1. INTRODUCTION

The world has seen an overall life expectancy increase over the last hundred years[1]. The purpose of this report is to propose an explanation for the life expectancy in 2019 for the world. A large set of data can be analysed in order to draw important conclusions and thus help simple and easy processes for making important decisions. The concerned work has been done on analyzing a data set containing World Development Indicators (WDI) taken from a primary world bank database. The work has been divided into different sections in which different methods such as Imputation, LinearRegression, Fitting model and others have been developed to perform predictions of data and draw important information from it.

2. PRELIMINARY ANALYSIS

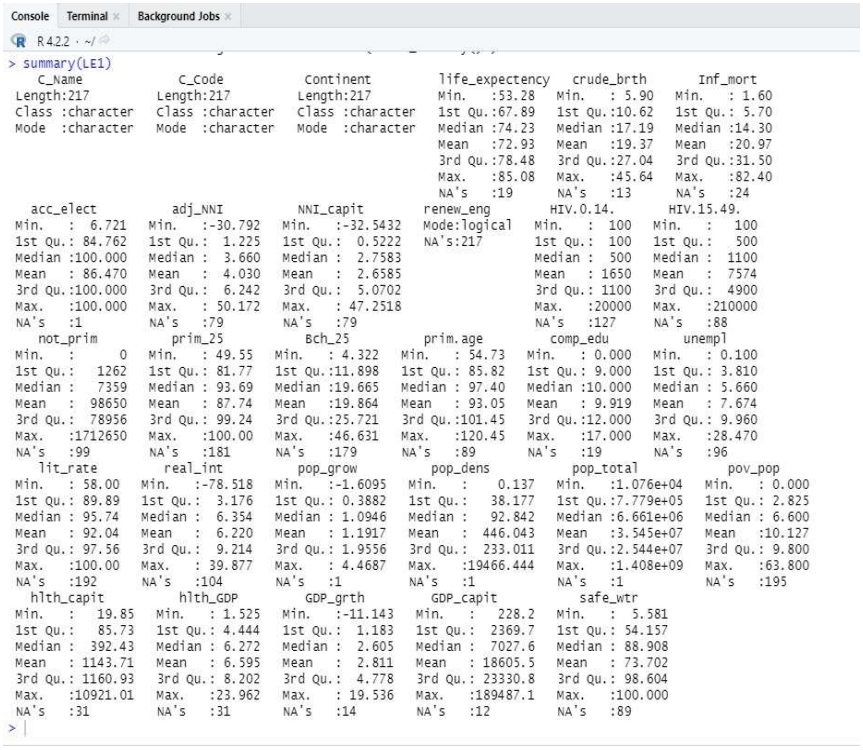
Exploratory Data Analysis (EDA) is a statistical approach or technique for analyzing data sets in order to summarize their important and main characteristics generally by using some visual aids.To describe this data set, we had to rely on statistics, in particular, descriptive statistics. Descriptive statistics is the process of evaluating data in such a way as to make them easily understandable[2].This figure illustrates the data set that has been used for the analysis. As can be seen from the data set, the given data set consists of 217 rows and 29 columns.

	C_Name	C_Code	Continent	life_expec	crude_brth	Inf_mort	acc_elect	adj_NNI	NNI_capit	renew_eng	HIV(0-14)	HIV(15-49)	not_prim	prim_25	Bch_25	prim=age	comp_edu	unempl	lit_rate	real_int	pop_grow	pop_dens	pop_total
2	Afghanistar	AFG	Asia	64.833	31.802	46.4	97.7	NA	NA	NA	200	1300	NA	NA	NA	84.33059	9	NA	NA	NA	2.313073	58.26939	38041757
3	Albania	ALB	Europe	78.573	11.62	8.6	100	0.146477	0.574018	NA	NA	100	3359	NA	NA	103.3227	9	11.47	NA	4.993848	-0.42601	104.1676	2854191
4	Algeria	DZA	Africa	76.88	23.583	20	99.5	2.938149	0.96647	NA	200	1400	12511	NA	NA	101.3646	10	NA	NA	8.507844	1.933983	18.0763	43053054
5	erican San	ASM	tralia/Ocea	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	-0.26902	276.56	55312
6	Andorra	AND	Europe	NA	7	2.5	100	NA	NA	NA	NA	NA	NA	NA	NA	NA	10	NA	NA	NA	0.179042	164.1404	77146
7	Angola	AGO	Africa	61.147	40.232	49.9	45.67031	1.12257	-2.10414	NA	6200	16000	NA	NA	NA	NA	6	NA	NA	-6.86647	3.242914	25.52763	31825299
8	ua and Bar	ATG	orth Ameri	77.016	15.107	5.6	100	NA	NA	NA	NA	NA	NA	NA	NA	NA	11	NA	NA	8.385803	0.861446	220.7159	97115
9	Argentina	ARG	outh Ameri	76.667	16.833	8.1	100	-4.6989	-5.64093	NA	100	5300	18518	NA	NA	98.54724	14	9.84	NA	10.82196	0.993397	16.42083	44938712
10	Armenia	ARM	Asia	75.087	13.646	10.2	100	6.360385	6.145091	NA	100	500	14928	NA	NA	93.13852	12	18.3	NA	10.95958	0.202624	103.8893	2957728

