AN ANALYSIS OF A DATA MODEL USING MULTIPLE LINEAR REGRESSION IN THE PREDICTION OF LIFE EXPECTANCY

MA317 GROUP COURSEWORK

GROUP 1
THARA JENI SAMALA ANTHONY RAJ
PRID: SAMAL11507
REGISTRATION NUMBER: 2200940

14 DECEMBER 2022

OUTLINE

- ✓Abstract
- ✓Introduction
- ✓Summary Of The Given Data Set
- ✓Descriptive Statistics Analysis Of The Given Data
- ✓Handling Missing Values In The Given Data Set
- ✓Investigating The Presence Of Multicollinearity
- ✓Finding The Best Model To Forecast Life Expectancy Of 2019
- ✓The Differences In Average Life Expectancy Across The Continents
- ✓Conclusion

ABSTRACT

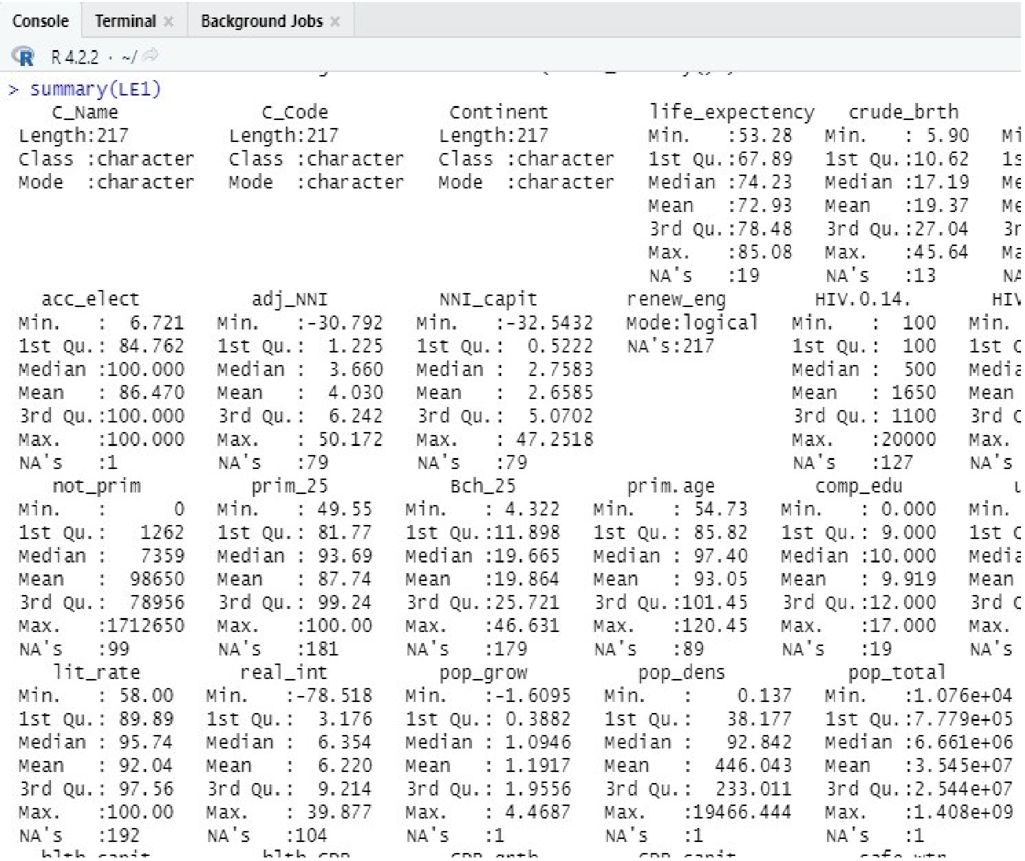
- In this report, a data collection from the World Development Indicators (WDI) is analyzed with the goal of suggesting the best model for predicting Life Expectancy in 2019, where a population's health and happiness can be summarized in terms of Life Expectancy.
- The project entails utilizing R to conduct an descriptive statistical analysis on the provided data set & the imputation methods were applied to handle the missing values of the given data set
- The method is used to build an interpret the fitted the model to the given data is Multiple Linear Regression.
- One of the method which plays vital role to predict the best model is Sequential Model Selection Methods such as Forward, Backward & Stepwise Selection.
- Finally, to compare the differences in Life Expectancy across a Continent, the ANOVA one way approach has been used.

INTRODUCTION

- The world has seen an overall life expectancy increase over the last hundred years.
- Here the given Life Expectancy data set is mentioned as LE1.
- 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 analysing 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, Linear Regression, Fitting model and others have been developed to perform predictions of data and draw important information from it.
- The values of the data set's mean, median, and mode are determined by a Measure of Central Tendency, whereas the values of the skewness, variance, standard deviation, and kurtosis are determined by a Measure of Variability, which is an numerical representation.

