Statistical Analysis of factors influencing Life Expectancy

***Abstract*--Even though many studies have been conducted in the past on factors impacting life expectancy, including demographic characteristics, income composition, and mortality rates, there are still many more to be done. The impact of vaccination and the human development index has hitherto been overlooked.**

***Index Terms*--The World Health Organization's (WHO) Global Health Observatory (GHO) data repository keeps track of health status and any other relevant parameters for all nations. For health data analysis, the data sets are made available to the public. The data on life expectancy and health variables for 193 nations were gathered from the same WHO data repository website, while comparable economic data was gathered from the UN website. Only the most representative critical elements were picked from all categories of health-related factors.**

I. Introduction

IN comparison to the previous 30 years, there has been great progress in the health sector, resulting in an improvement in human mortality rates, especially in developing countries. As a result, for this study, we looked at data from 193 nations from 2000 to 2015. Individual data files were combined into a single data collection. A cursory examination of the data revealed some missing numbers. We discovered no obvious flaws because the data sets came from WHO.

II. Assignment Focus Statement

We will be more over-focused on these three questions as overall analysis shows these three questions gain superiority in our progression of work in this document

1. How do **Infant and Adult mortality rates affect life expectancy**? Click here III. E.
2. Do **densely populated countries tend to have a lower life expectancy**? Click here III. E.
3. What is the **impact of schooling on the lifespan of humans**? Click here III. E.

I. Life Expectancy

II. Adult Mortality

III. Infant Deaths

IV. Alcohol Consumption

V. Percentage Expenditure

VI. Hepatitis B

VII. Measles

VIII. BMI

IX. Under Five Deaths

X. Polio

XI. Total Expenditure

XII. Diphtheria

XIII. HIV AIDS

XIV. GDP

XV. Population Thinness

XVI. Income Composition Of Resources

XVII. Schooling

III. EXPLORATORY DATA ANALYSIS

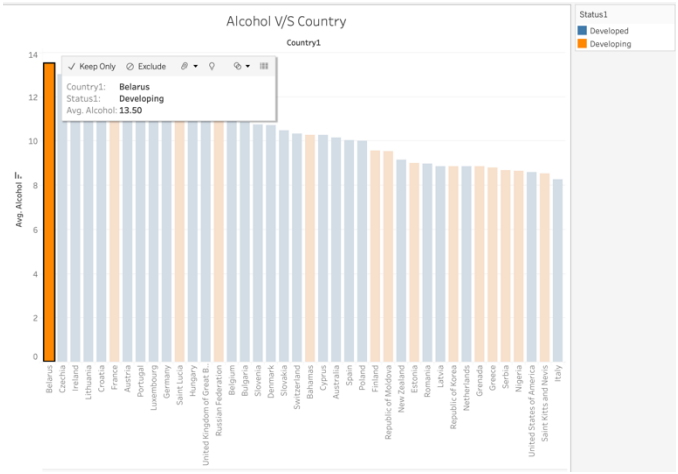
A. Health Factors Affecting Life Expectancy

Many associated factors contribute to affecting the Life Expectancy of people across the whole world.
It can be broadly categorized into three aspects First is Diseases (**Alcohol Consumption, Hepatitis B, Measles, Polio, Diphtheria, HIV AIDS**), and the Second is Population Overall Health such as death rates, body mass index, population thinness (**Infant Deaths, Under Five Deaths, BMI, Population Thinness**) and Lastly by Country's Net Income and Expenditures with education levels also being looked at (**Percentage Expenditure, Total Expenditure, GDP, Income Composition of Resources, Schooling**).

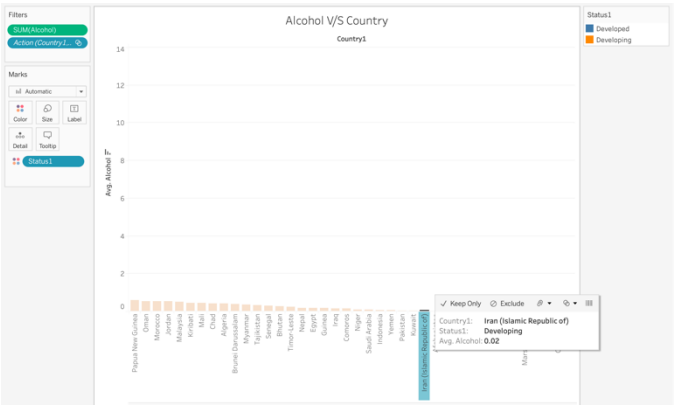
B. Diseases (The Year 2000 to 2015 Data)