Figure 1: Dataset

(Source: Excel file from Group coursework)

Descriptive statistics can be summarized visually with graphs and quantitatively with numbers. The numerical representation can be divided into two segments which are the measure of the central tendency of the feature and the measure of variability.



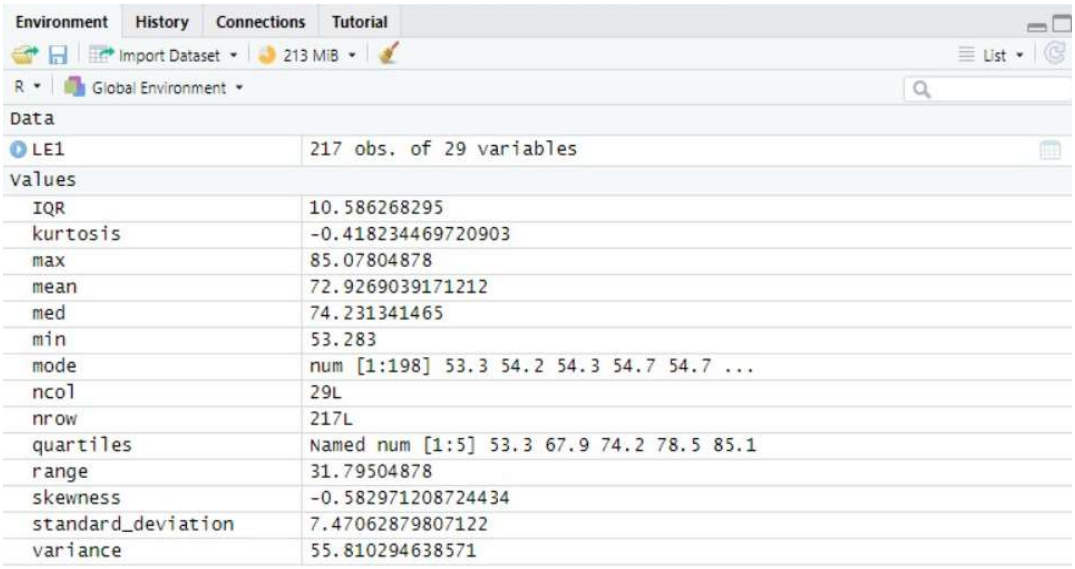
```
Console   Terminal   Background Jobs
R 4.2.2 · ~/

> summary(LE1)
    C_Name             C_Code           Continent         life_expectency   crude_brth       Inf_mort
 Length:217         Length:217         Length:217         Min.   :53.28   Min.   : 5.90   Min.   : 1.60
 Class :character   Class :character   Class :character   1st Qu.:67.89   1st Qu.:10.62   1st Qu.: 5.70
 Mode  :character   Mode  :character   Mode  :character   Median :74.23   Median :17.19   Median :14.30
                                                          Mean   :72.93   Mean   :19.37   Mean   :20.97
                                                          3rd Qu.:78.48   3rd Qu.:27.04   3rd Qu.:31.50
                                                          Max.   :85.08   Max.   :45.64   Max.   :82.40
                                                          NA's   :19      NA's   :13      NA's   :24
   acc_elect          adj_NNI          NNI_capit        renew_eng        HIV.0.14.       HIV.15.49.
 Min.   :  6.721   Min.   :-30.792   Min.   :-32.5432   Mode:logical   Min.   :  100   Min.   :  100
 1st Qu.: 84.762   1st Qu.:  1.225   1st Qu.:  0.5222   NA's:217       1st Qu.:  100   1st Qu.:  500
 Median :100.000   Median :  3.660   Median :  2.7583                  Median :  500   Median : 1100
 Mean   : 86.470   Mean   :  4.030   Mean   :  2.6585                  Mean   : 1650   Mean   : 7574
 3rd Qu.:100.000   3rd Qu.:  6.242   3rd Qu.:  5.0702                  3rd Qu.: 1100   3rd Qu.: 4900
 Max.   :100.000   Max.   : 50.172   Max.   : 47.2518                  Max.   :20000   Max.   :210000
 NA's   :1         NA's   :79        NA's   :79                        NA's   :127     NA's   :88
    not_prim          prim_25           Bch_25          prim.age         comp_edu          unempl
 Min.   :     0   Min.   : 49.55   Min.   : 4.322   Min.   : 54.73   Min.   : 0.000   Min.   : 0.100