SUMMARY OF LE1

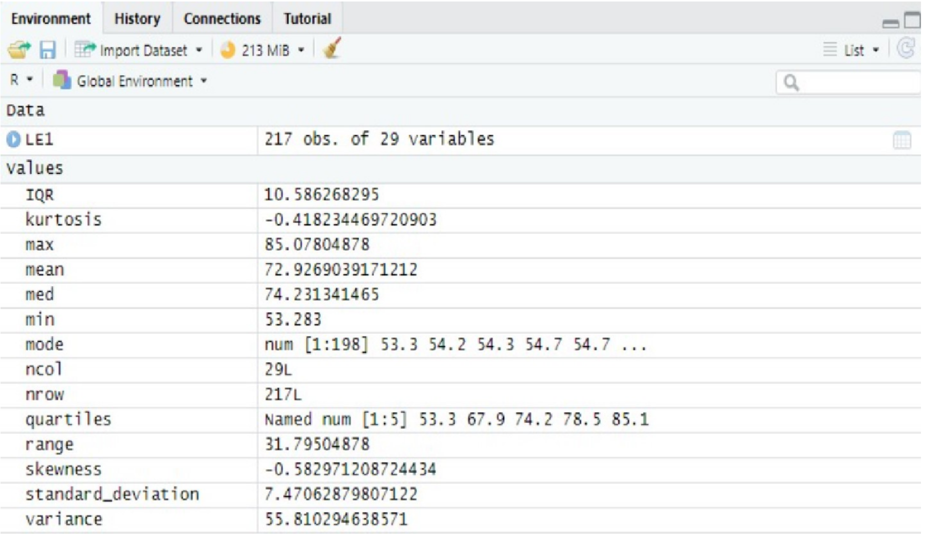
- ✓ The given data set consists of 217 rows & 29 columns.
- ✓ And it contains many null values, which are handled by some strategies called as Complete case analysis & Impute method.



```
> summary(LE1)
    C_Name              C_Code            Continent         life_expectency   crude_brth     
 Length:217         Length:217         Length:217         Min.   :53.28   Min.   : 5.90   Mi
 Class :character   Class :character   Class :character   1st Qu.:67.89   1st Qu.:10.62   1s
 Mode  :character   Mode  :character   Mode  :character   Median :74.23   Median :17.19   Me
                                                          Mean   :72.93   Mean   :19.37   Me
                                                          3rd Qu.:78.48   3rd Qu.:27.04   3r
                                                          Max.   :85.08   Max.   :45.64   Ma
                                                          NA's   :19      NA's   :13      NA
   acc_elect          adj_NNI            NNI_capit        renew_eng         HIV.0.14.        HIV
 Min.   :  6.721   Min.   :-30.792   Min.   :-32.5432   Mode:logical   Min.   :  100   Min.  
 1st Qu.: 84.762   1st Qu.:  1.225   1st Qu.:  0.5222   NA's:217       1st Qu.:  100   1st C
 Median :100.000   Median :  3.660   Median :  2.7583                  Median :  500   Media
 Mean   : 86.470   Mean   :  4.030   Mean   :  2.6585                  Mean   : 1650   Mean 
 3rd Qu.:100.000   3rd Qu.:  6.242   3rd Qu.:  5.0702                  3rd Qu.: 1100   3rd C
 Max.   :100.000   Max.   : 50.172   Max.   : 47.2518                  Max.   :20000   Max. 
 NA's   :1         NA's   :79        NA's   :79                        NA's   :127     NA's 
    not_prim           prim_25           Bch_25          prim.age         comp_edu          u
 Min.   :      0   Min.   : 49.55   Min.   : 4.322   Min.   : 54.73   Min.   : 0.000   Min. 
 1st Qu.:   1262   1st Qu.: 81.77   1st Qu.:11.898   1st Qu.: 85.82   1st Qu.: 9.000   1st C
 Median :   7359   Median : 93.69   Median :19.665   Median : 97.40   Median :10.000   Media
 Mean   :  98650   Mean   : 87.74   Mean   :19.864   Mean   : 93.05   Mean   : 9.919   Mean 
 3rd Qu.:  78956   3rd Qu.: 99.24   3rd Qu.:25.721   3rd Qu.:101.45   3rd Qu.:12.000   3rd C
 Max.   :1712650   Max.   :100.00   Max.   :46.631   Max.   :120.45   Max.   :17.000   Max. 
 NA's   :99        NA's   :181      NA's   :179      NA's   :89       NA's   :19       NA's 
    lit_rate         real_int          pop_grow          pop_dens            pop_total       
 Min.   : 58.00   Min.   :-78.518   Min.   :-1.6095   Min.   :    0.137   Min.   :1.076e+04  
 1st Qu.: 89.89   1st Qu.:  3.176   1st Qu.: 0.3882   1st Qu.:   38.177   1st Qu.:7.779e+05  
 Median : 95.74   Median :  6.354   Median : 1.0946   Median :   92.842   Median :6.661e+06  
 Mean   : 92.04   Mean   :  6.220   Mean   : 1.1917   Mean   :  446.043   Mean   :3.545e+07  
 3rd Qu.: 97.56   3rd Qu.:  9.214   3rd Qu.: 1.9556   3rd Qu.:  233.011   3rd Qu.:2.544e+07  
 Max.   :100.00   Max.   : 39.877   Max.   : 4.4687   Max.   :19466.444   Max.   :1.408e+09  
 NA's   :192      NA's   :104       NA's   :1         NA's   :1           NA's   :1          
   hlth_capit          hlth_GDP          GDP_grth          GDP_capit           safe_wtr
```

