Derek Norton

Braiden Bangalan

Elaine Bunten

Rohan Chinalachaiagari

**EECS 331 Final Project Report**

**Introduction**

The overall purpose of our project was to achieve further insights regarding the influence of different variables and their weight of effect on the outcome of life expectancy. As stated in our original Project Plan document: “For our project, we intend to use ML and data analysis techniques to determine what health factors weigh heaviest into the average lifespan by analyzing data from Asian countries.” By utilizing various data science analysis techniques available, we achieved a better understanding of the topic. After the initial data wrangling stage, the techniques used were linear regression, random forest clustering, and statistical inferencing to dissect the data and the relationships present.

**Data Wrangling**

The first challenge faced in the project was choosing what variables to pursue. What is relevant given the scope of the project? In the macro, we determined these in the project plan document as stated below:

*“What is the endgame of health sciences? The endgame is to optimize the length and quality of life for all people. There is a seemingly endless list of individual factors that play into long-term health, but put broadly those factors are internal and external. Internal factors include variables such as diet, exercise, and exposure to infections and diseases. External factors include, but are not limited to, sanitation, air quality”*

The data sources used for this were from reputable organizations primarily the World Health Organization and the University of Oxford as well as some other sources.

One key exclusion in the data is economic variables, this was done based on the reasoning that no human has ever died or fallen ill due to a lack of wealth, it’s the lack of wealth leading to a lack of access to needed resources that leads to poor outcomes. For this reason, variables such as GDP, income per capita, and any other economic metric were withheld.

The datafields selected for this project were:

-Air Quality (Total, Rural, Urban, Towns, Cities)

-Alcohol consumption (by sex)

-Child mortality percentage of under 5 years old (by sex)

-Calorie coefficient

-the difference in calories consumed amongst the population

-Causes of death (diseases, conditions, etc.)

-Population

-Food supply (kcal, protein g, fat g consumption)

-Population not drinking from quality water sources

-Obesity percentage of country

Our values to be predicted are

-Life Expectancy from birth (by sex)

-Life Expectancy once age 60 is reached (by sex)

