# Data Science Project Report

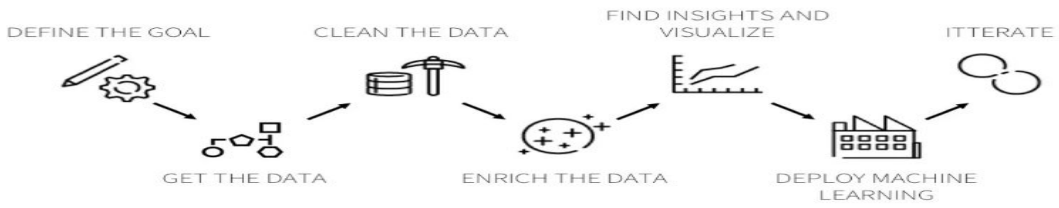
## Analysis of factors influencing Life Expectancy

Ridam Pal - PhD 19201
Sai Krishna Vamshi - 2016033
Jayanth Krishna - 2016038

## Motivation and Problem Statement

Recent trends have shown that life expectancy is increasing at a slow rate in various countries. Reason of such trends are not being yet stated or answered which have lead to different assumptions. We are trying to figure out the main factors which are influencing the life expectancy thereby trying to predict as well enhance the life expectancy rate in upcoming years. We will also try to analyse the data among developing and developed countries (India being primary concern) and try to draw contrast among these countries.

## Project Plan




- Collecting Data from various sources
- Explore and cleaning
- Creating a new dataset by analyzing various dataset from other sources.
- Build Visualizations.
- Create models which will predict the trends of life expectancy in upcoming years.
- Understanding Consequences of increasing Life Expectancy

## Data Collection

- **Data Sources**
  - Kaggle. https://www.kaggle.com/kumarajarshi/life-expectancy-who/data

○World bank open data. [https://data.worldbank.org](https://data.worldbank.org)

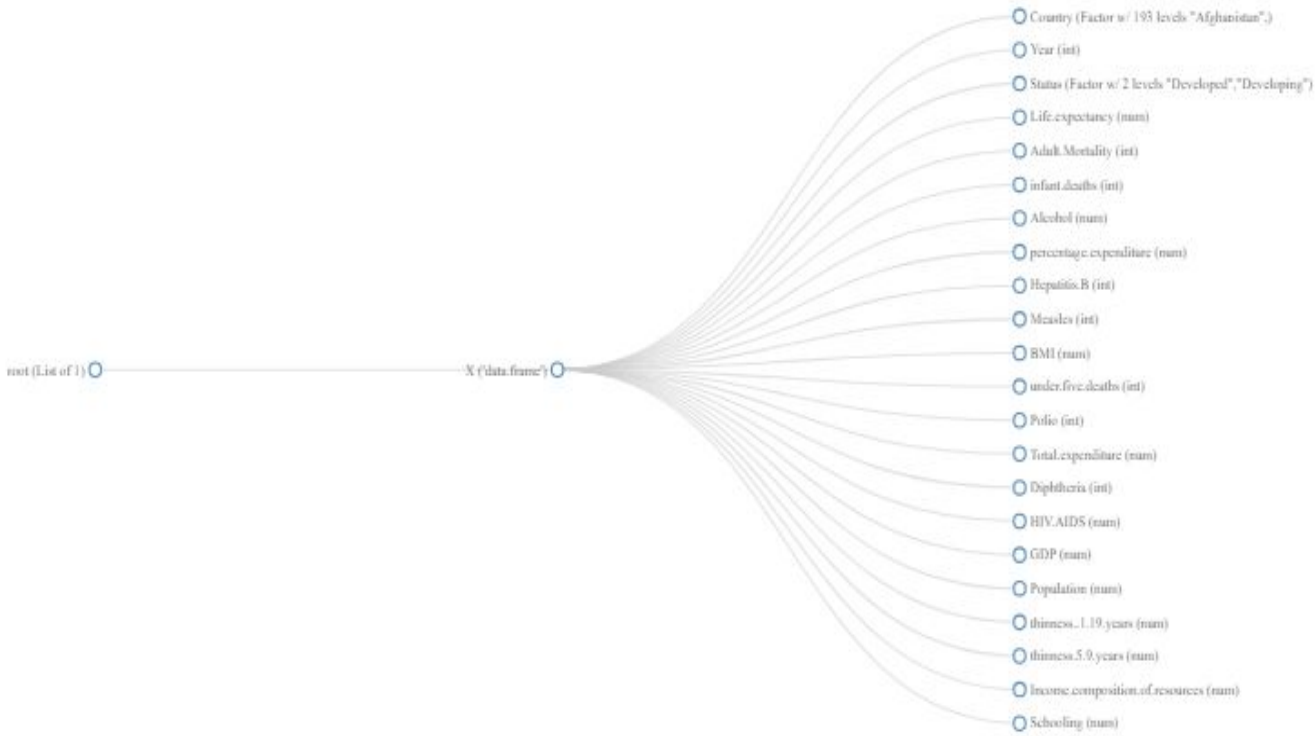
- **Data Curation, Acquisition and preprocessing of data**

○We have collected our main dataset from Kaggle.

○We have also collected some other data from World Bank Open Data of attributes/features which might affect Life Expectancy, thereby adding these attributes to create a dataset of our own.

