Life Expectancy

Scott Schirkofsky

MSDS Data Science Practicum II

Regis University

# Abstract

This project seeks to better understand the various factors that affect life expectancy and their relationship with each other and life expectancy itself. The project will also use various machine learning techniques to predict various life expectancies by country and other indicators and provide the metrics needed to evaluate the predictions.

Keywords: python, seaborn, numpy, plotly, pandas, matplotlib, scikit-learning, multiple linear regression, polynomial regression, decision tree, random forest, regression, logistic regression, machine learning

Contents

[Abstract 2](#_Toc70672903)

[Getting Started!! 4](#_Toc70672904)

[About the Data 4](#_Toc70672905)

[Approach 4](#_Toc70672906)

[Exploratory Data Analysis 4](#_Toc70672907)

[Imports and Loading Dataset 4](#_Toc70672908)

[Data Cleaning 6](#_Toc70672909)

[Dataset 7](#_Toc70672910)

[The features 7](#_Toc70672911)

[Column Headings 7](#_Toc70672912)

[Missing Values 8](#_Toc70672913)

[Inexplicit Nulls 9](#_Toc70672914)

[Explicit Nulls 11](#_Toc70672915)

[Outliers 13](#_Toc70672916)

[Data Exploration 18](#_Toc70672917)

[Univariate Analysis 19](#_Toc70672918)

[Visual Distributions 19](#_Toc70672919)

[Continuous Analysis 24](#_Toc70672920)

[More on Life Expectancy Features 29](#_Toc70672921)

[Modeling & Predictions (Machine Learning) 31](#_Toc70672922)

[Linear Regression 31](#_Toc70672923)

[Multiple Linear Regression 32](#_Toc70672924)

[Polynomial Regression 32](#_Toc70672925)

[Decision Tree Regression 33](#_Toc70672926)

[Random Forest Regression 34](#_Toc70672927)

[T-test 34](#_Toc70672928)

[Prediction/ Tune Model 34](#_Toc70672929)

[ROC Curves 36](#_Toc70672930)

[Summary 36](#_Toc70672931)

[References 37](#_Toc70672932)

[Coding and other valuable resources 37](#_Toc70672933)

Life Expectancy

# This PROJECT

Life expectancy is also regarded as the key metric for assessing population’s health and the Global Health Observatory indicates that life expectancy globally increased from 66.8 years in 2000 to 73.4 in 2019. Life expectancy is measured by the number of years a person is expected to live. Improving life expectancy for individuals at birth depends on identifying and improving factors that negatively affect life expectancy while stressing factors that improve life expectancy.

There are many factors that can affect a person’s life expectancy including a person’s sex, health factors, and other demographic data.

This project seeks to better understand the various factors that affect life expectancy and their relationship with each other and life expectancy itself. The project will also use various machine learning techniques to predict various life expectancy by country and other indicators and provide the metrics needed to evaluate the predictions.

# About the Data

The dataset being used in this project is comprised of data from all over the world from various countries aggregated by the World Health Organization (WHO).

The data contains 2938 rows and 22 columns and was collected from 193 countries over a period of 15 years from 2000 to 2015.

# Approach

This project will use Python and a variety of its libraries to explore and analyze the Life Expectancy dataset from the WHO.

* Data Cleaning & Exploration: Python, Pandas, NumPy, SciPy
* Visualization: Seaborn, Matplotlib, Dataframe, Plotly, and Tableau
* Modeling: Sklearn (Mean Squared Error, r2 Score, Confusion Matrix, Accuracy Score, Classification Report, Roc Curve)
* Deployment: Results will be published to the resources below:
  + Code & Documents - <https://github.com/schirko/Life_Expectancy-WHO_Data>
  + Blog - <https://sschirko.edublogs.org/>

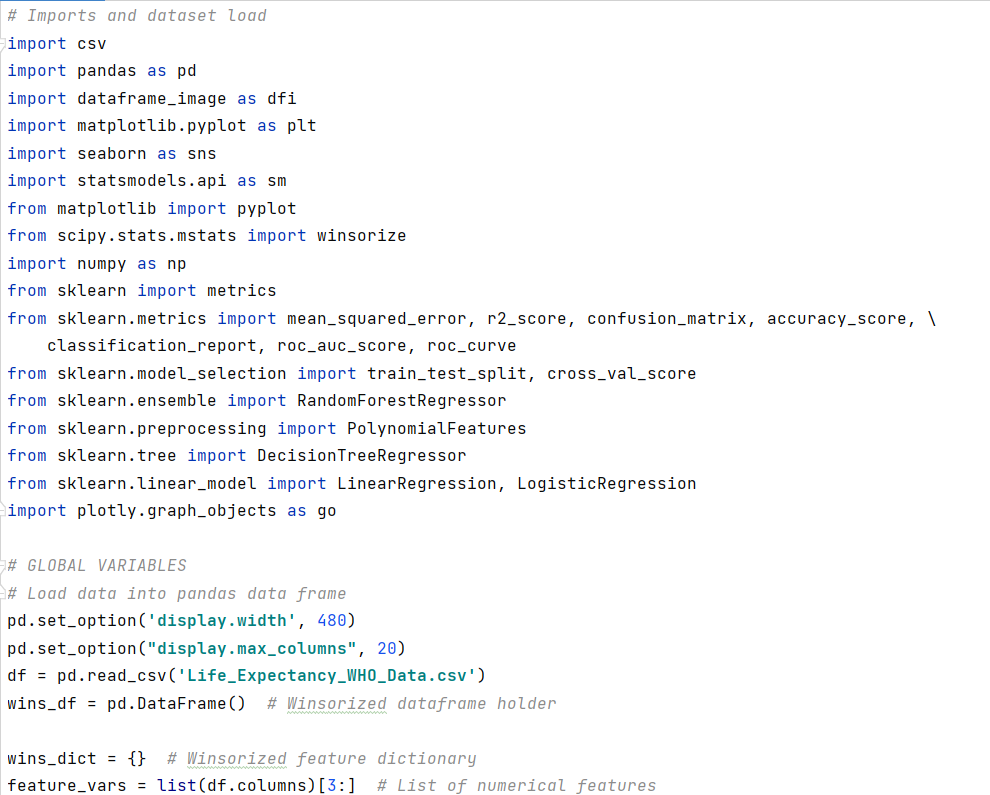
# GETTING STARTED!!

Alrighty then, let’s get started digging into the data. For this project I’m using PyCharm as my IDE along with various Python libraries.

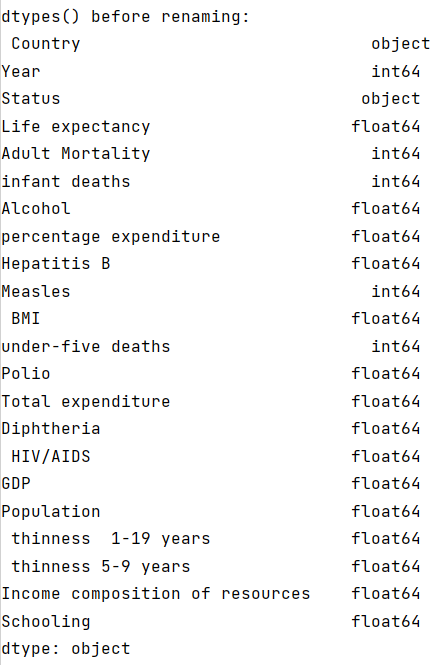
# Exploratory Data Analysis

## Imports and Loading Dataset

The image below show all the imports and setting of global variables for this project.

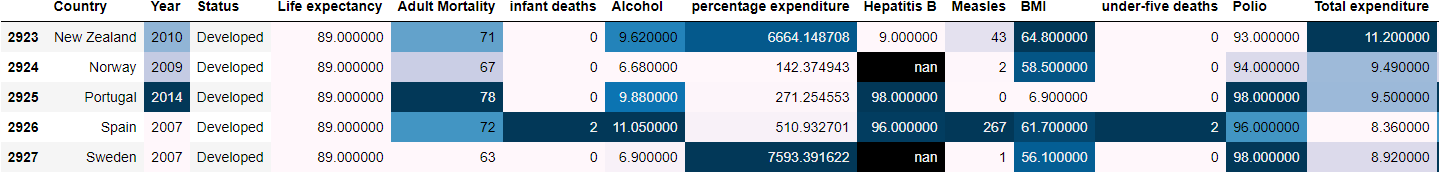


I used shape() to determine that the dataset has 2928 rows and 22 columns to begin with, and two of those columns (Country & Status) are categorical columns. Country is self-explanatory while status indicates the level or development a country has – Developed or Developing. Dtypes() used below provides us with a better view of the data and it’s messy column names. No worries, we’ll clean them up!



Head() and tail() give us a glimpse of the dataset and the features that it includes.

head()



tail()



With the data now loaded into a Pandas data frame we can begin the cleaning of the data.

## Data Cleaning

My primary goal in this project is to understand the factors that affect the life expectancy. So, the target variable will usually be life expectancy. Before I can explore the data, it must first be cleaned by detecting and removing null-values and treating outliers. Then I can move on to data exploration and analysis and begin to understand the data’s features.

A few things I will need to know about the dataset and it’s features in order to get it properly cleaned up. Each of these questions will be answered in the course of this project.

* What are the outliers for features?
* Are there missing values?
* What is the meaning of the variable?

### Dataset

This dataset was obtained from Kaggle challenges via the World Health Organization. It is comprised of data from around the world and includes many indicators for each country during the time frame of 2000-2015. The data essentially represents a times series for the countries and features included.

In general, I will be using life expectancy as the dependent variable and the other features as independent variables.

### The features



### Column Headings

As we saw from the initial quick look of the data, the string values for the column headings are fairly messy so we need to address this first. In the code below I replace spaces with underscores and shortened some of the text.



The new headings look much better and will make coding this project much easier and more readable!

|  |  |
| --- | --- |
| **Before** Renaming Columns | **After** Renaming Columns |

With the column names in order, let’s dig into missing values for the feature variables.

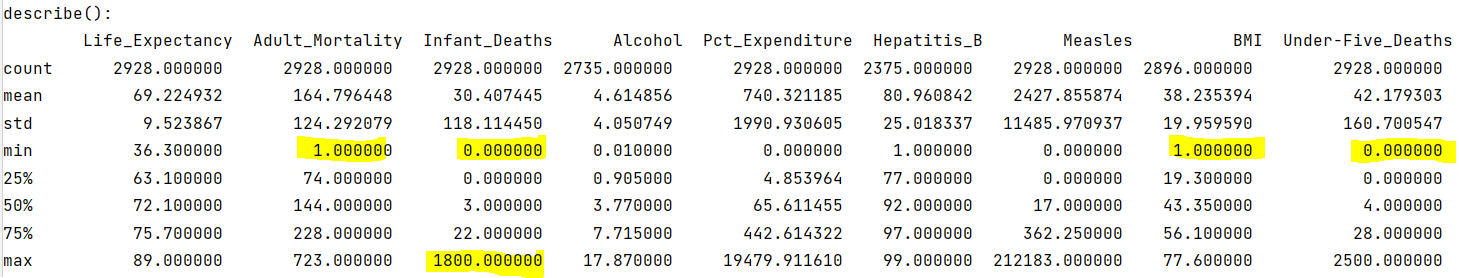
### Missing Values

Missing values can exist when we have nulls, but there can also be erroneous missing values when the values appear to be inexplicit. Inexplicit nulls may require a more discretion when deciding how to deal with them as they may still represent actual data.