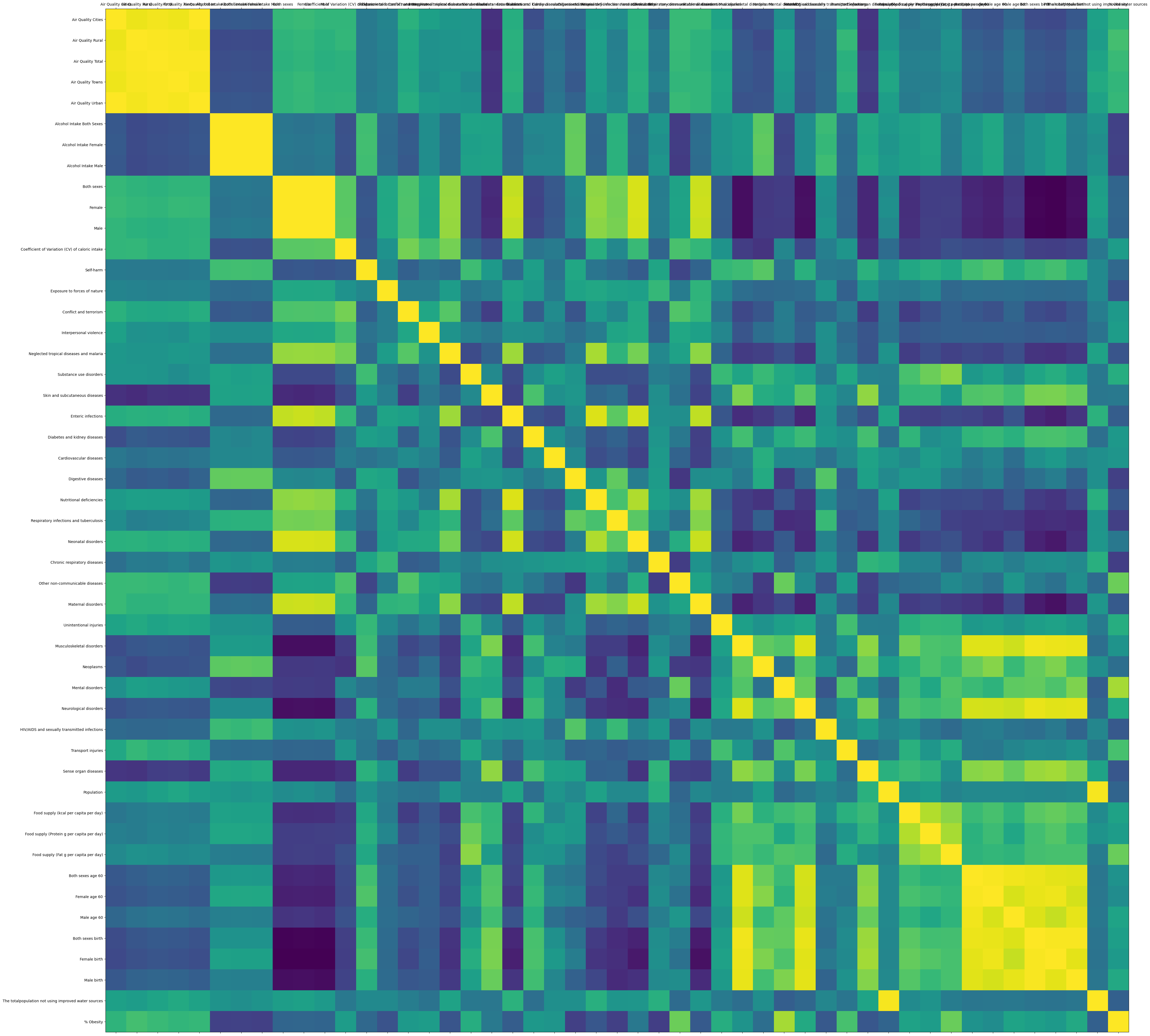
The cleaning of the data involved filtering unnecessary columns, renaming columns, pivoting data.

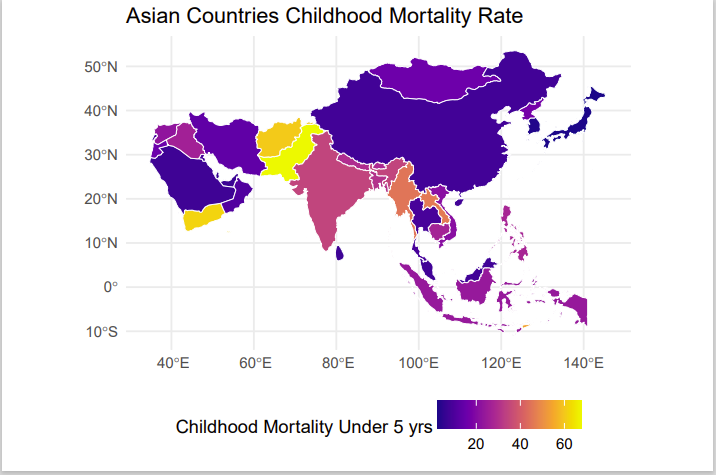
All data was merged based on the country parameter. Some countries were removed from the dataset due to a lack of presence of some of the values. Once the merging was completed 30 countries possessed all of the data values needed and were used for data manipulations.