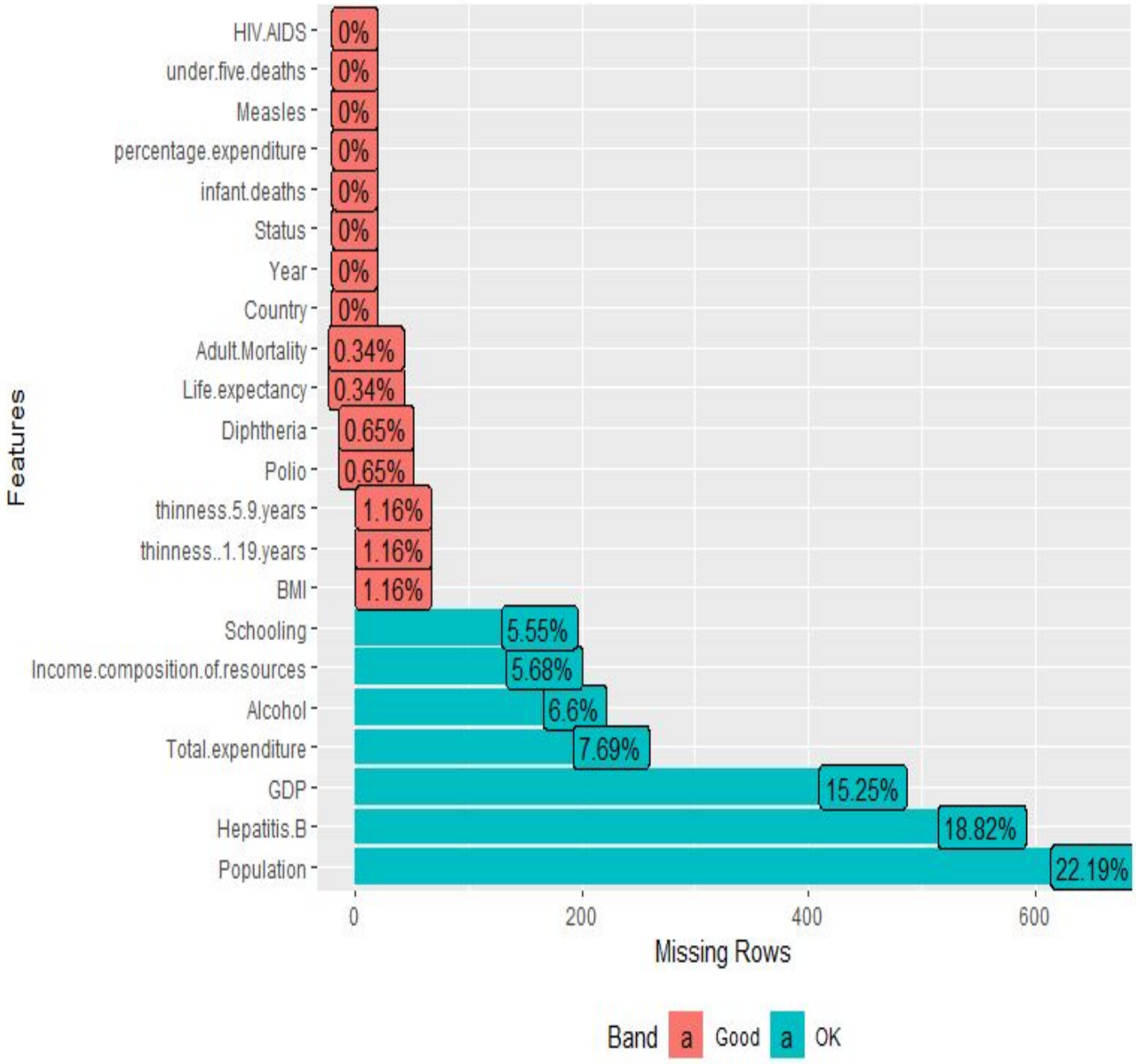
○For missing data in like GDP we referred wiki Pages and acquired data.