In the section below I will address missing values by finding nulls and dealing with them through removing, imputing or interpolating those that are concerning.

### Inexplicit Nulls

On way to look for values that stand out is to run describe() to generate descriptive statistics. This will provide us a summary which is much quicker than having to manually look at all the data.



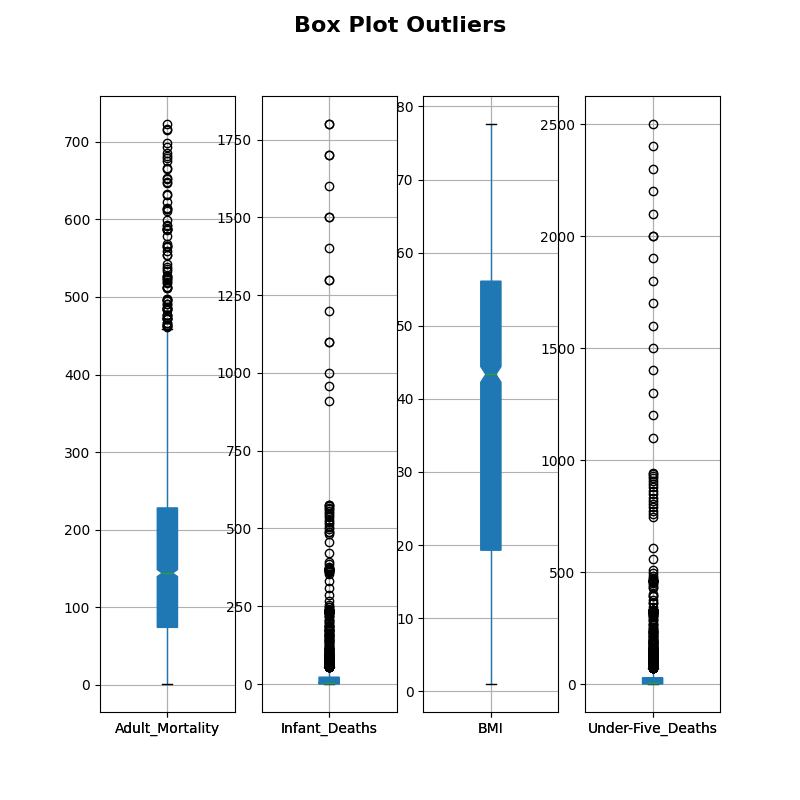
I’ve highlighted a few items that stand out to me as not being probable.

* Adult mortality has a 1 value for min.
* Infant Deaths has a min of 0 and max of 1800.
* BMI has min of 1.
* Under-Five\_Deaths has min of 0.

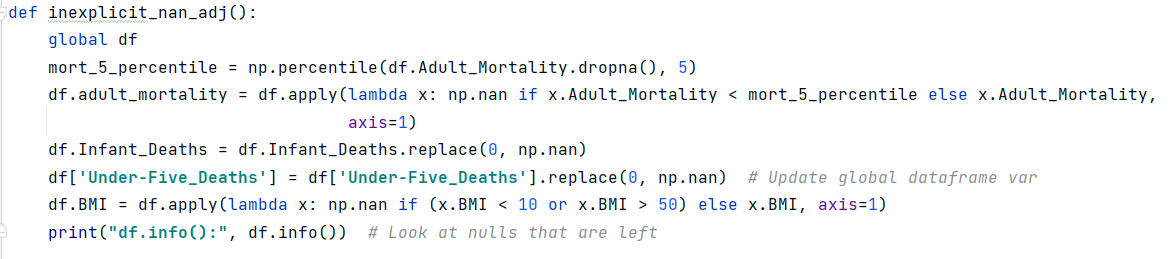
Let’s take a look at the features in question with boxplots.



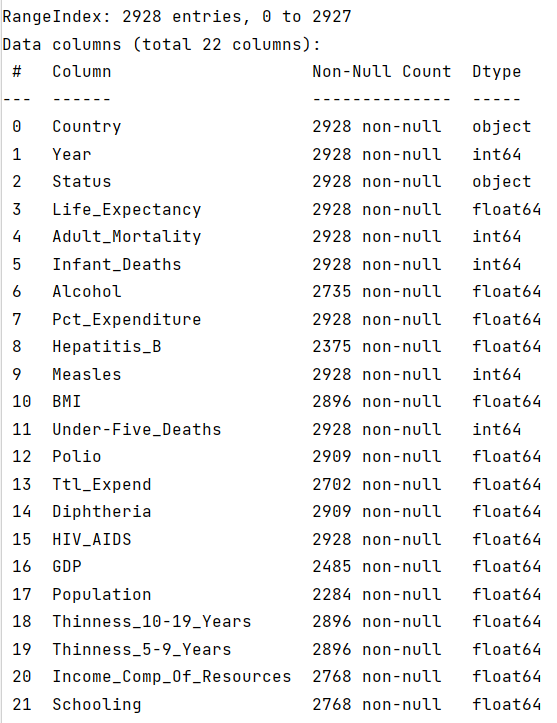
It appears from the plots that there are some outliers with Adult Mortality and BMI which we can deal with later, but there also seem to be errors resulting in nulls with infant deaths and Under Five Deaths.



Let’s deal with the outliers in BMI and Adult Mortality later with explicit nulls and address the nulls with Infant Deaths and Under Five Deaths.

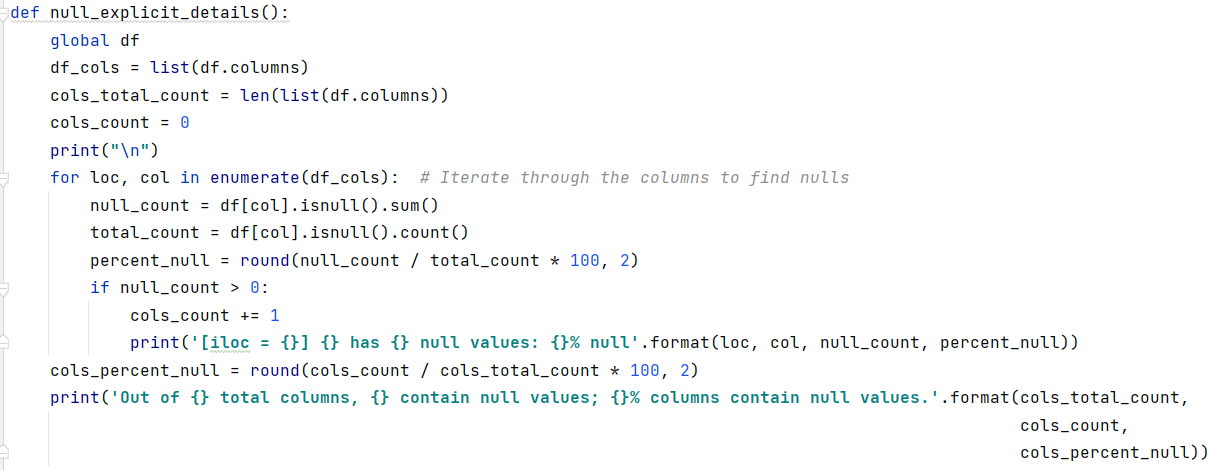


Info() below will help take a look at the results. At this point, the remaining missing values in the data should be attributable to explicit nulls.

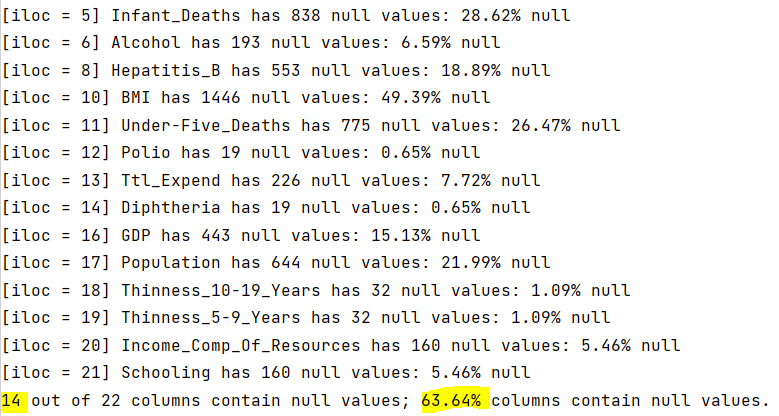


### Explicit Nulls

The data still has a significant amount null values so it will be necessary to look at the features in more detail to better understand where they are and their significance. So, let’s detail the nulls.



From the details below we can see that there are still 14 columns that contain nulls, that’s 63.64%! The details show us that we still have big contributing factors towards the remaining null values in the features BMI, Population, GDP, and Infant Deaths. We’ll deal with those as explicit nulls.



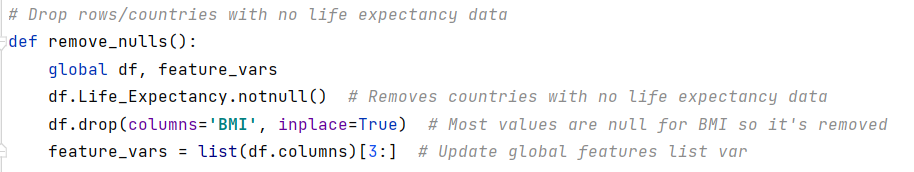
#### Dealing with the explicit nulls

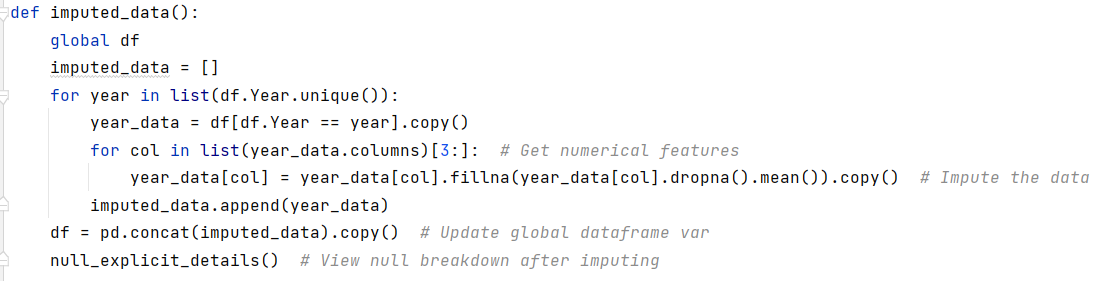
Because I am specifically dealing with life expectancy, there is no need to include data for any country that is missing the life expectancy values. Those countries will be removed.

