Term Project Report

Life Expectancy (WHO)

Statistical Analysis on factors influencing Life Expectancy



Group 5:

Nayera Kamel

Anna Sak

Jessica Cichon

Nicholas HoKoon

Farukh Jabeen

Zohaib Ratani

Table of Contents

1. Introduction: Objective 3
2. Data Preparation 3

Nulls - Life Expectancy and Adult Mortality 3

Nulls – Large percentage of missing values 4

Nulls – Small percentage of missing values 6

Unlikely Values – BMI < 10 6

1. Exploratory Analysis 7
2. Predictive Model 13
3. Conclusion 20
4. Reference 22

Appendix A – Data Cleansing – External Data Source Cleansing 22

# Introduction: Objective

The objective of this project is to analyze life expectancy data from the World Health Organization in order to determine trends and create a predictive model to predict a countries life expectancy. The dataset included various features that are related to life expectancy such as geographical locations, economics, and health-related factors. Among the variables in the dataset, it was found that some of them such as Adult Mortality, Income composition of resources, and Schooling were correlated with Life Expectancy. There were also other variables that were correlated with each other such as GDP & percentage expenditure, Income composition of resources & Schooling, etc. These correlations will be further explored in order to determine which variables are most related to life expectancy in order to create an accurate predictive model.

# Data Preparation

The main dataset used in the analysis and predictions was found on Kaggle and was originally from the World Heath Organization (WHO), which keeps track of the health status as well as many other related factors for all countries. The dataset had data from 193 countries from 2000 to 2015. It has been observed that in this time period, there has been a general increase in life expectancy, especially in developing countries. The dataset had 2938 datapoint and 22 features, these included: *Country, Year, Status, Life expectancy, Adult Mortality, infant deaths, Alcohol, percentage expenditure, Hepatitis B, Measles, BMI, under-five deaths, Polio, Total expenditure, Diphtheria, HIV/AIDS, GDP, Population, thinness 1-19 years, thinness 5-9 years, Income composition of resources,* and *Schooling*. The dataset contained some missing values, which will be explored in the following sections.

## Nulls - Life Expectancy and Adult Mortality

The missing data of *Life Expectancy* and *Adult Mortality* were from less known and small countries like the Marshal Islands, Palau, Tuvalu, etc. These countries also only had one year of data compared to the other countries, which had 16. Therefore, these countries were dropped from the final dataset and this had little effect on the final results. The countries that were dropped are:

1. Cook Islands

2. Dominica

3. Marshall Islands

4. Monaco

5. Nauru

6. Niue

7. Palau

8. Saint Kitts and Nevis

9. San Marino

10. Tuvalu

## Nulls – Large percentage of missing values

After checking the data-set, there are a large number of *nulls*, and in the case of the columns in Table 1, the *nulls* account for 15% or more of the data. It was decided to fill these columns from external data-sources in an attempt to make the model as accurate as possible.

Table 1

|  |  |  |
| --- | --- | --- |
| **Column** | **% Nulls** | **External Source** |
| **Hepatitis B** | 18.8 | https://data.worldbank.org/indicator/SH.IMM.HEPB |
| **GDP** | 15.5 | https://data.worldbank.org/indicator/NY.GDP.PCAP.CD |
| **Population** | 22.2 | https://data.worldbank.org/indicator/sp.pop.totl |

The data from the external sources also needed to be cleaned. It was noticed that the names of the countries used in the external dataset did not precisely match the names of the country in the primary dataset. They needed to be matched with each other, as shown in Table 4 in 0Appendix A – Data Cleansing – External Data Source Cleansing.

Additionally, when pulling year-by-year data from each of the external sources, it was discovered that for some of the specific years of interest, the external data-source also had *nulls*; however, there was data for other years that we were not interested in. The *nulls*, in this case, were filled using the average of the existing data.

As an example, for the variable *Hepatitis B*, in the external data-source, Algeria was missing values for the years 2000 to 2003 but had data for other years 2004 to 2015, as shown in Table 2.

Table 2 – External Data Source for Hepatitis B - Algeria

|  |  |
| --- | --- |
| **Year** | **Hepatitis B** |
| **2015** | 95 |
| **2014** | 95 |
| **2013** | 95 |
| **…** | … |
| **2006** | 8 |
| **2005** | 83 |
| **2004** | 81 |
| **2003** | null |
| **2002** | null |
| **2001** | null |
| **2000** | null |

See Table 5 in 0Appendix A – Data Cleansing – External Data Source Cleansing for the list of countries where *nulls* in the external data-source were filled using this method.

Specifically, in the case of Hepatitis B, the following had absolutely no data to work with, however, it was noticed these countries were generally well-developed European countries (except for Japan), and so the *nulls* were filled using the average of another similar country, the Netherlands:

* Denmark
* Finland
* Hungary
* Iceland
* Japan
* Norway
* Slovenia
* United Kingdom of Great Britain and Northern Ireland

Also, for GDP, it was found that the *Democratic People's Republic of Korea* had no data, and the *nulls* were filled using the only value available through the CIA World Factbook[[1]](#footnote-2).

## Nulls – Small percentage of missing values

The remaining columns had a small percentage of *nulls* in comparison to the overall data, which could be filled using the interpolation method, which creates values based on the existing data and does not have a significant impact on the model. Linear interpolation involves estimating a new value by connecting two adjacent known values with a straight line. Additionally, some of the columns had the same missing *nulls* as another column, as shown in Table 3.

Table 3

|  |  |  |
| --- | --- | --- |
| **Column** | **Matching Nulls Column** | **% Nulls** |
| Polio | Diphtheria | 0.65 |
| Total Expenditure |  | 7.69 |
| Diphtheria | Polio | 0.65 |
| Thinness 10-19 years | Thinness 5-9 years | 1.15 |
| Thinness 5-9 years | Thinness 10-19 years | 1.15 |
| Income composition of resources |  | 5.68 |
| Schooling |  | 5.54 |

We used interpolate function to fill the nulls in *Polio, Diphtheria, Total Expenditure, Income composition of resources, schooling, Thinness 10-19 years, and Thinness 5-9 years* as it is a very powerful function to fill the missing values. It uses various interpolation techniques to fill the missing values rather than hard coding the values. We used the linear method to fill the remaining null values in the life expectancy dataset where the parameter ‘method’ to ‘linear’ ignores the index and treats the values as equally spaced.  Also, we used forward direction to fill in any consecutive nulls. The interpolate function has manipulated all the remaining null values.

## Unlikely Values – BMI < 10

It was noticed that there are over 200 entries where the BMI is less than 10, which is near impossible to achieve in reality. These values were replaced with the average BMI for the respective country, excluding the values where BMI is less than 10.