- **Alcohol Consumption:**

Mean Max.: 13.50 Country: Belarus

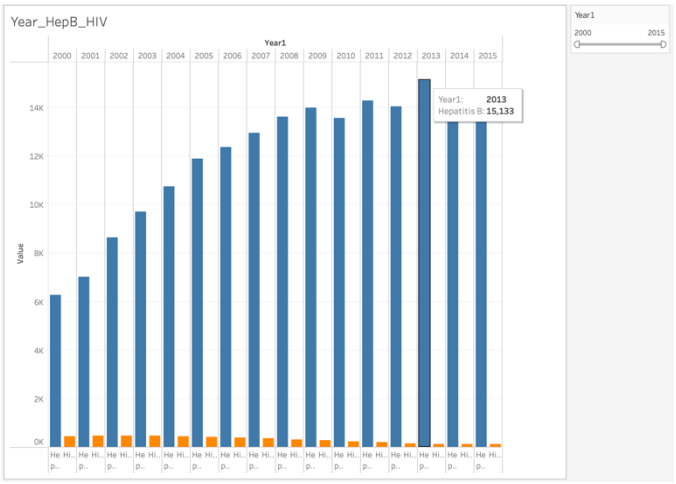


Mean Min.: 0.02 Country: Iran

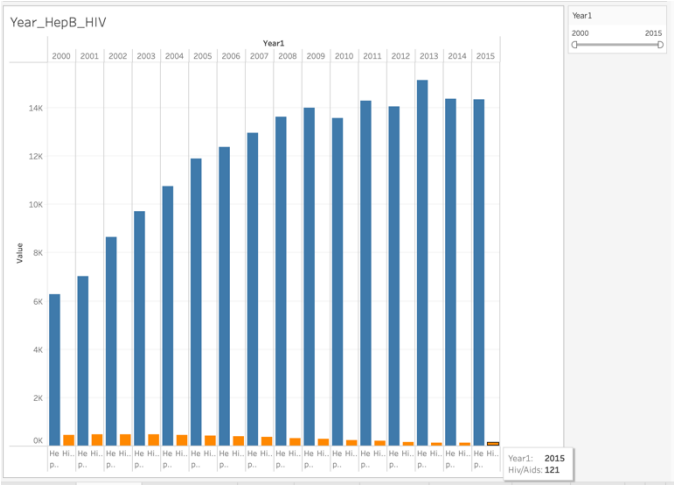


- **Hepatitis B:**

Count Max.:15,133 Year: 2013 Country Status: Developed

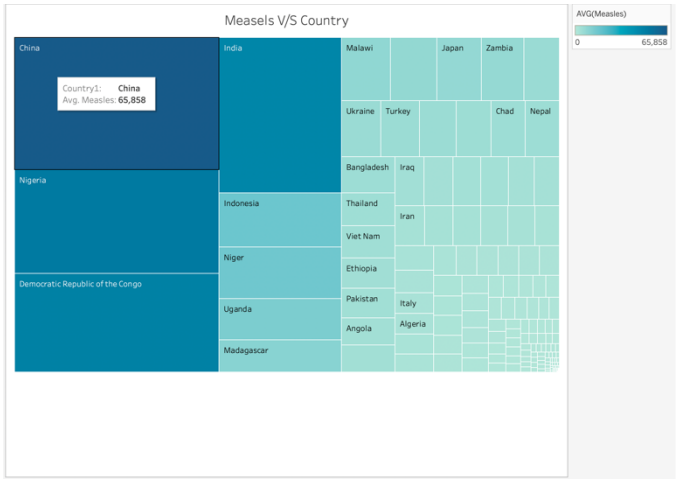


Count Min.: 121 Year: 2015 Country Status: Developing

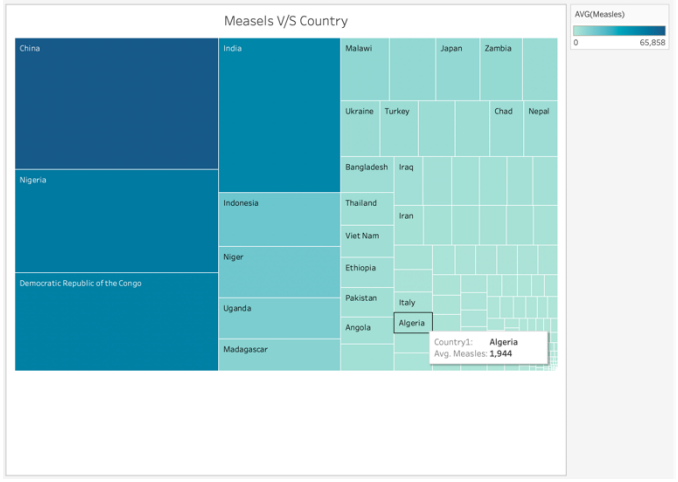