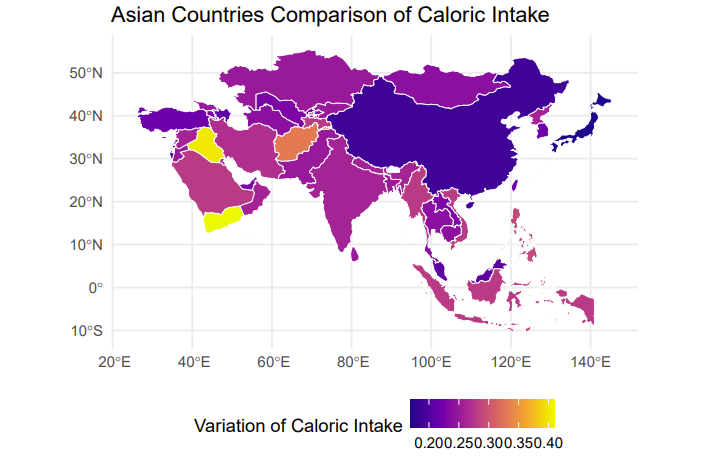
**Visualizations**

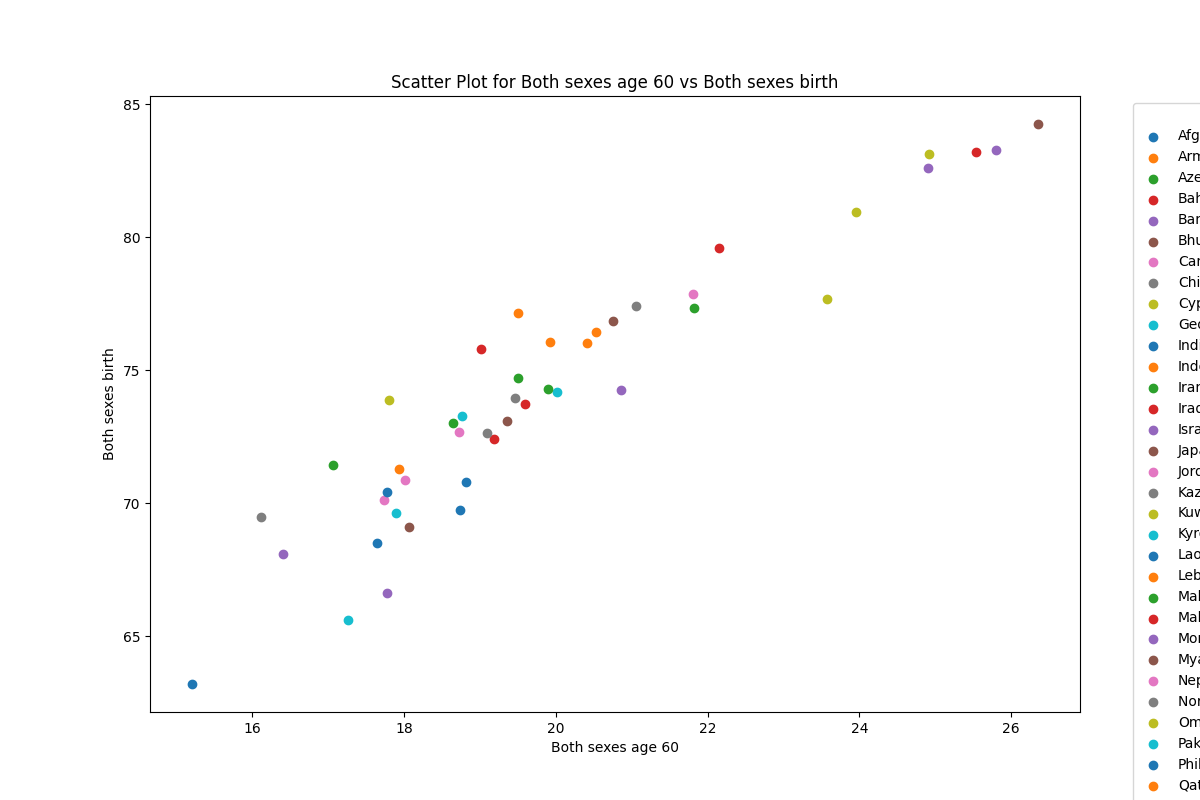
Using the data, we began by creating a variety of initial visualizations including correlation maps, geographical indicators, and circle graphs to create illustrations comparing the data present. This data is available in the visualization folder.