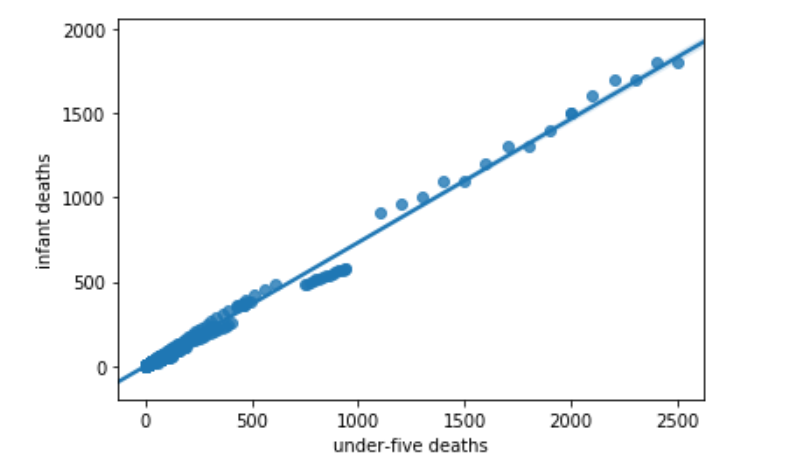
# Exploratory Analysis

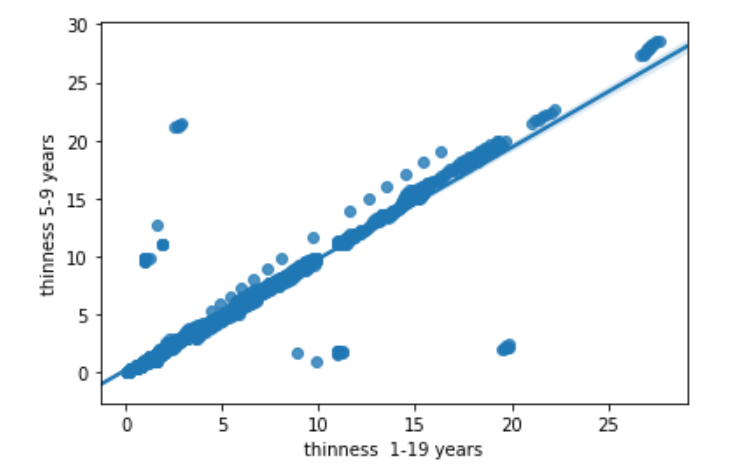
During the exploratory analysis, we looked at pair-wise correlations between all attributes in the dataset and there are some attributes that are highly correlated with each other, which could adversely impact the OLS model. In order to reduce collinearity, some features were combined.

Feature engineering was done for the following attributes:

* ***under-five deaths and infant deaths*** werefound to be almost perfectly linearly related, as shown in the plot below. Since the unit of measurement is also the same between the two, they can both be merged into an engineered feature called ***pct\_infantDeaths***.
* ***Thinness 10-19 years and thinness 5-9 years*** were found to be almost perfectly linearly related, as shown in the plot below. Since the unit of measurement is also the same between the two, they can both be merged into an engineered feature ***pct\_thinness.***

This decision was made based on the strong correlation of these variables, which are presented on the plots below:





From the initial data-exploration, it was determined that the following could be merged/engineered into a new attribute, as listed in Table 4. This was achieved by finding the mean of the attributes, then converting it to a percentage. A base of 1000 was used since the individual attributes were measured per 1000 people.

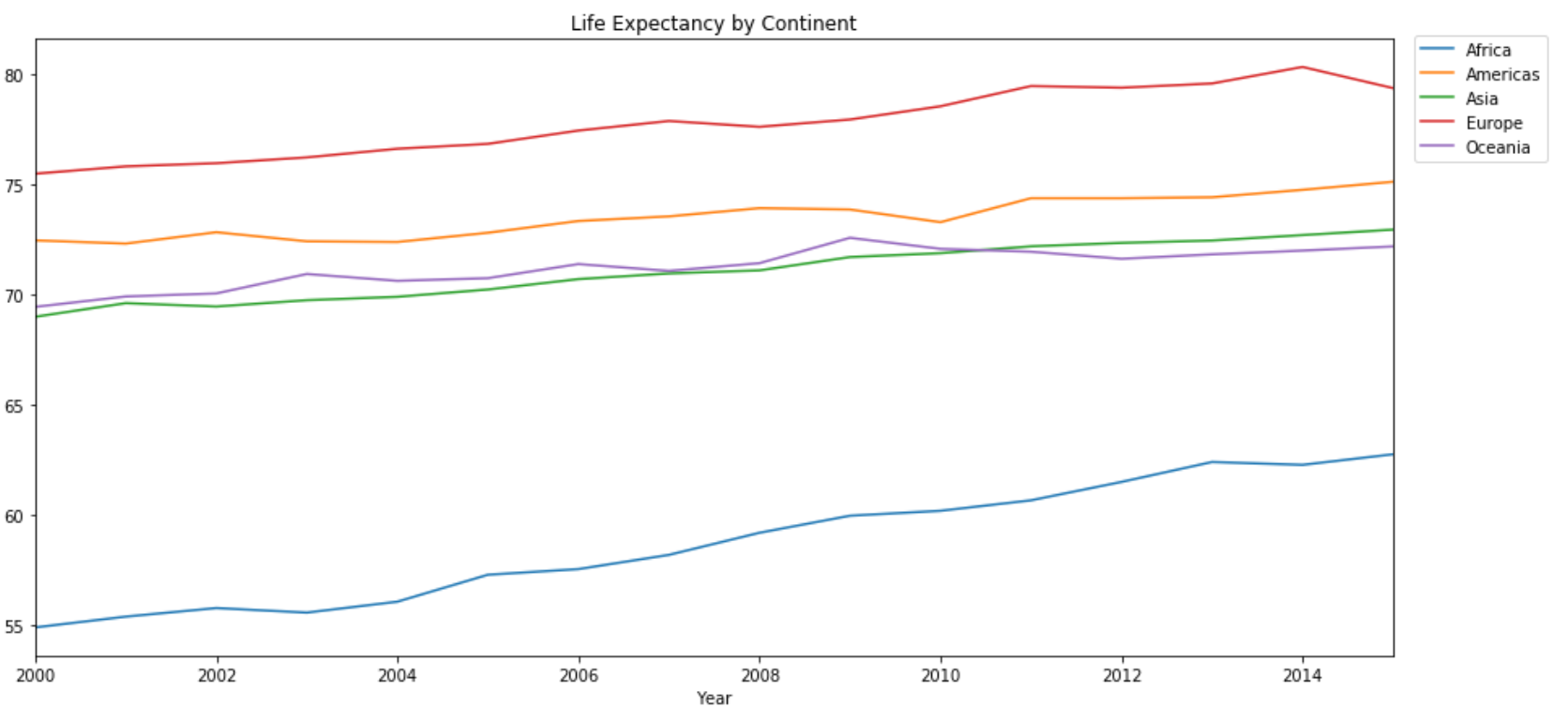
* Table 4 – Attribute Pairs merged into Engineered Attribute

|  |  |
| --- | --- |
| **Attribute Pair** | **Engineered Attribute** |
| under-five deaths/ infant deaths | pct\_infantDeaths |
| thinness 10-19 years/thinness 5-9 years | pct\_thinness |

***GDP*** and ***percentage expenditure*** – the unit of measurement is not the same between these, and so they cannot be easily merged into an engineered feature, therefore percentage expenditure was removed from the dataset.