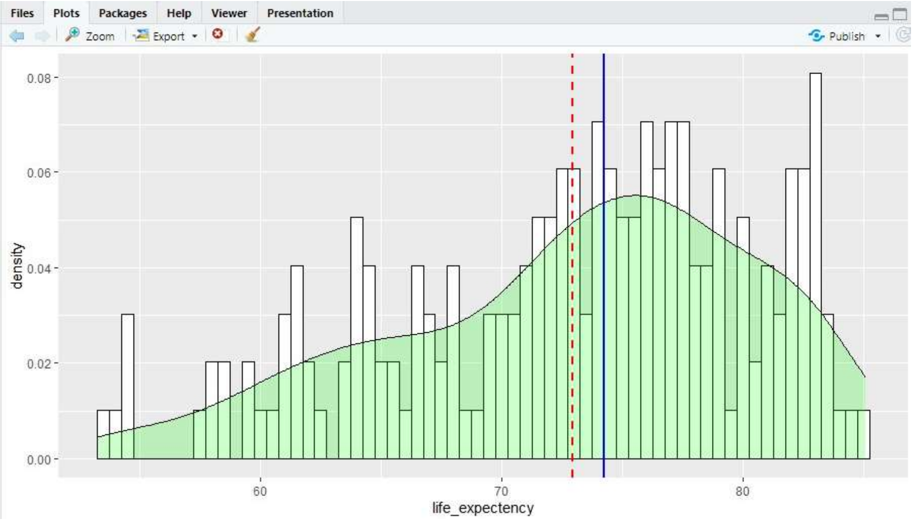
DESCRIPTIVE STATISTICS ANALYSIS OF LE1

Numerical Representation



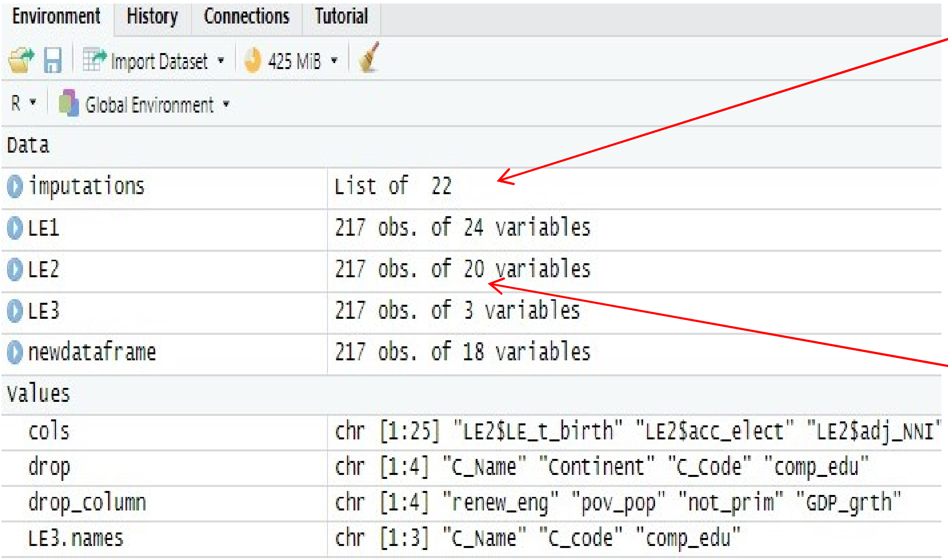
```
Data
LE1                  217 obs. of 29 variables
Values
  IQR                10.586268295
  kurtosis           -0.418234469720903
  max                85.07804878
  mean               72.9269039171212
  med                74.231341465
  min                53.283
  mode               num [1:198] 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
```