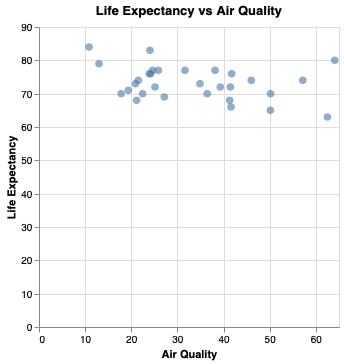
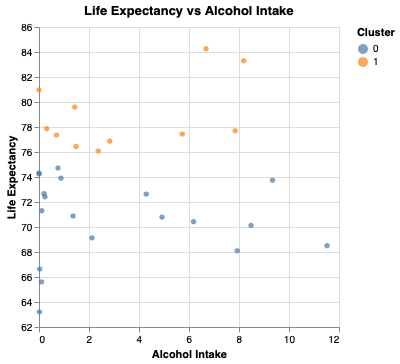
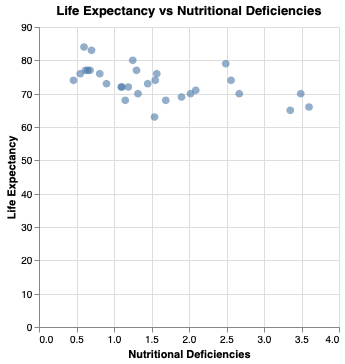
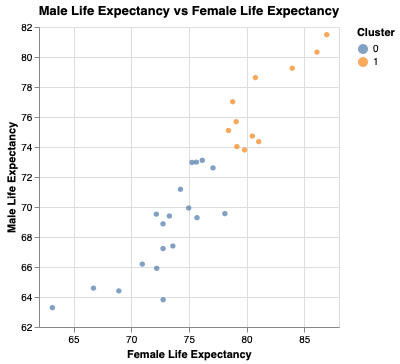
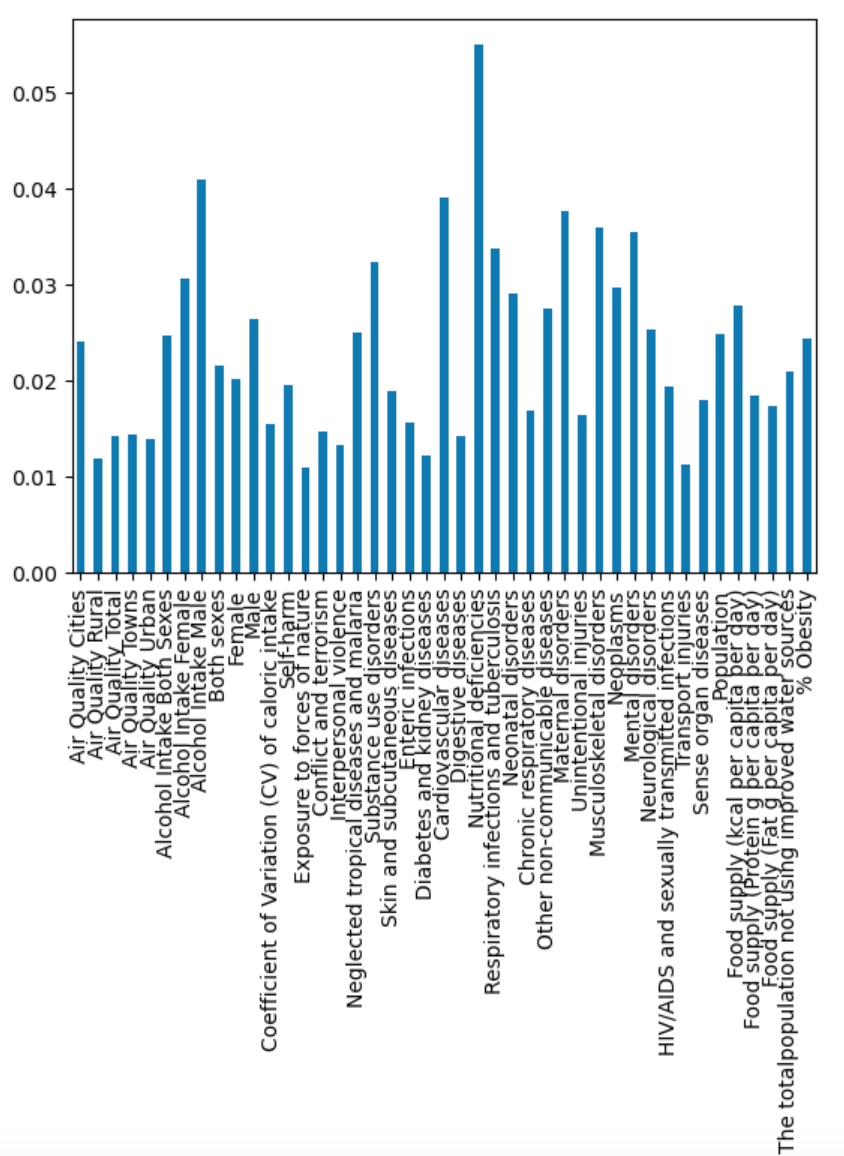
Below are some of the visualizations produced using the data.