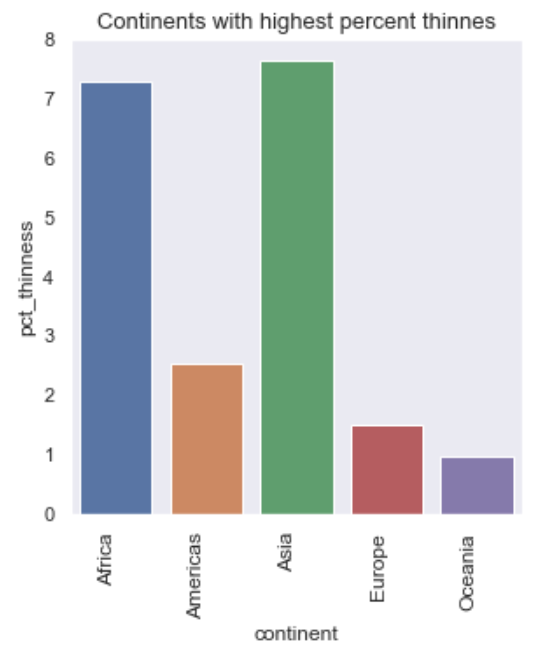
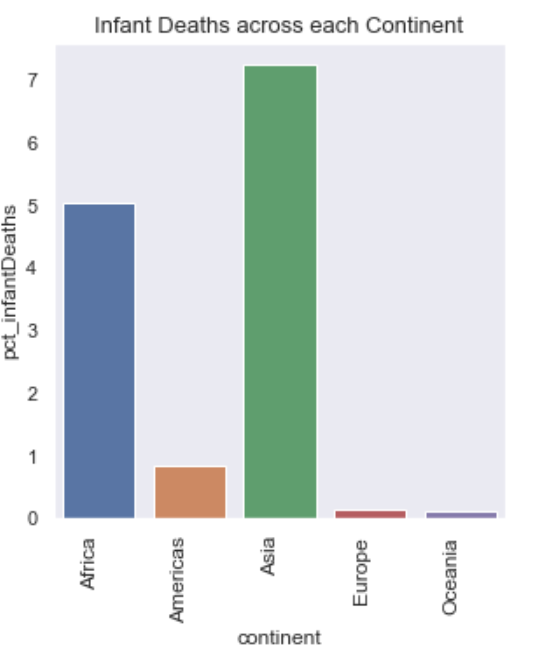
***Income composition of resources*** and ***schooling*** – the unit of measurement is not the same between these, and so they cannot be easily merged into an engineered feature, therefore income composition of resources was removed from the dataset.

During our analysis, we decided to add Continent data, which was sources from Kaggle to our dataset, to group all countries by continents. The additional variable, continent, helped to subset and group the data. In the plot below, we can see that Africa has the lowest Life Expectancy in comparison to other continents.



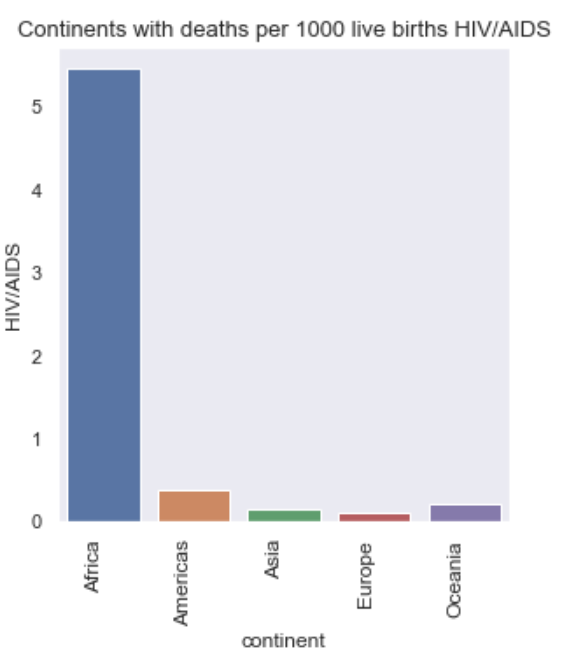
Following the data preparation by filling null values, feature engineering, and adding continent variable the data was explored in depth to uncover trends.

The first feature we would like to explore is *pct\_thinness.* On the following plots, we can see that Africa and Asia has the highest *pct\_thinness* (kids thinness) and *Infant death* rate among the other continents. The reason of the highest *pct\_thinness* and *Infant death* rates in Asia continent could be explained by the fact that India is one of the Asian countries with the largest population which has a big impact on all Asian continent.

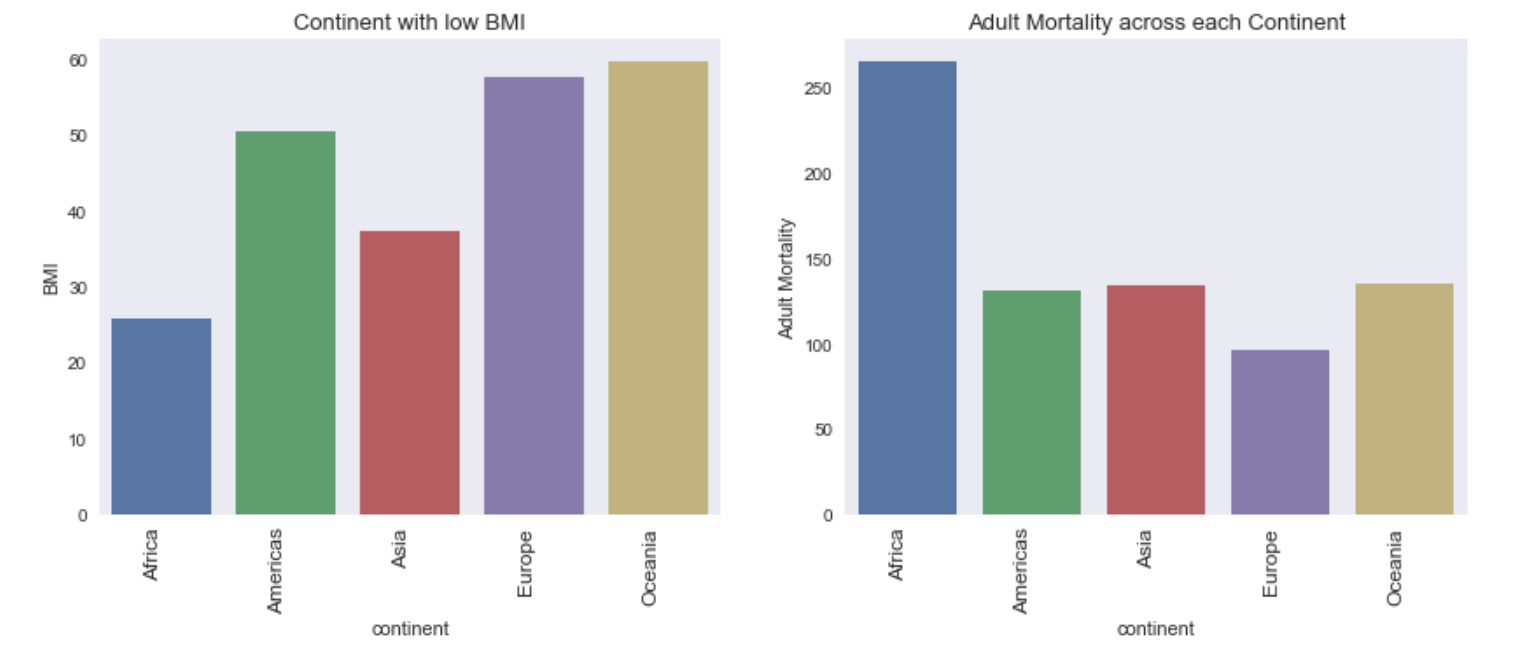
 

We can observe a strong relationship between *pct\_thinness* (kids thinness) and *Infant death* variables: Africa and Asia have the highest rate of *pct\_thinnes* and *Infant Deaths.*