- **Data Representation**



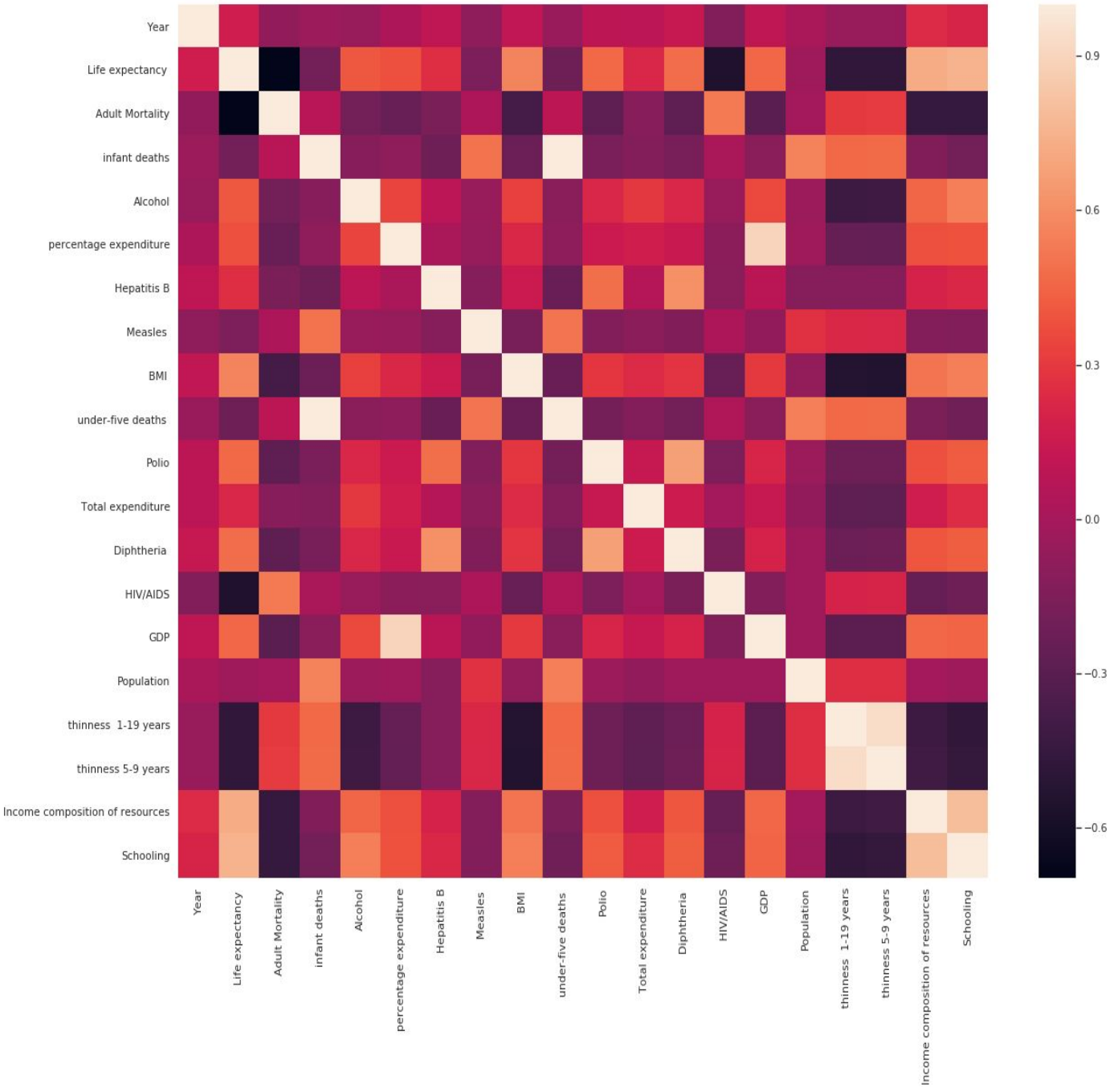

**All features of the final Data**

## ●Missing Data



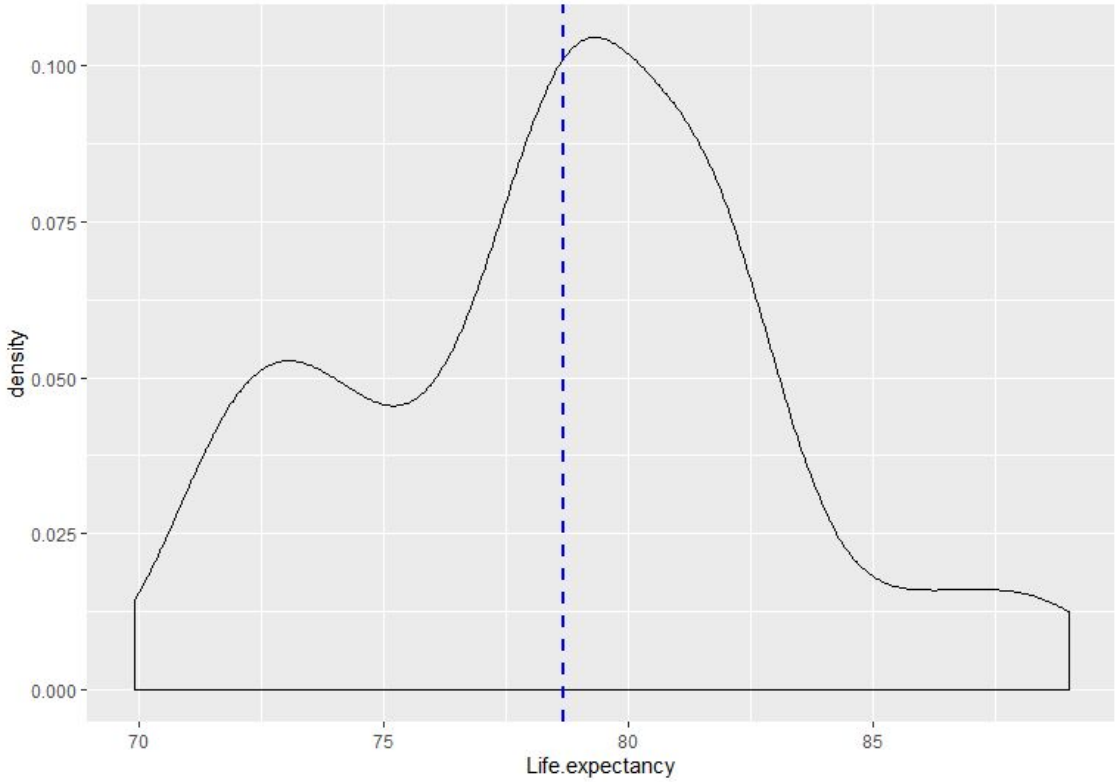

**Missing Data Distribution**

## ●Correlation Matrix

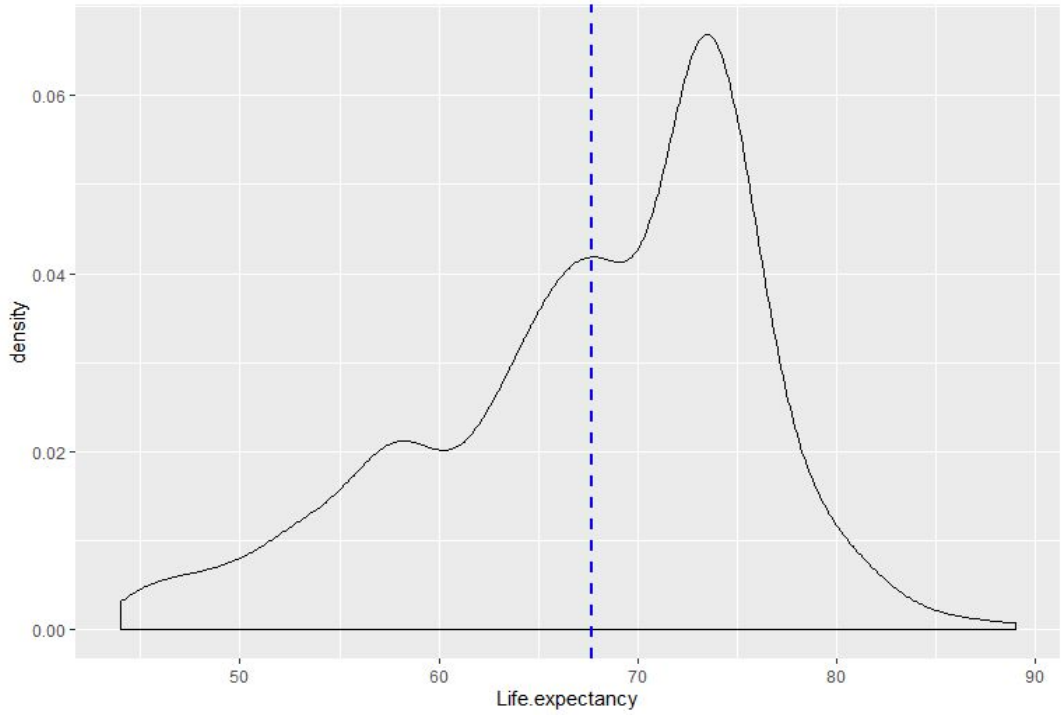



Correlation among all features

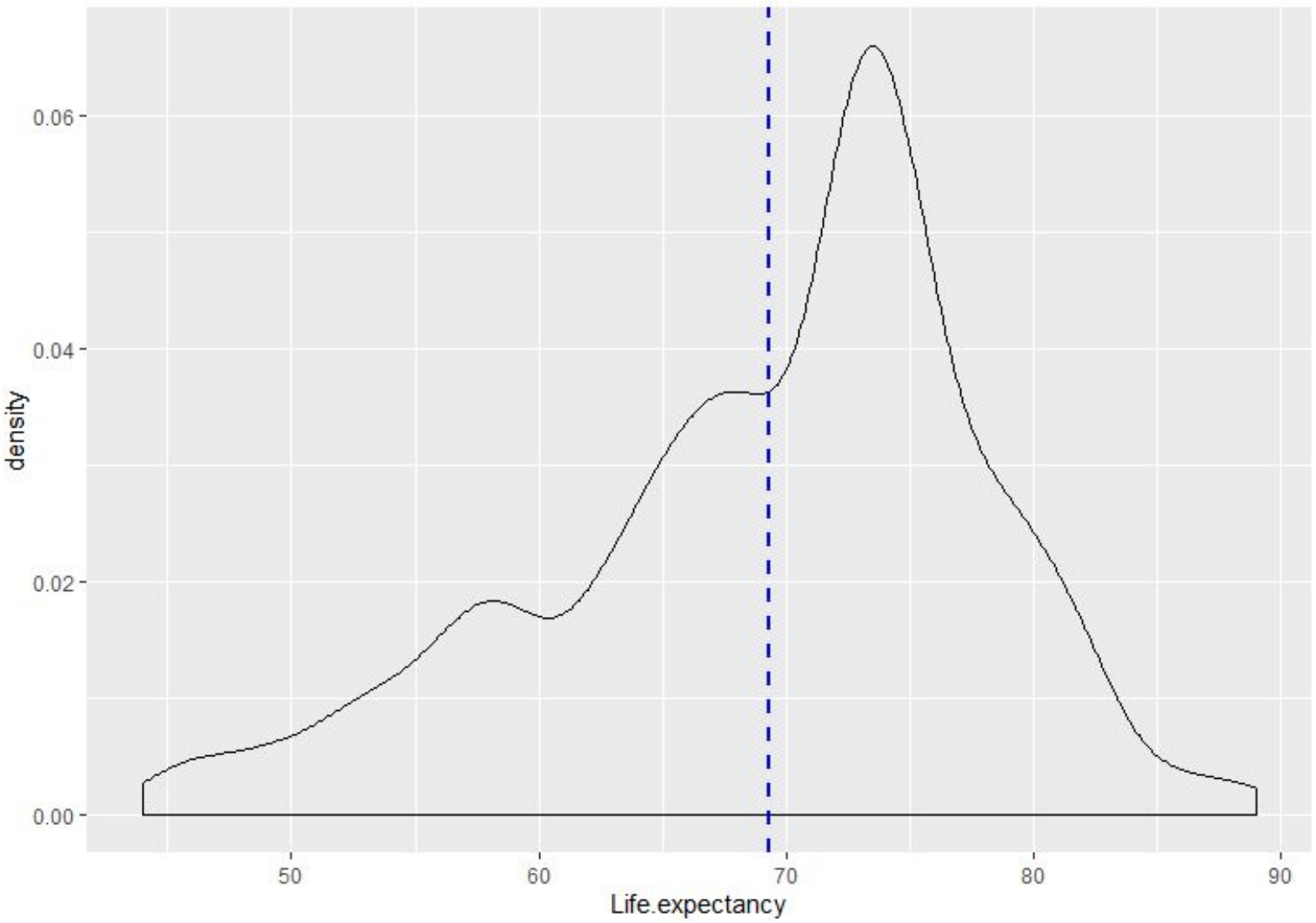
# Exploratory Data Analysis



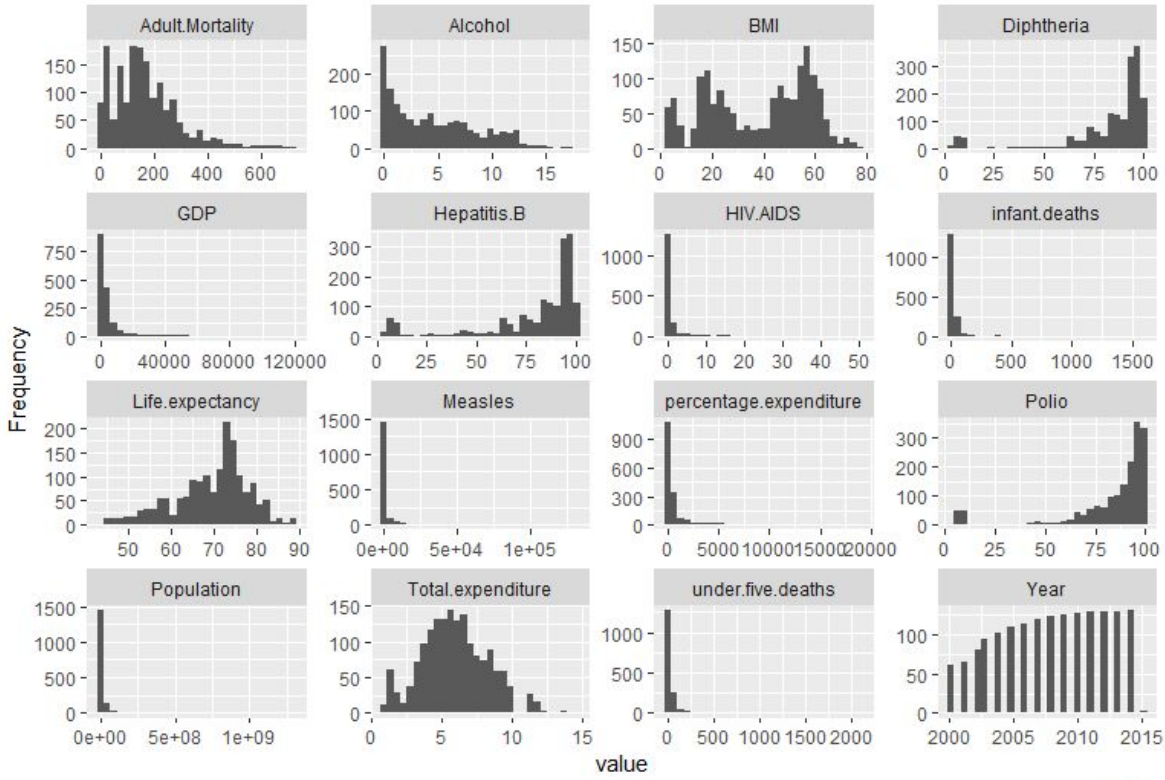

**Developed Countries Life Expectancy Distribution**



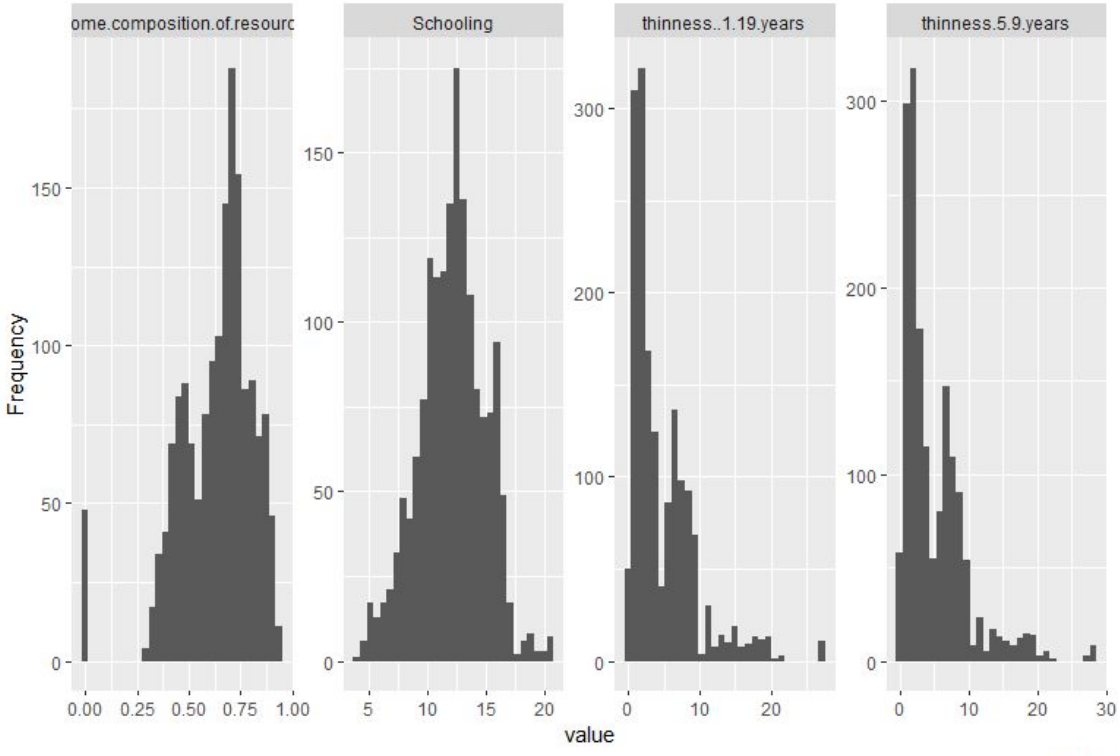

**Developing Countries Life Expectancy Distribution**




**Overall Life Expectancy Distribution**



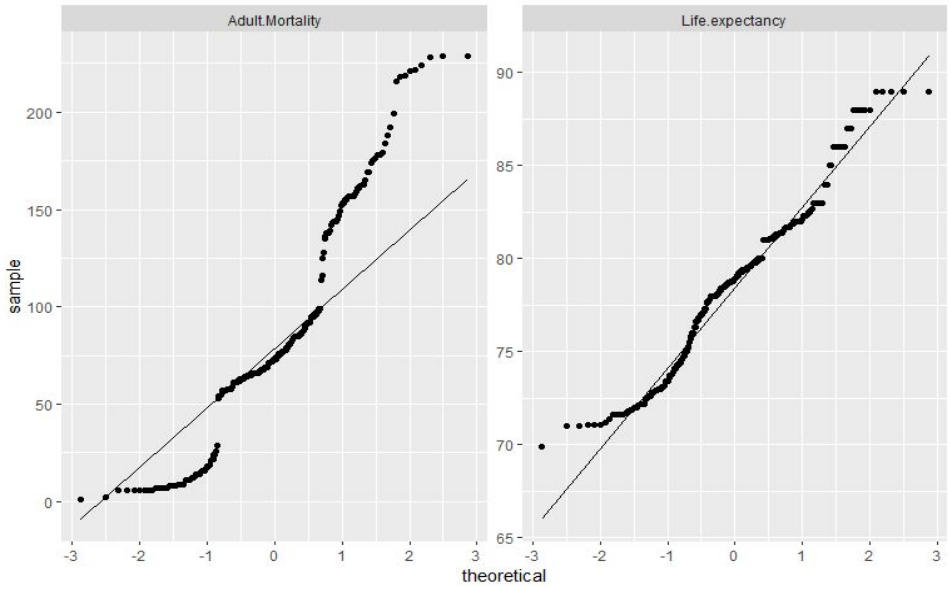

Page 1



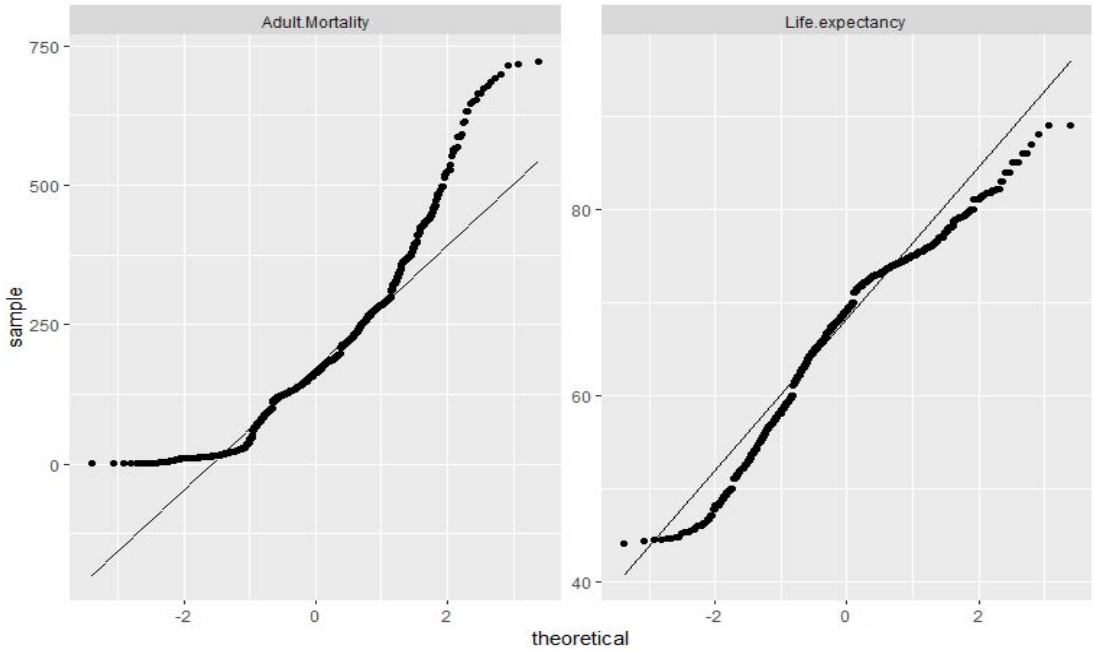

Page 2

**Frequency Distributions of all features**

## ●Quantile - Quantile Distribution



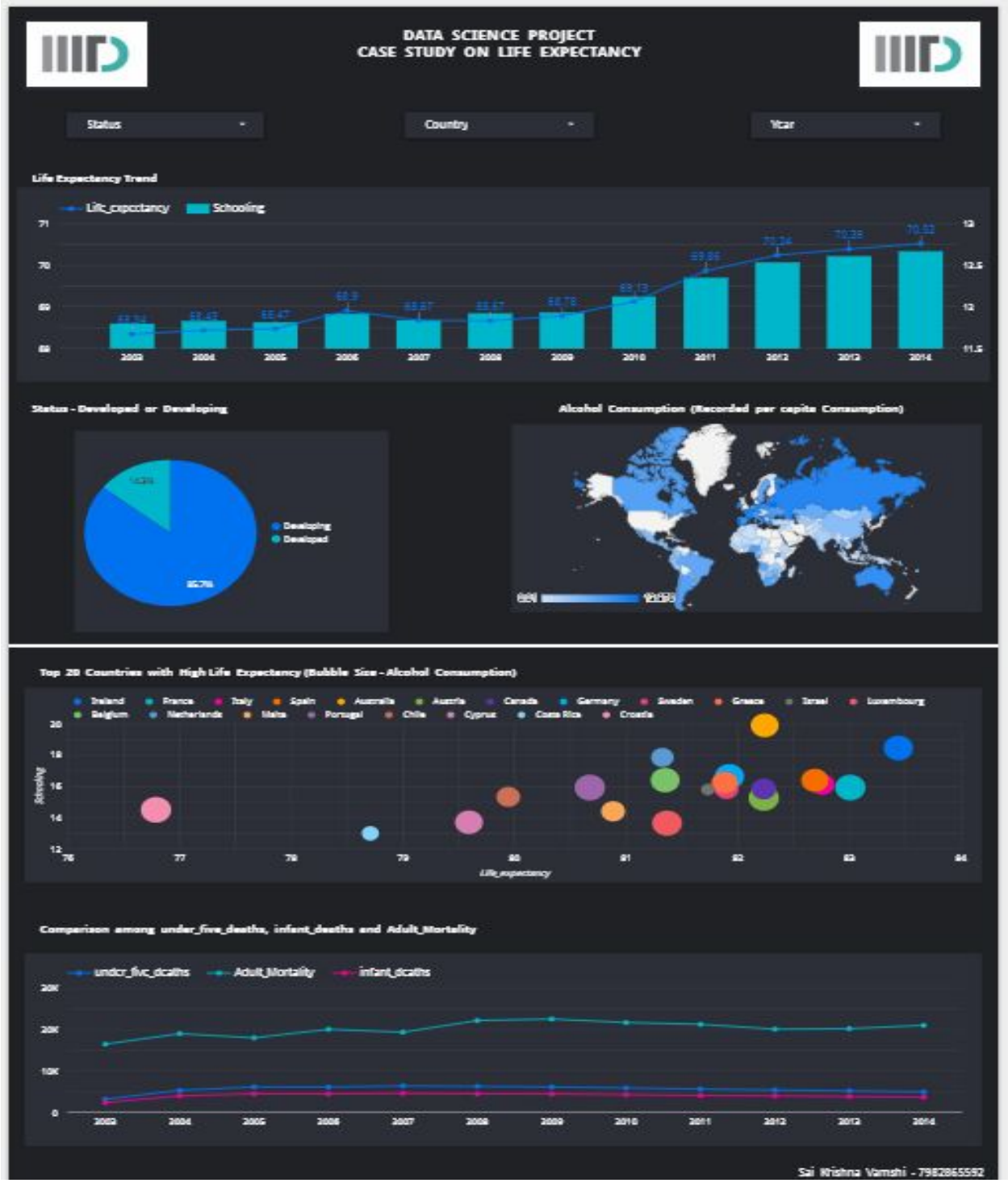