Graphical Representation



HANDLING MISSING VALUES IN LE1

Cleaned Dataset with no null values



```
Data
imputations     List of  22
LE1             217 obs. of 24 variables
LE2             217 obs. of 20 variables
LE3             217 obs. of 3 variables
newdataframe    217 obs. of 18 variables
Values
  cols          chr [1:25] "LE2$LE_t_birth" "LE2$acc_elect" "LE2$adj_NNI"
  drop          chr [1:4] "C_Name" "Continent" "C_Code" "comp_edu"
  drop_column   chr [1:4] "renew_eng" "pov_pop" "not_prim" "GDP_grth"
  LE3.names     chr [1:3] "C_Name" "C_code" "comp_edu"
```

Strategies Handled For Missing Values

- The columns with 75% of null values are removed since they don't contribute to the model, where will be 24 columns reduced from 29 columns.
- There are still found some columns with null values, those columns should be filled using the "*Predictive Mean Matching Method*".
- Some columns which are not important were manually removed.
- There will now be 18 columns, as opposed to 29, which is a '*newdataframe'* created to hold the new, null-free data.

INVESTIGATING THE PRESENCE OF MULTICOLLINEARITY

Investigating The VIF

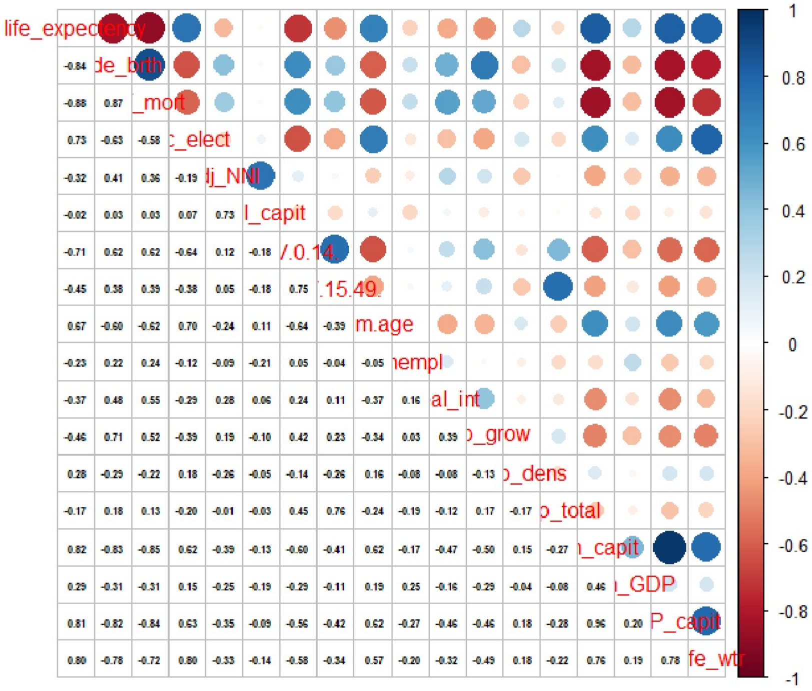
- The pairwise correlations figure shown that there are some large correlations so there is evidence for Collinearity between some predictors, since the *Variance Inflation Factor (VIF)* indicates that the variables have a high correlation with each other by the following conditions:

  VIF >10 indicates high correlation with each other.

  VIF < 4 is considered good for a model.

- So further investigating to find the best model by considering necessary variables to be included in the presence of one of these predictor variables, where by *Analysis Of Variance Table* preferring the smaller model.