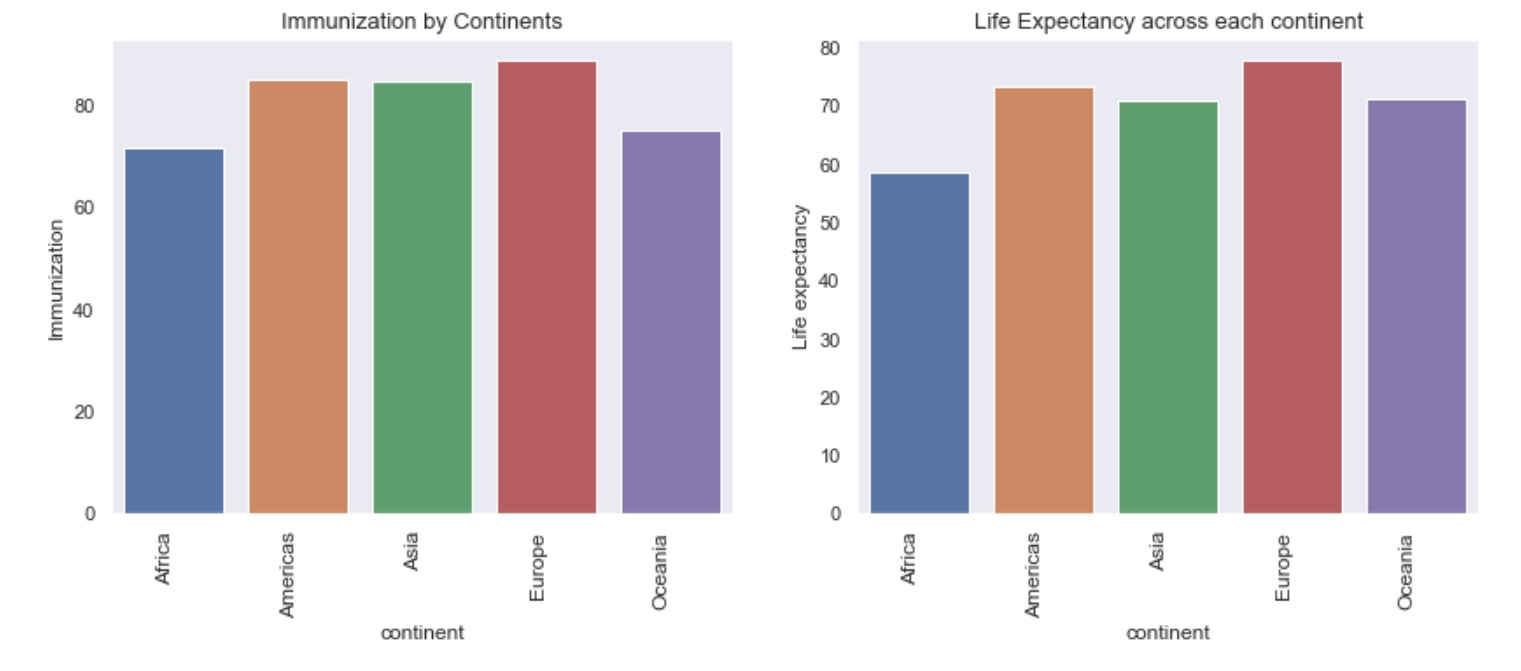
The following exploration of the HIV/AIDS variable showing us that the highest rate of Deaths per 1000 live births HIV/AIDS (0-4 years) is presented in African continent as well.



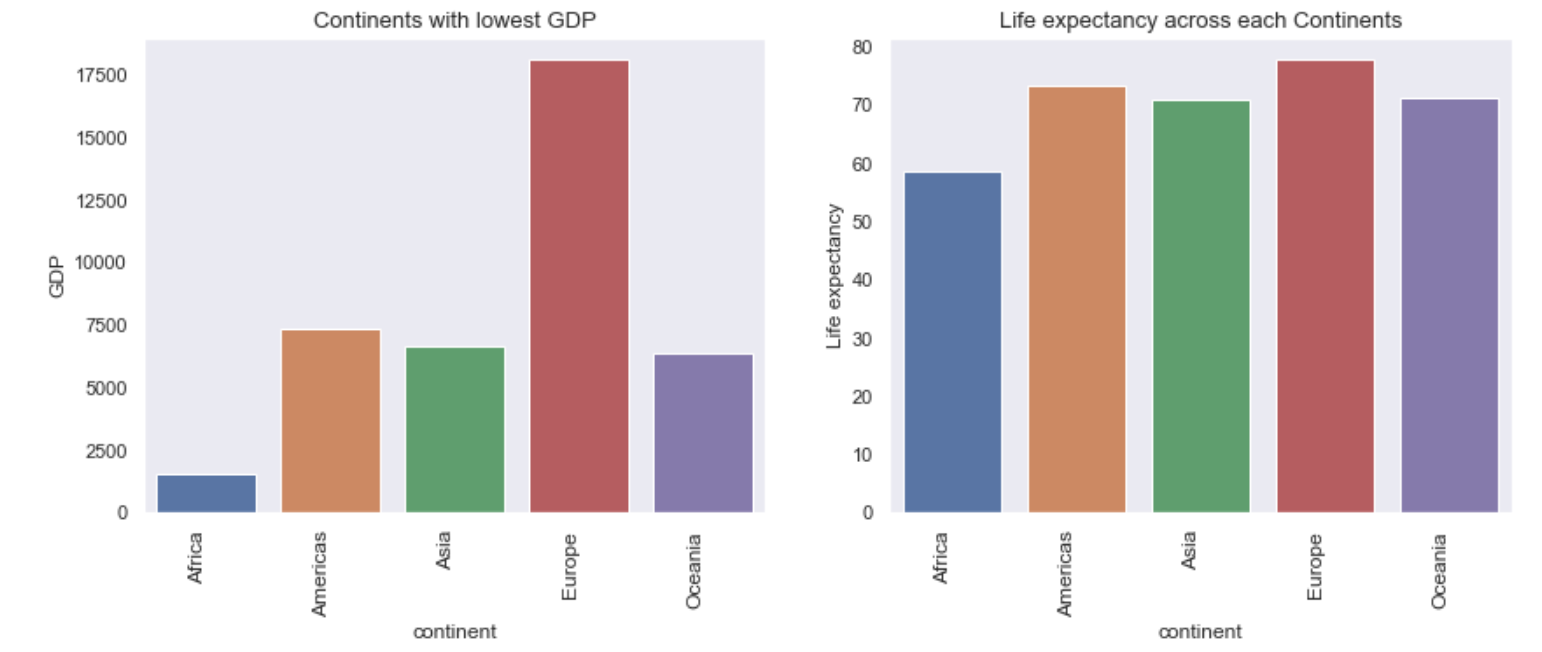
Exploring BMI variable, we can see on the following plots that the lowest rate of BMI is in African Countries at the same time Adult Mortality is high in Africa continent as well.



