**MA317 – Group Coursework**

**Group :39**

**Registration Number : 2106458**

**The registration numbers of other group member (1)2111142**

**(2)2105732**

**(3)2111177**

**(4)2111060**

Table of Contents

[Table of Contents 2](#_Toc99211660)

[Abstract 3](#_Toc99211661)

[Introduction 4](#_Toc99211662)

[Preliminary analysis 5](#_Toc99211663)

[1. Descriptive statistics 5](#_Toc99211664)

[2. Dealing with Missing values 10](#_Toc99211665)

[3. Collinearity Analysis 12](#_Toc99211666)

[4. Analysis for life expectancy 14](#_Toc99211667)

[5. Design of Experiments 20](#_Toc99211668)

[Conclusion 22](#_Toc99211669)

[References 23](#_Toc99211670)

[Individual Contribution 23](#_Toc99211671)

[Appendix 24](#_Toc99211672)

# Abstract

For the track development different indicators are initiated in overall countries and in over time. The method of world bank’s world development indicators WDI database occupies with 1300 time to accommodate the socio-economic indicators and it results needed input towards the global trade analysis projects GTAP, further it enables to occupy the growth of socio-economic scenarios. Some of the indicators saved towards the database has been derived to be the highest socio-economic indicators in the level of national based. Some of the often-applied GDP and components through the population. Some of the indicators was derived to be the annual based time process which would be back towards the 1960. The highest number of the indicators were raised directly through the world bank, (in mean while the coordination through some international agencies related to the international monetary fund IMF through united nations), but some of the related numbers are merged through the external database.

# Introduction

The international database was derived to be the world development indicators database and been enhanced to grow and manage the world banks in each sector. The WDI will consists of, for example the data from energy are derived through the international energy agency IEA, the data of labor derived through the international labor organization ILO and data related to agricultural are through the agriculture organization FAO and united nations food. As per the requirement, WDI would be applied on the one-stop shopping comparing with the particular database (Pritwani, et al., 2018) system. Further, through method of open access towards the WDI the API is complicated, and further the data extraction methods are processed through the less cost. The data of WDI can be effectively applied on the different assessing structural in order to modify over the time, for example cohorts’ population, it’s been compared through the GDP agricultural and employment through various indicators those are matched towards the millennium development goals MDGs.

The development of the indicator code 11 was done through the world bank and been initiated as the small description in means of indicator. Further, that is divided through various five methods. The first method was derived to be the topic of ‘SP’ which indicates the population and social indicators. For the overall population the indicators have derived as the SP.POP.TOTL. And the indicators of GDP were elaborated towards the topic national income NY. In the second method, national income was derived for example the GNP or GDP was derived. The third method was that it initiates the income indicator based, for example per capita PCAP through which overall for market price MKTP was initiated. The fourth method was that the indicator which denotes the rate pricing of regime, dollar-based cost, current local currency and purchasing power parity PPP or some of the fixed process in particular year all these groups of fourth. For example, the NY.GDP.MKTP.KD.12 were derived through the constant dollars for GDP in the exchanges of market rates.

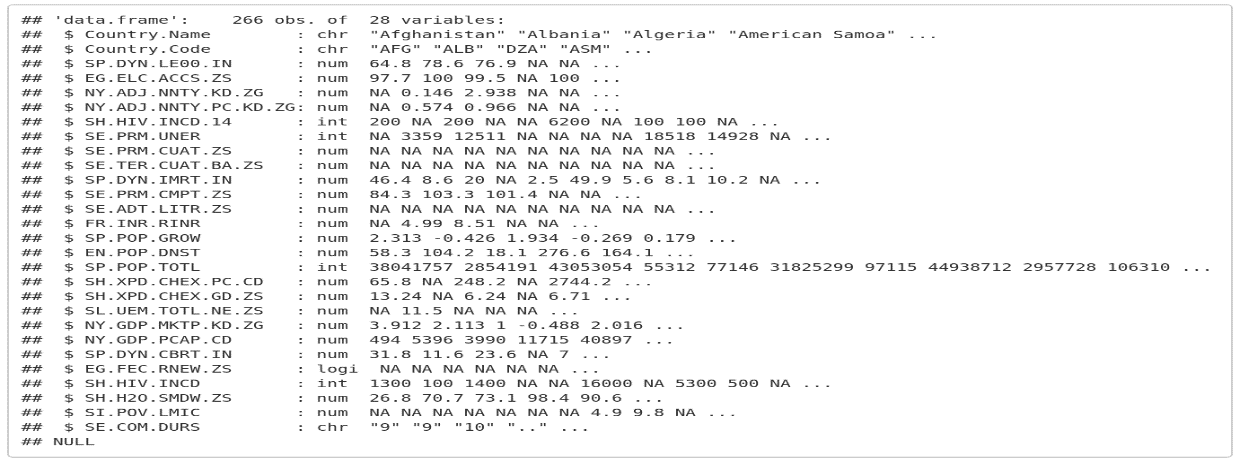
# Preliminary analysis

# Descriptive statistics

|  |  |  |  |
| --- | --- | --- | --- |
| **Code** | **Indicator Name** | **Code** | **Indicator Name** |
| **SP.DYN.LE00.IN** | life\_expectancy | **EN.POP.DNST** | Population\_density |
| **EG.ELC.ACCS.ZS** | Electricity | **SP.POP.TOTL** | total\_population |
| **NY.ADJ.NNTY.KD.ZG** | Net\_National\_income | **SH.XPD.CHEX.PC.CD** | health\_expenditure\_capita |
| **NY.ADJ.NNTY.PC.KD.ZG** | Net\_national\_income\_capita | **SH.XPD.CHEX.GD.ZS** | health\_expenditure |
| **SH.HIV.INCD.14** | childeren\_affected\_with\_HIV | **SL.UEM.TOTL.NE.ZS** | unemployment |
| **SE.PRM.UNER** | children\_out\_of\_school | **NY.GDP.MKTP.KD.ZG** | GDP\_annual |
| **SE.PRM.CUAT.ZS** | Educational\_attainment\_primary | **NY.GDP.PCAP.CD** | GDP\_per\_capita |
| **SE.TER.CUAT.BA.ZS** | Educational\_attainment\_bachelor | **SP.DYN.CBRT.IN** | Birth\_rate |
| **SP.DYN.IMRT.IN** | Mortality\_rate | **EG.FEC.RNEW.ZS** | renewablw\_energy |
| **SE.PRM.CMPT.ZS** | Primary\_completion | **SH.HIV.INCD** | Adults\_HIV |
| **SE.ADT.LITR.ZS** | Literacy\_rate | **SH.H2O.SMDW.ZS** | safely\_drinking\_water |
| **FR.INR.RINR** | Real\_interest\_rate | **SI.POV.LMIC** | poverty |
| **SP.POP.GROW** | Population\_growth | **SE.COM.DURS** | compulsory\_education |

