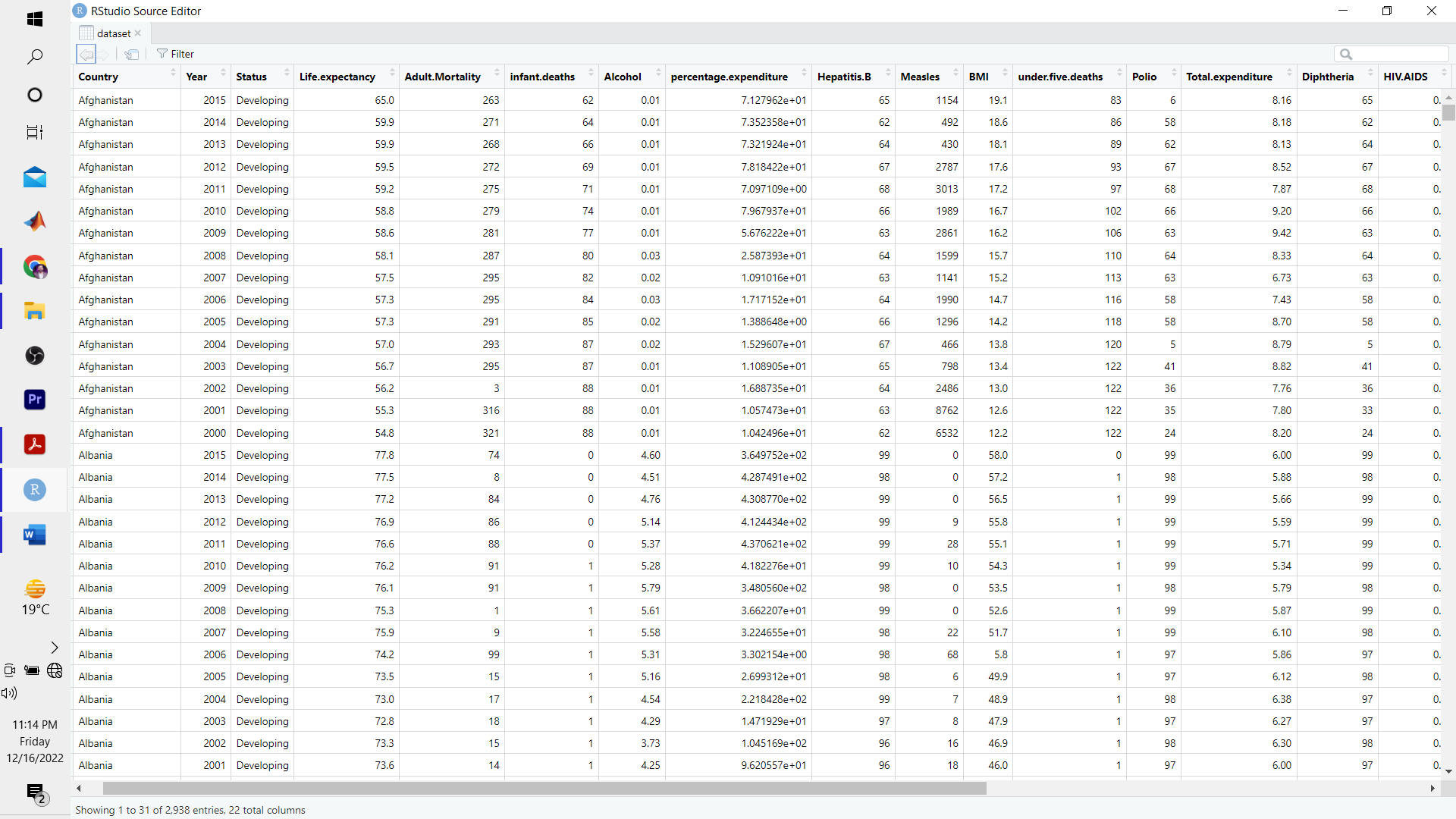
**Project Report**

**Research Question 1: Explore the Data**

1. **Overview of Dataset**



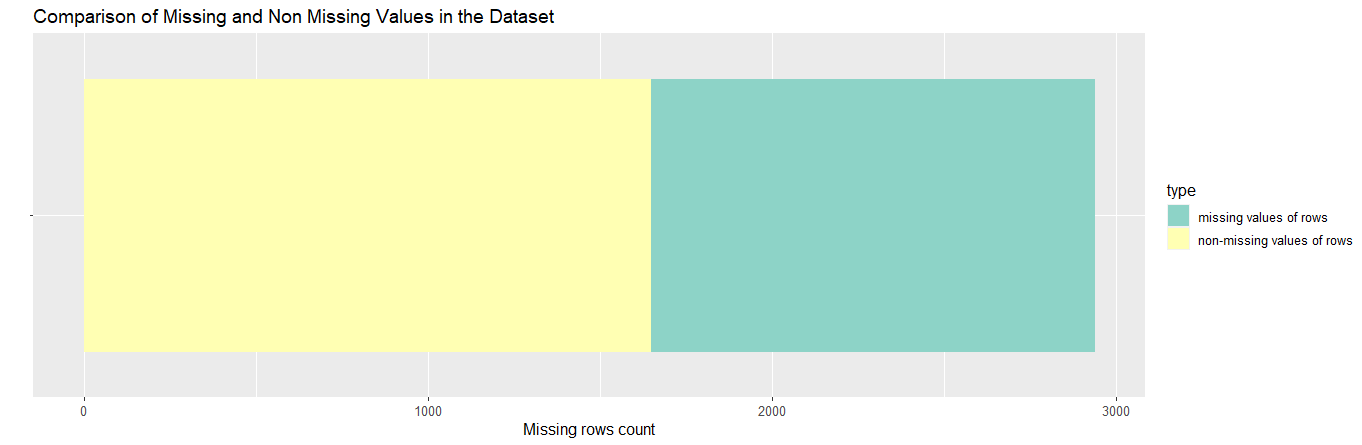
Given dataset is based on 2938 rows (observations) and 22 variables.

As per my observations, 22 variables can be categorized as