- **Measles:**

Mean Max.: 65,858 Country: China

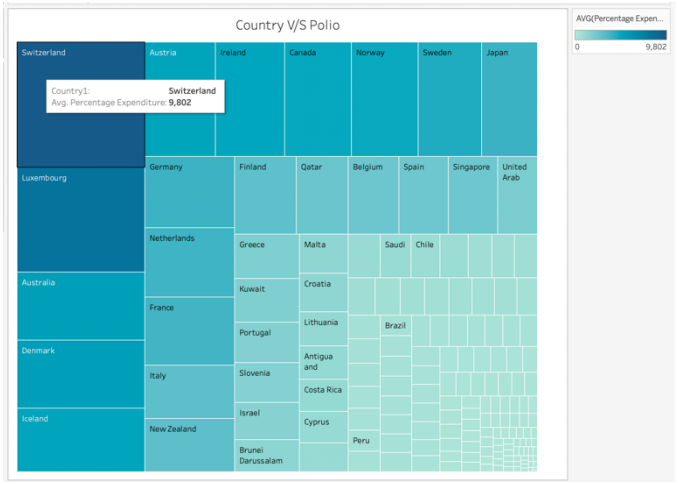


Mean Min.: 1,944 Country: Algeria

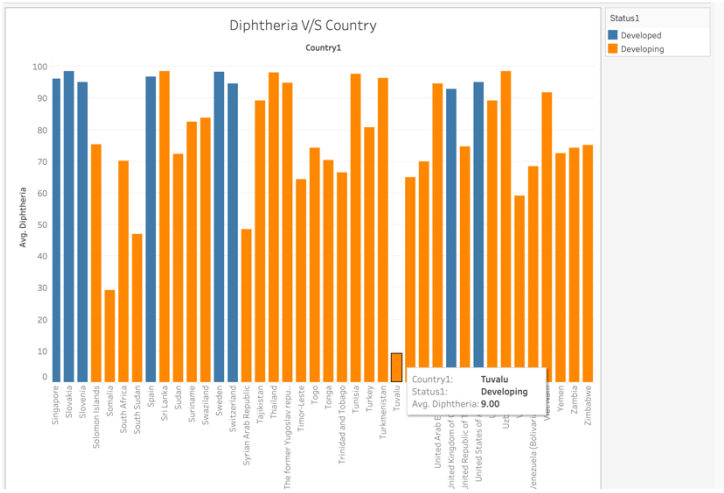


- **Polio:**

Mean Max.: 9,802 Country: Switzerland

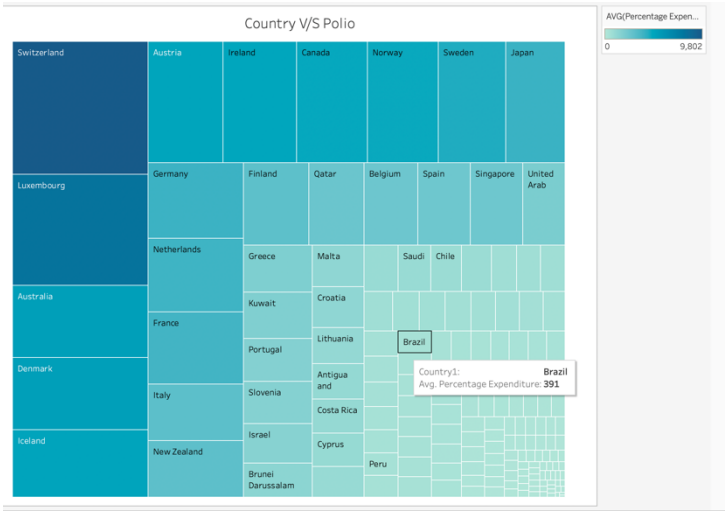


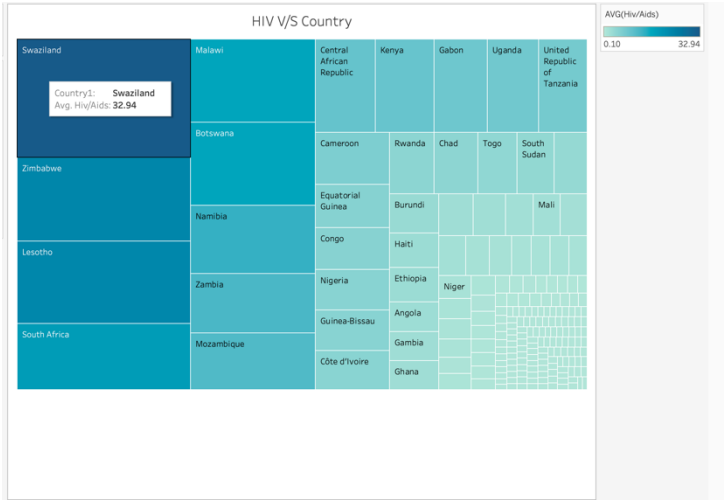
Mean Min.: 391 Country: Brazil



- **HIV AIDS:**