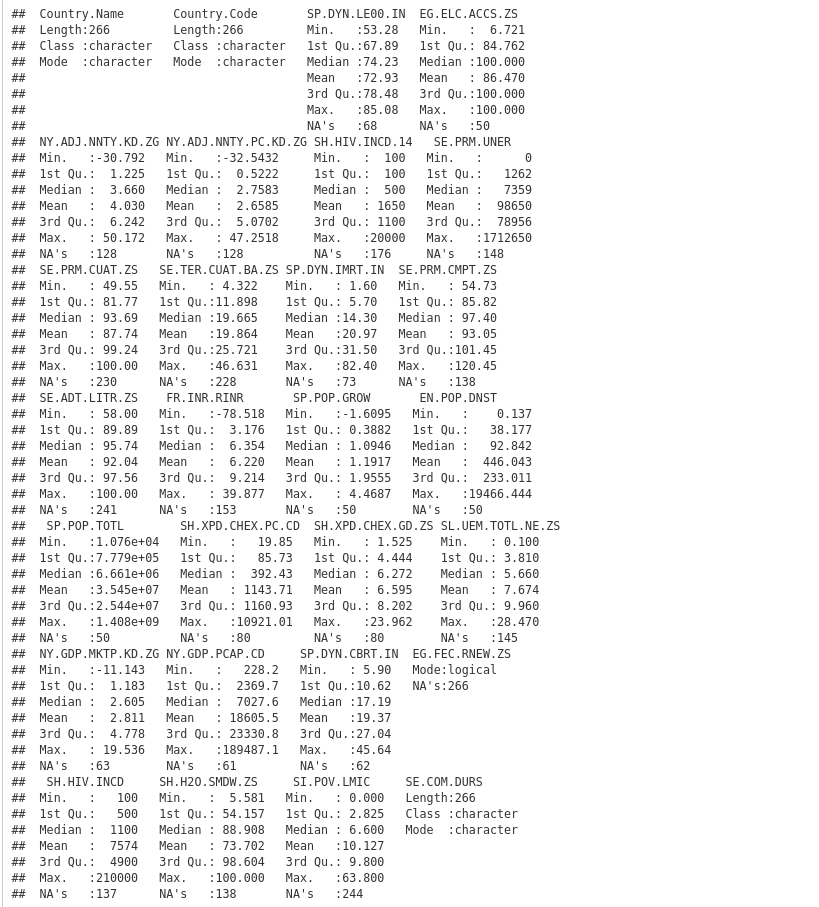
**Table.1. Dataset includes the following world bank indicator variable**

Different 26 variables were initiated on the sets of code, according to the convenience the names are modified through particular names in the above table.



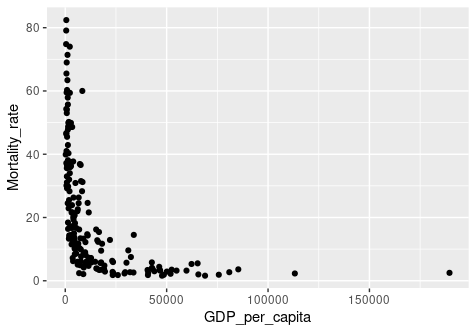
**Fig.1 Dataset structure**

Given dataset processes with 266 variables and observation results, which denotes the WDI width and length. Length and width of the WDI are numeric and integer variables and the Country name and code are characters (Akanmu & Hassan, 2020).



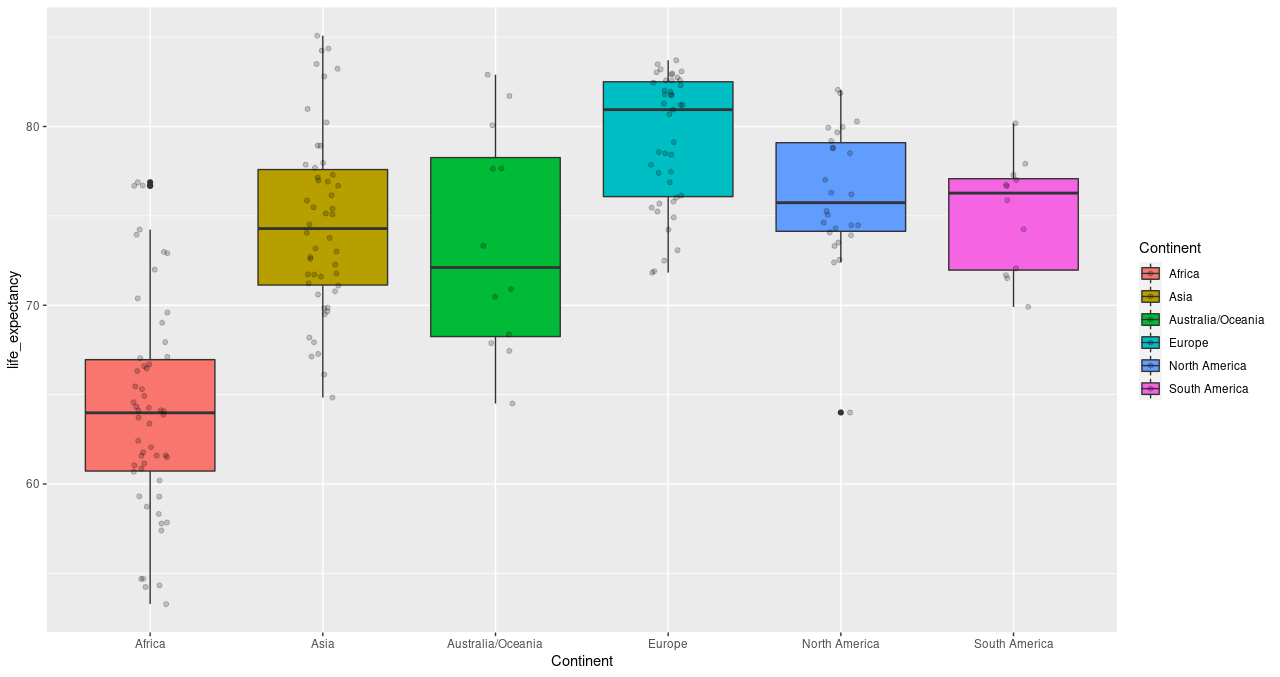
**Fig.2. Summary of dataset**

The mortality rates and GDP capita was initiated on the above figure, it enables to interpret the mortality rate for raising, GDP per capita raises down.

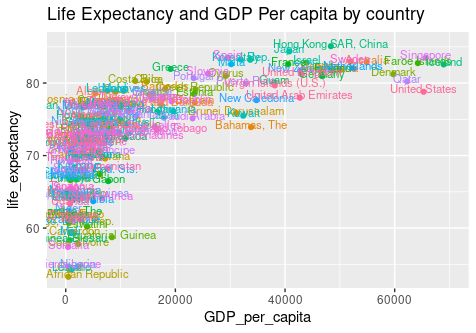


**Fig.4. GDP per capita vs Mortality rate**

In the above analysis, life expectancy and GDP per capita was analyzed, which helps to integrate the raises of country’s GDP through the raise of life expectancy of people.