 1st Qu.:  1262   1st Qu.: 81.77   1st Qu.:11.898   1st Qu.: 85.82   1st Qu.: 9.000   1st Qu.: 3.810
 Median :  7359   Median : 93.69   Median :19.665   Median : 97.40   Median :10.000   Median : 5.660
 Mean   : 98650   Mean   : 87.74   Mean   :19.864   Mean   : 93.05   Mean   : 9.919   Mean   : 7.674
 3rd Qu.: 78956   3rd Qu.: 99.24   3rd Qu.:25.721   3rd Qu.:101.45   3rd Qu.:12.000   3rd Qu.: 9.960
 Max.   :1712650   Max.   :100.00   Max.   :46.631   Max.   :120.45   Max.   :17.000   Max.   :28.470
 NA's   :99        NA's   :181      NA's   :179      NA's   :89       NA's   :19       NA's   :96
    lit_rate         real_int          pop_grow          pop_dens          pop_total          pov_pop
 Min.   : 58.00   Min.   :-78.518   Min.   :-1.6095   Min.   :    0.137   Min.   :1.076e+04   Min.   : 0.000
 1st Qu.: 89.89   1st Qu.:  3.176   1st Qu.: 0.3882   1st Qu.:   38.177   1st Qu.:7.779e+05   1st Qu.: 2.825
 Median : 95.74   Median :  6.354   Median : 1.0946   Median :   92.842   Median :6.661e+06   Median : 6.600
 Mean   : 92.04   Mean   :  6.220   Mean   : 1.1917   Mean   :  446.043   Mean   :3.545e+07   Mean   :10.127
 3rd Qu.: 97.56   3rd Qu.:  9.214   3rd Qu.: 1.9556   3rd Qu.:  233.011   3rd Qu.:2.544e+07   3rd Qu.: 9.800
 Max.   :100.00   Max.   : 39.877   Max.   : 4.4687   Max.   :19466.444   Max.   :1.408e+09   Max.   :63.800
 NA's   :192      NA's   :104       NA's   :1         NA's   :1           NA's   :1           NA's   :195
   hlth_capit         hlth_GDP          GDP_grth         GDP_capit          safe_wtr
 Min.   :   19.85   Min.   : 1.525   Min.   :-11.143   Min.   :   228.2   Min.   :  5.581
 1st Qu.:   85.73   1st Qu.: 4.444   1st Qu.:  1.183   1st Qu.:  2369.7   1st Qu.: 54.157
 Median :  392.43   Median : 6.272   Median :  2.605   Median :  7027.6   Median : 88.908
 Mean   : 1143.71   Mean   : 6.595   Mean   :  2.811   Mean   : 18605.5   Mean   : 73.702
 3rd Qu.: 1160.93   3rd Qu.: 8.202   3rd Qu.:  4.778   3rd Qu.: 23330.8   3rd Qu.: 98.604
 Max.   :10921.01   Max.   :23.962   Max.   : 19.536   Max.   :189487.1   Max.   :100.000
 NA's   :31         NA's   :31       NA's   :14        NA's   :12         NA's   :89
>
```