Countries with no life expectancy data:

* Cook Islands, Dominica, Marshall Islands, Monaco, Nauru, Niue, Palau, Saint Kitts and Nevis, San Marino, Tuvalu

It also looks like the BMI feature has half null values, so I do not see the value in keeping this feature. It will be removed.



Because this data is a times series which can be regarded as a sequence of values, imputation can be used to estimate new values for the missing values some of these features contain, specifically the big null value offenders.   
  


If we take another look at the null details, we see that we are null free. Join me in a sigh of relief. Kind of feels like cleaning out a closet, doesn’t it?!



We’ve cleaned up the column heads, missing values, dealt with implicit and explicit nulls - outliers are next.

### Outliers

In this section I will look into outliers by detecting them, visualizing them, and Winsorizing them. Wins-a-what? Winsoring the data will help minimize the influence of the outliers in the data and we know from the boxplots that at very least the adult mortality and BMI features have outliers.

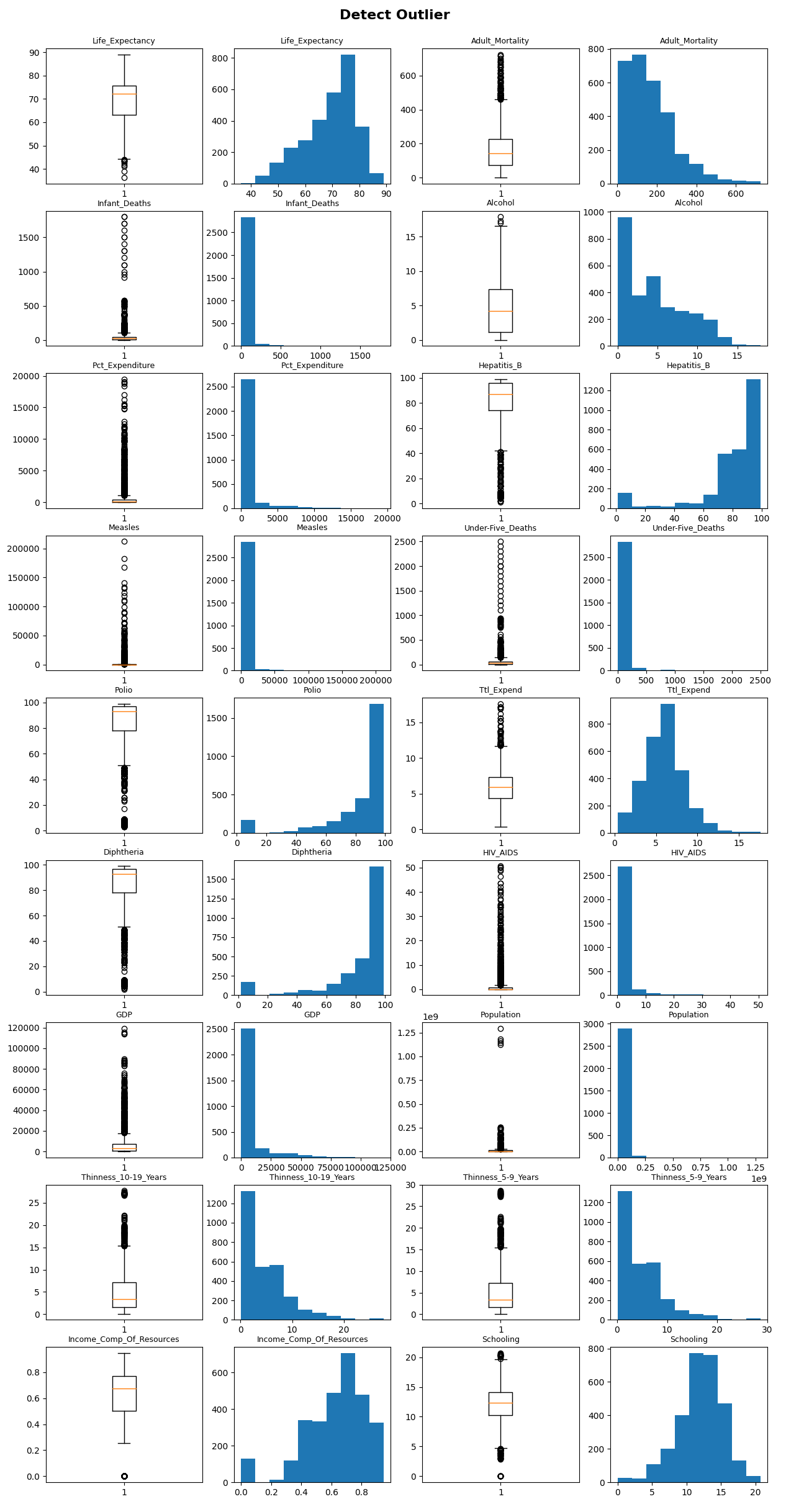
Let’s go outlier hunting for outliers!

#### Outlier Detection

I think the best first step here is to expand on the original box plots done for BMI and Adult Mortality by plotting histograms and boxplots for each feature.



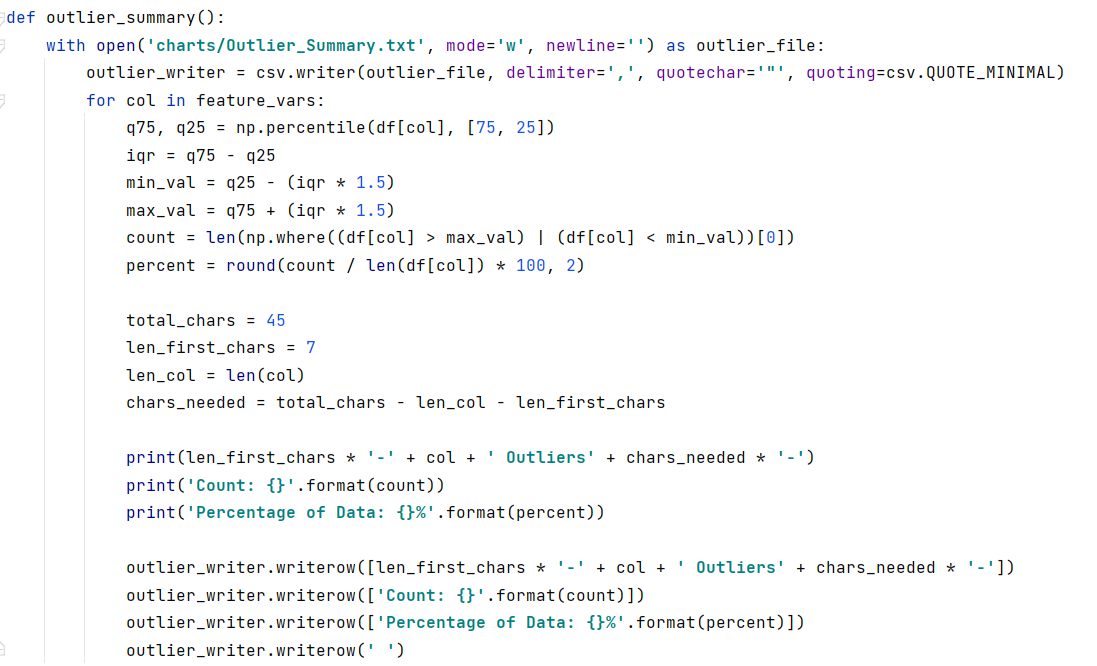
The box plots below are one big image making it a bit lengthy, but I think its value makes it worth it the look.



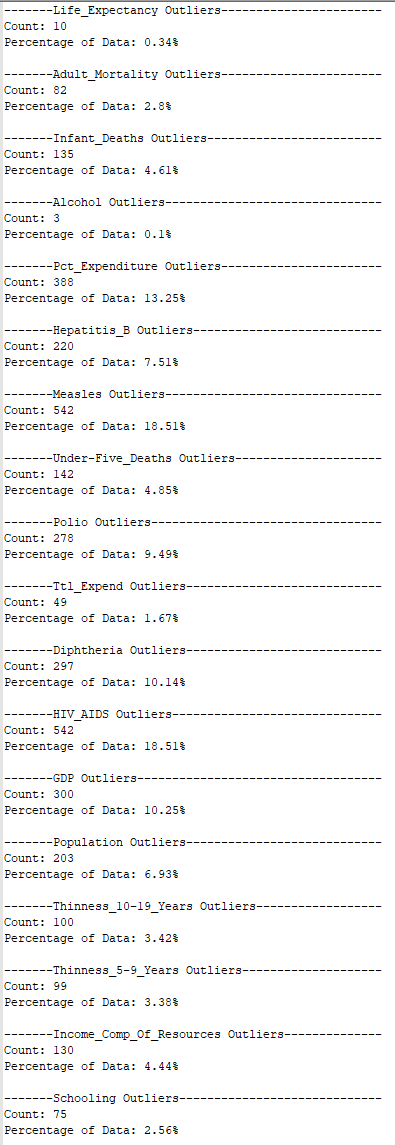
That’s a lot to unpack! We can see that outliers are not uncommon and that’s a problem. The biggest concern to me is that life expectancy has an outlier issue, especially given that this is our target feature.

But let’s stay focused.

I would like to know the outlier details by feature so I will start there and apply Tukey’s rule which is a statistical method for identify what constitutes an outlier.



The outlier summary shows a decent number of outliers that I will need to address.



##### Working with the Outliers

Because there are a number of features with outlier issues and each one has its own unique issues, the method I will use to address outliners is to Winsorize the data. This method will help to minimize that influence that outliers have on each variable. Winsorizing the data will enable me to set upper and lower limits on each feature. This limits the outlier’s severity and influence on the data.

The parameters for winsorizing will be stored in a dictionary for easy access later in the project.