Mean Max.: 32.94 Country: Swaziland

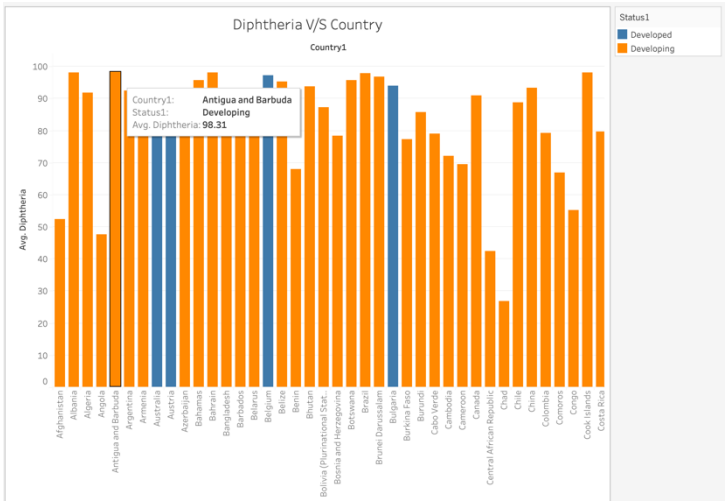


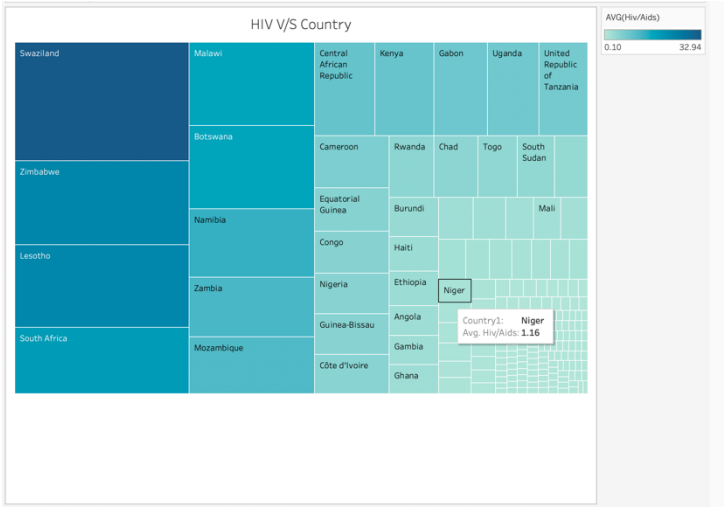


- **Diphtheria:**

Mean Max.: 98.31 Country: Antigua and Barbuda

Mean Min.: 1.16 Country: Niger



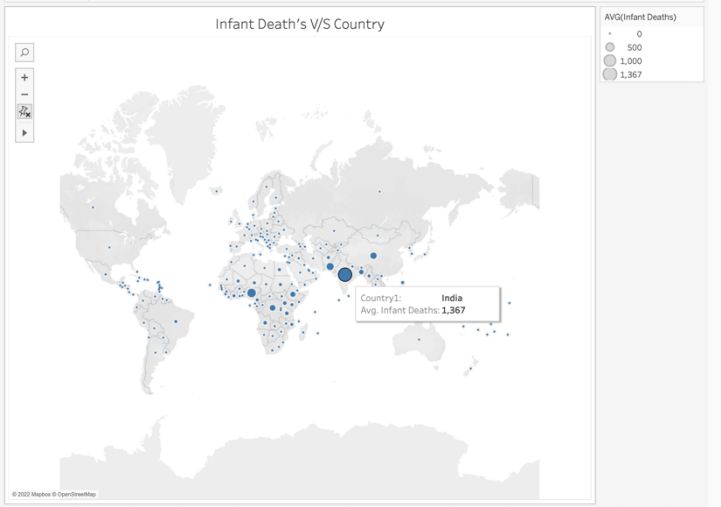


Mean Min.: 9.00 Country: Tuvalu

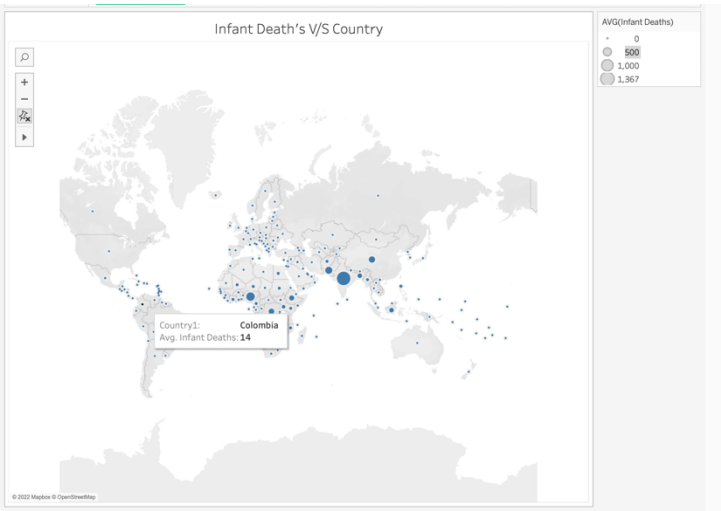
C. Population Over All Health (Diet Associations)