Figure 2: Summary of LE1

(Source: Created in RStudio)

A measure of central tendency determines the value of the mean, median and mode of the data set whereas a measure of variability determines the value of the measures of skewness, variance, standard deviation, and kurtosis.



Environment | History | Connections | Tutorial

Import Dataset · 213 MiB · List

R · Global Environment

Data	
LE1	217 obs. of 29 variables
Values	
IQR	10.586268295
kurtosis	-0.418234469720903
max	85.07804878
mean	72.9269039171212
med	74.231341465
min	53.283
mode	num [1:198] 53.3 54.2 54.3 54.7 54.7 ...
ncol	29L
nrow	217L
quartiles	Named num [1:5] 53.3 67.9 74.2 78.5 85.1
range	31.79504878
skewness	-0.582971208724434
standard_deviation	7.47062879807122
variance	55.810294638571

Figure 3: Numerical representation

(Source: Created in RStudio)

The graphical histogram plot with density and life expectancy is obtained using ggplot.

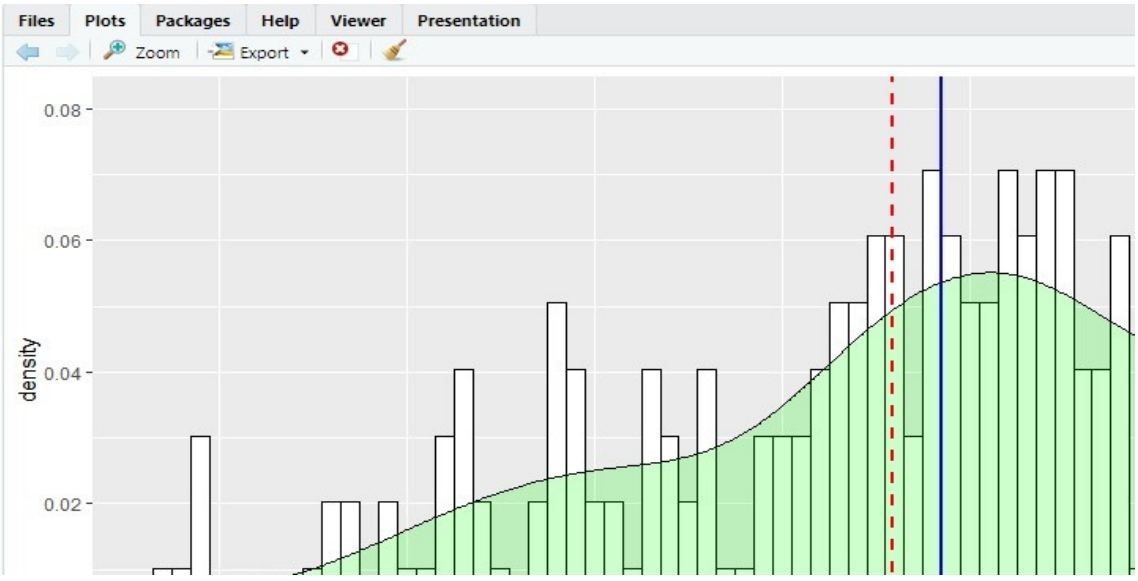


Figure 4: Graphical representation

(Source: Created in RStudio)

Upon analyzing the data set, it is revealed that it contains many null values. When null values are present in a data set, the resulting insights may be of lower quality. The solution to this problem is to remove the columns with more than 75% missing data. Since they don't contribute to the model, they should be removed.

After removing all columns containing 75% missing data, there will still be columns that contain null values. To solve this issue, all remaining missing data should be filled with the predictive mean matching imputation method. Predictive Mean Matching(PMM) is a semi-parametric imputation approach[2]. It is identical to the regression approach, with the exception that for each missing value, a new value is produced using a donor observation whose predicted values for the variable are closest to the predicted value for the missing value from the simulated regression model.

Instead of 24 columns, there will now be 20 columns.Some columns will have all null values, so those columns will be dropped manually and there will only be 18 columns left. The new data without null values will be stored in a new data frame.

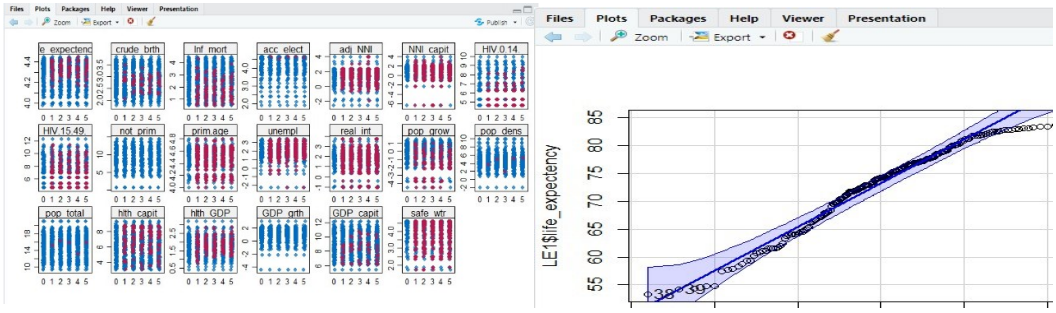


Figure 5: Cleaned dataset with no null values & QQ plot (LE $ Life Expectancy)

(Source: Created in RStudio)

3. ANALYSIS

In a multiple regression model, multicollinearity occurs when more than one explanatory variable is highly linearly related. When an independent variable is exactly linearly combined with other variables, it is said to be perfect multicollinearity [3]. A multicollinear system can occur as a result of dummy variables being inserted or misused. Other reasons might be the use of derived variables, where one variable is derived from another and taking variables that have a very high correlation between each other or that are similar in nature or provide similar information.