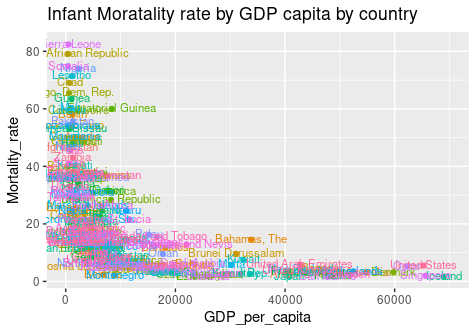


**Fig.5. Life expectancy boxplot over continents**

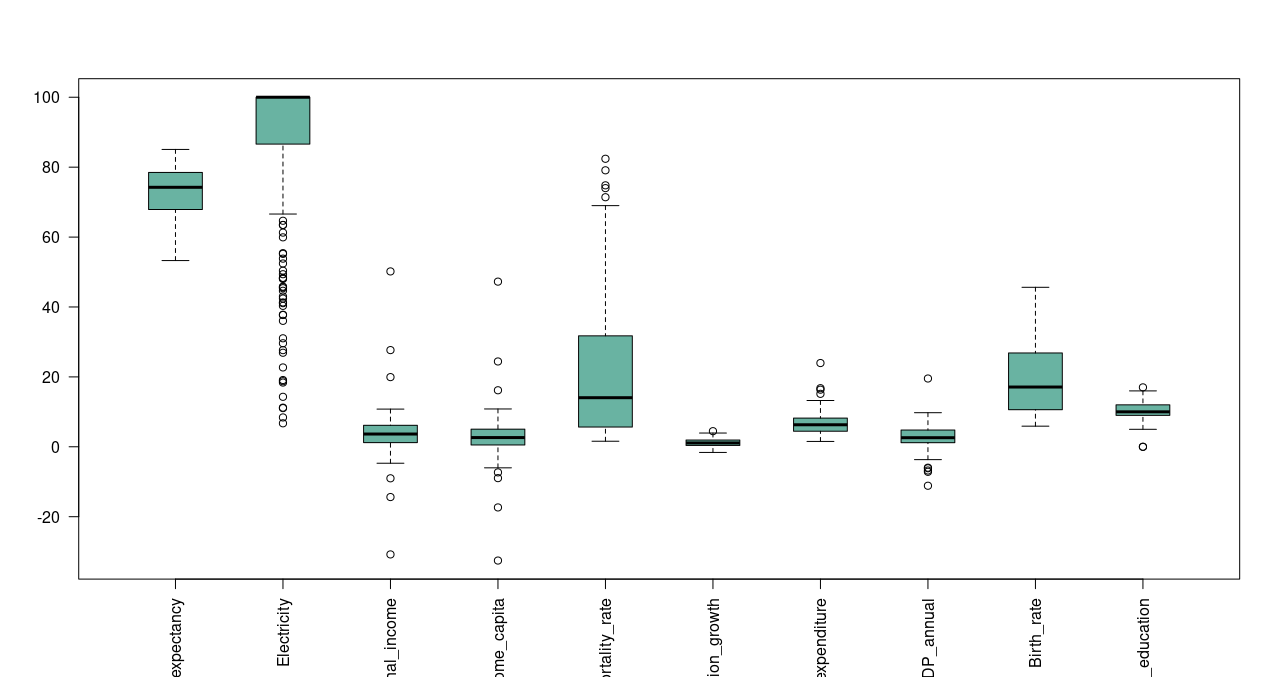


**Fig.6. Life expectancy and GDP per capita by country**

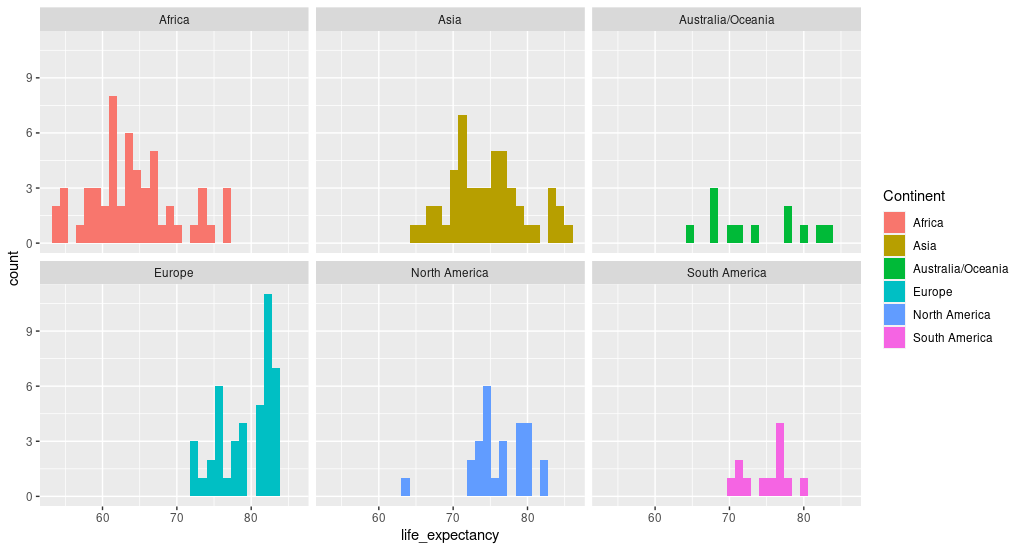
In above analysis, statistics regarding the mortality rate and GDP for capita was derived which helps to integrate the infant mortality for raises of GDP capita through raise.



**Fig.7. Infant Mortality rate by country GDP per capital**

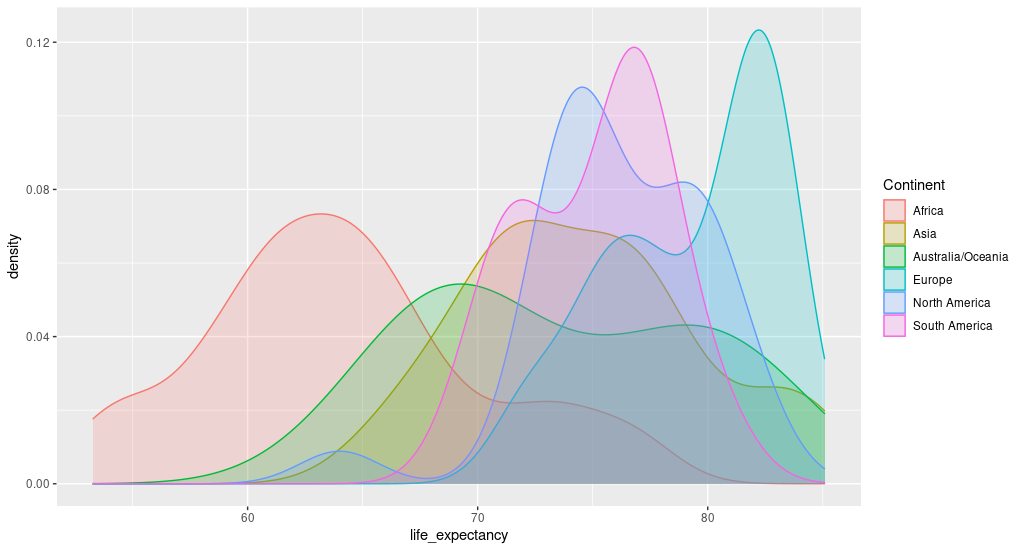
****

**Fig.8. boxplot for varaibles other than count**



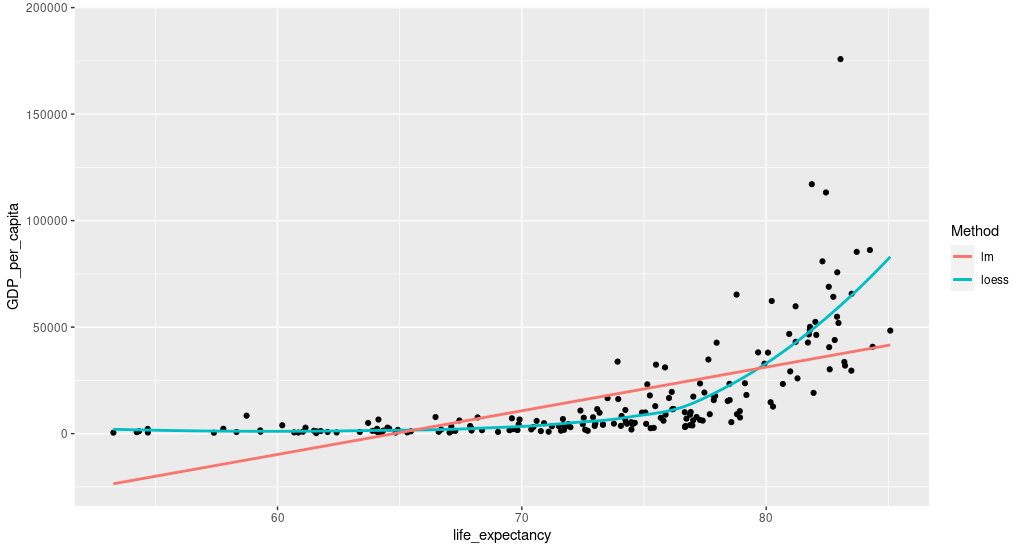
**Fig.9. histogram for life expectancy over continents**

The above graph illustrates the histogram for life expectancy in all over the continents, some of the six continents are illustrated.