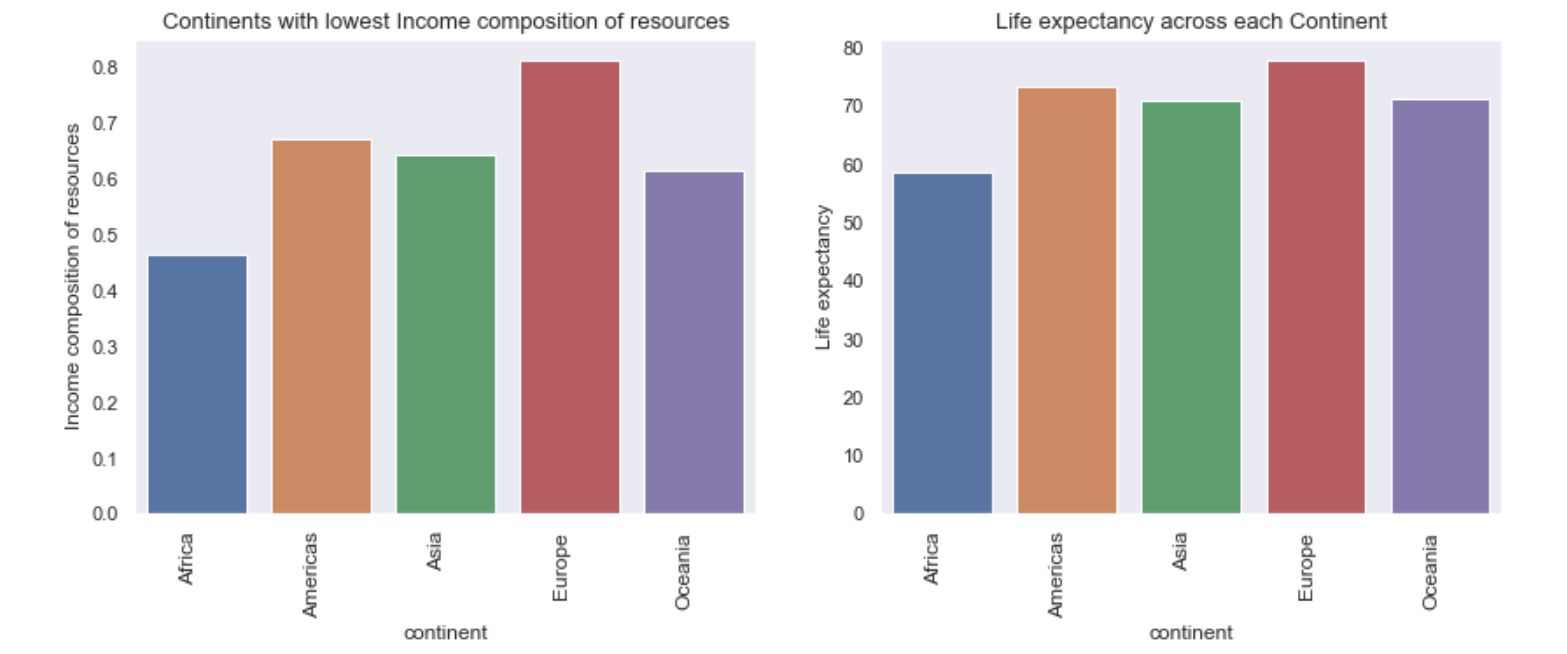
On the next step of our analysis, we will check continents with the lowest Engineered *Immunization* variable and *pct\_infantDeaths*. The following plots are showing that the African countries have the lowest immunization rate and the lowest Life Expectancy rate respectively.



Going further, we can look at the plots representing continents with the lowest GDP rate and Life Expectancy. On the plots, we can notice that African continent has the lowest GDP rate as well as the Lowest Life Expectancy.



Another confirmation we can get from the plots representing continents with the lowest Income Composition of resources and Life Expectancy. We can see on the plots below that the same African continent have Lowest Income composition of resources and Life Expectancy.



Based on the analysis performed, Africa has the lowest Life Expectancy rate for several reasons:

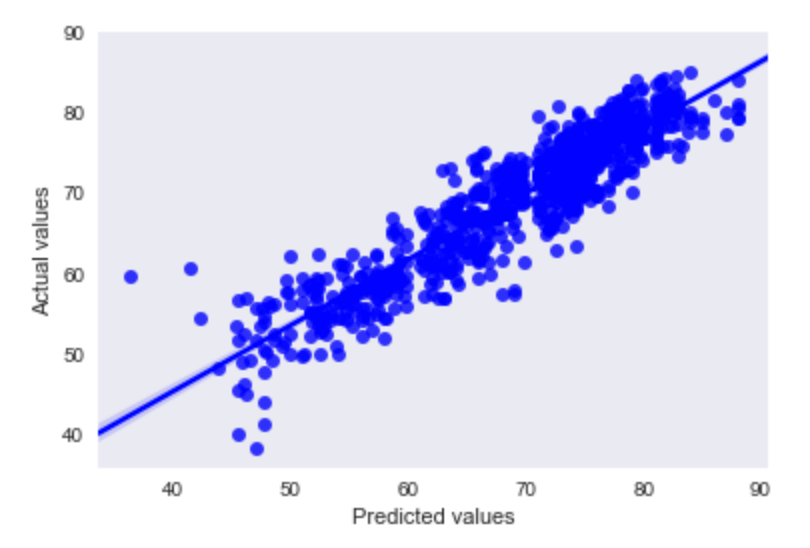
* + Kids thinness which is increasing the kid’s deaths
  + HIV/AIDS rate which is increasing kid’s deaths
  + Low level of immunization which leads to the early deaths
  + Lowest GDP
  + Lowest Income Composition of resources

# Predictive Model

The predictive model was built to predict the life expectancy for different countries based on the historical data for 16 years.

Two algorithms were used for building predictive models:

* Linear Regression
* Random Forest Regression

As multicollinearity was observed in the final data set, it was decided to remove the following variables which are highly correlated with each other: 'Income composition of resources,' 'percentage expenditure,' and ‘region\_code’. Afterward, the data was split into test set and training set with the ratio of 30% to 70%, respectively. The training set contains a known output and the model learns on this data in order to be generalized to other data later on. We have the test dataset (or subset) in order to test our model’s prediction on this subset. The data was splitted randomly into test and training sets and then fitted a regression model to the training data, made predictions based on this data and tested the predictions on the test data. The OLS results can be seen in the following figure. The R2 value is close to 0.997, which is very close to 1(The R² is a “number that indicates the proportion of the variance in the dependent variable that is predictable from the independent variable(s)”. Basically, how accurate is our model) The Kurtosis value is 4.018, which is higher than expected. The condition number is also quite high, which indicated there is still some multicollinearity.

|  |  |  |  |
| --- | --- | --- | --- |
| **OLS Regression Results** | | | |
| **Dep. Variable:** | Life expectancy | **R-squared:** | 0.997 |
| **Model:** | OLS | **Adj. R-squared:** | 0.997 |
| **Method:** | Least Squares | **F-statistic:** | 3.507e+04 |
| **Date:** | Thu, 05 Dec 2019 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 22:39:20 | **Log-Likelihood:** | -5419.8 |
| **No. Observations:** | 1971 | **AIC:** | 1.088e+04 |
| **Df Residuals:** | 1952 | **BIC:** | 1.098e+04 |
| **Df Model:** | 19 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **Year** | 0.0275 | 0.000 | 71.881 | 0.000 | 0.027 | 0.028 |
| **Adult Mortality** | -0.0156 | 0.001 | -16.366 | 0.000 | -0.017 | -0.014 |
| **Alcohol** | -0.1346 | 0.034 | -3.905 | 0.000 | -0.202 | -0.067 |
| **Measles** | -1.666e-05 | 1.24e-05 | -1.340 | 0.180 | -4.1e-05 | 7.73e-06 |
| **BMI** | 0.0336 | 0.006 | 5.668 | 0.000 | 0.022 | 0.045 |
| **Total expenditure** | -0.0484 | 0.039 | -1.238 | 0.216 | -0.125 | 0.028 |
| **HIV/AIDS** | -0.3787 | 0.021 | -17.712 | 0.000 | -0.421 | -0.337 |
| **GDP** | 6.442e-05 | 7.71e-06 | 8.352 | 0.000 | 4.93e-05 | 7.95e-05 |
| **Population** | 5.286e-10 | 1.6e-09 | 0.331 | 0.741 | -2.61e-09 | 3.66e-09 |
| **Schooling** | 0.8832 | 0.039 | 22.775 | 0.000 | 0.807 | 0.959 |
| **pct\_infantDeaths** | -0.0147 | 0.009 | -1.641 | 0.101 | -0.032 | 0.003 |
| **pct\_thinness** | -0.0192 | 0.031 | -0.615 | 0.538 | -0.080 | 0.042 |
| **country\_code** | 0.0005 | 0.000 | 1.494 | 0.135 | -0.000 | 0.001 |
| **Immunization** | 0.0517 | 0.005 | 10.542 | 0.000 | 0.042 | 0.061 |
| **Developing** | -2.2917 | 0.364 | -6.297 | 0.000 | -3.005 | -1.578 |
| **Americas** | 6.0098 | 0.347 | 17.308 | 0.000 | 5.329 | 6.691 |
| **Asia** | 4.2867 | 0.284 | 15.107 | 0.000 | 3.730 | 4.843 |
| **Europe** | 5.3165 | 0.429 | 12.388 | 0.000 | 4.475 | 6.158 |
| **Oceania** | 2.6717 | 0.459 | 5.824 | 0.000 | 1.772 | 3.571 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 40.841 | **Durbin-Watson:** | 2.025 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 85.201 |
| **Skew:** | -0.015 | **Prob(JB):** | 3.15e-19 |
| **Kurtosis:** | 4.018 | **Cond. No.** | 4.61e+08 |