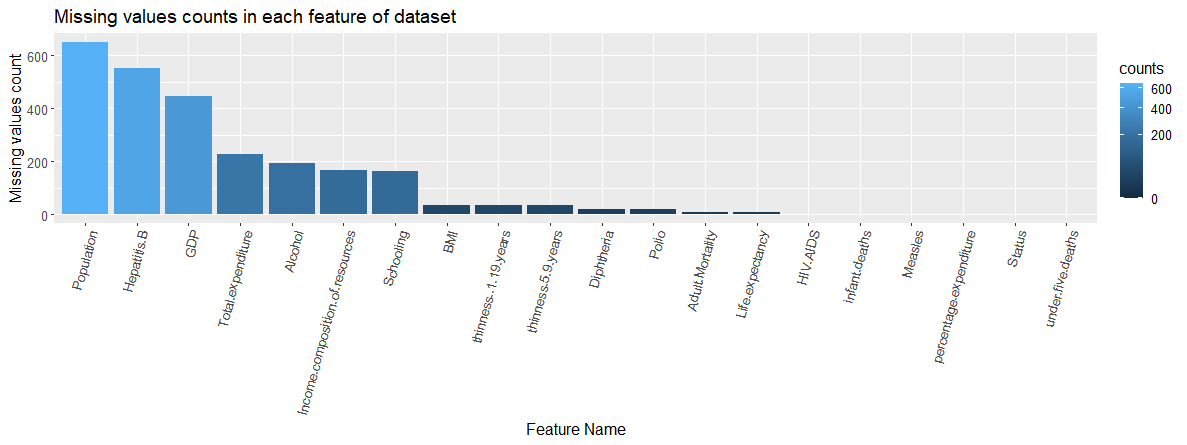
* Dependent Variable: Life Expectancy
* Categorical Features: Stats
* Removed from Analysis: Year and Country are to be removed from analysis
* Continues and Independent Variable: Rest of the features belong to this category.