Developed Countries QQ Plot




Developing Countries QQ Plot

# Visualization using Google Data Studio(Online Tool)

**Link -** *https://datastudio.google.com/reporting/1WAojPm4w8LZM2z1xNd8gMC3jvXJte4JD/page/iSq5*



# Correlation of Attributes with Life Expectancy

Comparing Correlations among All Countries(World), Developed Countries, Developing Countries, India, China, USA

| | World Life Expectancy | Developed Life Expectancy | Developing Life Expectancy | India Life Expectancy | China Life Expectancy | USA Life Expectancy |
|---|---|---|---|---|---|---|
| Life expectancy | 1.000000 | 1.000000e+00 | 1.000000 | 1.000000 | 1.000000e+00 | 1.000000e+00 |
| Adult Mortality | -0.696359 | -4.854888e-01 | -0.660836 | -0.274277 | 4.754545e-01 | -9.015021e-01 |
| infant deaths | -0.196557 | -5.476379e-02 | -0.166474 | -0.994640 | -9.984629e-01 | -9.179026e-01 |
| Alcohol | 0.404877 | -2.877694e-01 | 0.203429 | 0.963908 | 8.770610e-01 | 9.097630e-01 |
| percentage expenditure | 0.381864 | 3.503151e-01 | 0.344402 | 0.521725 | 9.375067e-03 | NaN |
| Hepatitis B | 0.256762 | -8.618966e-02 | 0.253645 | 0.792038 | 8.242457e-01 | 1.396073e-01 |
| Measles | -0.157586 | 3.780051e-02 | -0.141788 | 0.125425 | -4.387001e-01 | 4.910216e-01 |
| BMI | 0.567694 | -4.396246e-02 | 0.555682 | 0.998254 | 2.805363e-01 | 6.596335e-01 |
| under-five deaths | -0.222529 | -4.795308e-02 | -0.195344 | -0.998021 | -9.984355e-01 | -9.116829e-01 |
| Polio | 0.465556 | 1.834187e-02 | 0.436644 | 0.984122 | 9.085077e-01 | 5.705619e-01 |
| Total expenditure | 0.218086 | 7.159335e-02 | 0.094317 | 0.467931 | 6.312584e-01 | 9.446202e-01 |
| Diphtheria | 0.479495 | -1.960463e-02 | 0.459877 | 0.526268 | 9.250667e-01 | 4.297444e-02 |
| HIV/AIDS | -0.556556 | -9.364703e-15 | -0.570596 | -0.869180 | 1.391951e-15 | 1.762851e-14 |
| GDP | 0.461455 | 3.534512e-01 | 0.389506 | 0.763034 | 5.096245e-01 | NaN |
| Population | -0.021538 | 7.962044e-02 | 0.000417 | -0.007515 | -2.084723e-01 | NaN |
| thinness 1-19 years | -0.477183 | -5.880775e-01 | -0.366642 | -0.964594 | -9.914010e-01 | -2.028131e-01 |
| thinness 5-9 years | -0.471584 | -5.965299e-01 | -0.359685 | -0.994588 | -9.768958e-01 | -4.043722e-01 |
| Income composition of resources | 0.724776 | 7.240268e-01 | 0.644114 | 0.997996 | 9.860143e-01 | NaN |
| Schooling | 0.751975 | 3.952101e-01 | 0.688119 | 0.991049 | 9.696810e-01 | NaN |