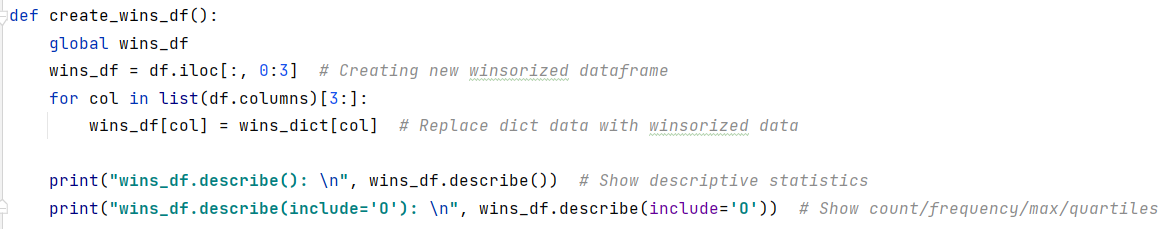


The boxplots below are illustrations of successful winsorization for life expectancy and GDP. These two are just shown as examples of the winsorization results for all numerical features.

|  |  |
| --- | --- |
|  |  |

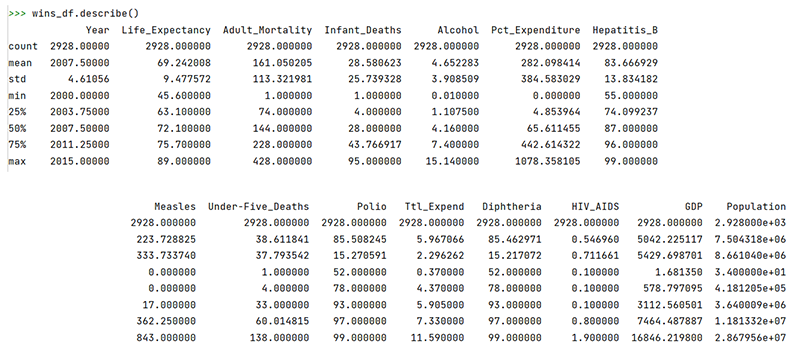
## Data Exploration

Not that we are into the exploration phase, a new dataframe will be created from the winsorized data.

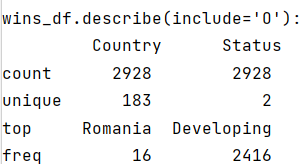


### Univariate Analysis

A univariate analysis is about looking into variables one at a time. To do this I will first use describe() to show descriptive statistics, histograms for our continuous features and barplots for the categorical features.



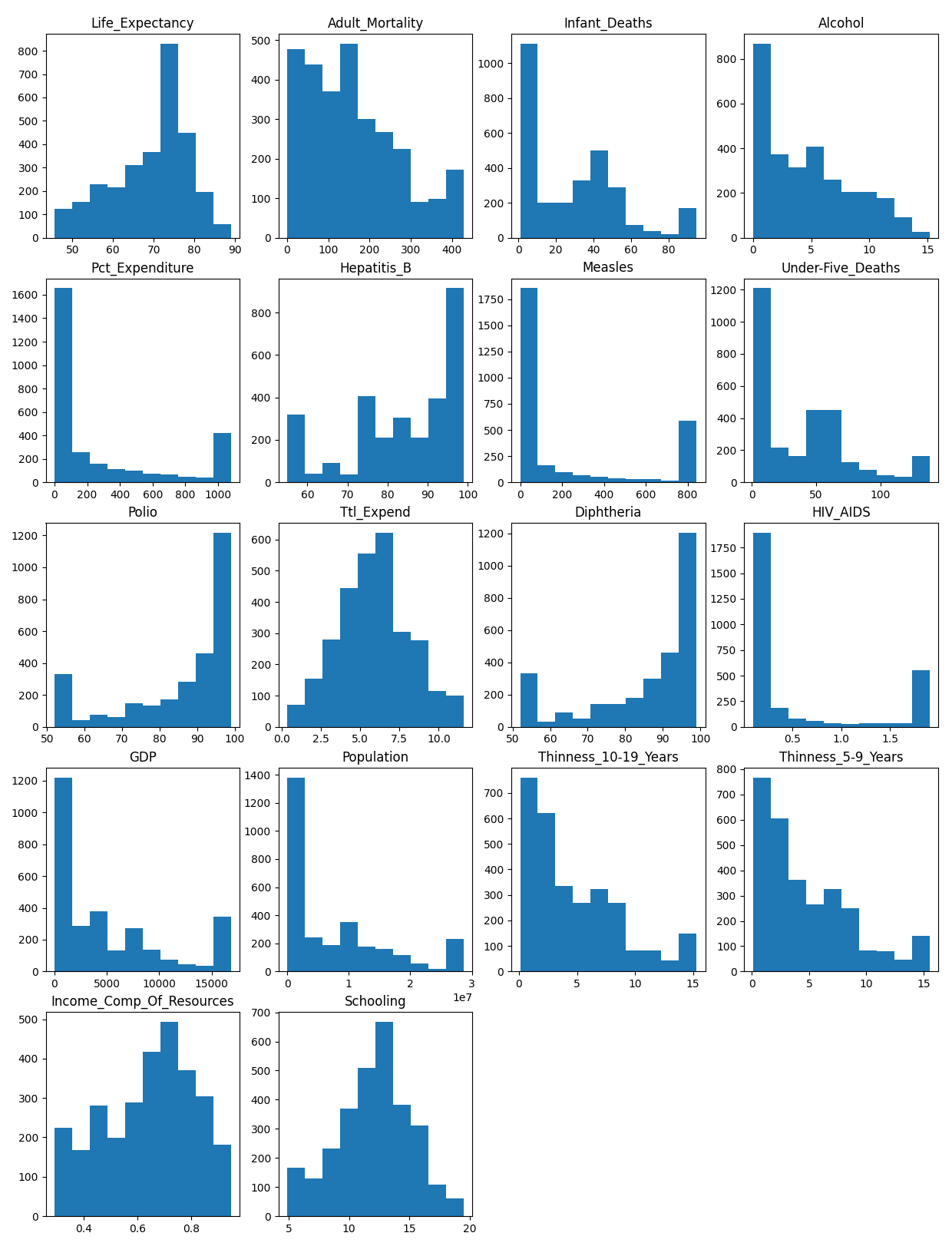
Here I run another describe() but with a parameter to show count, unique, top, and frequency.



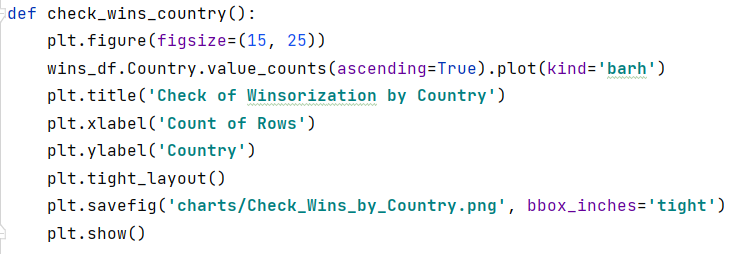
### Visual Distributions

Next, histograms are used for our continuous features.





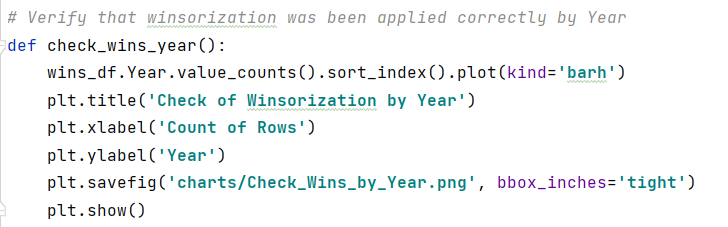
It appears as if the winsorization has had a big impact on many features, and not so much with other features. Overall, the data looks much tighter for each feature.

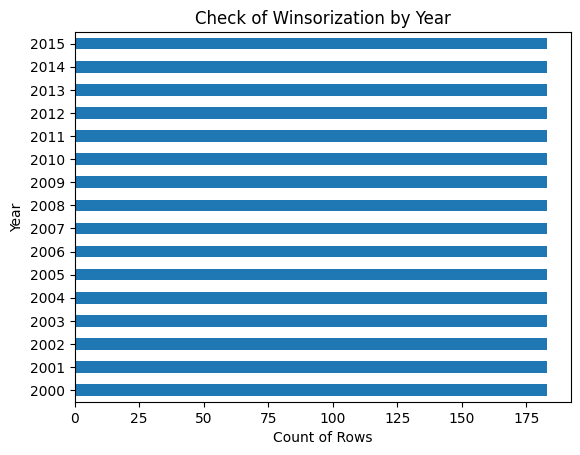


This chart is not visually stimulating; however, it shows that all of the remaining countries have 16 rows worth of data. Good to see that all the countries are being represented in the data!

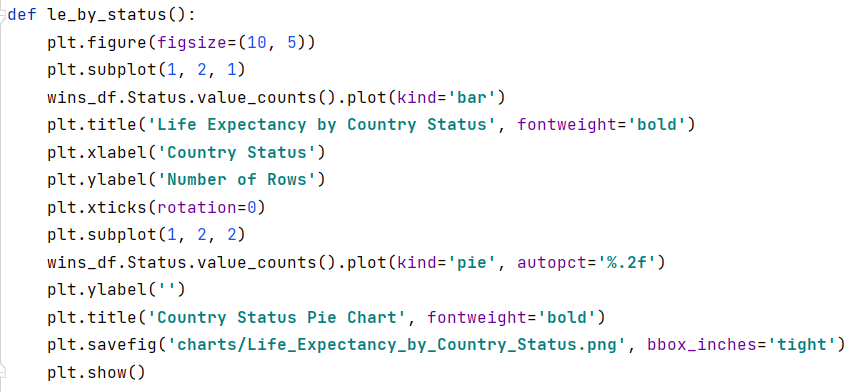


This is another plot that doesn’t dazzle the eyes but show us that each year has the same number of rows of data. Another score!

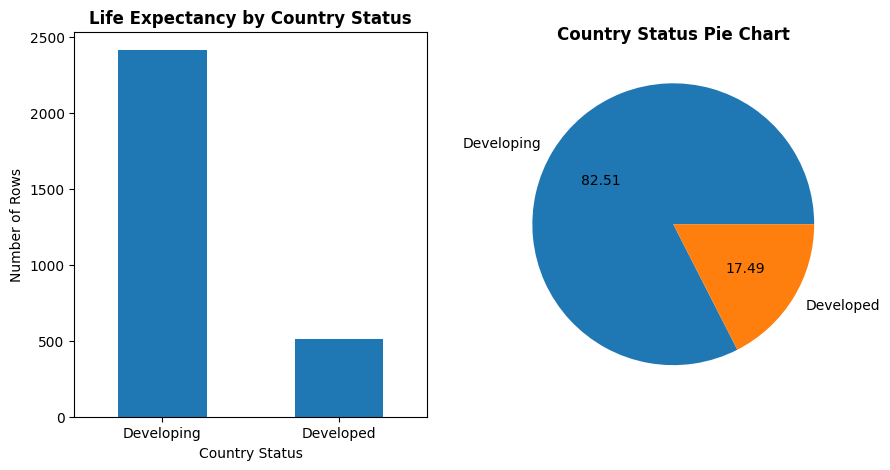




Now let’s look at the volume of data from the standpoint of a developing or developed country status.



What’s significant about these next two charts is the amount of data that comes from the developing countries versus developed countries. The pie chart shows that 82.51% of all the data comes from countries which have the developing Status!