- **Infant Deaths:**

Mean Max.: 1,367 Country: India

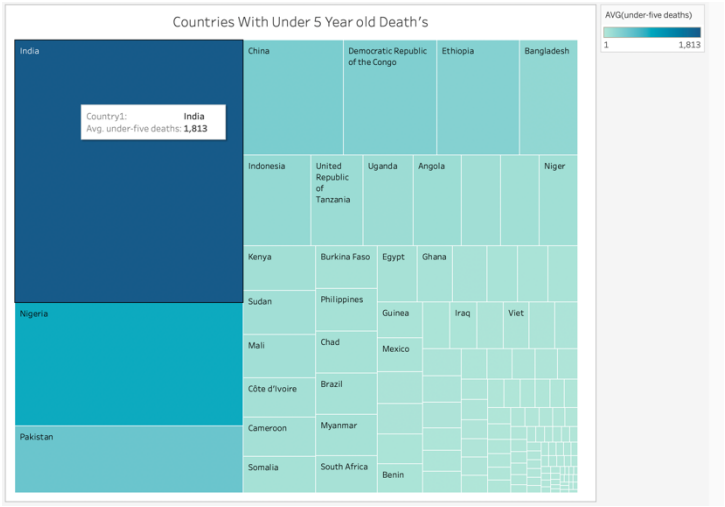


Mean Min.: 14.00 Country: Colombia

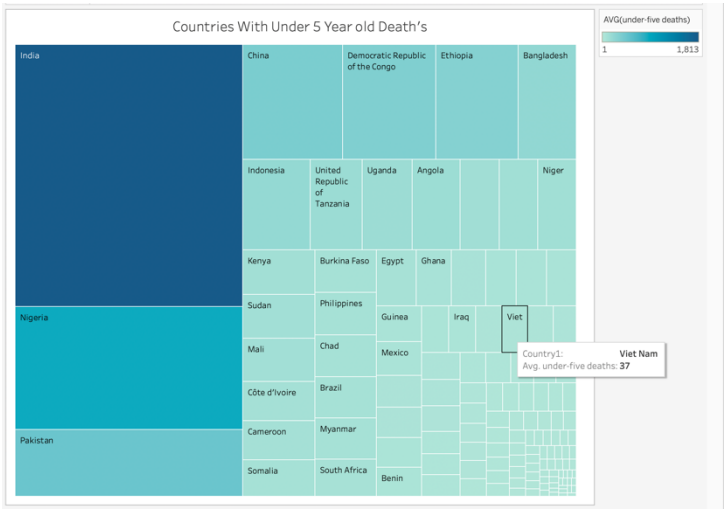


- **Under Five Deaths:**

Mean Max.: 1,813 Country: India

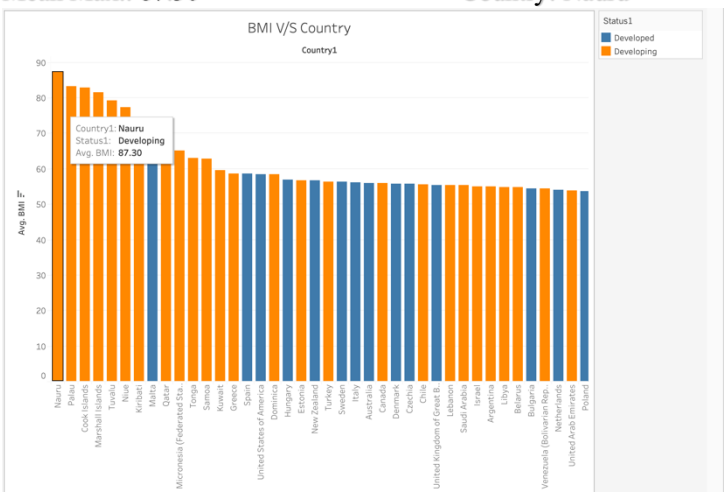


Mean Min.: 14.00 Country: Viet Nam

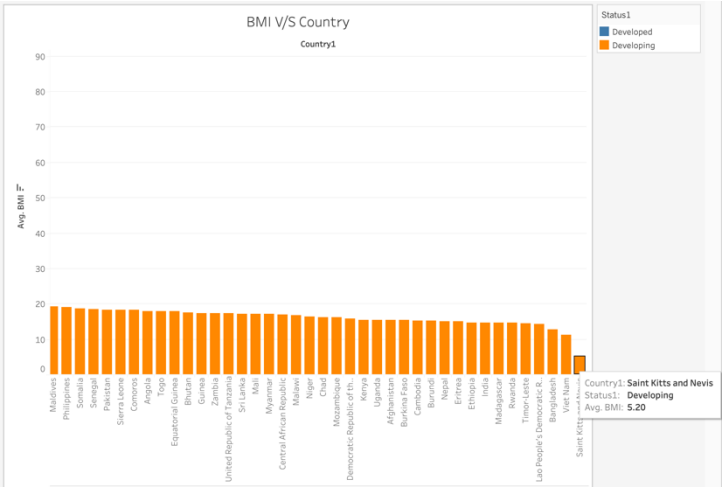


- **BMI:**

Mean Max.: 87.30 Country: Nauru