In the following plot, you can see that the predicted vs. the actual values have a good linear relationship.

The residual errors and well as the quantile-quantile plot can be seen below. The residual error plot is normally distributed with equal variance and the normal quartile to quartile plot also looks normal and shows linear relationship.

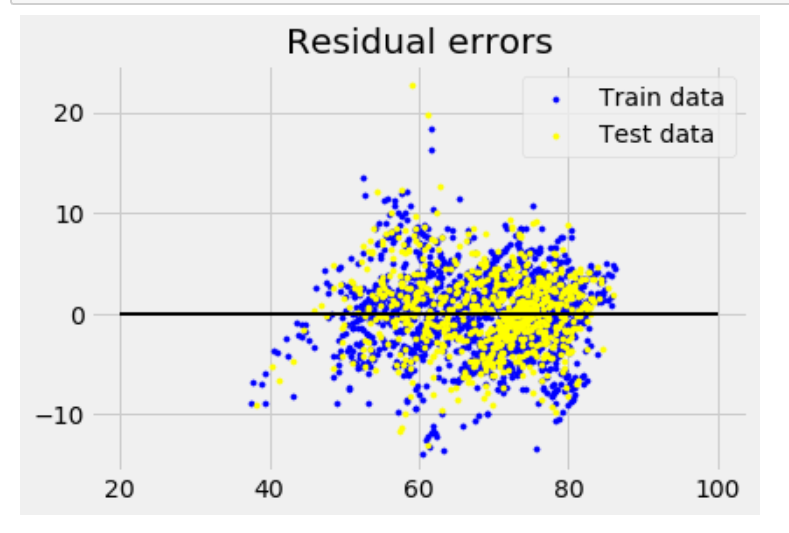
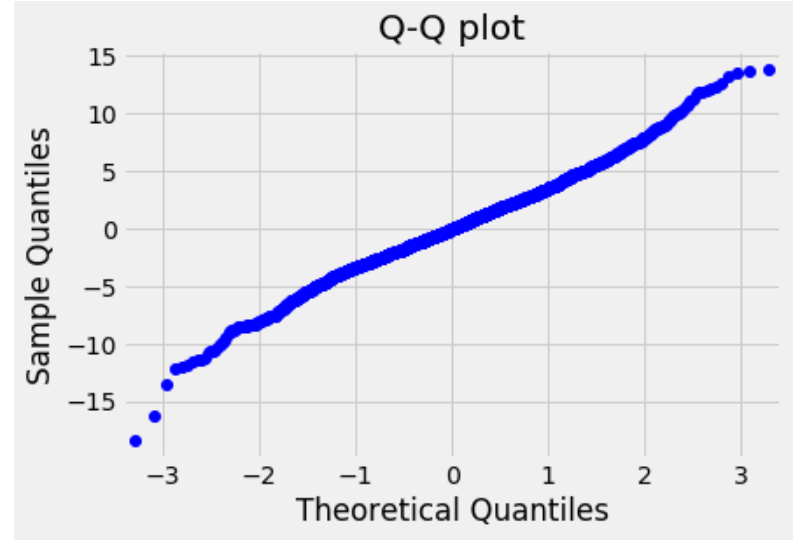
 

Figure: Residual error plot Figure: Quantile plot

The results reveal that the p value is less than the alpha Level. A common Alpha Level for test is 0.05, means we will reject the Null hypothesis. Next we will Study the individual p values to find out which of the individual variables are statistically significant and then we will run linear regression again on those variable. To check our data for multicollinearity, we import VIF (variance\_inflation\_factor), and after applying the code, we discover that we have removed all variables that have a VIF > 5.

Another algorithm we used in our model is **Random Forest Regression**. The results we received are quite good. R square close to 1, RMSE = 1.95, which is good and tells us that correlation between variables and Life Expectancy is excellent. MAE 1.18 degrees, which is very low, and the Accuracy score shows that the algorithm has learned 98% on the training data without cross validation and with cross validation of 88%, the value is 94 % on the test data, predicted and actual values have linear relationship with response variable. On the following Learning Curve Plot we can see that the prediction was made with a high percentage of accuracy. Two curves have converged shows that the performance error is stable and constant for both the training and testing sets.