# Insights from Data Correlation and Visuals

**Complete Data**

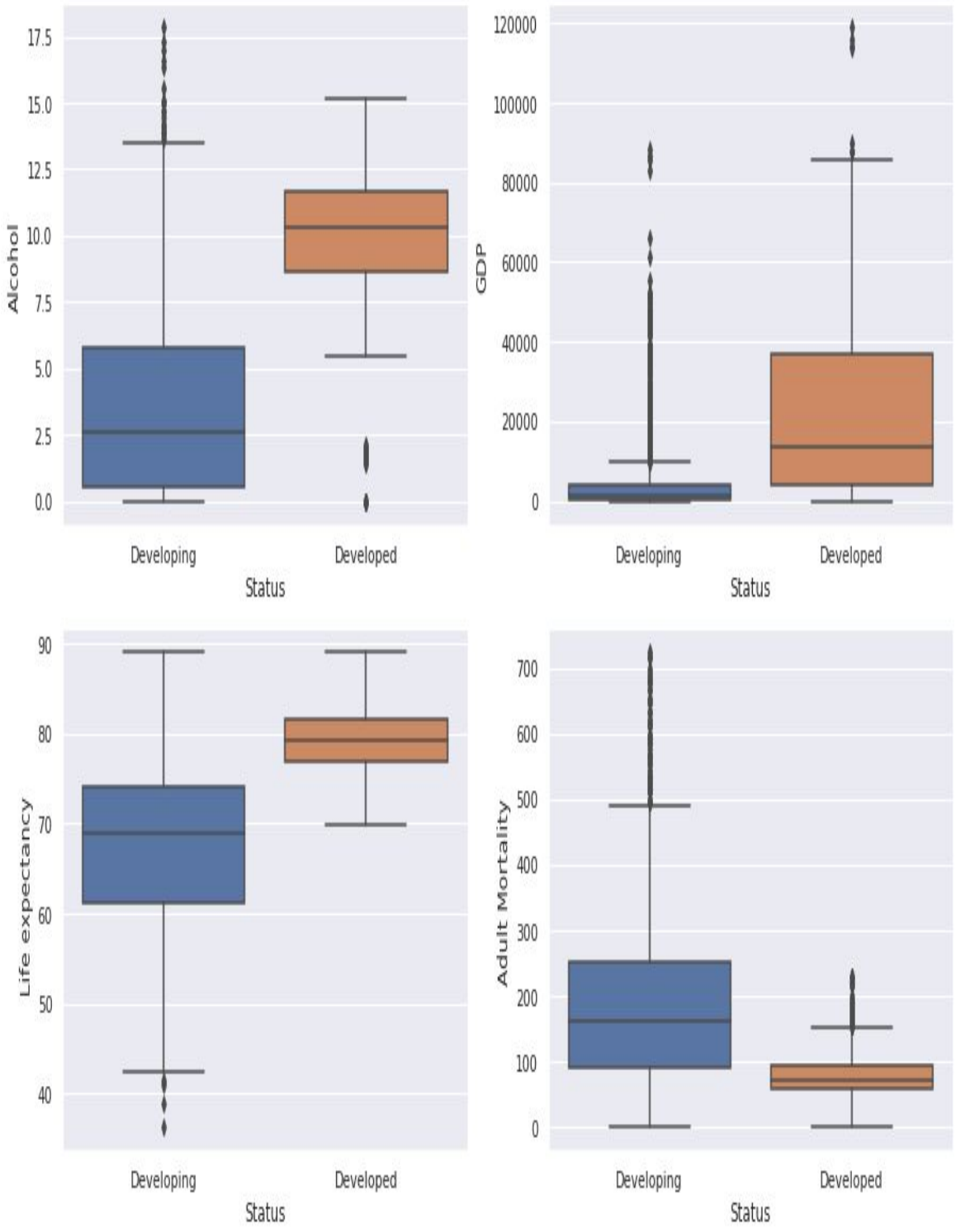
- Life Expectancy - Schooling → Significantly high Positively Correlated (**0.751975**)
- Life Expectancy - HIV/AIDS → Significant Negative correlation (**-0.556556**)
- Life Expectancy - Adult Mortality → Significant Negative correlation (**-0.696359**)

**Indian Data**

- Life Expectancy - Schooling → Very strongly Positive Correlated (**0.9910**)
- Life Expectancy - Alcohol → Very strongly Positive Correlated (**0.9639**)
- Life Expectancy - Infant Deaths → Very strongly Negative Correlated (**-0.9946**)