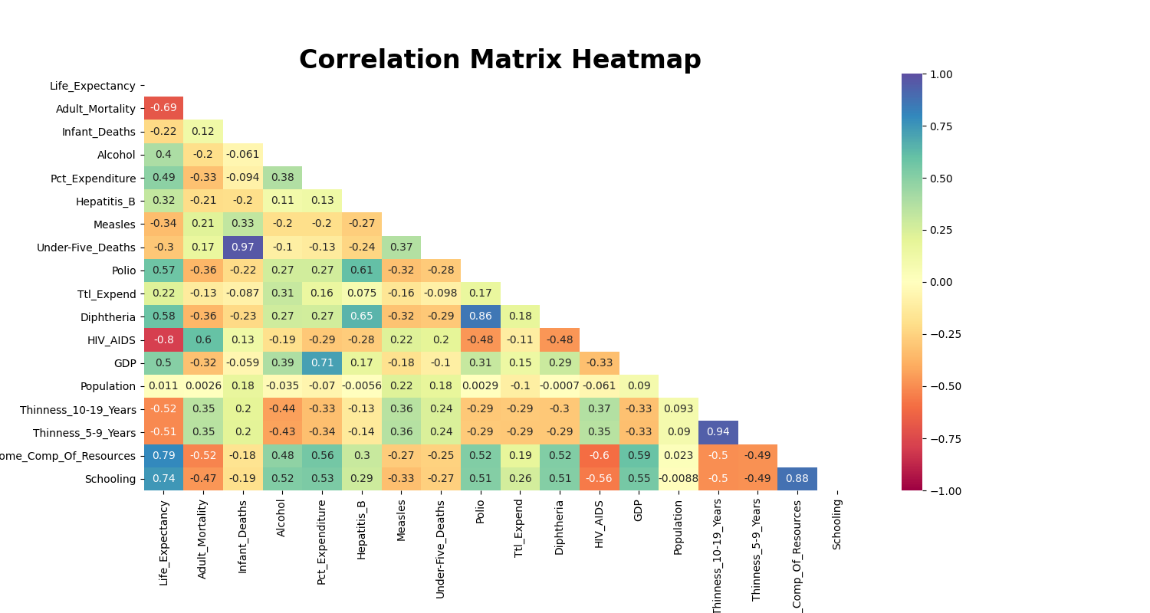


### Continuous Analysis

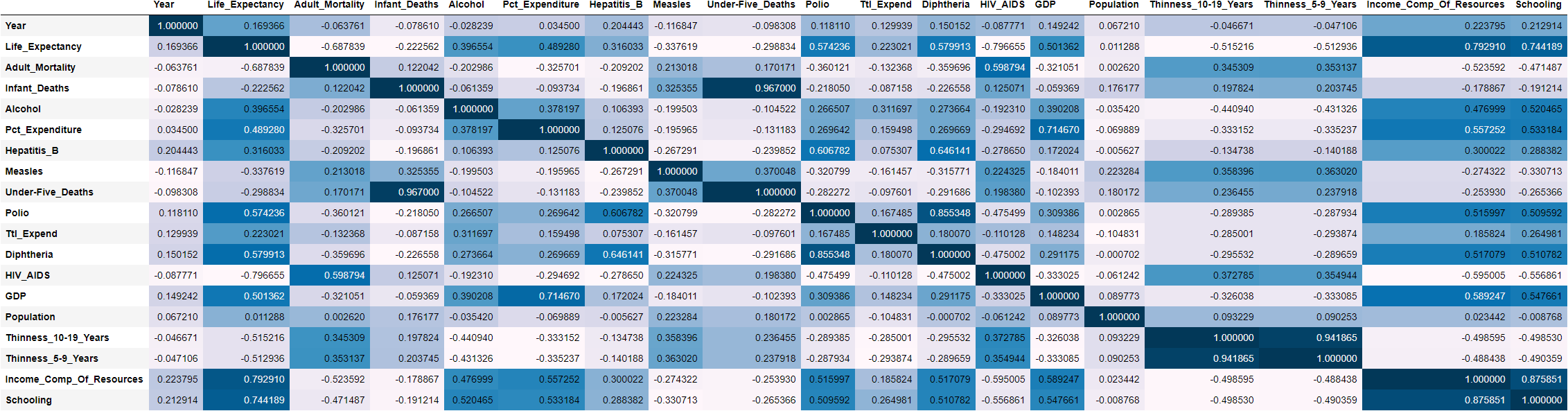
In this section I look at the continuous features compared to each other as well as their relationship to the life expectancy target feature.



The two heatmaps below are quite beneficial as they layout the correlations in a manner that is very easy to read. Same data is used in both heatmaps, just a different way to illustrate the correlations.



**Correlation Matrix Heatmap**

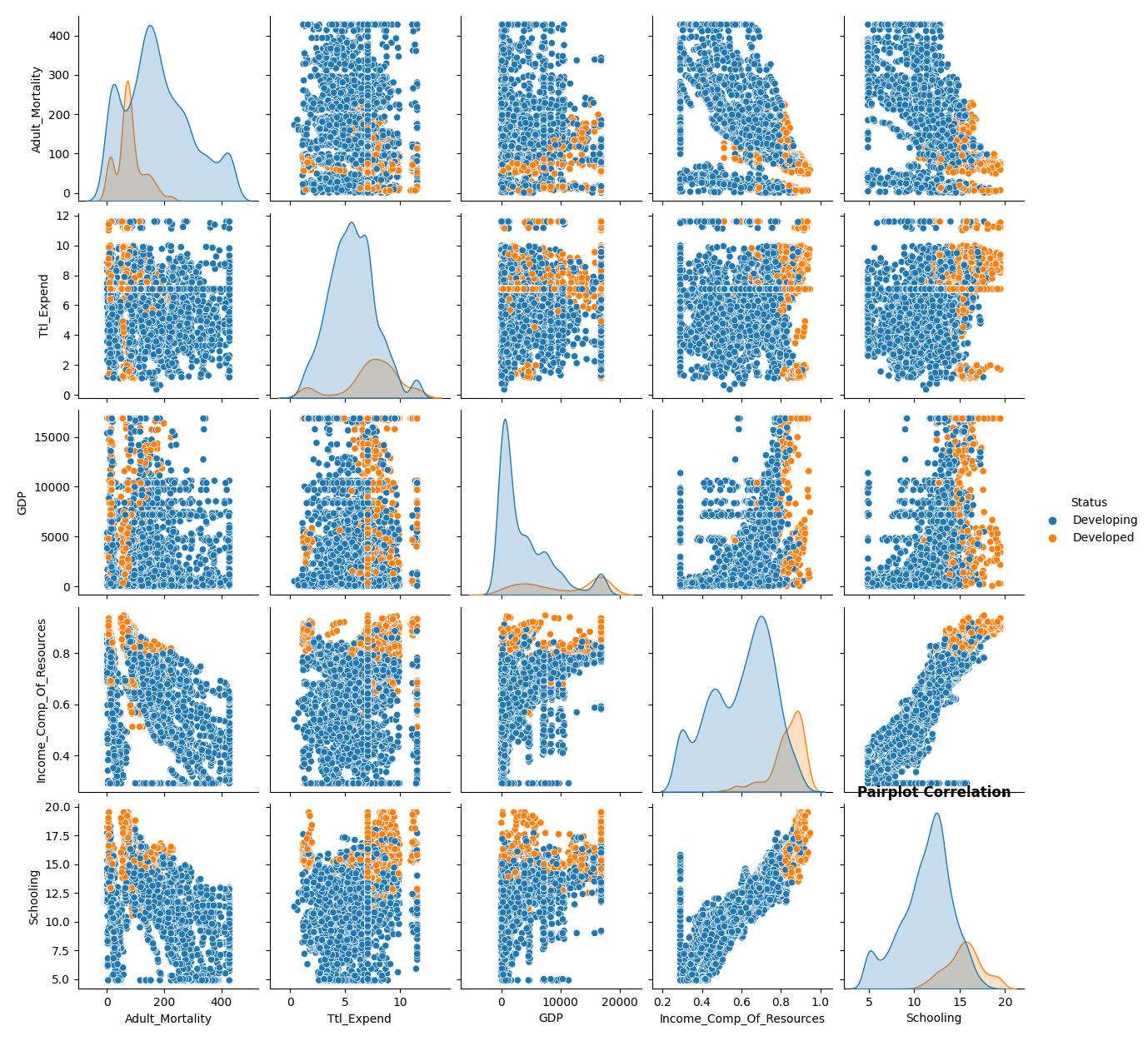


A number of observations can be made including the somewhat high correlation between life expectancy, the target variable, and Adult Mortality, HIV/AIDS, income of composition resources, and schooling.

Here are a few other observations:

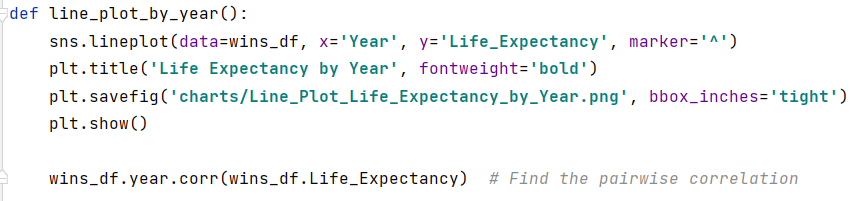
* Infant deaths and Under Five deaths are extremely highly correlated
* GDP and Percentage Expenditure are fairly highly correlated
* Diphtheria and Polio vaccine rate are highly positively correlated
* Hepatitis B vaccine rate is positively correlated with Polio
* Hepatitis B vaccine rate is positively correlated with and Diphtheria vaccine rates
* Life Expectancy has almost no correlation to Population.
* Schooling and Income Composition of Resources are highly correlated

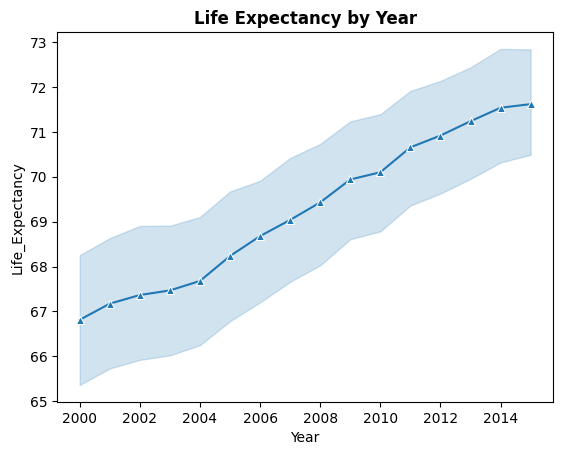
For additional perspectives on correlation the chart below contains a matrix of scatter plots showing correlation among the numerical features.