1. **Exploratory Data Analysis of Dataset**

* In this report, I have performed data cleaning, exploratory analysis and applied linear regression model by considering life expectancy as a main continues feature.
* In this dataset, country and status were two categorial features.
* In the data cleaning process, I have removed year and country as these were not relevant to the analysis.
* Two estimate the missing data, a statistical analysis has been performed and I observed that 43.87% of the rows were based on the missing data.
* A visual representation has been given in the figure below.

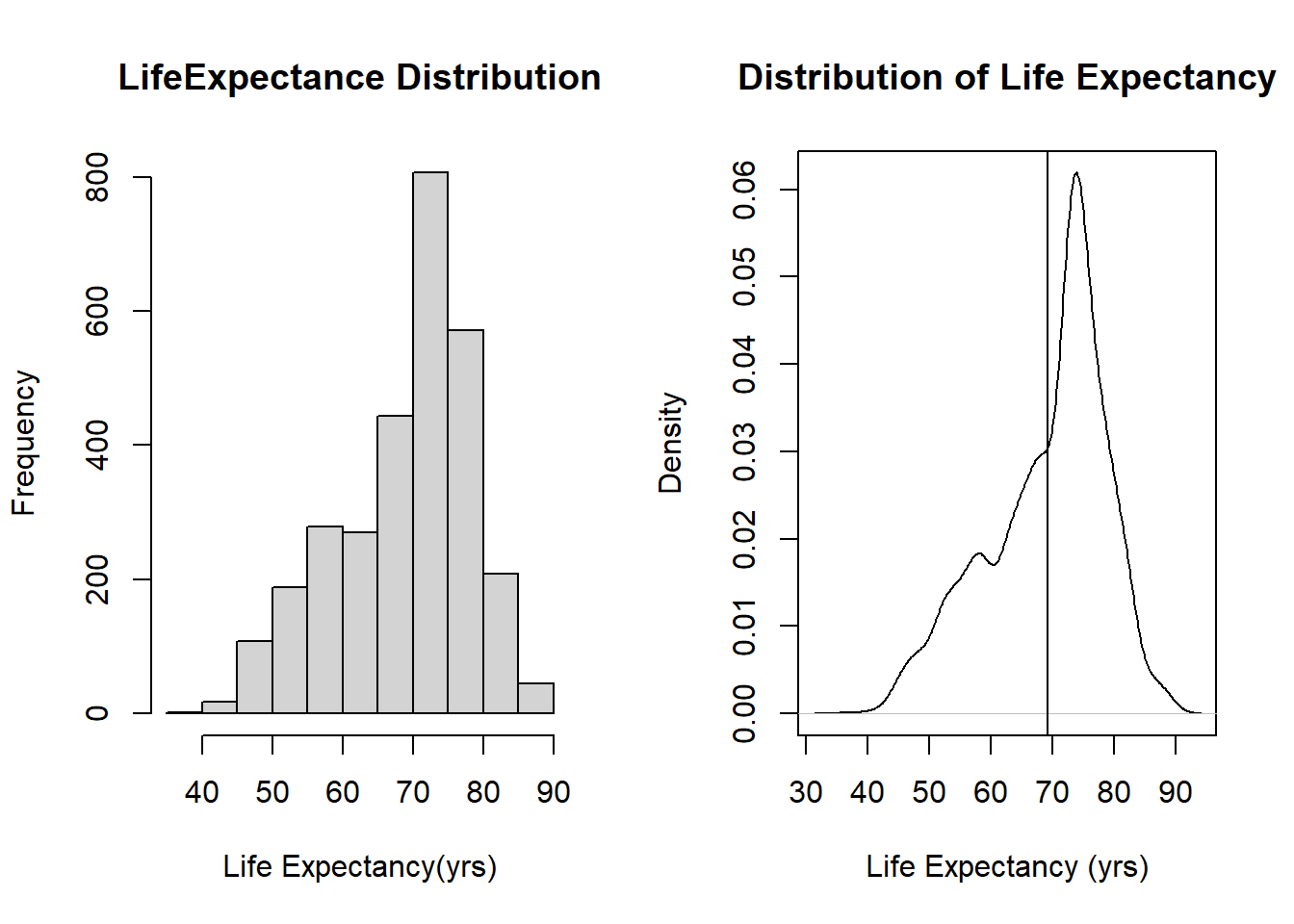


* To further understand the missing values, below chart provides the overall, missing values per feature.