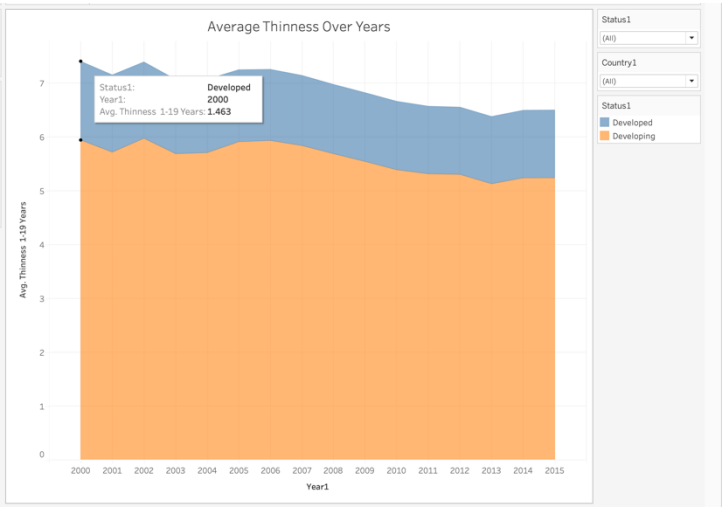


Mean Min.: 5.20 Country: Saint Kitts and Nevis

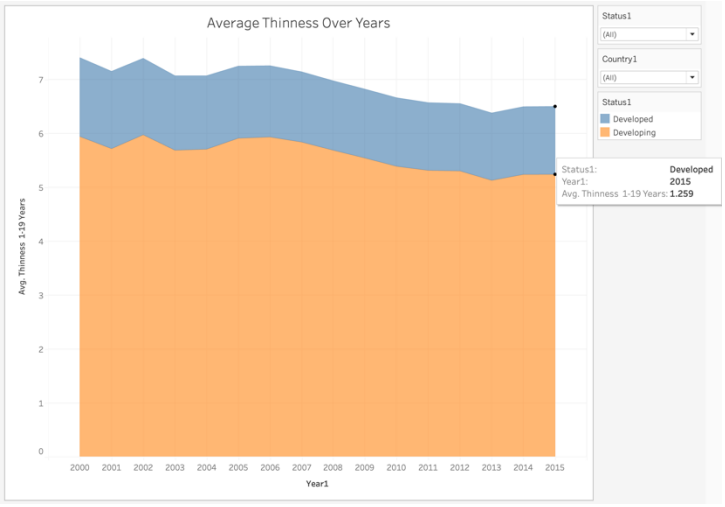


- **Population Thinness:**

Mean Max.: 1.463 Year of Record: 2000



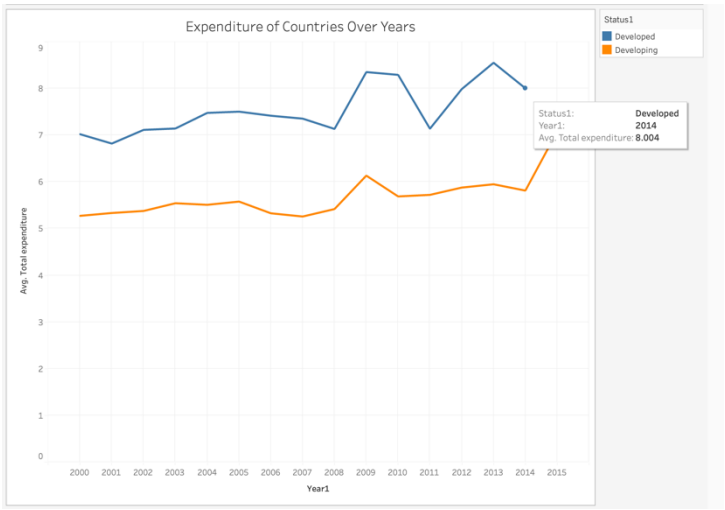
Mean Min.: 1.259 Year of Record: 2015



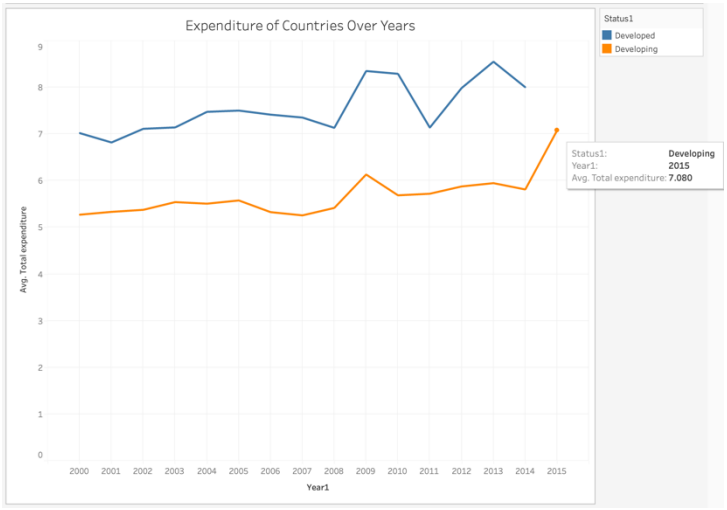
D. Country's Net Income and Expenditures concerning Education increment.