Pairwise Correlations between predictor variables



FINDING THE BEST MODEL TO FORECAST LIFE EXPECTANCY OF 2020

Applying Multiple Linear Regression

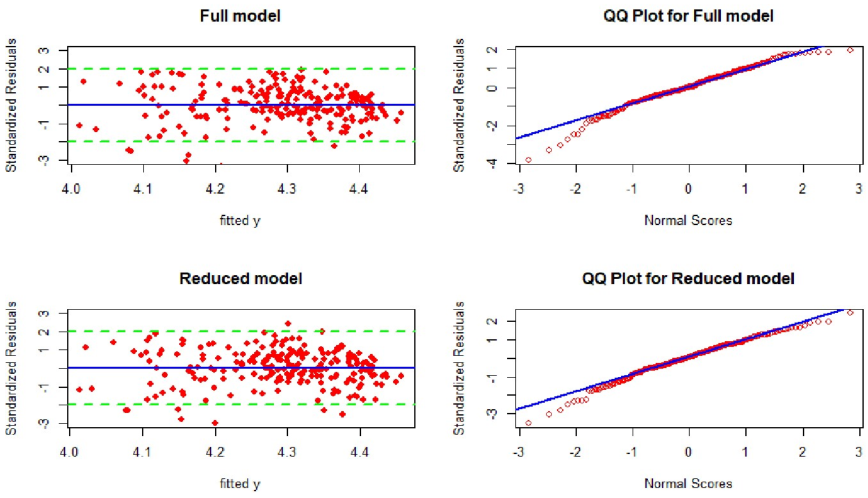
- Comparing the Full model with the proposed reduced model.
- Calculating the AIC and Mallow's Cp,

  where AIC_full = -828.7816,

  AIC_reduced = -827.3107,

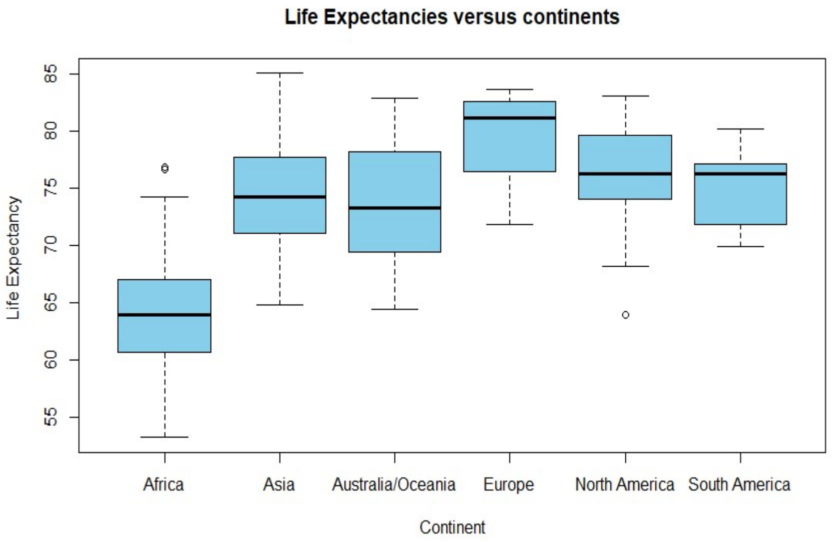
  & Cp = 12.49639 which is much higher than the number of variables in the model (=5+1 (intercept)) .
- The better fitting model is the one with the lowest value, which in this case is the reduced model.
- Therefore ,concluding that the Reduced Model is a good option.

Full Model VS Reduced Model

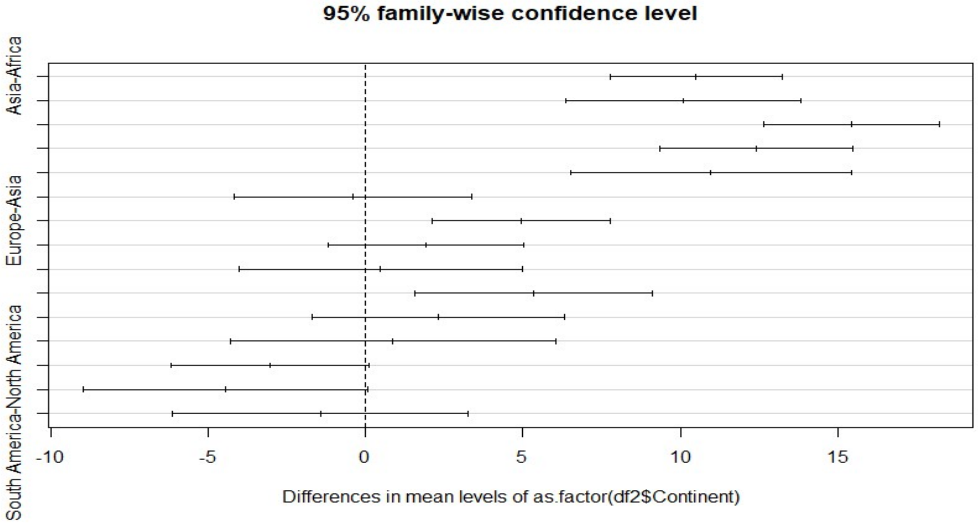


THE DIFFERENCES IN AVERAGE LIFE EXPECTANCY ACROSS THE CONTINENTS

Life Expectancy Versus Continents



Differences Of Average Life Expectancy Across Continents



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.

THANK YOU

References are not included in slides, since the slides are exceeding the limit.