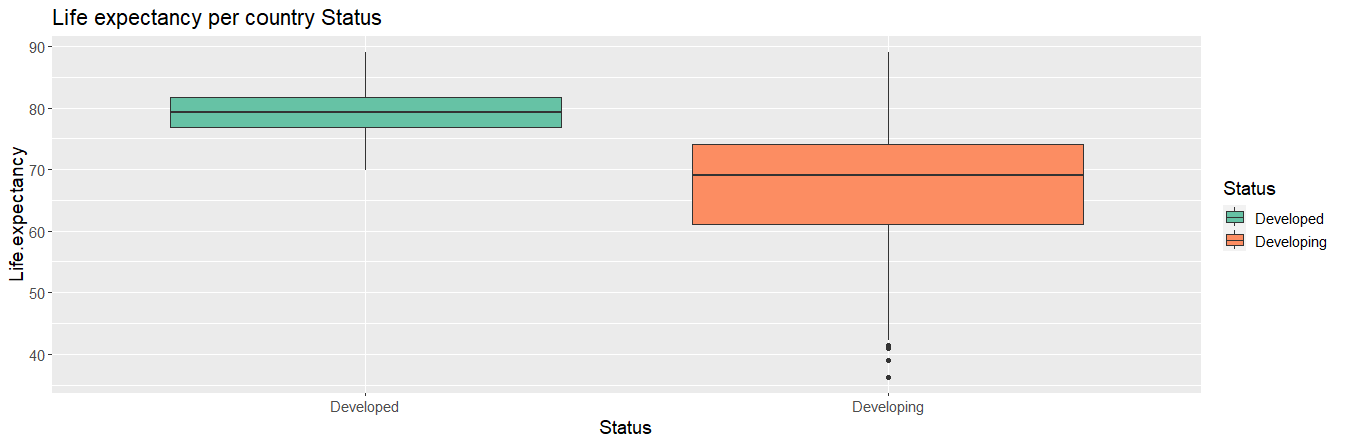
1. **Life Expectancy Distribution Trends**

* In this analysis, life expectancy is my core focus, so lets visualize the frequency distribution of this specific feature.
* From figure below it can be observed that it is not perfectly distributed.
* It is left-skewed.
* To standardize this, scaling can be applied.

****

1. **Categorical Variables Visualization**

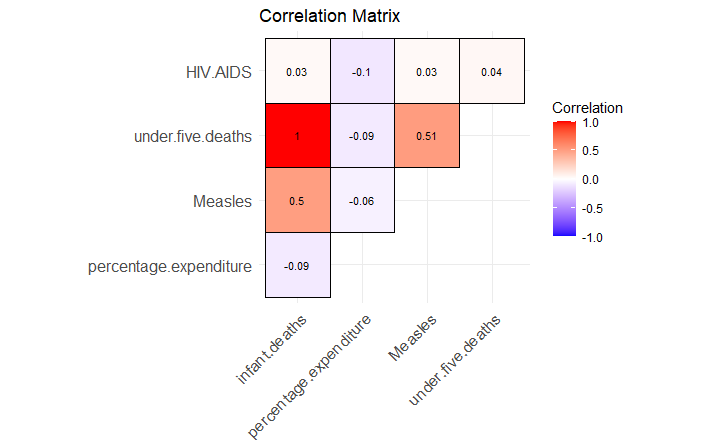
* I need to look at the life expectancy boxplots for both developed and developing countries so that I can get a better understanding of the relationship between the two.