- **Total Expenditure:**

Mean Max.: 8.004 Year of Record: 2014
Country Status: Developed

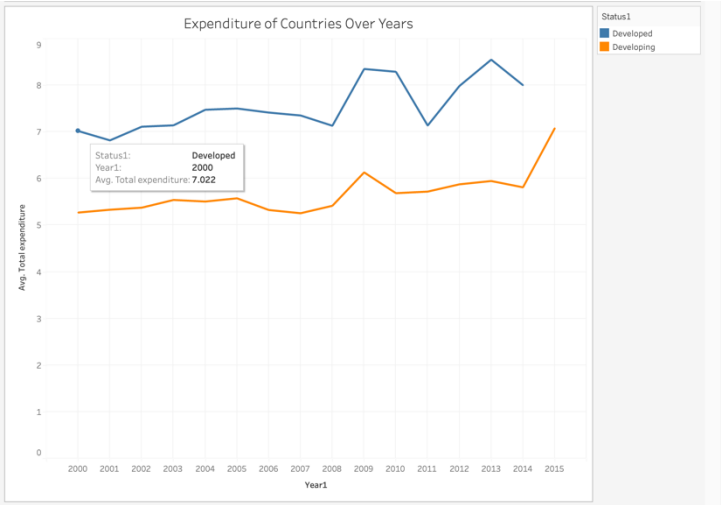


Mean Max.: 1.463 Year of Record: 2015
Country Status: Developing

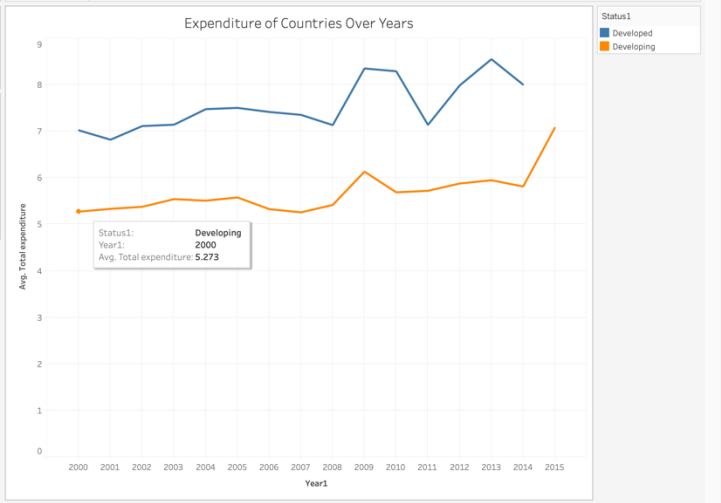


{Please Scroll Down}

Mean Min.: 7.022 Year of Record: 2000
Country Status: Developed

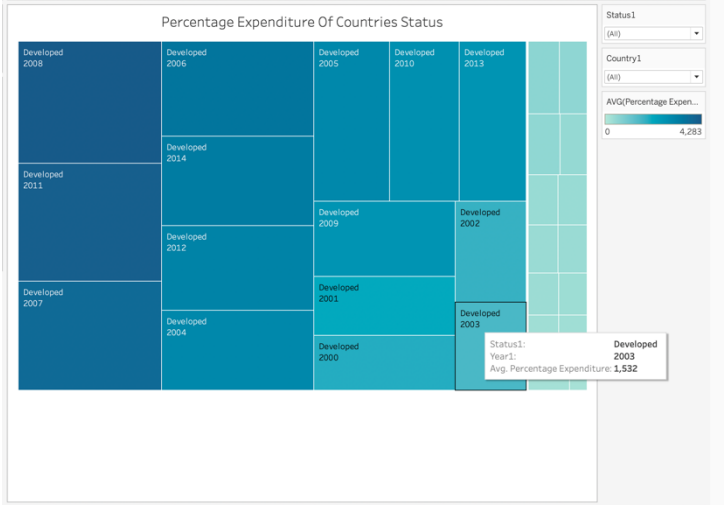


Mean Min.: 7.022 Year of Record: 2000
Country Status: Developing

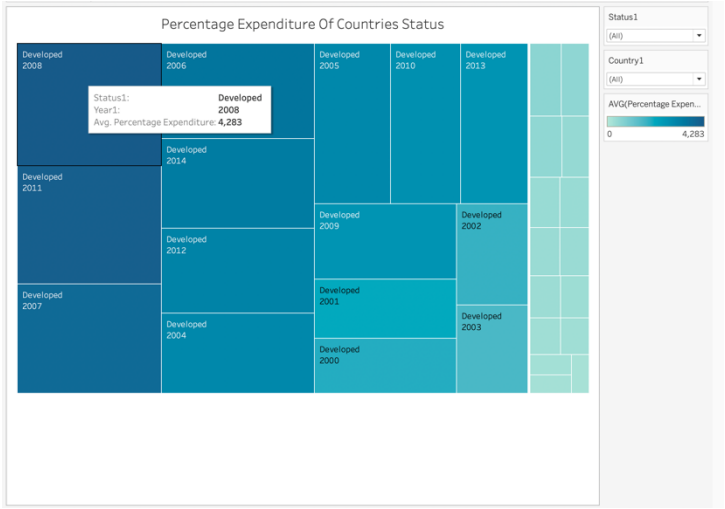


- **Percentage Expenditure:**

Mean Min.: 1,532 Year of Record: 2003
Country Status: Developed

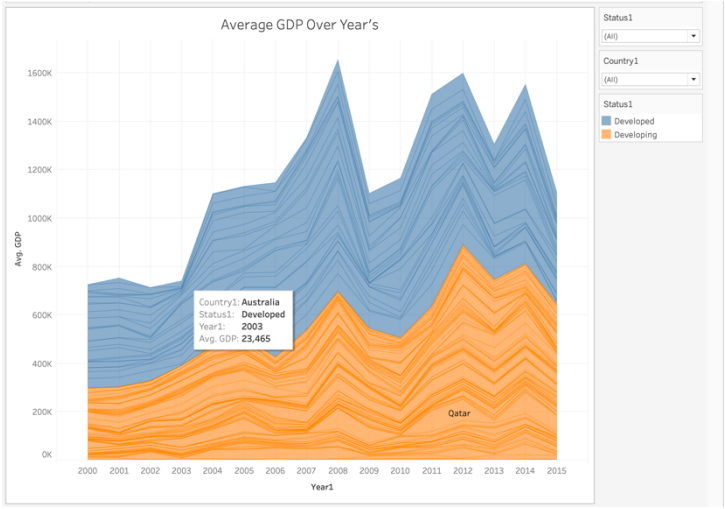