It might not be a problem to have moderate multicollinearity. Severe multicollinearity is problematic, too, as it can raise the variance of coefficient estimates and make such estimates highly sensitive to even little model modifications.

The regression coefficients may significantly vary from sample to sample, which is another possibility. Statistically significant variables may emerge differently with different samples.

The use of tolerance or a variance inflation factor is an alternative approach (VIF) by

$$VIF = 1 / Tolerance$$

$$VIF = 1/ (1 – R\ square)$$

The VIF of over 10 indicates that the variables have high correlation among each other. Usually, VIF value of less than 4 is considered good for a model[3].

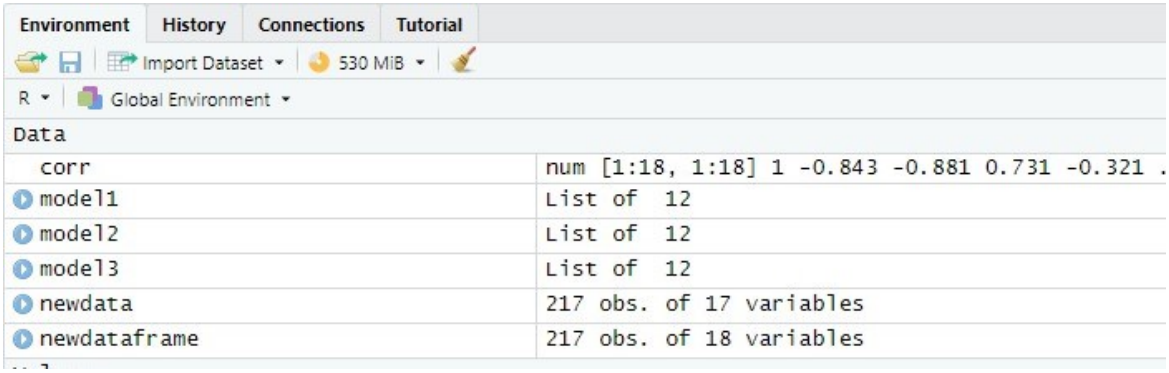


Figure 6 : VIF values

(Source: Created in RStudio)

Multicollinearity reduces the statistical power of the analysis, has the potential to lead to the coefficients changing sign, and makes it more challenging to define the appropriate model. It is also difficult to determine which variables are statistically significant since some variables will provide similar outputs [3].

The presence of multicollinearity among the independent or explanatory variables can be identified by several different factors.The first and simplest method is to generate a pair-wise correlation plot between the various variables[8]. Multicollinearity may be present if there are significant fluctuations in regression coefficients after the addition or removal of new independent or explanatory variables.

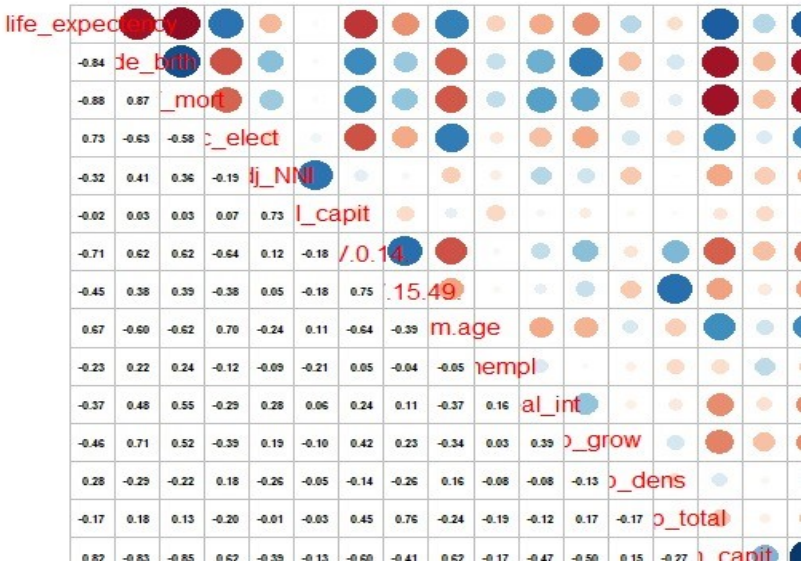


Figure 7: Pairwise correlation

(Source: Created in RStudio)

4. DISCUSSION

Using the Imputation approach's cleaned data, selecting the best model to forecast life expectancy for 2020. Investigate the full model first, which includes each variable. After then, use the full model to find the summary and examine the 'star-gaze'. It would appear that Crude_brth, adj_NNI, acc_elect, NNI_capit, HIV.0.14, pop dense, Inf mort, real int, pop total, pop_grow, prim.age, and HIV.15.49 are the crucial variables [8]. Note that this includes the intercept value, which is insignificant, as none of the other values are statistically different from 0. When running the Anova command on the complete model, the Analysis of the Variance Table, which computes the sum of the squares for each variable, can be produced.