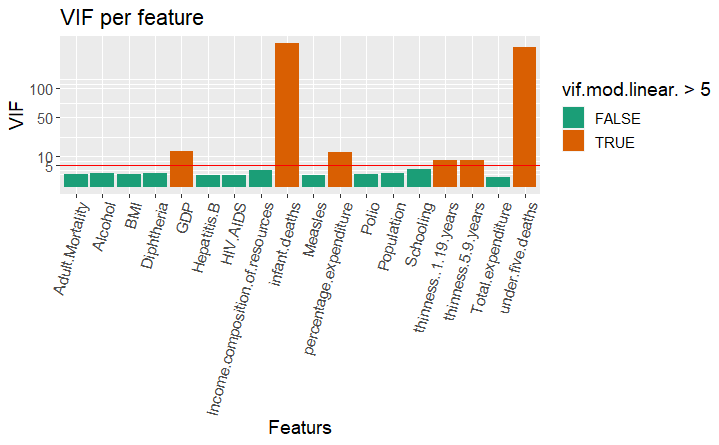
* Since Life expectancy is a categorical variable that appears to be a good predictor given the disparity between industrialised and developing nations.

1. **Correlation of Variables**

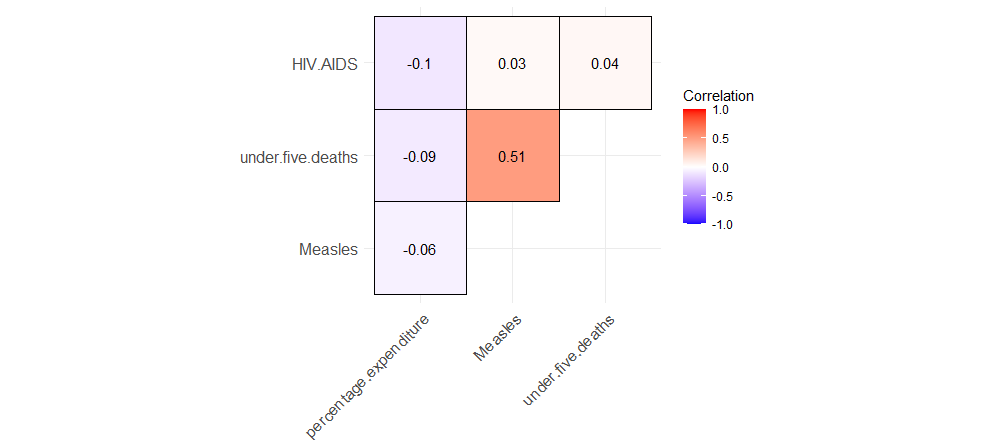
* Examining the connection between the features and the target will help me decide which ones can best function as predictors in the model. Collinearity is something we want to avoid in constructing our model, and the correlation between features and themselves can show that it is likely to occur.



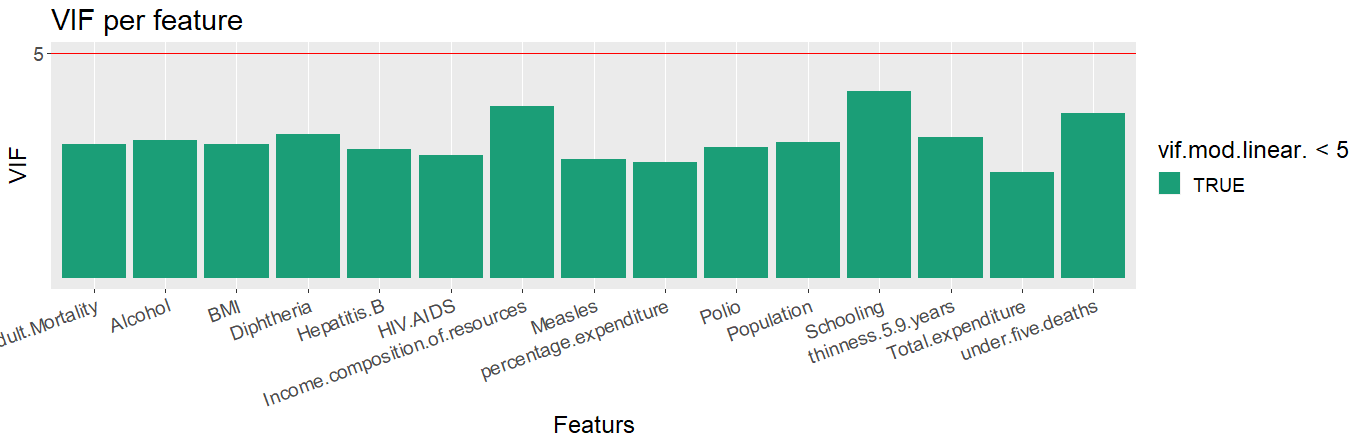
* Above matrix shows the correlation values of infant.deaths and under.five.death=1
* In order to determine which features, have the potential to serve as reliable predictors for the model, it is necessary for me to investigate the correlation that exists between the features and the target. Collinearity is something we want to avoid in constructing our model, and the correlation between features and themselves can show that it is likely to occur.



