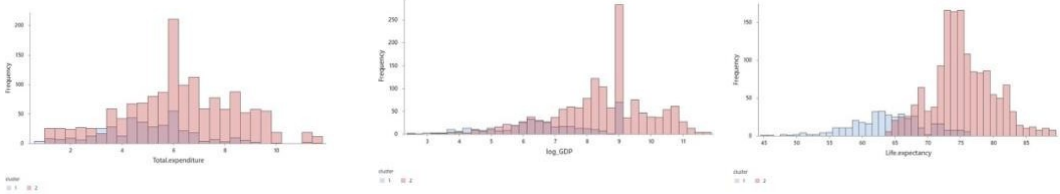


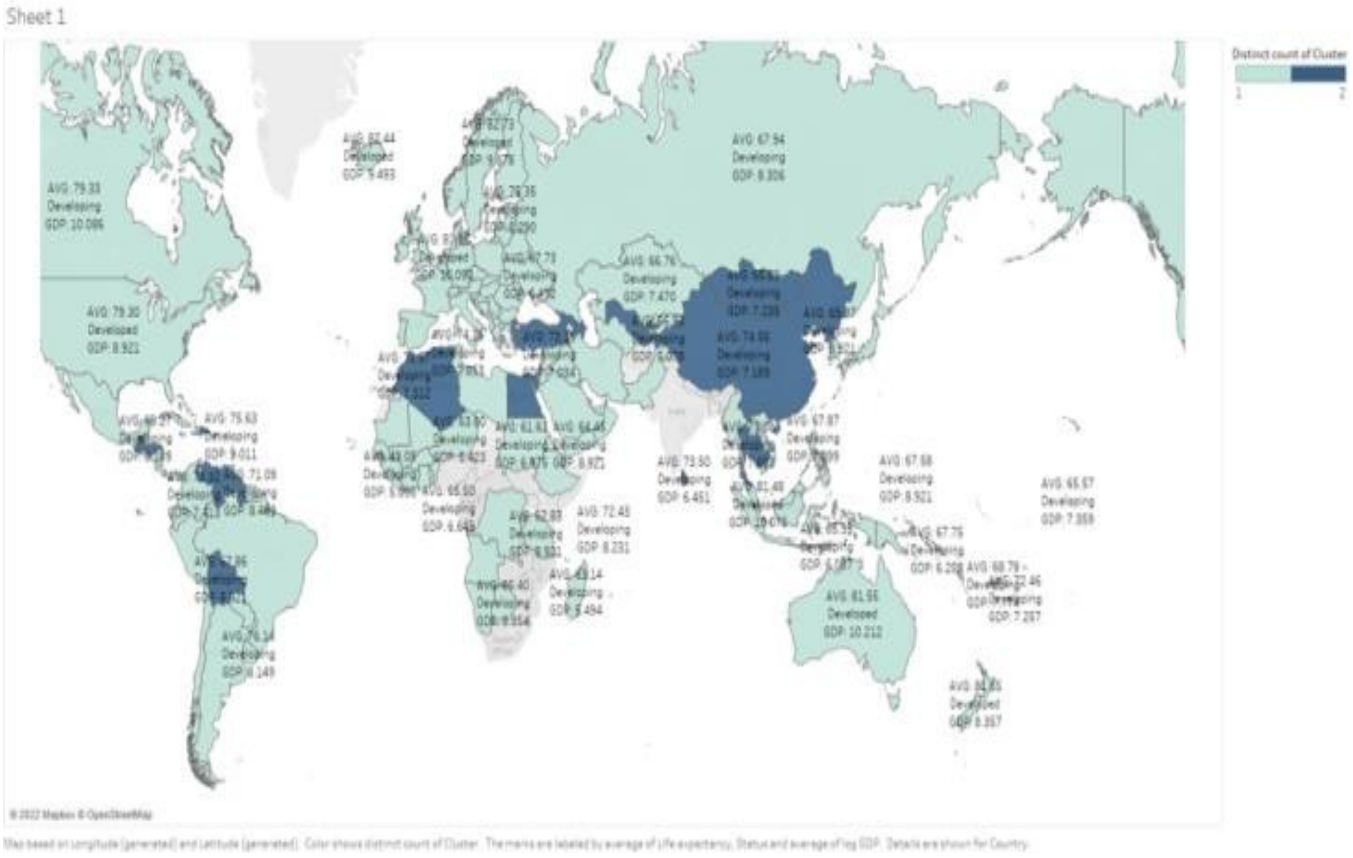
```
> dens = dbscan(ds, eps=0.5, MinPts = 3)
> dens
dbscan Pts=1854 MinPts=3 eps=0.5

   0
1854
```

Dimension 3 is our preferred option since it reduces stress by 12.44 percent.
We can see from the Shepard diagram that there is a step-line that represents a decent fit, but it is not readily evident due to the vast amount of data points.

Clustering based on density on 1854 observations of life expectancy data. First, we scaled our data set and used clustering with an epsilon value of 0.5 and a minimum points value of 3 on the data.
Only one cluster was returned by the algorithm, and it contained all our data. To get a solid idea of life expectancy, we'd like our clusters to be more evenly distributed not only in one cluster.
It would seem DBSCAN is not the optimal clustering algorithm for this dataset.

We then used K-mean clustering once more, this time selecting k=2 as the best option. For assessment, we combined the k-mean cluster with the original dataset.
We used life to display data. The dataset's expectancy, log.GDP, cluster, and country properties. According to statics and graphs, we can't cluster the data since we found observations in both clusters that correspond to virtually the same range; hence we can't declare that data from a certain range belongs to a specific cluster.





We experimented with several clustering algorithms. We've concluded that we won't be able to cluster this data using any approach. The average life expectancy of humans is depicted on a globe map based on GDP and nation status.

Conclusion:

By, applying PFA and LDA, we can group the variables into their respective groups, and it make the data easier to interpret and understand.

Schooling has a significant impact on life expectancy.

Countries with higher income composition of resources for human development have a better life expectancy.

After performing the clustering on life Expectancy Data and we observe that Life expectancy in developed countries is more than that of developing countries.