**Fig.10. Density plot for life expectancy over continents**

The above graph initiates the density plot in all over continents, for the seven different continents the graph is done.



**Fig.11.** **Life expectancy density country wise**

The above figure shows the life expectancy density for the country basis and method was initiated.

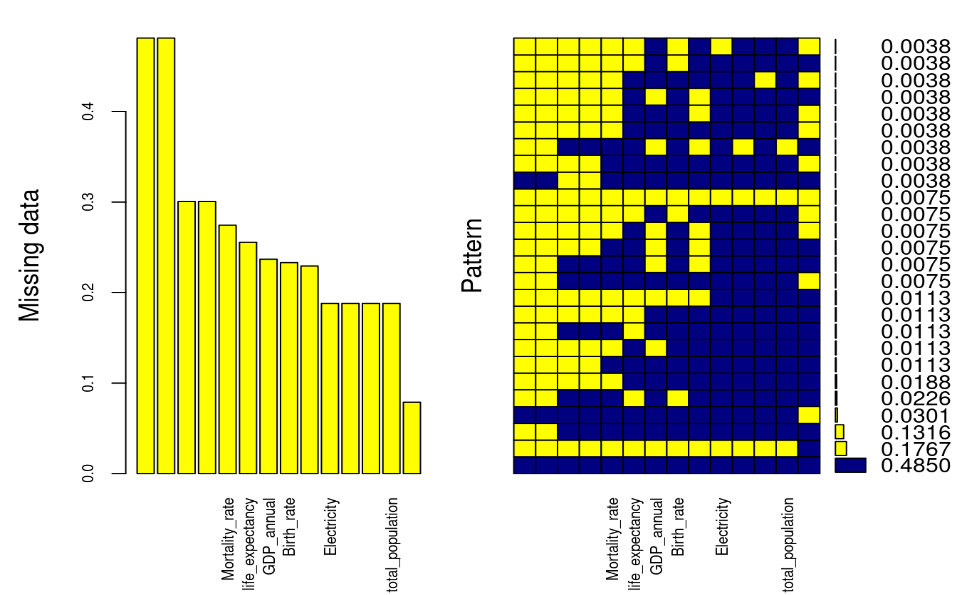
# 2. Dealing with Missing values

**1. Checking missing values in dataset**

|  |  |  |  |
| --- | --- | --- | --- |
| **Indicator** | **Total NA** | **Indicator** | **Total NA** |
| **life\_expectancy** | 68 | **Educational\_attainment\_bachelor** | 228 |
| **Electricity** | 50 | **Mortality\_rate** | 73 |
| **Net\_National\_income** | 128 | **Primary\_completion** | 138 |
| **Net\_national\_income\_capita** | 128 | **Literacy\_rate** | 241 |
| **childeren\_affected\_with\_HIV** | 176 | **Real\_interest\_rate** | 153 |
| **children\_out\_of\_school** | 148 | **Population\_growth** | 50 |
| **Educational\_attainment\_primary** | 230 | **Population\_density** | 50 |
| **health\_expenditure** | 80 | **total\_population** | 50 |
| **Unemployment** | 145 | **health\_expenditure\_capita** | 80 |
| **GDP\_annual** | 63 | **renewablw\_energy** | 266 |
| **GDP\_per\_capita** | 61 | **Adults\_HIV** | 137 |
| **Birth\_rate** | 62 | **safely\_drinking\_water** | 138 |
| **compulsory\_education** | 21 | **poverty** | 244 |

**Table 2.2 Number of NA in each indicator**

Its indicators like poverty, renewable energy is completely NA and 14 out of 28 indicators having more than 50% of NA are removed.



**Figure.12. Missing values in each Indicators**

The above figure initiates the indicator which has less than 50% of data related to NA and some patterns.

The random forest algorithm has implemented the Miss Forest. This is the method of imputation in non-parametric which are suitable for different types of variables. The explicit assumptions could not be made through the non-parametric method those are related to the functional method if y any kinds of arbitrary function. As the compensation (Ali, et al., 2012), it enables to estimate the f those could be closely related to the data points by not progressing the impractical. In small words, it enables to grow the random forest model for different variables. Further, it applies the enhanced model in order to predict the missing values in the variable through help of the gained values. It accesses the OOB which is out of bag derived through the imputation method for error estimation. Further, it results the large level of controls towards the process of imputation. It has some of the equivalent options in order to reverse the OOB separately in different variables rather of aggregating the whole data matrix in it. It helps to closely look for the data based on accurate model observation how much it has imputed the values for different variables (Biau, 2012).

The NRMSE is been derived as the normalized mean square error. It helps to initiate the errors those are observed through the continuous values of imputing. The PFC portion of falsely classified, helps to initiate the different errors those are derived through values of imputing methods. This helps in running variables, those are implemented through the 2% of errors. This could be developed through tuning of values those are mtry and ntree parameters methods. The mtry indicates the different variable numbers those are randomly applied towards the sample values for splitting. The ntree initiates the groups of trees in order to develop the forest.

# 3. Collinearity Analysis

Splitting dataset of 80% as training and 20% as testing dataset. After imputation, computing linear regression with life expectancy as y variable and rest as x variable. RMSE we got was 1.78 and R-square 0.918. While check collinearity

|  |  |  |  |
| --- | --- | --- | --- |
| **Indicator** | **VIF** | **Indicator** | **VIF** |
| **Electricity** | 3.628383 | **Net\_National\_income** | 84.42845 |
| **Net\_national\_income\_capita** | 86.129617 | **Mortality\_rate** | 5.18791 |
| **Population\_growth** | 5.185404 | **Population\_density** | 1.624714 |
| **total\_population** | 1.056881 | **health\_expenditure\_capita** | 4.875857 |
| **health\_expenditure** | 1.663265 | **GDP\_annual** | 1.827891 |
| **GDP\_per\_capita** | 4.687688 | **Birth\_rate** | 9.962839 |
|  |  | **compulsory\_education** | 1.245202 |

**Table.3. VIF value of Indicator**

VIF has the minimum potential of one (absence of multicollinearity). A VIF score greater than 5 or 10 implies a significant level of co-integration, as a rule of thumb (Daoud, 2017).

For analyzing the forecast error attributable to a lack of model fit estimation must be made. A regression’s residual mean square error can be divided into two parts, and they are:

* Based on the lack of compatibility
* As a result of random variability

The design must be scrapped and a new one should be created if the majority of the mistake is because of the shortage of compatibility rather than experimental errors. Only simple linear regression will be compatible with the deficit of fit F test. Furthermore, the data must include repeated occurrences, which means duplicating at least several of the predictor x values. Only databases with a large number of repetitions should be used for this test.

|  |  |
| --- | --- |
| https://lh5.googleusercontent.com/Vi0oZHLurDgW-acLxAbJ8uh0PA1UxuqjTjZ5kWGv_kmNXJtiaT0epj6ZhFrr8b4pL6FN4wwRC7Q69K7U0vOBz8rsGv8a5z6utrfiz5nxgHg--mmxoZu6LxZyGd8d1u6reXxtfAxo           Fig 13(a)  Residual Fit Spread Plot | https://lh4.googleusercontent.com/rvlZqf1K1cOQhsevCpV68Gl_P0fXTpeR4ljvbB0-EslU9950ZmLyLMkN9355HHQPNLDtedbn1Fih7YhZerl1eZ5y5X_8CmZ_BYC6zQWCP2UKrjDlrUbNb0dCIRcnAw9K05ph_rhv       Fig.13(b)  Observed vs Predicted Plot |