* Here three features are showing higher correlated pairs.
* It is required to eliminate those pairs.
* GDP needs to be omitted.
* There is a high degree of collinearity between the features baby deaths and under-five deaths, as seen by their aberrant VIFs, which are much higher than 5. Because of this, we have to get rid of one of the two options. Given that infant.deaths has a greater VIF than the other categories, we have decided to exclude it.
* I have created a newer version of data in code as well for this.



* Now, it can be observed that there are no suspicious feature.
* Correlation of Measles and under.five.deaths is 0.51
* Correlation of HIV.AIDS and percentage expenditure is -0.1
* Correlation of HIV.AIDS and under five deaths is 0.04
* Lets recheck the VIF per feature values.



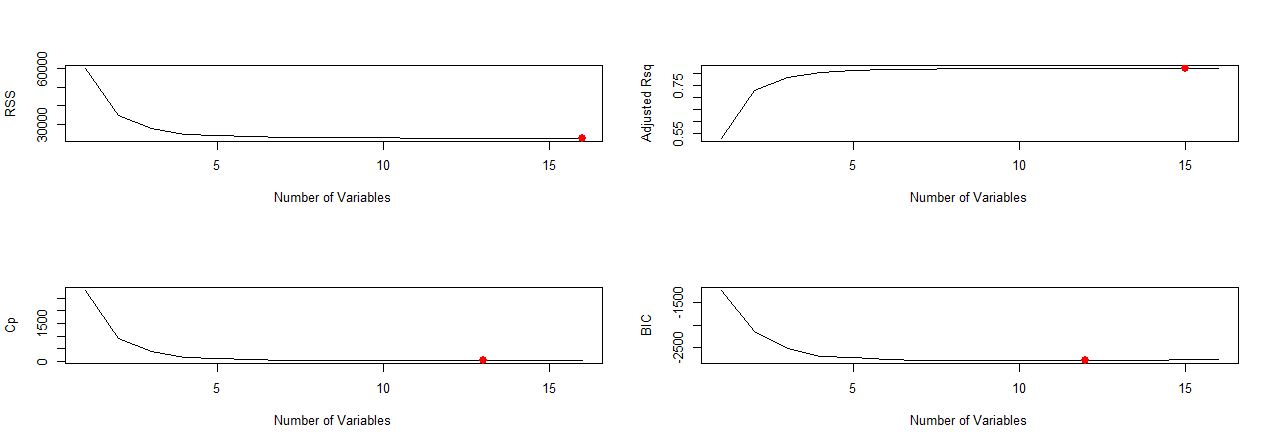
* It can be observed that there are no abnormalities present in the dataset.
* Here, all the exploratory data analysis has been completed. Lets observe how model building is based on these findings of the datasets.

**2. Model Building**

The act of selecting a subset of relevant features or variables from a larger data collection in order to use them in the construction of models is referred to as "feature selection." In order to determine which set of characteristics is the most important, we are going to apply three different regression subset methods.