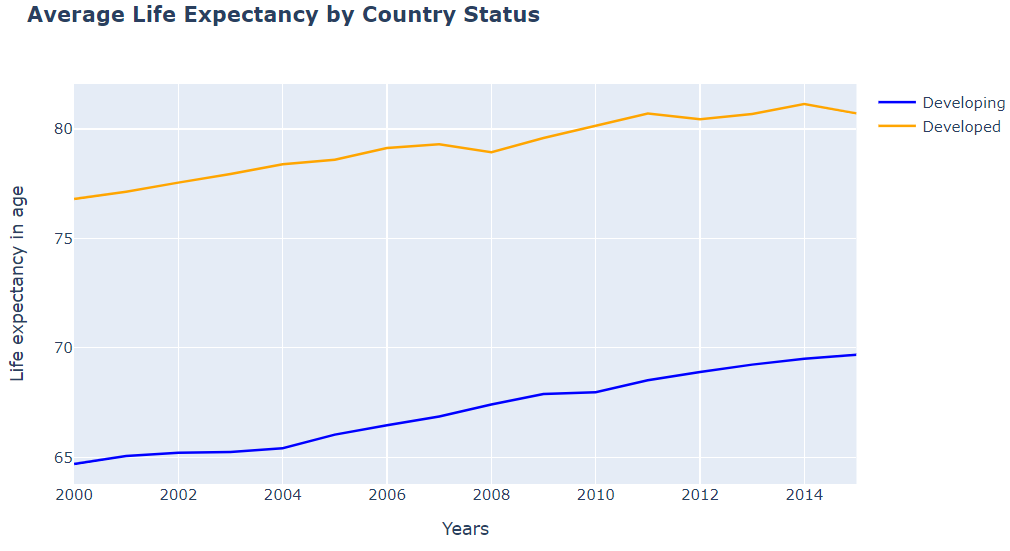


#### Categorical Data

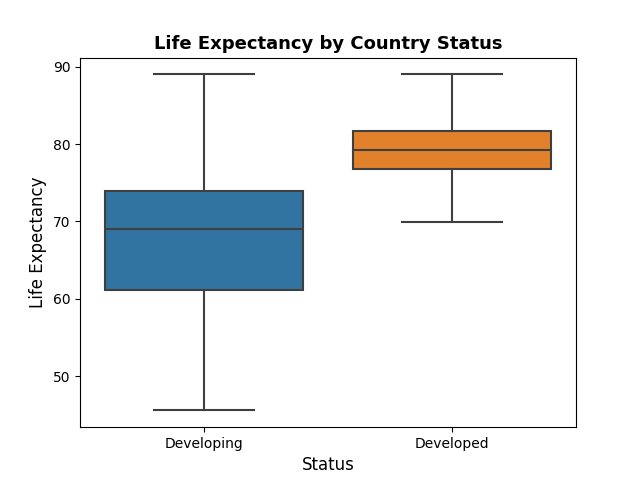
How has life expectancy changed over the years? Let’s look at a line plot showing a time series of life expectancy from 2000 to 2015.



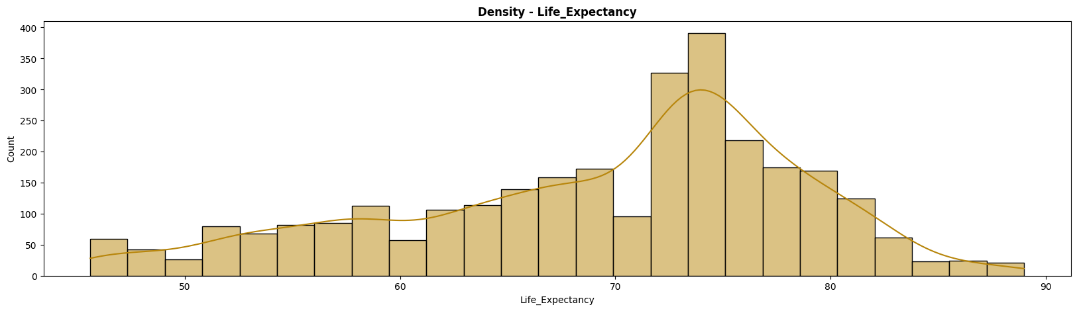




There’s definitely a positive trend with life expectancy over the 15 years of data provided. The next chart below are reminder that the majority of our data comes from developing countries, yet the lowest life expectancy comes from the same developing countries.

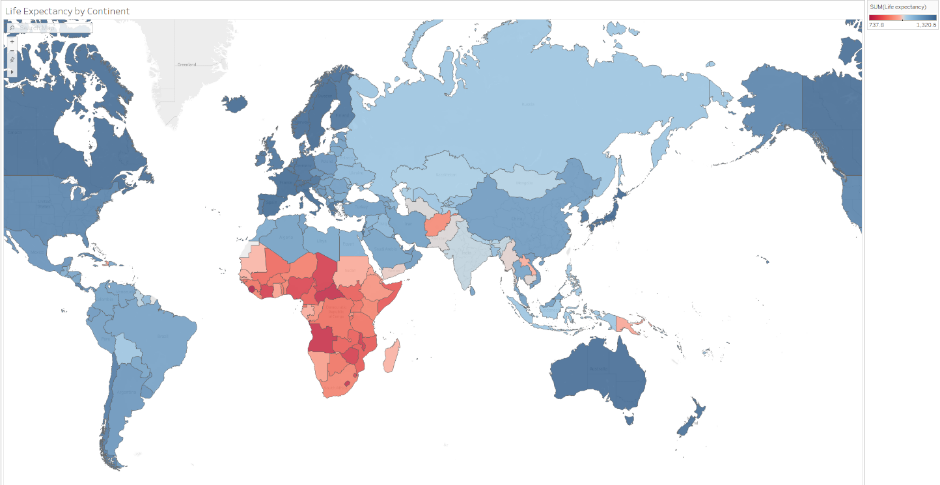


For additional information to think about I’ve done a density plot of life expectancy which shows that highest density in the lower to mid-70s.



We know where the correlation exist, how about any geographical correlation?

We see that the majority of “correlation” comes from less developed countries, how about seeing that on map so we can better understand where these developing countries are located. I imported the dataset into Tableau because of Tableau’s ability to associate each country with a continent and then provide a visually pleasing and informative chart. With a few exceptions we can see that the majority of the countries with the status of developing are located in Africa.

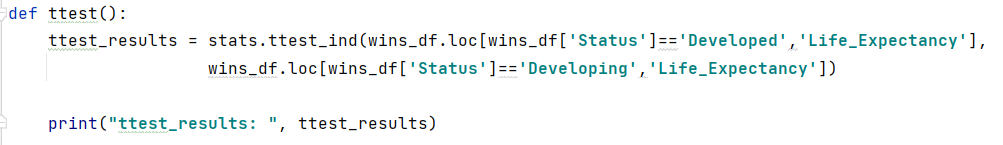


We can also take a look at the “top” and “bottom” lists by mean and median to visually inspect the countries that stand out at the top and bottom. As expected, based on the information above has indicated, the “bottom” lists are filled with African countries. While This project does not do a deep dive into the features that represent root causes of lower or higher life expectancy, future projects might use this project as a spring board to do that deep dive.

|  |  |
| --- | --- |
| **Top Life 15 Expectancy  by Country (MEAN)** | **Top Life 15 Expectancy  by Country (MEAN)** |

|  |  |
| --- | --- |
| **Bottom 15 Life Expectancy  by Country (MEAN)** | **Bottom 15 Life Expectancy  by Country (MEDIAN)** |

To see if there is a statical difference between developed and developing I’ll run a t-test to test the hypothesis that there is a significant statistical difference.



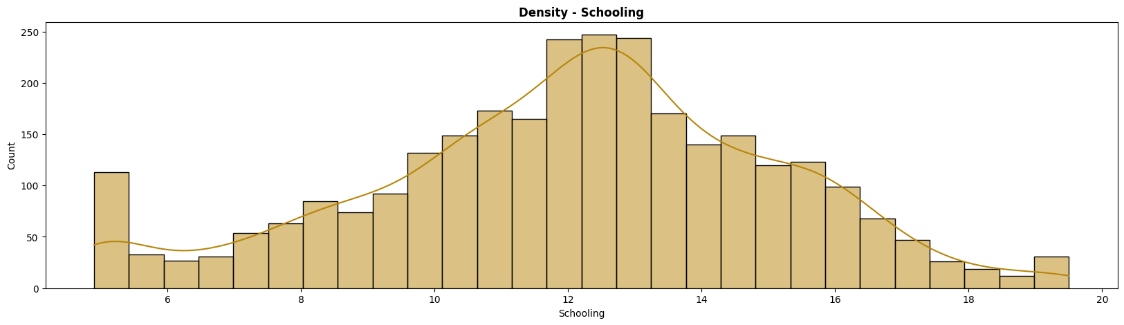
The results of the t-test shows a p-value of 1.478 indicating that there is a significant difference between developing and develop countries.

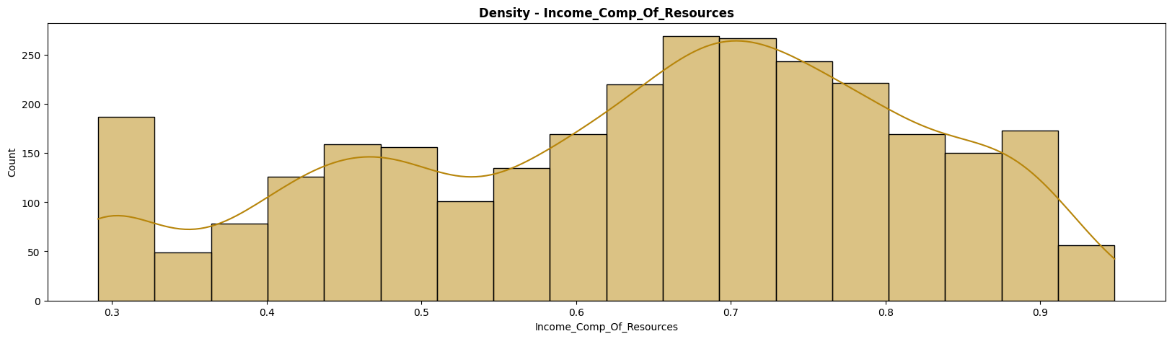


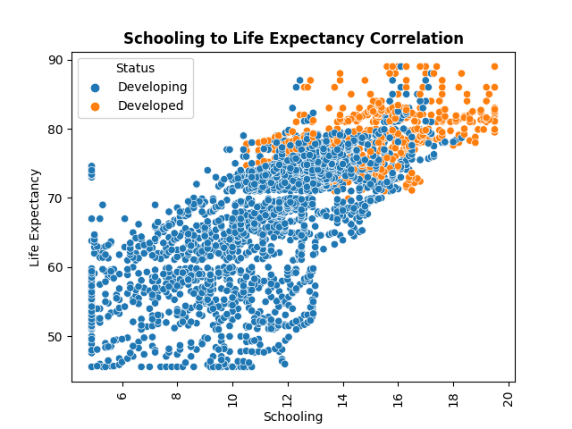
### More on Life Expectancy Features

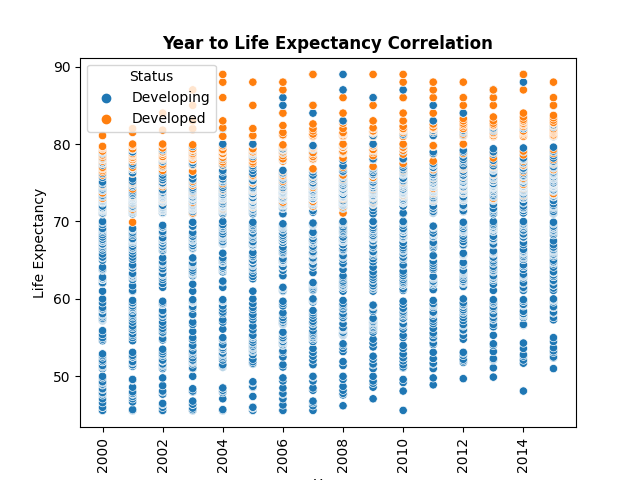
I wanted to analyze two of the independent numerical two features to get an idea of their density and correlation to Life Expectancy and decided to do this with Schooling and Income of Resources. IoR represents how productive resources are used in a country and Schooling represents number of years in school. Both have shown to be correlated to Life Expectancy so let’s look a bit more at them.