The above analysis initiates the plots for diagnostics of regression.

|  |  |
| --- | --- |
| https://lh5.googleusercontent.com/OK0rZnDwQ78KsWxCtJiddTvVdZVCBEOxXLFII7Lp5YdS7wKcGYCed9oGBhFoy_DXmlWWUZog7NFnyiAS3eyaGfa9SCQIYOtXaYN4wSGl8dqadpRsSVCs28cR0WwDe_KDBpAwWOez | https://lh5.googleusercontent.com/HwW6KAHrcs8RzmZcBTC3TUyufbeK4WJbthreznCc9R8q5IxD96E-fuYgC9DIsXbsE5IKMkchtal-jeF4wII6Gv64pBsb9F6P8IqK5qYNycnTC-FD7zum0lXxYEkvGj2E3wEUVQS_ |
| https://lh5.googleusercontent.com/0gfL3XIWolAxmE48p8UD7DJY5dybT9eyzocxmw-JPktxYByZABkti7J31uhjG4pSeQbxsVHJQG0Ip-rvcz5VaKlpz_aEPuox2y7zuOlhUiv-bjg0sNYXFmKh5bYOAkfamOW57uOx                                          Fig.14.  Panel of plots for regression diagnostics | |

The communication for the residuals in life expectancy through some variables are shown in above figure.

|  |  |  |
| --- | --- | --- |
| https://lh6.googleusercontent.com/QxKJ-RE1pJzxeJ9WrRGnLxxG6TczEZ8lCiPW-lf-zo7z3mtgZUp2ew3dPsIIVPM4IYRcYr2HPDrh6UC92yxTzYb76hE6xZZjj0lzyEOPhY8vLgDVzVb--ukoG9ds_W4zzrJMj5ZZ | https://lh3.googleusercontent.com/w25125Uf7C9a1zUODmHB4mHq9eJkARPhj_0Ci_WORDjvMOy6XNQ7cTxq5cfLeQIM0eLXeBRq3njYf2iFbXyfoWzzcWsQUcPHzjp4aa-6gLYm7XvIeI4g-xIoGoHTplp6tbghz81r | https://lh5.googleusercontent.com/pKMgsHmSp91WaE-lOjSRZmYyrmYFUtYZBc6pMQe_aweBMCN1LrjxnnbEwZMQbwl4ARjIS1zFTTi3q60R9eOHBDZJfzzYoBB5cVa6JaxoSNWrfhBp-UcTWHaUc-7D2FwsDuvHWvYY |
| **Fig.15.   Residual Plus Component Plot** | | |

The baseline plus element plot shows whether the connection between Y and X axis is non-linear and can propose data adjustments.

# 4. Analysis for life expectancy

Stepwise regression involves adding and subtracting predictions over and over, to the prediction model, in order to obtain a lower set of variables in the data set leading to a more efficient model, which is a model that minimizes predictive error.

There are three ways to Stepwise regression

1. Forward selection, the first of which has no predictions in the model, repeatedly adds the most offered predictions, and stops when improvements are no longer statistically significant.
2. Backward which starts with all the predictions in the model (full model), repeatedly removes the small predictive predictions, and stops if you have a model where all predictions are mathematically important.
3. Stepwise selection, which is a combination of forward and backward selection. You start without predicting, and then add in order the most helpful predictions. After adding each new variant, remove any variables that no longer offer model improvement

**Both model output**

**Model summary**

|  |  |  |  |
| --- | --- | --- | --- |
| R | 0.963 | RMSE | 1.977 |
| R-squared | 0. 928 | Coef. var | 2.690 |
| Adj. R-Squared | 0. 923 | MSE | 3. 908 |
| Pred R-Squared | 0. 914 | MAE | 1.506 |

 RMSE: Root Mean Square Error

 MSE: Mean Square Error

 MAE: Mean Absolute Error

**ANOVA**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Sum of squares** | **DF** | **Mean square** | **F** | **Sig** |
| Regression | 8254.024 | 10 | 825.402 | 211.199 | 0.0000 |
| Residual | 644.849 | 165 | 3.908 |  |  |
| Total | 8898.873 | 175 |  |  |  |

**Parameter estimates**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Beta** | **Std. error** | **Std. beta** | **T** | **Sig** | **Lower** | **Upper** |
| (Intercept) | 73.779 | 1.967 |  | 37.502 | 0.000 | 69.894 | 77.663 |
| Mortality\_rate | -0.190 | 0.019 | -0.479 | -9. 975 | 0.000 | -0.227 | -0.152 |
| GDP\_per\_capita | 0.000 | 0.000 | 0.107 | 2.803 | 0.006 | 0.000 | 0.000 |
| Birth\_rate | -0.193 | 0.052 | -0.259 | -3.720 | 0.000 | -0.295 | -0.091 |
| Population\_growth | 0.671 | 0.234 | 0.104 | 2.870 | 0.005 | 0.209 | 1.133 |
| health\_expenditure | 0.141 | 0.064 | 0.057 | 2.200 | 0.029 | 0.014 | 0.267 |
| Electricity | 0.027 | 0.011 | 0.092 | 2.341 | 0.020 | 0.004 | 0.049 |
| Population\_density | 0.000 | 0.000 | 0.051 | 2.289 | 0.023 | 0.000 | 0.001 |
| Health\_expenditure\_capita | 0.000 | 0.000 | 0.085 | 2.047 | 0.042 | 0.000 | 0.001 |
| Real\_interest\_rate | 0.037 | 0.018 | 0.045 | 2.070 | 0.040 | 0.002 | 0.072 |
| safely\_drinking\_water | 0.024 | 0.013 | 0.100 | 1.873 | 0.063 | -0.001 | 0.050 |

**Table.4. Summary of Both direction**

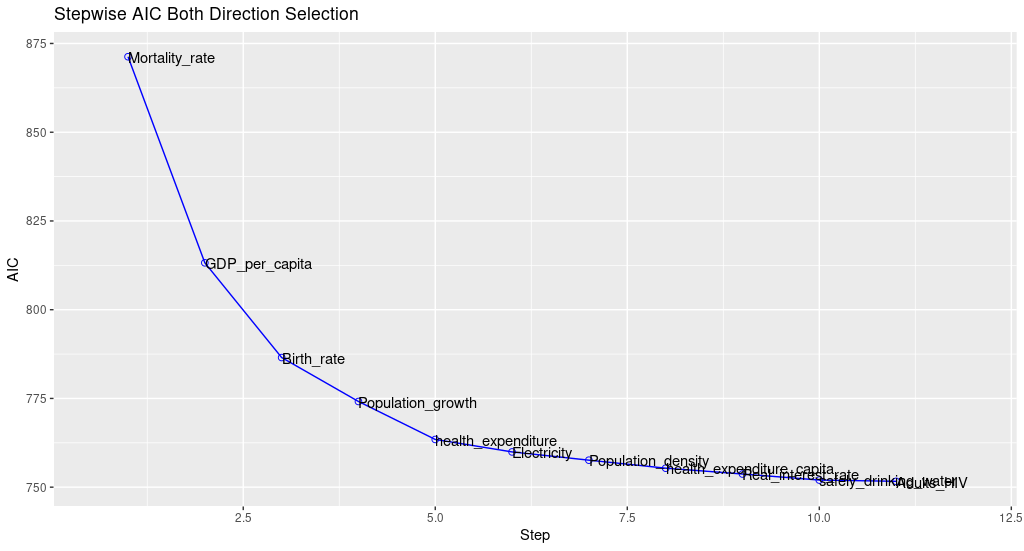


Fig.16.stepwise AIC both direction

**Backward Model Output**

**Model summary**

|  |  |  |  |
| --- | --- | --- | --- |
| R | 0.965 | RMSE | 1.947 |
| R-squared | 0. 932 | Coef. var | 2.649 |
| Adj. R-Squared | 0. 925 | MSE | 3. 792 |
| Pred R-Squared | 0. 915 | MAE | 1.439 |