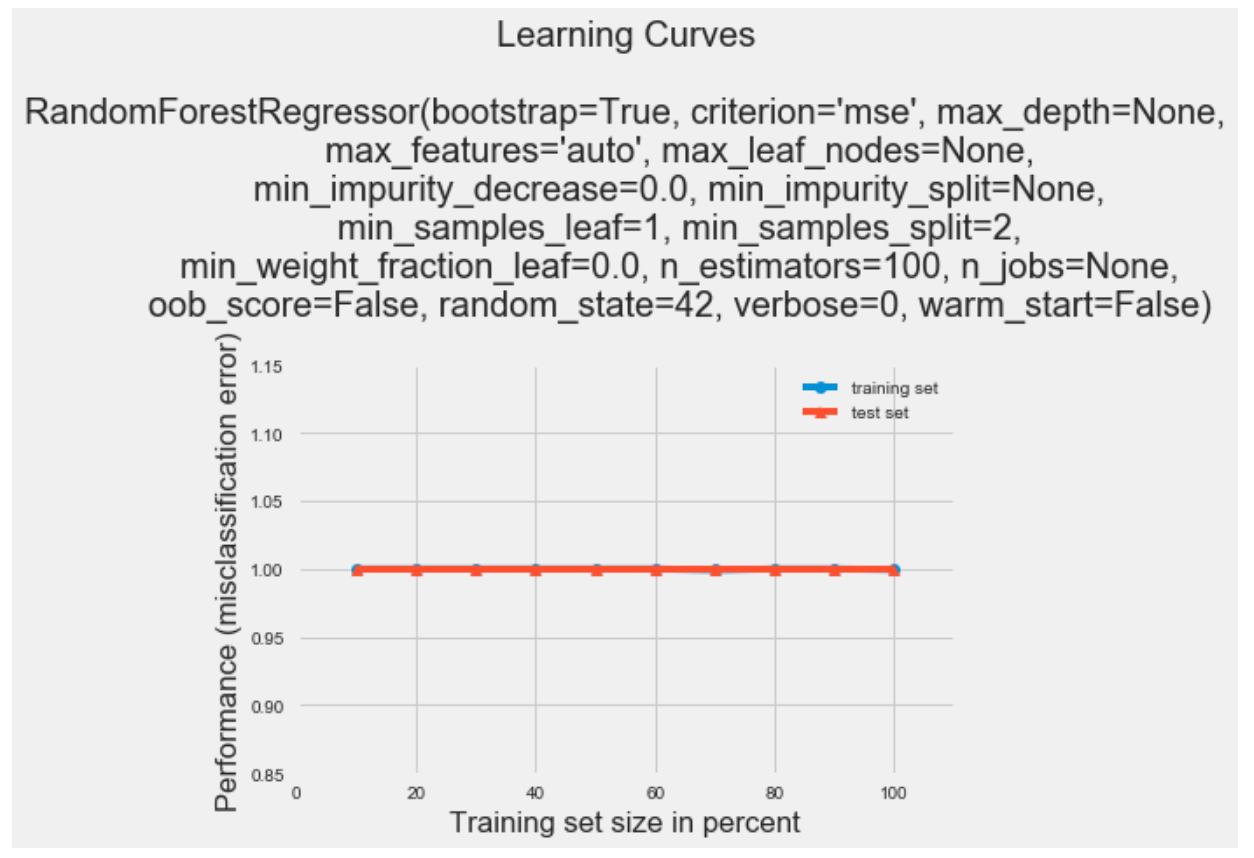


Figure: Learning Curve of training and test sets.

Important features mean the features that are more closely related with dependent variable and contribute more for variation of the dependent variable. On the following Plot, we can see the essential features from the random forest, and it is concluded that *HIV/AID*, *adult morality, schooling, BMI, pct\_thinness, pct\_infantDeaths, and schooling* have a strong influence and closely related to life expectancy as compared to the other features.

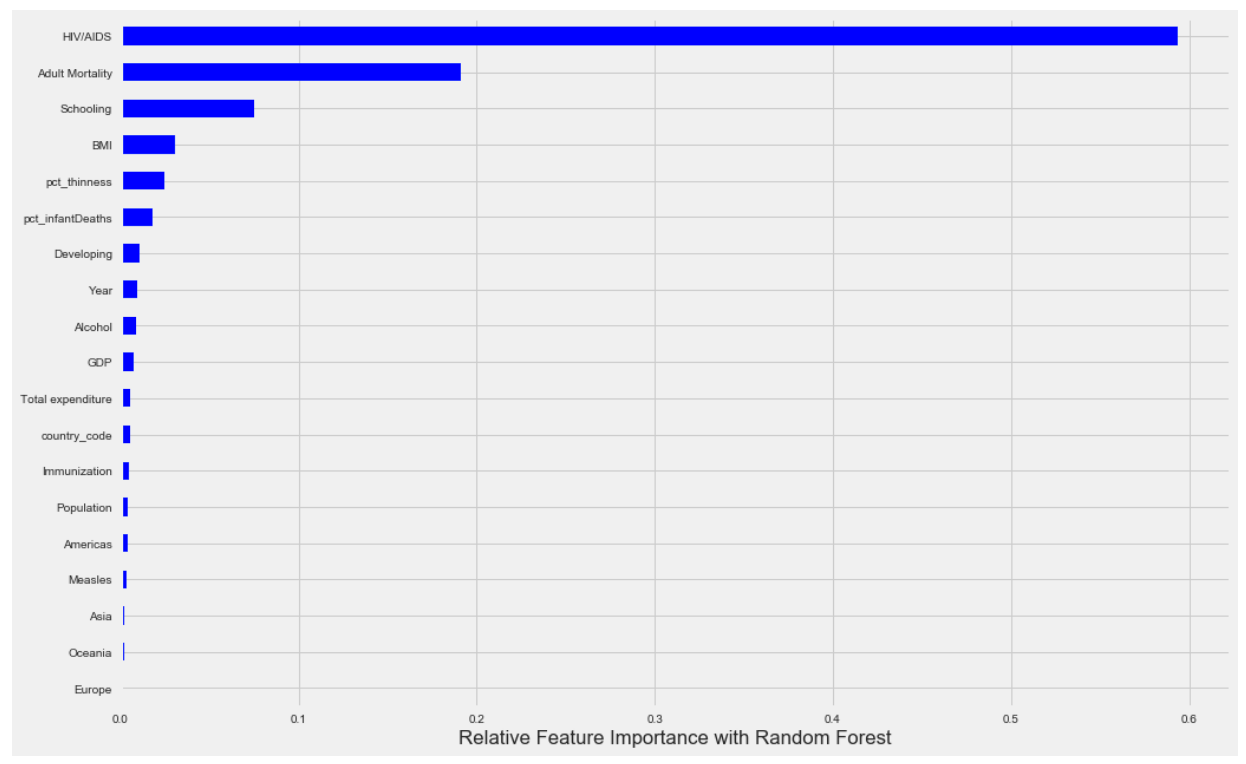
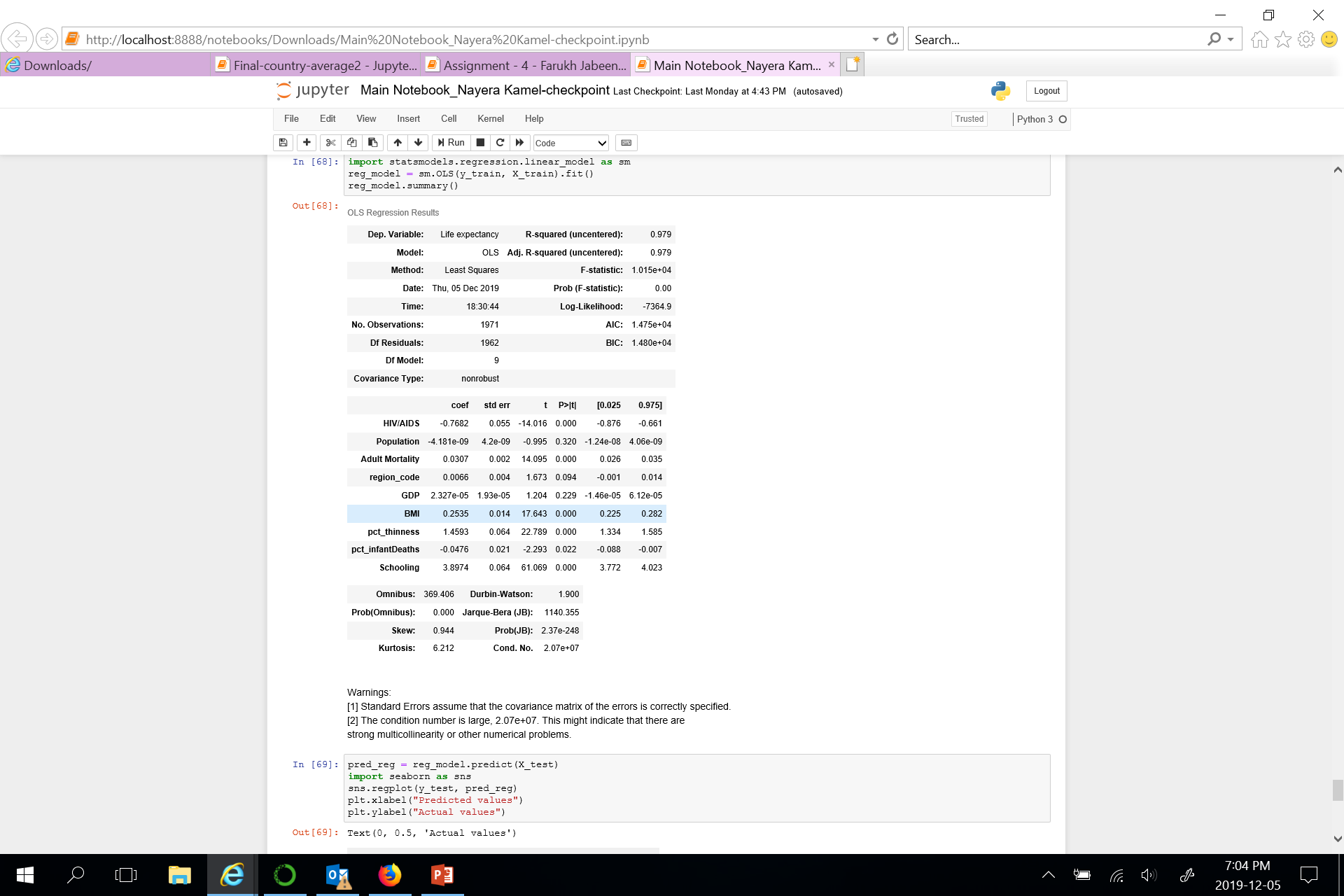