**1. Best Subset Method:**The first methos is the best subset method. Its results are based on these parameters

* RSS: Residual Sum of Square
* ADJR2: Adjusted Rseq
* CP: Mallows Cp
* BIC: Bayesian Information Criterion

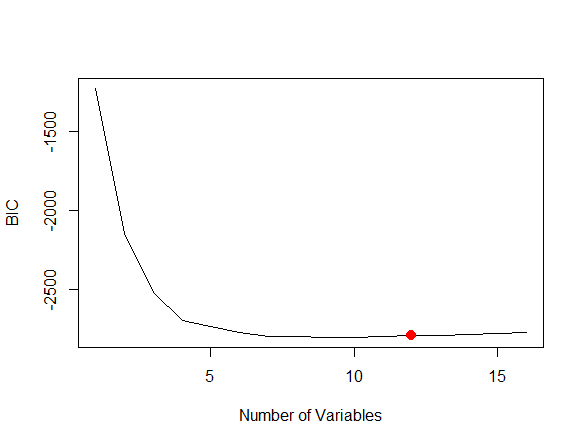


**Results:**

The various metrics each proposed a unique number of variables that would make the optimal subsets. Its results indicate that these parameters shows minimum number of suitable variables as given below:

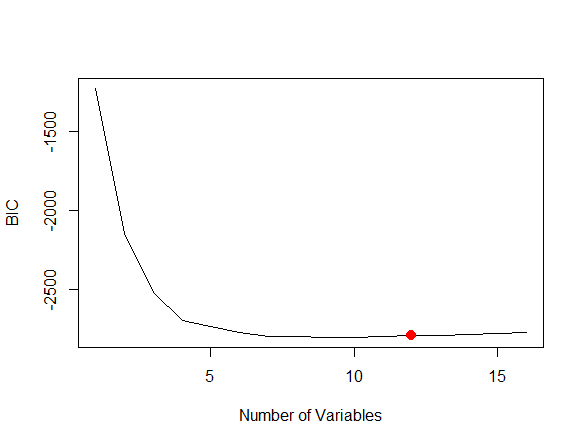
* RSS: 16
* ADJR2: 15
* CP: 13
* BIC: 12

**2. Forward inclusion Method:**

* It has recommended the minimum number of variables to be 12.

**3. Backward elimination Method:**

It has also shown the 12 minimum number of variables.



1. **Linear Regression Model and Results Evaluation**

**Results:**

* Residual standard error: 3.661 on 1131 degrees of freedom (908 observations deleted due to missingness)
* Multiple R-squared: 0.823, Adjusted R-squared: 0.8205
* F-statistic: 328.7 on 16 and 1131 DF, p-value: < 2.2e-16

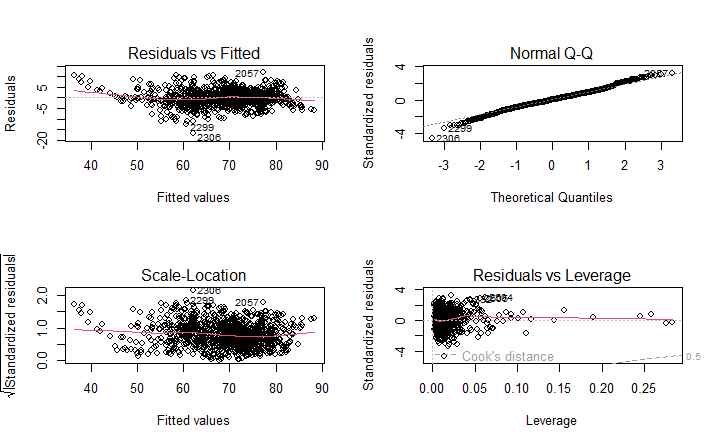
For the comprehensive model, the linearity assumptions are satisfied. It seems clear that the following assumptions about linear regression are approximately true:

**The Residuals vs Fitted plot:** The residual plot does not display any discernible patterns. It is reasonable to believe that there is a linear connection between the factors being predicted and the outcomes.