 RMSE: Root Mean Square Error

 MSE: Mean Square Error

 MAE: Mean Absolute Error

**ANOVA**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Sum of squares** | **DF** | **Mean square** | **F** | **Sig** |
| Regression | 8292.137 | 15 | 552.809 | 145.779 | 0.0000 |
| Residual | 606.736 | 160 | 3.792 |  |  |
| Total | 8898.873 | 175 |  |  |  |

**Parameter estimates**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Beta** | **Std. error** | **Std. beta** | **T** | **Sig** | **Lower** | **Upper** |
| (Intercept) | 73.799 | 2.011 |  | 36.643 | 0.000 | 69.727 | 77.672 |
| Electricity | 0.032 | 0.012 | 0.112 | 2.620 | 0.010 | 0.008 | 0.057 |
| Net\_National\_income | 0.386 | 0.244 | 0.319 | 1.584 | 0.115 | -0.095 | 0.868 |
| Net\_national\_income\_capita | -0.351 | 0.257 | -0.285 | -1.363 | 0.175 | -0.859 | 0.1 |
| children\_out\_of\_school | 0.000 | 0.000 | 0.037 | 1.193 | 0.235 | 0.000 | 0.000 |
| Mortality\_rate | -0.189 | 0.019 | -0.478 | -10. 058 | 0.000 | -0.226 | -0.152 |
| Real\_interest\_rate | 0.047 | 0.019 | 0.057 | 2.499 | 0.013 | 0.010 | 0.084 |
| Population\_growth | 0.500 | 0.308 | 0.077 | 1.652 | 0.106 | -0.108 | 1.107 |
| Population\_density | 0.000 | 0.000 | 0.040 | 1.753 | 0.081 | 0.000 | 0.001 |
| health\_expenditure\_capita | 0.000 | 0.000 | 0.084 | 2.045 | 0.043 | 0.000 | 0.001 |
| health\_expenditure | 0.153 | 0.065 | 0.062 | 2.376 | 0.019 | 0.026 | 0.281 |
| GDP\_annual | -0.077 | 0.071 | -0.035 | -1.085 | 0.280 | -0.217 | 0.063 |
| GDP\_per\_capita | 0.000 | 0.000 | 0.102 | 2.666 | 0.008 | 0.000 | 0.000 |
| Birth\_rate | -0.216 | 0.052 | -0.290 | -4.114 | 0.000 | -0.320 | -0.112 |
| Adults\_HIV | 0.000 | 0.000 | -0.040 | -1.698 | 0.091 | 0.000 | 0.000 |
| Safely\_drinking\_water | 0.021 | 0.013 | 0.085 | 1.579 | 0.116 | -0.005 | 0.046 |

**Table.5.Summary of Backward direction**

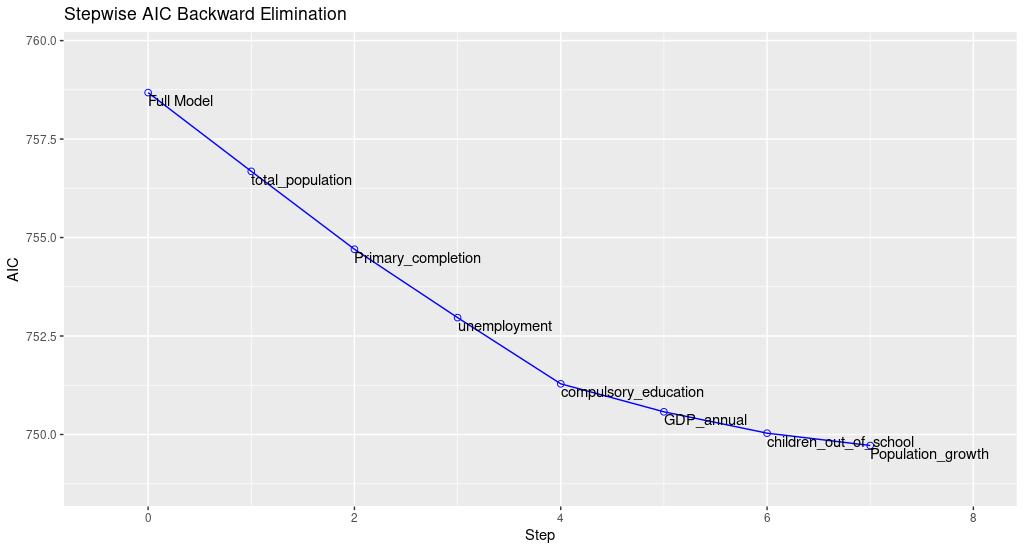


Fig.17. stepwise AIC Backward direction

**Forward Model Output**

**Model summary**

|  |  |  |  |
| --- | --- | --- | --- |
| R | 0.965 | RMSE | 1.947 |
| R-squared | 0. 932 | Coef. var | 2.649 |
| Adj. R-Squared | 0. 925 | MSE | 3. 792 |
| Pred R-Squared | 0. 915 | MAE | 1.439 |

 RMSE: Root Mean Square Error

 MSE: Mean Square Error

 MAE: Mean Absolute Error

**ANOVA**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Sum of squares** | **DF** | **Mean square** | **F** | **Sig** |
| Regression | 8292.137 | 15 | 552.809 | 145.779 | 0.0000 |
| Residual | 606.736 | 160 | 3.792 |  |  |
| Total | 8898.873 | 175 |  |  |  |

**Parameter estimates**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Beta** | **Std. error** | **Std. beta** | **T** | **Sig** | **Lower** | **Upper** |
| (Intercept) | 73.799 | 2.011 |  | 36.643 | 0.000 | 69.727 | 77.672 |
| Mortality\_rate | -0.189 | 0.019 | -0.478 | -10. 058 | 0.000 | -0.226 | -0.152 |
| GDP\_per\_capita | 0.000 | 0.000 | 0.102 | 2.666 | 0.008 | 0.000 | 0.000 |
| Birth\_rate | -0.216 | 0.052 | -0.290 | -4.114 | 0.000 | -0.320 | -0.112 |
| Population\_growth | 0.500 | 0.308 | 0.077 | 1.652 | 0.106 | -0.108 | 1.107 |
| health\_expenditure | 0.153 | 0.065 | 0.062 | 2.376 | 0.019 | 0.026 | 0.281 |
| Electricity | 0.032 | 0.012 | 0.112 | 2.620 | 0.010 | 0.008 | 0.057 |
| Population\_density | 0.000 | 0.000 | 0.040 | 1.753 | 0.081 | 0.000 | 0.001 |
| health\_expenditure\_capita | 0.000 | 0.000 | 0.084 | 2.045 | 0.043 | 0.000 | 0.001 |
| Real\_interest\_rate | 0.047 | 0.019 | 0.057 | 2.499 | 0.013 | 0.010 | 0.084 |
| Safely\_drinking\_water | 0.021 | 0.013 | 0.085 | 1.579 | 0.116 | -0.005 | 0.046 |
| Adults\_HIV | 0.000 | 0.000 | -0.040 | -1.698 | 0.091 | 0.000 | 0.000 |
| children\_out\_of\_school | 0.000 | 0.000 | 0.037 | 1.193 | 0.235 | 0.000 | 0.000 |
| GDP\_annual | -0.077 | 0.071 | -0.035 | -1.085 | 0.280 | -0.217 | 0.063 |
| Net\_National\_income | 0.386 | 0.244 | 0.319 | 1.584 | 0.115 | -0.095 | 0.868 |
| Net\_national\_income\_capita | -0.351 | 0.257 | -0.285 | -1.363 | 0.175 | -0.859 | 0.158 |