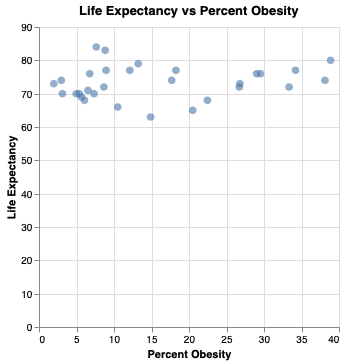
Correlation grid









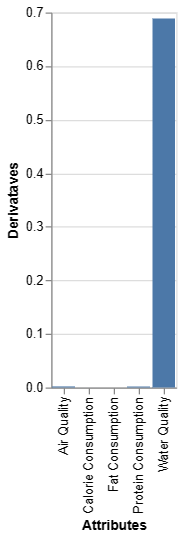


**Linear Regression**

For the data manipulations done Linear Regression was chosen over KNN Regression due to better accuracy of the model using Linear Regression.

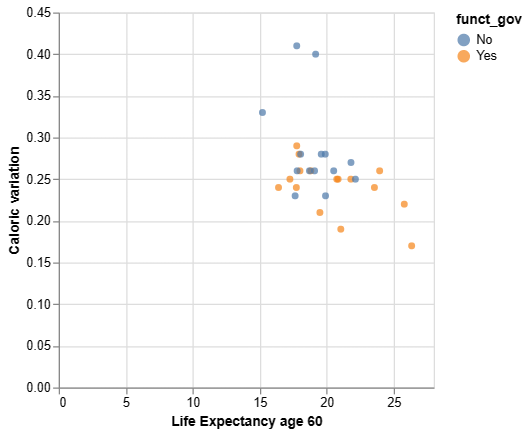
In the first block of code the variables used as the X variable for training the model were reduced from the original data set to include both sexes in data fields where both sexes were separated. After creating a pipeline and training the model, a new dataset was created containing a similar dataframe structure but values from countries outside of Asia. The countries selected were Chile, Belgium, and Chad. The model was able to predict life expectancy with a degree of accuracy.

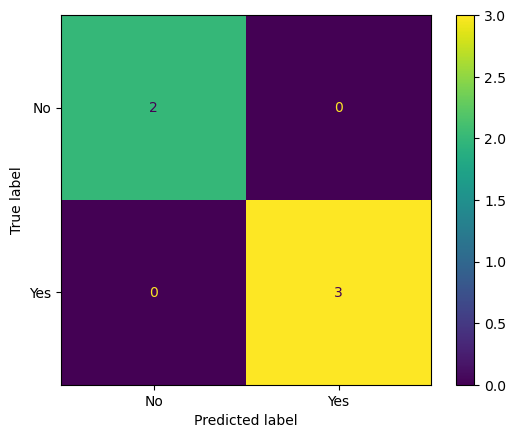
Next, the dataset from Chile was taken and used to evaluate the model itself. The assumption being some variables have a greater influence on life expectancy than others. This was tested by finding the change in life expectancy when the value of a variable was changed incrementally. The values of interest in this case were water quality, air quality, calorie consumption, protein consumption, and fat consumption. After finding the rate of change water quality was the mosty influencing variable according to the linear regression model(1).