**The Scale-Location plot:** The values of the residuals are distributed uniformly across the ranges of the predictors

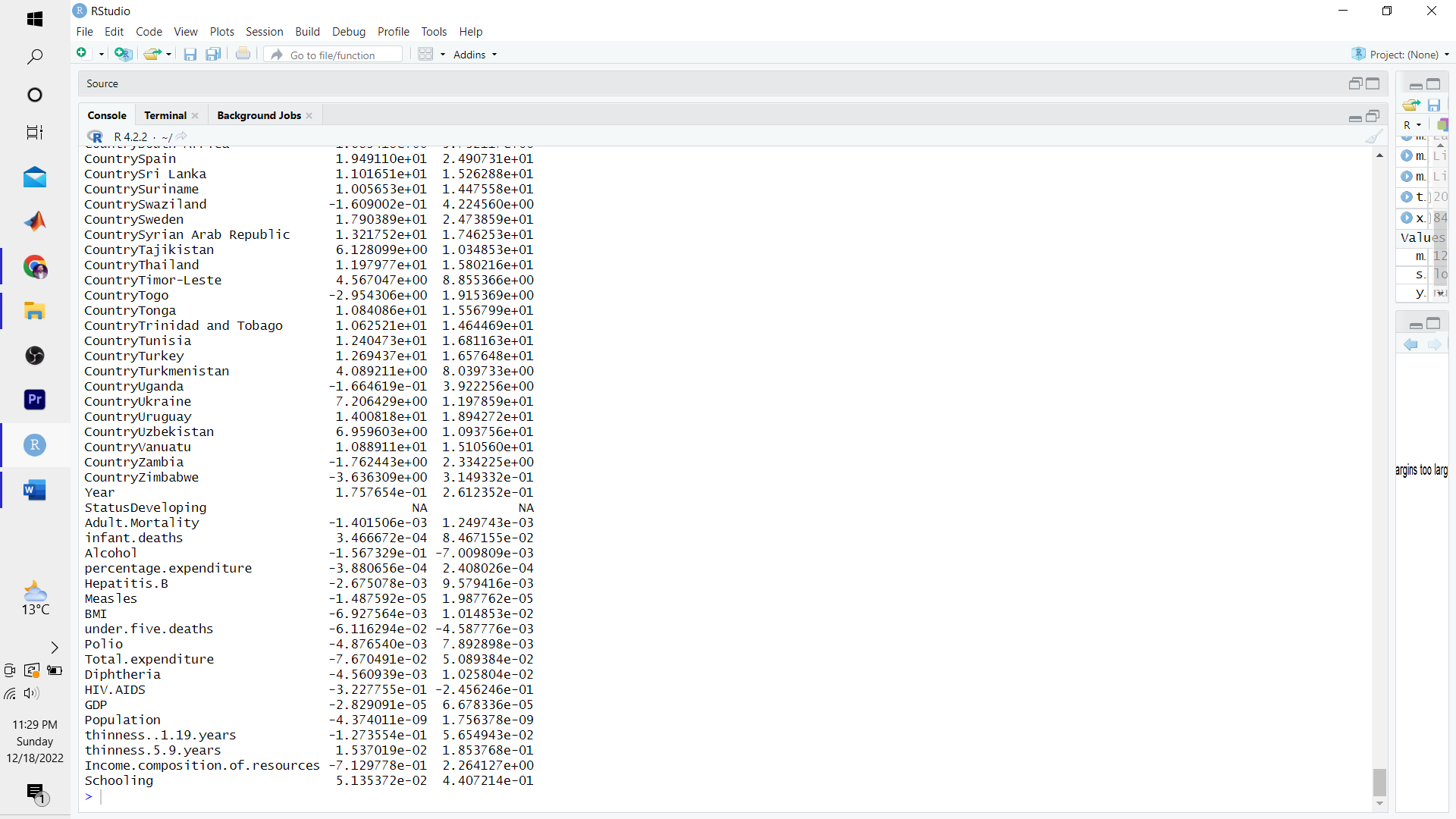
**The Normal Q-Q plot:** In the event that all of the points are distributed along the reference line in an equal manner, it is safe to assume that the data possesses the normality of the residuals.

**The Residuals vs Leverage:** Any location inside the boundary of 0.5 cook distance will have an effect. On the other hand, there is nothing of note to mention here. (based on the assumption of linearity)



* **Confidence Interval Results:**

Parameter 2.5% 97.5%



* **View Model**