**Table.6. Summary of Forward direction**

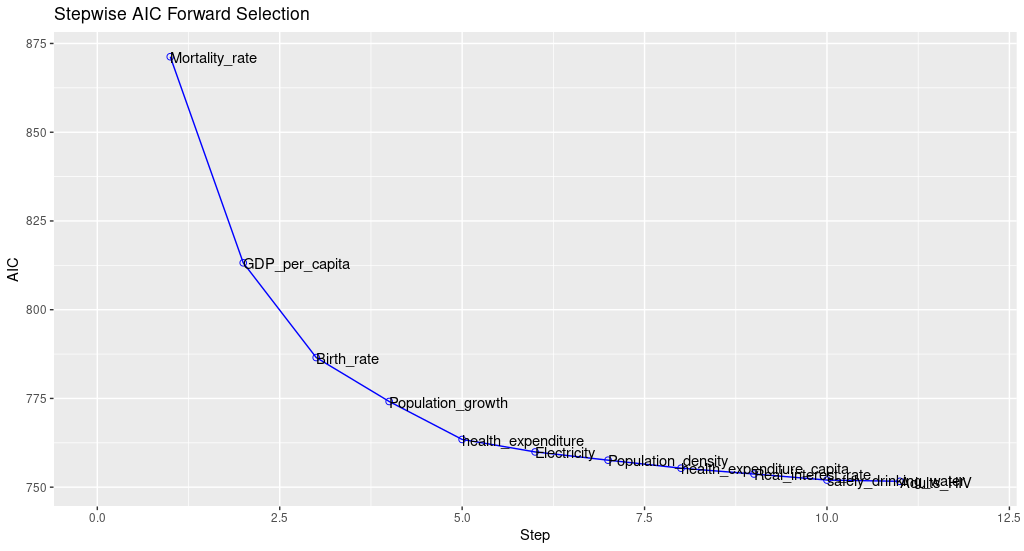


Fig.18..Stepwise AIC Forward direction

Both forward and backward direction regression provides same model while both direction provides different model.

# 5. Design of Experiments

Hence, in order to make the effective analysis the life\_expectancy\_date1.csv was applied. Whereas it processes with the 219 observations through variables with 5. Some of the variables which contains the NA values were effectively removed in order it will emerge with the error through computation. Hence, the one way of ANOVA was highly helpful for the determination and to find the difference between the countries given.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Degree of Freedom** | **Sum of square** | **Mean square** | **F value** | **P- Value** |
| Continent | 5 | 6414 | 1283 | 53.49 | <2e-16 |
| Residuals | 191 | 4581 | 24 |  |  |

**Table 4: ANOVA one way summary**

|  |  |  |  |
| --- | --- | --- | --- |
| **Continent** | **Count** | **Mean** | **Sd** |
| Africa | 50 | 64.1 | 5. 93 |
| Asia | 50 | 74.6 | 5.07 |
| Australia / Oceania | 18 | 73.6 | 6.21 |
| Europe | 48 | 79.3 | 3.63 |
| North America | 34 | 76.2 | 3.83 |
| South America | 12 | 75.1 | 3.15 |

**Table 5: ANOVA one way summary**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Diff** | **lwr** | **P-adj** |
| Asia – Africa | 10.507249 | 7.7 | 0 |
| Australia/Oceania -Africa | 9.4624783 | 4.96 | 0 |
| Europe -Africa | 15.174134 | 12.29 | 0 |
| North America -Africa | 12.063989 | 8.69 | 0 |
| South America -Africa | 10.9808564 | 6.48 | 0 |
| Australia/Oceania – Asia | -1.04 | -5.57 | 0.9856 |
| Europe - Asia | 4.6668848 | 1.73 | 0.0001 |
| North America- Asia | 1.5566894 | -1.85 | 0.7766 |
| South America- Asia | 0.4736068 | -4.05 | 0.9999 |
| Europe -Australia/Oceania | 5.711656 | 1.108 | 0.005 |
| North America- Australia/Oceania | 2.6014606 | -2.3188 | 0.65 |
| South America- Australia/Oceania | 1.5 | -4.23 | 0.97 |
| North America -Europe | -3.11 | -6.27 | 0.113 |
| South America – Europe | -4.193 | -8.79 | 0.09 |
| South America -North America | -1.08 | -6.003 | 0.9 |

**Table 5: one way turkey HSD**

Finally with the end of experiment it was gained that the p-values were <2e-16, it was observed with the lesser values through 0.05 significance level. Hence, we mitigated the null hypothesis in order it was analyzed the variance through method of life expectancy considering with various countries.

# Conclusion

Thus, given analysis is enhanced with proper materials and the outputs are shown with the effective result. In the proposed report five different analysis and experiments were taken, which induced through variety of representations. In the first part of the report the analysis for descriptive statistics were undertaken for numerical and graphical representations. In the other part the missing values predictions were taken, the part introduces the appropriate method for missing values processing. in other section the increases of collinearity were done for the estimations of variance, in which the particular method was applied to life expectancy data. In the other section the suitable model was proposed in order to effectively predict the 2019 life expectancy. Finally, using the dataset of life expectancy effective experimental design was undertaken in order to revise the differences of life expectancies around the countries.

# References

Akanmu, D. M. & Hassan, M. g., 2020. A preliminary analysis modeling of the relationship between quality management practices and sustainable performance. *Quality Management Journal ,* 27(1), pp. 37-61.

Ali, J., Khan, R., Ahmad, N. & Maqsood, I., 2012. Random Forests and Decision Trees. *IJCSI International Journal of Computer Science,* 9(5), pp. 1-30.

Biau, G., 2012. Analysis of a Random Forests Model. *Journal of Machine Learning Research,* Volume 13, pp. 1-33.

Daoud, J., 2017. Multicollinearity and Regression Analysis. *Journal of Physics Conference Series,* 949(1), pp. 1-30.

Kakkeri, R. B., 2017. Distributed Cable Harness Tester. *International Research Journal of Engineering and Technology,* 4(5), pp. 1-10.

Pritwani, K., Wasley, K. & Woo, J., 2018. Spark Big Data Analysis of World Development Indicators. *Global Journal of Computer Science and Technology: C Software & Data Engineering,* 18(3), pp. 1-11.

# Individual Contribution :

|  |  |  |
| --- | --- | --- |
| Student Name | Student ID | Contribution |
| Gurrapu Venkat | 2106458 | Q1 complete answer, Q4 help,Q2 Help, Report writing, PPT Writing |
| Ambarapu Renuka | 2111177 | Q5 complete answer, PPT Help |
| Arun Annaveni | 2105732 | Q2 complete answer, PPT Help |
| Ravichandran Karthikeyan | 2111060 | Q4 complete answer, PPT Design |
| Preeti Yadav | 2111142 | Q3 complete Answer, PPT design, Report Help |

# Appendix :

**Source code**

setwd('/C:/MA317/')

library(ggplot2)

library(maptools) # For wrld\_simpl polygons

library(missForest)

library(sf)

library(VIM)

library(tidyverse)

library(caret)

library(olsrr)

library(dplyr)

library(mice)

#--------------------------------question - 1 --------------------------------------------

# Analyse using descriptive statistics (both graphical and numerical representations)

# and R the Life Expectancy data1.csv dataset.

in.csv = read.csv('/C:/MA317/data/Life\_Expectancy\_Data1.csv')

# table column name

print(colnames(in.csv))