The density plots represent all countries regardless of Status, but we’ll use that as the basis for our conclusions. It would certainly be useful to break them down with respect to Status in my next project. The scatter plots are crystal clear in illustrating the correlation between Schooling and Life Expectancy and Income of Resources and Life Expectancy. WOW! It almost seems like a miracle to live past 80 years old in a developing country.







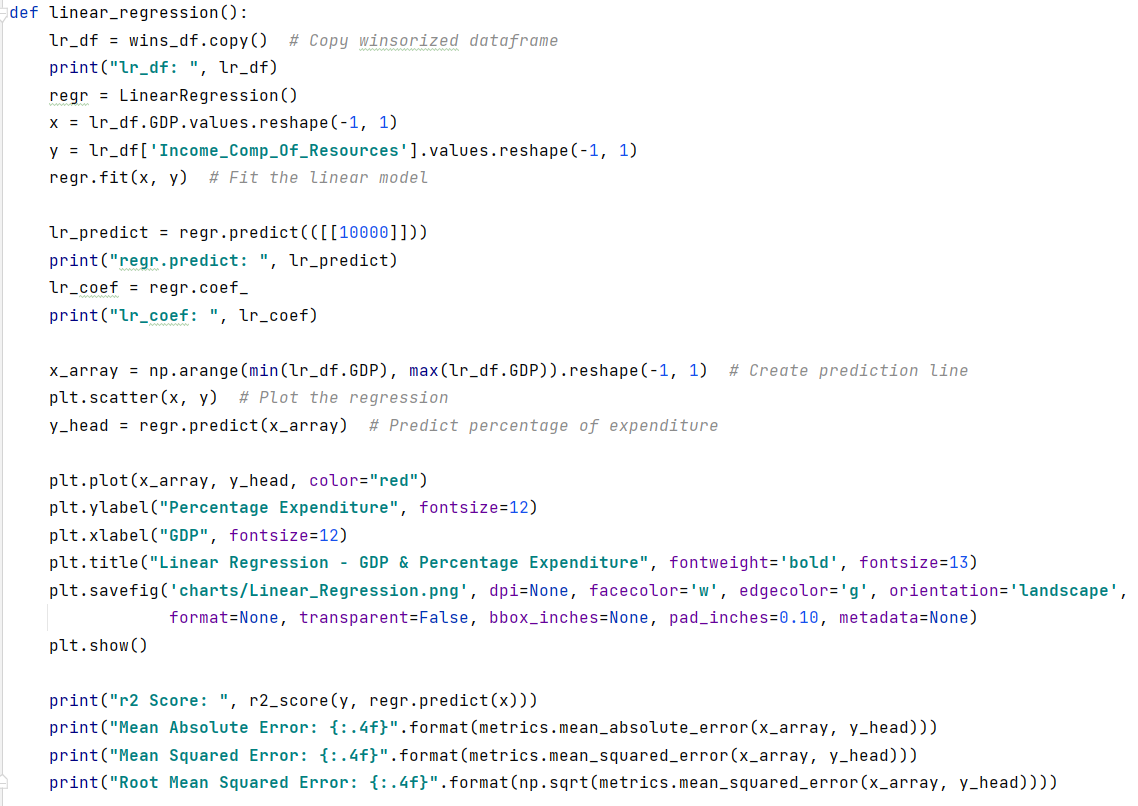


# Modeling & Predictions (Machine Learning)

So here we are, ready to model and predict! In this section I will walk through linear regression, multiple regression, polynomial regression, decision tree, random forest, and logistical regression.

## Linear Regression

For our modeling and predicting, let’s start off with linear regression to model the relationships between our independent and dependent variables. I’ll do this with the GDP and Income Composition of Resources.

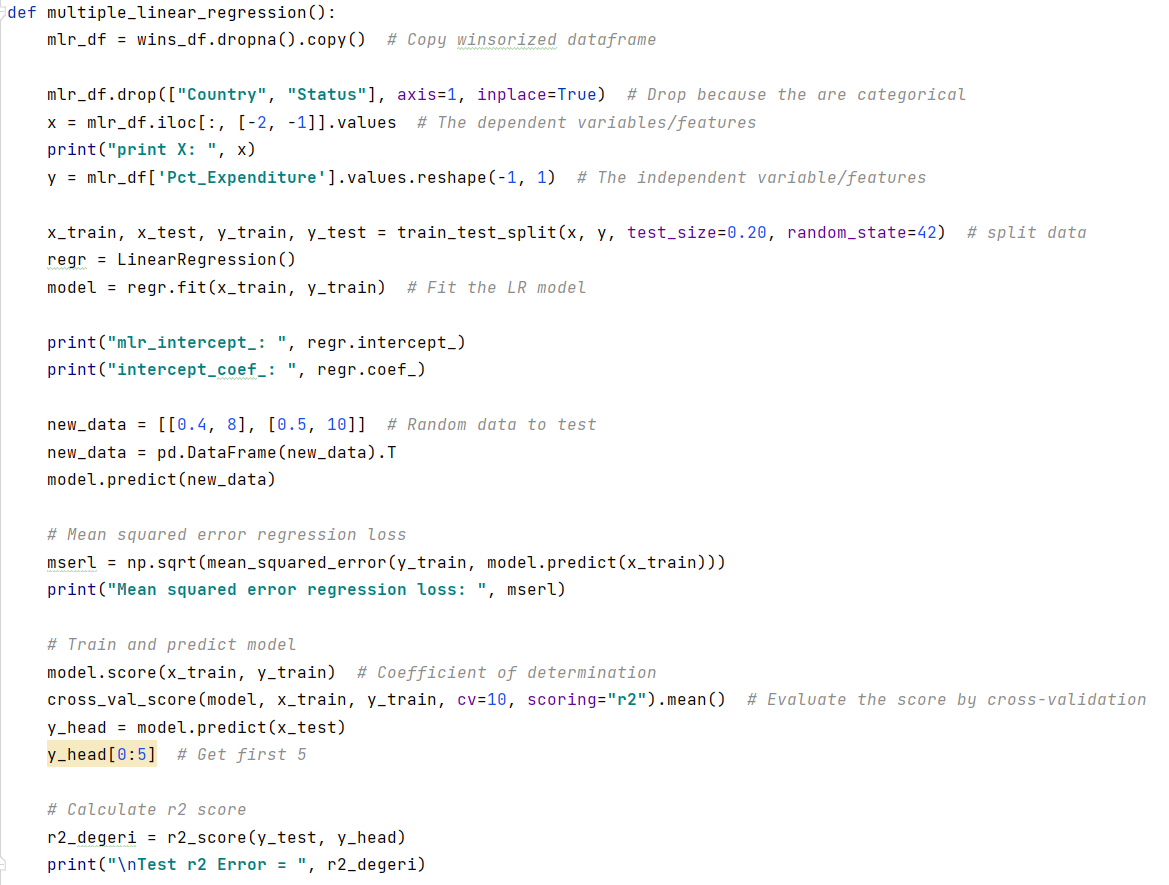


The function above fits the linear regression model, calculates a few statistics, and plots the regression.

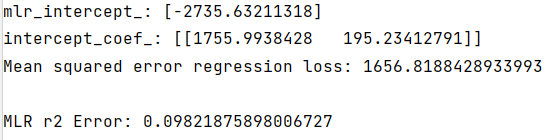
|  |  |
| --- | --- |
|  |  |

## Multiple Linear Regression

With multiple linear regression I would like to look at Life Expectancy, the independent variable.



Results from multiple linear regression prediction are below. The correctness of the model is good. It’s the difference between the training error and testing error.



## Polynomial Regression

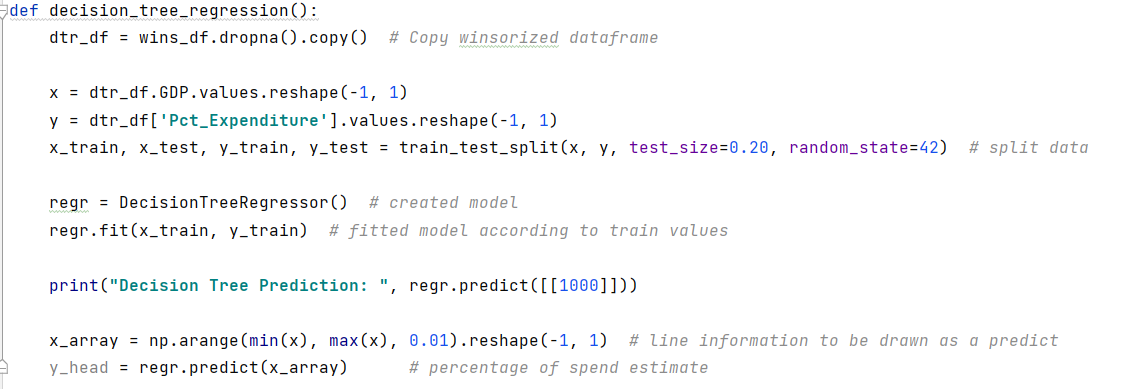
Polynomial regression can be beneficial in some cases where the data may have a a curvilinear relationship between the target variable and the independent variables.



The results from the polynomial regression prediction are below.   