Figure: Feature importance with Random Forest

The final stage in our prediction model is to create the Regression Model with Important Features. Before applying the OLS we split the data into Train and Test with the ratio 70:30 percent, respectively. After using OLS on our data set, we get the following results:



The above OLS results shows that R square value equal to 0.97 is quite good as close to 1 and the condition number decreased, while, F-statistics: 1.015 with p-value 0.00, less than 0.05 alpha value and the JB (Jarque-Bera) test also shows that relationship between dependent and independent variable is linear.

The residual plots below also confirm the unbiased fit because the data points fall randomly around zero and follow a normal distribution. And the fitted line regression plot below suggests that this model fits the data.

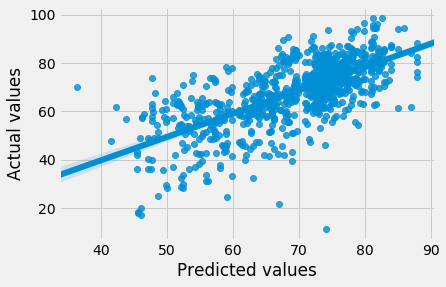
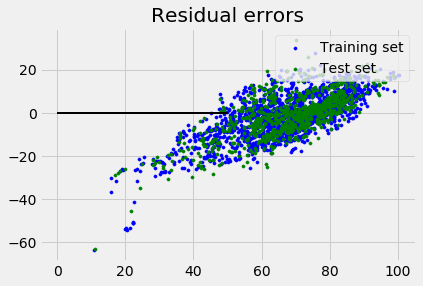


Figure: Regression Plot Figure: Residual Plot

# Conclusion

Life Expectancy data set had lots of missing values in the important features such as GDP, Population and Hepatitis B which were filled out from external data sources. The small percentage of missing values for such variable as Alcohol, BMI, Income composition of resources and others were treated by interpolation method and filled with mean values. During our analysis feature engineering needed to be done for some attributes such as thinness 1-19 year & thinness 5-9 years and under-five deaths & infant deaths. In order to have a better understanding of the data we added a new “continent” variable to the data set. It helped us to make a deeper analysis by continents with visualization plots.

After a deeper analysis we came to the conclusion that African continent has the lowest Life Expectancy rate for the following reasons:

• Kids thinness which is increasing the kid’s deaths

• HIV/AIDS rate which is increasing kid’s deaths

• Low level of immunization which leads to the early deaths

• Lowest GDP

• Lowest Income Composition of resources

After cleansing and exploring the data predictive model was built using 2 different algorithms:

• Linear Regression

• Random Forest Regression

In order to get better results some variables such as 'Income composition of resources', 'percentage expenditure', and 'region\_code' were removed as they were highly correlated with other variables. Multicollinearity test was done using VIF (variance\_inflation\_factor) code. Determination of important features was done using plot visualization. And the residual errors plot shows that the residual errors were randomly scattered. Two predictive models that were built showed quite good results. R squared was close to 1, skewness is low, kurtosis is a bit higher. Plots with the Actual vs Predicted values show that it was Linear relationship between two of them.

# Reference

## Appendix A – Data Cleansing – External Data Source Cleansing

Table 5 – External mapped to Primary data-source country names

|  |  |
| --- | --- |
| **Country Name**  **(External Source)** | **Country Name**  **(Primary Source)** |
| Bahamas, The | Bahamas |
| Bolivia | Bolivia (Plurinational State of) |
| Cote d'Ivoire | Côte d'Ivoire |
| Congo, Rep. | Congo |
| Congo, Dem. Rep. | Democratic Republic of the Congo |
| Czech Republic | Czechia |
| Korea, Dem. People’s Rep. | Democratic Peoples Republic of Korea |
| Egypt, Arab Rep. | Egypt |
| Gambia, The | Gambia |
| Iran, Islamic Rep. | Iran (Islamic Republic of) |
| Kyrgyz Republic | Kyrgyzstan |
| Lao PDR | Lao Peoples Democratic Republic |
| Micronesia, Fed. Sts. | Micronesia (Federated States of) |
| Korea, Rep. | Republic of Korea |
| Moldova | Republic of Moldova |
| St. Lucia | Saint Lucia |
| St. Vincent and the Grenadines | Saint Vincent and the Grenadines |
| Slovak Republic | Slovakia |
| North Macedonia | The former Yugoslav republic of Macedonia |
| United Kingdom | United Kingdom of Great Britain and Northern Ireland |
| Tanzania | United Republic of Tanzania |
| United States | United States of America |
| Venezuela, RB | Venezuela (Bolivarian Republic of) |
| Vietnam | Viet Nam |
| Yemen, Rep. | Yemen |

Table 6 – External data-source attribute and countries where nulls were filled with average

|  |  |
| --- | --- |
| **Attribute** | **Country** |
| **Hepatitis B** | Algeria, Angola, Antigua and Barbuda, Argentina, Australia, Azerbaijan, Bahamas, Bangladesh, Barbados, Benin, Bosnia and Herzegovina, Burkina Faso, Burundi, Côte d'Ivoire, Cabo Verde, Cambodia, Cameroon, Canada, Central African Republic, Chad, Chile, Comoros, Congo, Croatia, Czechia, Democratic People's Republic of Korea, Democratic Republic of the Congo, Djibouti, Equatorial Guinea, Eritrea, Estonia, Ethiopia, Gabon, Ghana, Grenada, Guatemala, Guinea, Guinea-Bissau, Guyana, Haiti, India, Ireland, Jamaica, Kenya, Lao People's Democratic Republic, Lesotho, Liberia, Madagascar, Malawi, Mali, Malta, Mauritania, Montenegro, Mozambique, Myanmar, Namibia, Nepal, Netherlands, Niger, Nigeria, Pakistan, Panama, Paraguay, Peru, Russian Federation, Rwanda, Saint Lucia, Saint Vincent and the Grenadines, Sao Tome and Principe, Senegal, Serbia, Sierra Leone, Somalia, South Sudan, Sri Lanka, Sudan, Suriname, Sweden, Switzerland, Tajikistan, The former Yugoslav republic of Macedonia, Timor-Leste, Togo, Trinidad and Tobago, Turkmenistan, Uganda, United Republic of Tanzania, Uzbekistan, Viet Nam, Zambia |
| **GDP** | Eritrea, Iraq, Sao Tome and Principe, Somalia, South Sudan, Syrian Arab Republic, Venezuela (Bolivarian Republic of) |
| **Population** | Eritrea |

1. https://www.cia.gov/library/publications/the-world-factbook/geos/kn.html [↑](#footnote-ref-2)