Again star-gazing implies that crude_brth, Inf_mort, acc_elect, NNI_Capit,HIV.0.14, unempl, real_int, pop_grow, pop_dense and pop_total may be important in the model. The next step would be comparing the full and the reduced models [8].

To determine whether each model has a fair residual distribution, need to examine the residuals. The next step would be plotting the standardised residuals against fitted values and QQ plots for standardised residuals against normal scores for full model and reduced model.

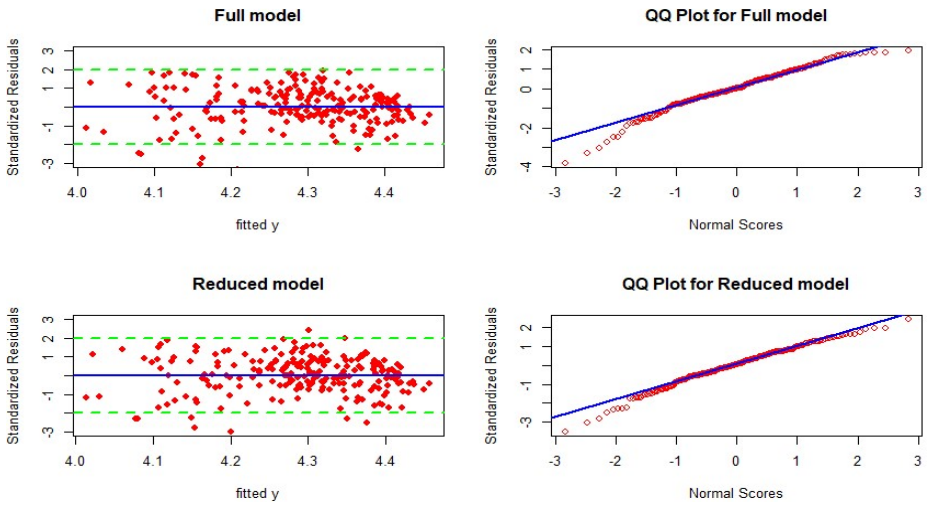


Figure 8: Full Model vs Reduced Model

(Source: Created in RStudio)

All the plots show a reasonable fit for the residuals. There is evidence of outliers in both plots of 'residuals v. fitted y', the QQ plots have two clear outlying values, It is seen that the influential observations from the 'Residuals vs Leverage' plot but in, general the fit of the models looks alright. Another way to choose between the reduced and the full model is to use one of the available model selection criteria. calculating the AIC and Mallow's Cp. The AIC can be easily found using the command 'AIC (. . .)' .

The better fitting model is the one with the lowest value, which in this case is the reduced model with AIC_full = -828.7816 compared with AIC_reduced = -827.3107. To check whether the full model is a viable fit by calculating Mallow's Cp, using the 'ols_mallows_cp' function. To use this function the full model would enter first, followed by the reduced model [8].

If the given value is close to, or smaller than, the number of predictor variables in the submodel then it is an acceptable model. Here, we have Cp = 12.49639 which is much higher than the number of variables in the model (=5+1 (intercept)) and therefore it is concluded that the full model is not a good option.

In order to obtain the best linear model which predicts life expectancy need to find the best subset of features out, the model which includes all the predictor variables is not the ideal one as we have found evidence of collinearity. It might be necessary to consider every single sub model and decide which of those are good models rather than just randomly trying. This can be calculated using the 'leaps' command from the 'leaps' package. This will consider each submodel and calculate their respective Mallow's Cp value. First, the entire model needs to be run once more with a small modification. The design matrix X and our y variables will then be defined. The Cp score using the 'leaps' command can be calculated repeatedly.The 'leaps' command will save the best 10 submodels for each value of p ,the package uses p to refer to the number of parameters in the submodel, which is defined as |M|. By plotting the value of Cp against |M|, the good models can be determined by considering those below the line Cp = |M|.

From the plot (Fig.4 in appendix), it is obtained that there aren't many 'good' submodels for the data as there are only a few points below the line. These have 14,15,16 or 17 parameters[8].