## Decision Tree Regression

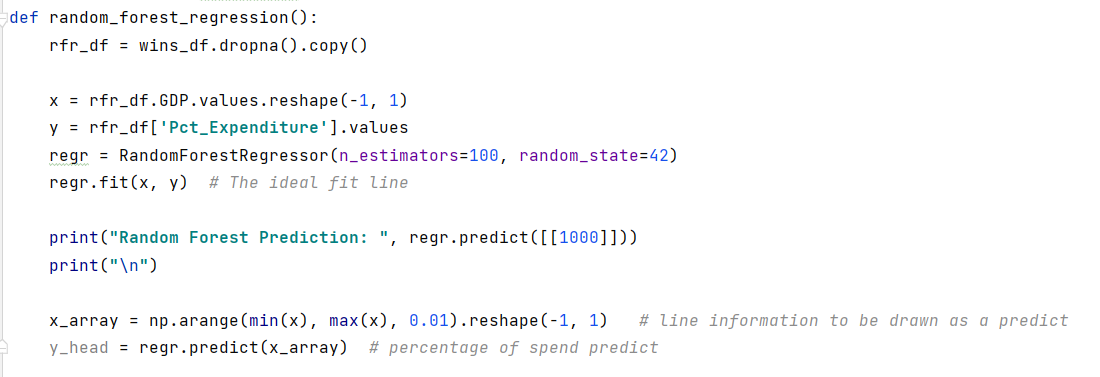
This function is a look at Percent Expenditure estimation of a Country with "GDP" value of 1000.



The results from the decision tree prediction are below and they are impressive – 98%. Another victory!  
  
v

## Random Forest Regression

Random forest regression uses decision tree logic and may be helpful in predicting average.

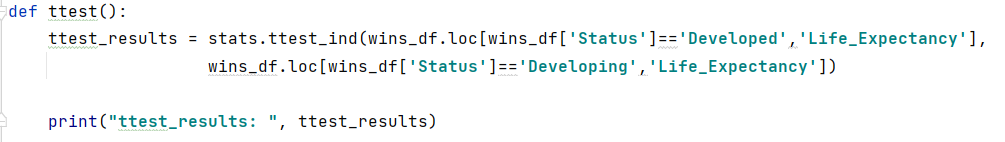


The results from the random forest prediction are below. 73.34 represents the Expenditure percentage estimation of a country with a GDP value of 1000. Not as impressive as our previous prediction.



## T-test

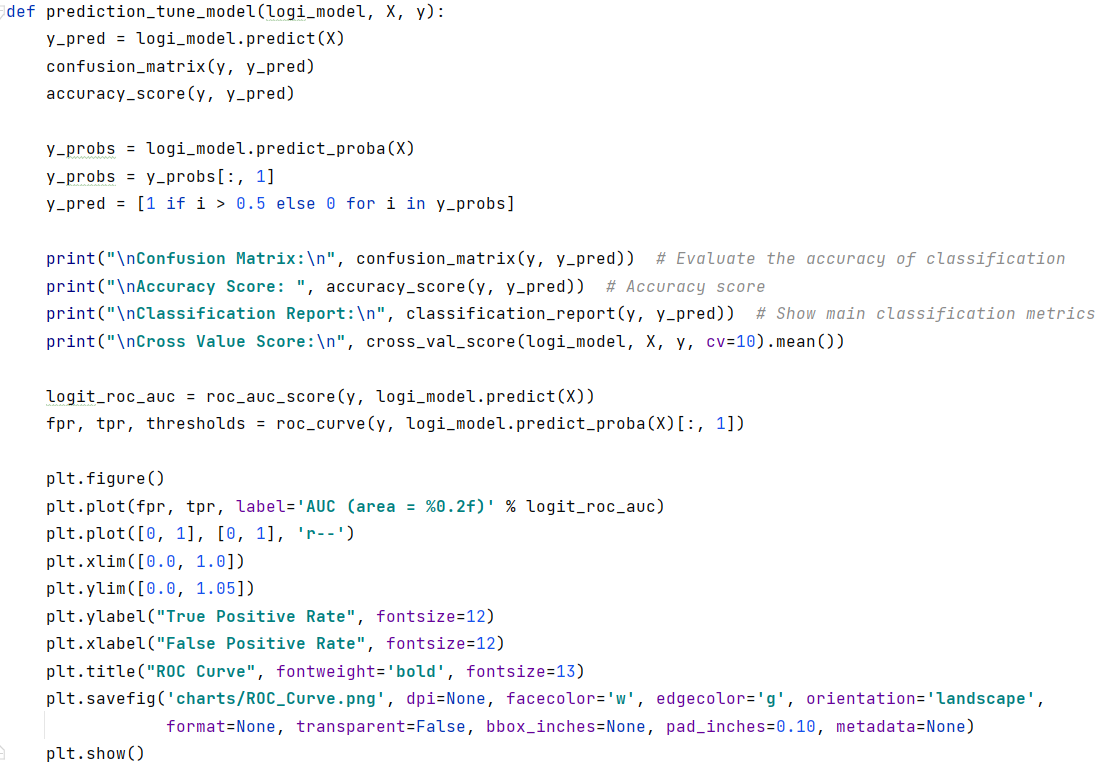
We can see that developed countries have a higher mean Life Expectancy. We can use a t-test comparison to see if the difference is significant.



The t-test results show that there is a significant difference between developing and developed countries when it comes to life expectancy.



## Prediction Accuracy

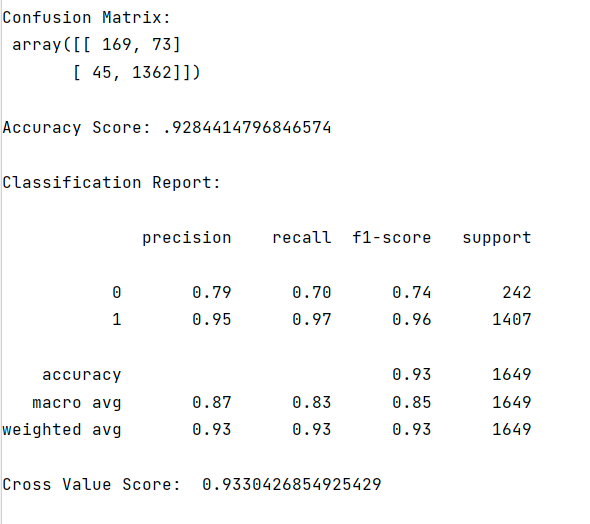


The confusion matrix allows us to see the high number of True Negatives (169) and True Positives (1362), greatly outnumbering the False Positives (73) and False Negatives (45).

The model has proved to be 92.8% accurate. I’d say that’s a success!

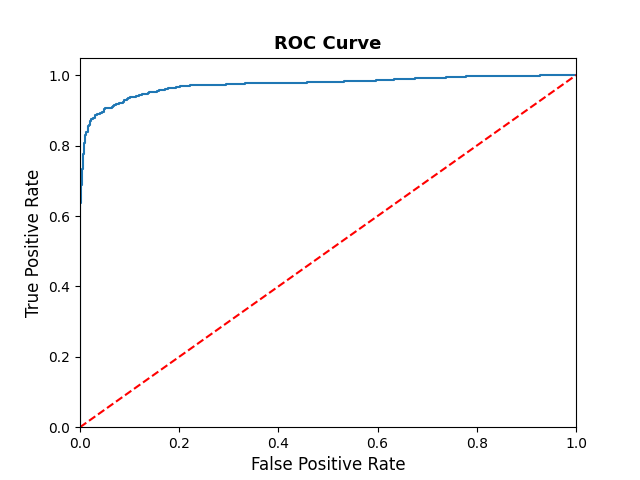
An f1-score of .96 is also impressive as the closer it is to 1, the better the precision and recall for the data.

We can also see that the cross value score (uses k-fold) tells us that with this data the model scored 93.3%! What’s nice here is the consistency in prediction accuracy scoring.



### ROC Curve

The ROC curve performed on the model to visualize the True False Positives and True False Negatives. The higher the curve, the better the model is at predicting, so the model I have created is an excellent predictor.



# Summary

The dataset began with 21 variables that were processed down to 12 independent variables (features) that describe Life Expectancy, the dependent variable.

The data was cleaned by first cleaning up the headings, then detecting and dealt with missing values, both inexplicit and explicit. The data was also imputed and winsorized to address various missing values and outliers. From this a model was born.

A number of machine learning methods were applied to the model including Linear Regression, Multiple Regression, Decision Tree regression, Random Forest Regression, and t-test.

Accuracy metrics like Confusion Matrix, Accuracy Score, Classification Report, and Cross Value Score were run, and the model scored an excellent 93% accuracy.

While this project was more focused on predictions and the exploratory data process, future projects would benefit from what was learned here.

Additional questions to be answered in future projects:

* What is the impact of disease in developed versus developing nations?
* What are the impacts of vaccines on Life Expectancy in developing nations?
* Which immunizations have the highest correlation to greater life expectancy?
* What 5 features would be most impactful on life expectancy in developing countries?

# References

**Coding and other valuable resources**

I wish I could acknowledge each piece of code I used for inspiration but there are far too many to list. So, I’ve listed a few of the more useful resources I used below.

|  |  |
| --- | --- |
| **Topic** | **Sources** |
| Regression | * <https://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html#sphx-glr-auto-examples-linear-model-plot-ols-py> * <https://www.kaggle.com/mathchi/life-expectancy-who-with-several-ml-techniques#Logistic-Regression-Model> |
| Winsorize | * [https://www.kaggle.com/philbowman212/life-expectancy-exploratory-data-analysis/](https://www.kaggle.com/philbowman212/life-expectancy-exploratory-data-analysis/notebook#Section-3:-Feature-Engineering) |
| Matpltlib | * <https://stackoverflow.com/questions/46664082/python-how-to-save-statsmodels-results-as-image-file> |
| EDA | * [https://www.kaggle.com/philbowman212/life-expectancy-exploratory-data-analysis/](https://www.kaggle.com/philbowman212/life-expectancy-exploratory-data-analysis/notebook#Section-3:-Feature-Engineering) * <https://towardsdatascience.com/an-extensive-guide-to-exploratory-data-analysis-ddd99a03199e> |
| Training & Testing Model | * <https://www.kaggle.com/rishavchowdhury0123/life-expectancy-eda-and-prediction#Training-and-Testing-the-Model> * <https://www.kaggle.com/mathchi/life-expectancy-who-with-several-ml-techniques#Logistic-Regression-Model> |
| Plotting | * <https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51> * <https://stackabuse.com/matplotlib-box-plot-tutorial-and-examples/> * <https://thecleverprogrammer.com/2021/01/06/life-expectancy-analysis-with-python/> |
| Predicting | * <https://www.datasciencesociety.net/using-machine-learning-to-explain-and-predict-the-life-expectancy-of-different-countries/> * <https://www.kaggle.com/mathchi/life-expectancy-who-with-several-ml-techniques#Logistic-Regression-Model> |