# table dimension

print(dim(in.csv))

# table summary

print(summary(in.csv))

# table data type

print(str(in.csv))

indicator\_name =c('life\_expectancy','Electricity','Net\_National\_income','Net\_national\_income\_capita',

'childeren\_affected\_with\_HIV','children\_out\_of\_school','Educational\_attainment\_primary','Educational\_attainment\_bachelor',

'Mortality\_rate','Primary\_completion','Literacy\_rate','Real\_interest\_rate',

'Population\_growth','Population\_density','total\_population','health\_expenditure\_capita',

'health\_expenditure','unemployment','GDP\_annual','GDP\_per\_capita','Birth\_rate',

'renewablw\_energy','Adults\_HIV','safely\_drinking\_water','poverty','compulsory\_education')

#changing column name eith indicator name

colnames(in.csv)[-c(1:3)] = indicator\_name

print(colnames(in.csv))

as.data.frame(summary(in.csv))[-1]

output<-capture.output(summary(in.csv), file=NULL,append=FALSE)

output\_df <-as.data.frame(output)

View(output\_df)

print(str(in.csv))

ggplot(subset(in.csv), aes(x = GDP\_per\_capita, y = Mortality\_rate)) +

geom\_point()

in.csv %>%

ggplot(aes(x=Continent,y=life\_expectancy, fill=Continent)) +

geom\_boxplot() + geom\_jitter(width=0.1,alpha=0.2)

ggplot(in.csv, aes(x = life\_expectancy)) +

geom\_histogram(aes(fill = Continent)) +

facet\_wrap(~Continent)

ggplot(in.csv, aes(life\_expectancy, fill = Continent, colour = Continent)) +

geom\_density(alpha = 0.2, na.rm = TRUE)

ggplot(in.csv, aes(life\_expectancy, GDP\_per\_capita)) +

geom\_point() +

geom\_smooth(aes(colour = "loess"), method = "loess", se = FALSE) +

geom\_smooth(aes(colour = "lm"), method = "lm", se = FALSE) +

labs(colour = "Method")

in.csv %>% select(where(~mean(is.na(.))< 0.4))%>% select(is.numeric)%>%

select(-starts\_with(c("total\_population","GDP\_per\_capita","Population\_density",'Adults\_HIV',

'health\_expenditure\_capita','children\_out\_of\_school')))%>%boxplot(las= 2, col="#69b3a2",

boxwex=0.5)

ggplot(subset(in.csv), aes(x = GDP\_per\_capita, y = life\_expectancy, color = Country.Name)) +

geom\_point() + ggtitle("Life Expectancy and GDP Per capita by country")+

geom\_text(aes(label = Country.Name), size=3, nudge\_y = 0.4) +

scale\_x\_continuous(limits = c(0, 70000))+theme(legend.position="none")

ggplot(subset(in.csv), aes(x = GDP\_per\_capita, y = Mortality\_rate, color = Country.Name)) +

geom\_point() + ggtitle("Infant Moratality rate by GDP capita by country")+

geom\_text(aes(label = Country.Name), size=3, nudge\_y = 0.4) +

scale\_x\_continuous(limits = c(0, 70000))+theme(legend.position="none")

#------------------Question-2-----------------------------------------------

#Dealing with missing value

print(colnames(in.csv))

print(str(in.csv))

#change str type of compulsory education character to integer

in.csv$compulsory\_education = as.numeric(in.csv$compulsory\_education)

#Check missing values introduced in the data

df\_na = data.frame("colsu" = colSums(is.na(in.csv)))

print(df\_na)

#selecting column which having less than 70% of NA

df\_mis = in.csv %>% select(where(~mean(is.na(.))< 0.5))

#Check missing values introduced in the data

print(summary(df\_mis))

# Removing categorical data

print(colnames(df\_mis))

df\_no\_na <- subset(df\_mis, select = -c(Country.Name,Country.Code,Continent))

df\_plot <- aggr(df\_no\_na, col=c('navyblue','yellow'),

numbers=TRUE, sortVars=TRUE,

labels=names(df\_no\_na), cex.axis=.7,

gap=3, ylab=c("Missing data","Pattern"))

set.seed(300)

df.imp <- missForest(df\_no\_na)

print(head(df.imp$ximp,3))

print(head(df\_no\_na,3))

print(df.imp$OOBerror)

impu\_df = df.imp$ximp

#-------------------------------Question - 3------------------------------------

#dealing with coolinearity

print(colnames(impu\_df))

set.seed(123)

training.samples <- impu\_df$life\_expectancy %>%

createDataPartition(p = 0.8, list = FALSE)

train.data <- impu\_df[training.samples, ]

test.data <- impu\_df[-training.samples, ]

model1 = lm(life\_expectancy~.,data = train.data)

# Make predictions

predictions <- model1 %>% predict(test.data)

# Model performance

data.frame(

RMSE = RMSE(predictions, test.data$life\_expectancy),

R2 = R2(predictions, test.data$life\_expectancy)

)

# The smallest possible value of VIF is one (absence of multicollinearity). As a rule of thumb,

# a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity.

car::vif(model1 )

# Build a model excluding the tax variable

model2 <- lm(life\_expectancy ~. -Net\_National\_income - Net\_national\_income\_capita - GDP\_per\_capita -Birth\_rate,

data = train.data)

# Make predictions

predictions <- model2 %>% predict(test.data)

# Model performance

data.frame(

RMSE = RMSE(predictions, test.data$life\_expectancy),

R2 = R2(predictions, test.data$life\_expectancy)

)

# It can be seen that removing the these variable

ols\_vif\_tol(model2)

# n practice it is common to say that any VIF greater than 5

# is cause for concern. So in this example we see there is a huge multicollinearity

# issue as many of the predictors have a VIF greater than 5.

ols\_eigen\_cindex(model2)

ols\_coll\_diag(model2)

ols\_plot\_resid\_fit\_spread(model2)

ols\_correlations(model2)

ols\_plot\_obs\_fit(model2)

ols\_plot\_diagnostics(model2)

ols\_plot\_comp\_plus\_resid(model2)

#--------------------------QUSETION - 4-----------------------------

library(MASS)

# Fit the full model

full.model <- lm(life\_expectancy ~., data = train.data)

# Stepwise forward regression model

for.reg <- ols\_step\_forward\_p(full.model, details = TRUE)

plot(for.reg)

for.aic = ols\_step\_forward\_aic(full.model)

plot(for.aic)

# Stepwise backward regression model

back.reg <- ols\_step\_backward\_p(full.model, details = TRUE)

plot(back.reg)

back.aic = ols\_step\_backward\_aic(full.model)

plot(back.aic)

# Stepwise both regression model

both.reg <- ols\_step\_both\_p(full.model, details = TRUE)

plot(both.reg)

both.aic = ols\_step\_both\_aic(full.model)

plot(both.aic)

print(data.frame(for.reg$model$coefficients))

print(data.frame(back.reg$model$coefficients))

print(data.frame(both.reg$model$coefficients))

#-----------------Question -5---------------------------

print(colnames(in.csv))

print(nrow(in.csv))

print(head(in.csv))

in.csv$Continent = as.factor(in.csv$Continent)

levels(in.csv$Continent)

group\_by(in.csv, Continent) %>%

summarise(

count = n(),

mean = mean(life\_expectancy, na.rm = TRUE),

sd = sd(life\_expectancy, na.rm = TRUE)

)

res.aov <- aov(life\_expectancy ~ Continent, data = in.csv)

summary(res.aov)

TukeyHSD(res.aov)

# diff: difference between means of the two groups

# lwr, upr: the lower and the upper end point of the confidence interval at 95% (default)

# p adj: p-value after adjustment for the multiple comparisons.