From this, we can tell that the 'best' model according to the Cp value is one which includescrude_brth, inf_mort, acc_elect, adj_NNI, NNI_Capit, HIV.0.14., HIV.15.15., prim.age, real_int, pop_grow, pop_dense, pop_total, hlth_capit, safe_wtr. This model has corresponding Cp = 12.27643; this is slightly larger than the number of parameters in the model, 12, but it is close. It is saved as a multiple linear model.

Finally, the corresponding AIC value can be calculated and compared to the full and reduced models from earlier.This value -834.4804 is greater than that for the full model (-828.7816) and the reduced model (-827.3107), hence it is concluded that reduced model is a better fitting model.

An experimental model showing the differences of average life expectancy across the continents, one-way ANOVA method is used. The Life Expectancy data and Continentsare plotted based on their group mean values[8].

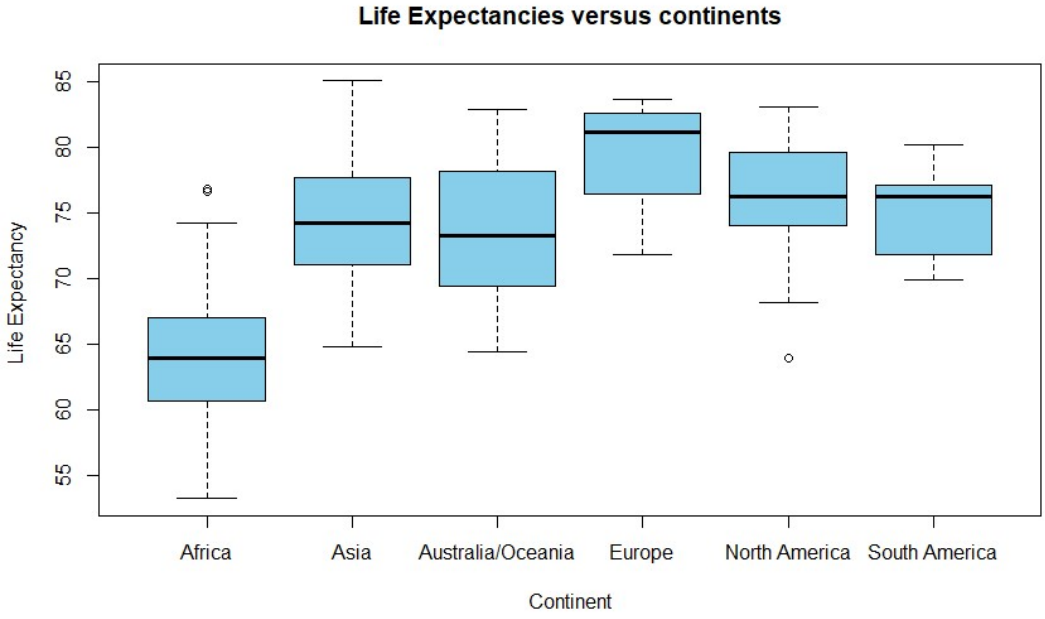


Fig9: Data showing Life Expectancy versus Continents

(Source: Created in RStudio)

To analyse using a single factor Continent one-way ANOVA method is used to investigate whether there are differences in the average life expectancy across the continents. The summary of annova1way is obtained as follows:

```
Console  Terminal  Background Jobs
R 4.2.2 · ~/
> anova1way<-aov(df2$SP.DYN.LE00.IN~as.factor(df2$Continent),data=df2)
> summary(anova1way)
                      Df Sum Sq Mean Sq F value Pr(>F)
as.factor(df2$Continent)   5   6931  1386.2   58.55 <2e-16 ***
Residuals                211   4996    23.7
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 1: Summary of ANOVA one-way

(Source: Created in RStudio)

The Shapiro_Wilk normality test is performed to check the normality of datasets of continent.The resultant gives an observation as there is differences in the average life expectancy as p_value is lesser than the significance value.

```
Console  Terminal  Background Jobs
R 4.2.2 · ~/
> #boxplot of Life Expectancy vs Continent
> boxplot(df2$SP.DYN.LE00.IN~df2$Continent,main='Life Expectancies versus continents', xlab='Continent', col="sky blue", ylab
 = "Life Expectancy ",)
> df2$residuals<-anova1way$residuals
> shapiro.test(df2$residuals)

        Shapiro-Wilk normality test

data:  df2$residuals
W = 0.9911, p-value = 0.2063
```

Table 2: Shapiro-wilk normality test

(Source: Created in RStudio)

To check the variance measures Levene's test is being performed. Levine's Test for Homogeneity of Variance (center = median) is given by

```
Console  Terminal  Background Jobs
R 4.2.2 · ~/
Levene's Test for Homogeneity of Variance (center = median)
       Df F value  Pr(>F)
group   5   2.773 0.01894 *
      211
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 1: Levene's Test for Homogeneity of Variance

(Source: Created in RStudio)

With the observed results, we can conclude that null hypothesis cannot be true. As the p_value is lesser than the F value we can conclude the differences of life expectancy across continents exists[6].