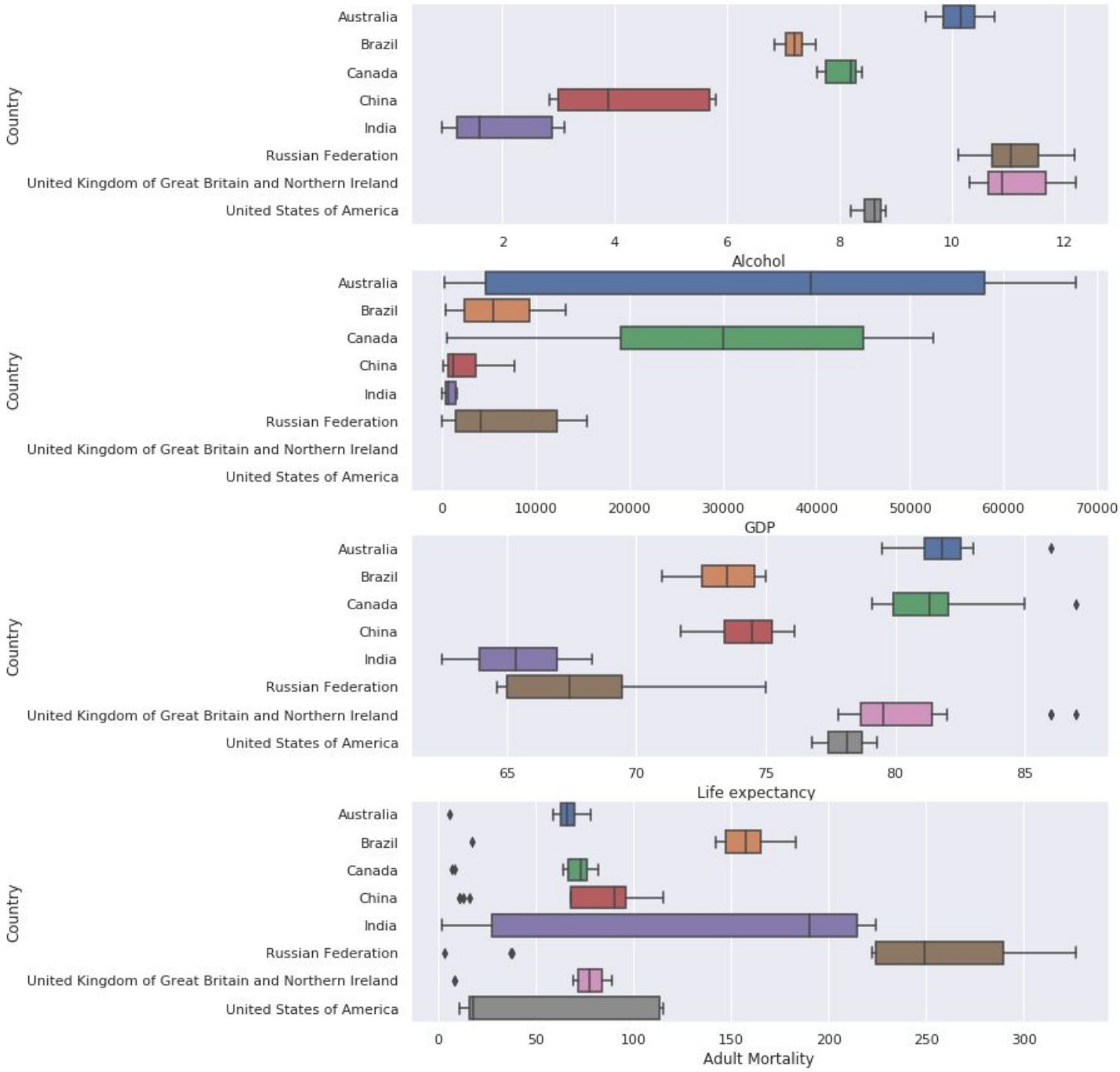
**USA Data**

- Life Expectancy - Adult Mortality → Very strongly Negative Correlated (**-0.9015**)
- Life Expectancy - Total Expenditure → Very strongly Positive Correlated (**0.9446**)




**Box Plots of Developed vs Developing**

Alcohol Consumption of Developed Countries greater than Developing Countries because of which alcohol consumption shows positive correlation with Life Expectancy.




Comparing Data Distribution among Australia, Brazil, Canada, India, Russian Federation, UK, USA

## Deploying Machine Learning and Simple Neural Network for Life Expectancy Prediction

We employed three Machine learning models. And we did train test split on the data. Models we employed are Random Forest, Linear Regression and Deep Neural Network. we used mean absolute error to compare our models.

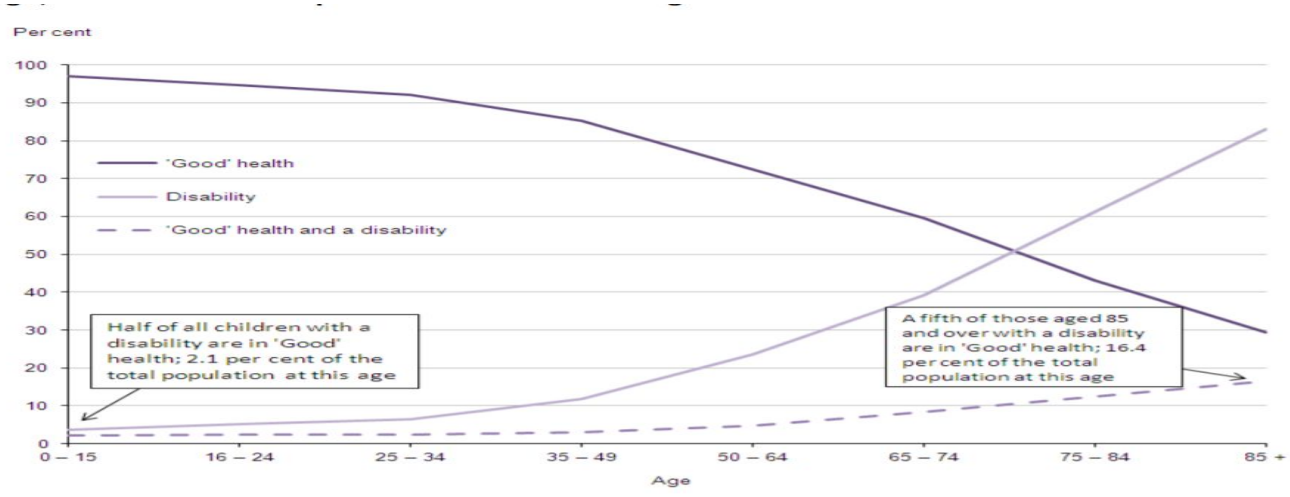
| **Prediction Models** | **MAE** |
|---|---|
| Random Forest | 1.382 |
| Linear Regression | 3.869 |
| Deep Neural Network | 2.071 |

Hence we conclude that **Random forest** works best for our data to predict the life expectancy.

## Consequences of Increasing Life Expectancy

- If we expect to improve quantity rather than quality of life it is harming our economy.
- With the improvement of quantity of life alone has increased the extent of ageing and age-related diseases.

***"…years are being added to our lives, life is not being added to our years: the extra years are being added at the very end of our lives and are of poor quality "***




Percentage of people in England reporting Disability or Good Health at different Ages
(Img Source - *http://www.ons.gov.uk/ons/dcp171776_353238.pdf )*

## Contributions

### Ridam Pal

- Exploratory Data Analysis
- Insights and Visuals from EDA
- Data resourcing and curation
- Feature engineering and Feature selection

### Sai Krishna Vamshi

- Found Insights using Correlation of attributes with Life Expectancy.
- Made a Dashboard on Google Data Studio for Visualization.
- Applied Deep Neural Network for predicting Life Expectancy.

### Jayanth Krishna

- Features Extraction for applying ML using Correlation Matrix and KBest.
- Deploying Machine Learning for predicting Life Expectancy.

**GitHub Link for Code and Data - https://github.com/jayanthkrishna/DSC_Project**

## References


[o] *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4328740/*

[1] *https://www.kaggle.com/kumarajarshi/life-expectancy-who*

[2] *http://teachersinstitute.yale.edu/curriculum/units/1998/7/98.07.02.x.html*

[3] *https://www.coursera.org/specializations/jhu-data-science*

[4] *https://en.wikipedia.org/wiki/Economy_of_the_United_States*