**Classification**

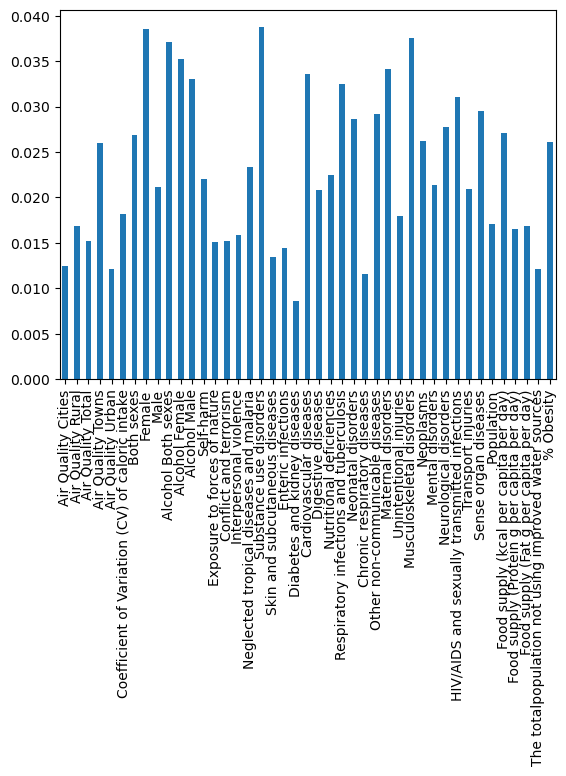
For the use of the classification model the curiosity arose if from the data in the dataframe we could predict if a country had a functioning system of government based on the quantitative data. The data imported Oxford University’s functioning government index in the year 2019 exclusively. Initially the data was visualized comparing two variables with the classification data serving as the color (2) The initial test of the model is similar to the methods used to initially test the accuracy of the linear regression model. Following this, a confusion matrix produced represents the accuracy recall and precision of the model(3).

(2)

(3) 

**Clustering**

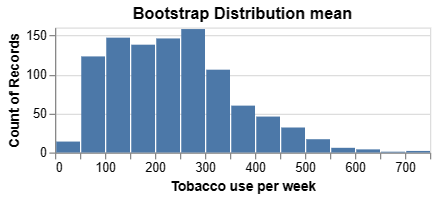
The random forest clustering model yielded some important insights into the data itself and determined that the most important health factors were nutritional deficiencies, male alcohol intake, and cardiovascular diseases (4). Other means of clustering were used, however, due to the limited amount of data available no correlations or insights could be derived.

(4) 

**Statistical Inferencing**

For statistical inferencing, data from a WHO survey regarding tobacco usage in Bangladesh was used to derive statistical insights from one of the countries with the heaviest usage of tobacco products in the world.

The survey surveyed households and asking a variety of questions about the lives of the participants, tobacco usage in particular. The portion of the survey focused on to derive data was the field of questions regarding tobacco product usage per week. The cryptic looking data field represent different delivery methods for tobacco (hookah, cigarettes, hand rolled cigarettes, cigars, etc.) Bootstrapping was used and through that method it was determined that on average, amongst tobacco users, tobacco products are used 237 times a week on average according to survey participants(5). This equates to roughly once every 30 minutes.

(5)

**Conclusions**

Through the use of data science applications, the important variables determined to an extended life expectancy are primarily water quality, and adequate nutrition. For recommendations for the future, the data can be applied on a national and international scale, but also at an individual level. While this study was limited to Asian countries, the people of Asia share 99.9% of their DNA with people in Greenland, or Brazil, or Nigeria, or Memphis, Tennessee. Major dietary deficiencies would influence any person in a comparable way. Adequate nutrition and quality water are the variable important to longevity according to this project.