A pictorial representation in the differences of average life expectancy across continents is shown below.

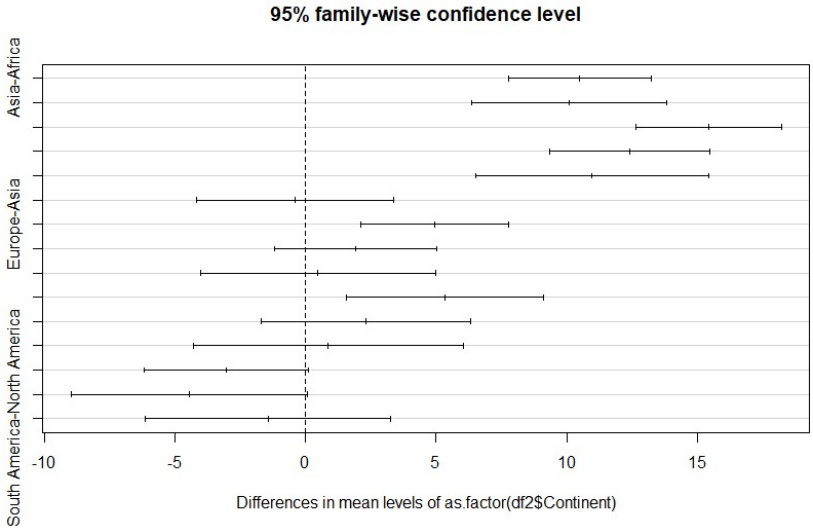


Fig 10: Data showing Life Expectancy versus Continents

(Source: Created in RStudio)

5. CONCLUSION

In analyzing the Life Expectancy Data Set and determining the 'best' fitting model for predicting Life Expectancy in 2019 ,some of the following methods are approached to find the best fitting model such are Predictive Mean Matching Method (PMM), Variance Inflation Factor (VIF), Mallow's cp, Multiple Linear Regression, Sequential model selection methods, Shapiro-Wilk normality test, Levene's test, Bartlett test of homogeneity of variances and others.

With the observed results, we can conclude the best fitting model for predicting Life Expectancy in 2019 is the Multiple Linear Regression Model and the differences of life expectancy across continents exists. So suggesting the best linear model as the multiple linear regression model to predict life expectancy for 2020.

REFERENCES

1. Max Roser, Esteban Ortiz-Ospina and Hannah Ritchie (2013), "Life Expectancy", published online at OurWorldInData.org., retrieved from: 'https://ourworldindata.org/life-expectancy' [Online Resource].
2. Top 100 R Tutorials: Step by Step guide, Listen Data, all rights reserved © 2022 RSGB Business Consultant Private Limited, URL: https://www.listendata.com/p/r-programming-tutorials.html
3. R-Bloggers, Dealing with the Problem of Multicollinearity in R, posted on August 15, 2018, Perceptive Analytics in R-bloggers, URL: "https://www.r-bloggers.com/2018/08/dealing-with-the-problem-of-multicollinearity-in-r/".
4. Lathan Liou, William Joe, Abishek Kumar, S.V. Subramanian, Inequalities in life expectancy, An analysis of 201 countries, 1950-2015, Social Science & Medicine, Vol-253, 2020, ISSN 0277-9536, https://doi.org/10.1016/j.socscimed.2020.112964.
5. Colin D Mathers, RituSadana, Joshua A Saomon, Christopher JL Murray, Alan D Lopez, Healthy life expectancy in 191 countries, 1999, The Lancet, Vol 357, Issue 9269, 2001, pg 1685-1691, ISSN 0140-6736, URL: "https://doi.org/10.1016/S0140-6736(00)04824-8".
6. Kabir, Mahfuz, "Determinants of Life Expectancy in Developing Countries, The journal of Developing Area, Vol. 41, No. 2, 2008, pg 185-204.JSTOR, URL: "http://www.jstor.org/stable/40376184. Accessed 13 Dec. 2022".
7. An Introduction to R, W.N.Venables, D.M.Smith and The R Development Code Team, second edition, May 2009, published by Network Theory Limited, ISBN: 0-9546120-8-6, "https://moodle.essex.ac.uk/mod/resource/view.php?id=772325".
8. Modelling Experimental Data module (MA317) – Lab notes, Dr. Stella Hadjiantoni, 2022-2023, The University of Essex.