Mean Min.: 1,532 Year of Record: 2008
Country Status: Developed

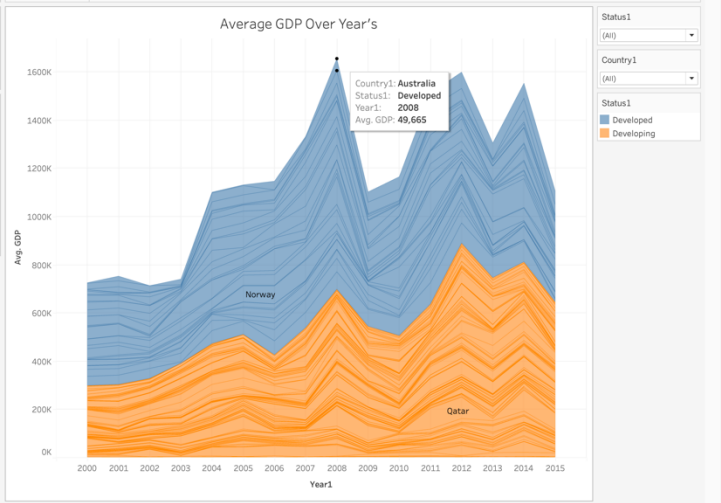


- **GDP:**

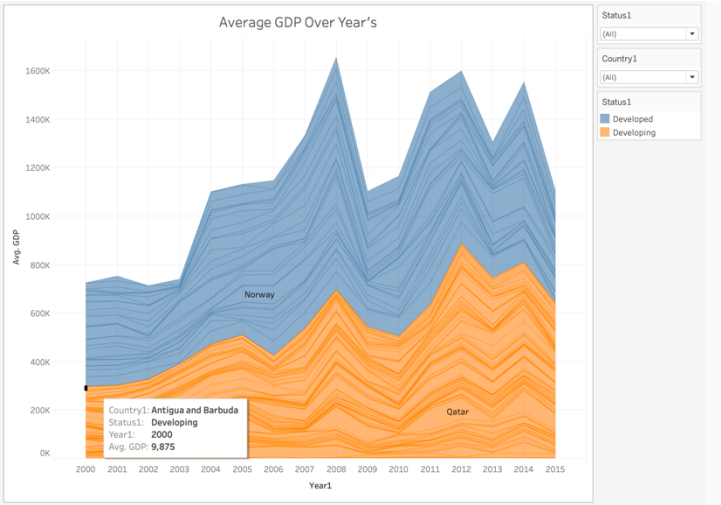
Mean Min.: 23,465 Year of Record: 2003
Country Status: Developed Country: Australia



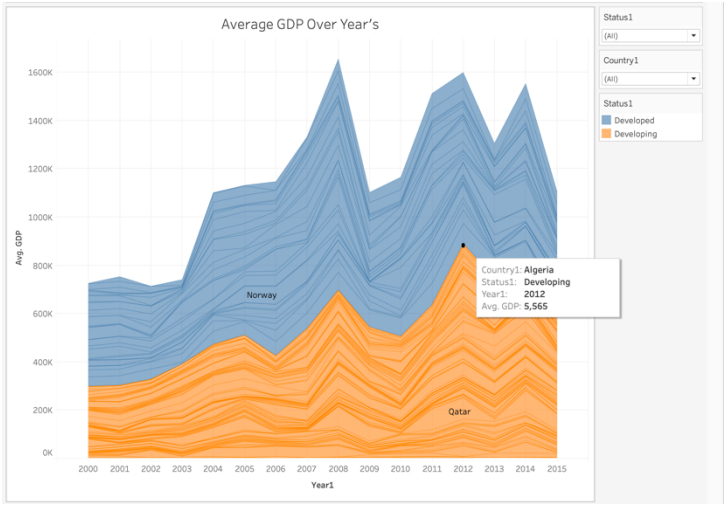
Mean Max.: 49,655 Year of Record: 2008
Country Status: Developed Country: Australia



Mean Min.: 9,875 Year of Record: 2000
Country Status: Developing Country: Antigua and Barbuda

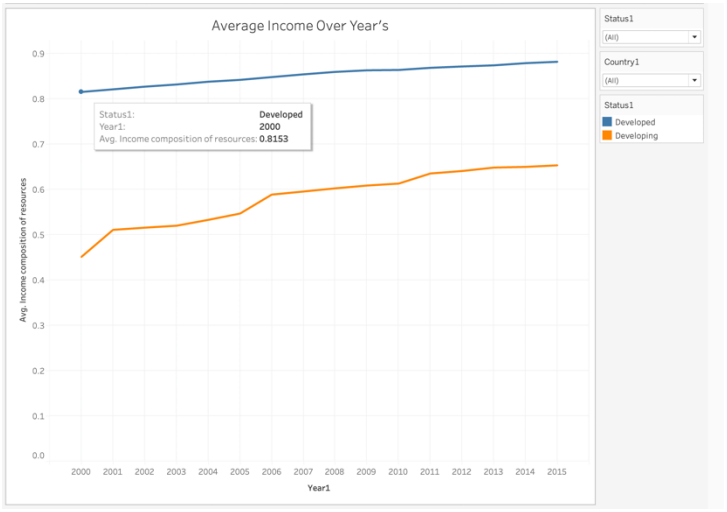


Mean Max.: 5,565 Year of Record: 2012
Country Status: Developing Country: Algeria

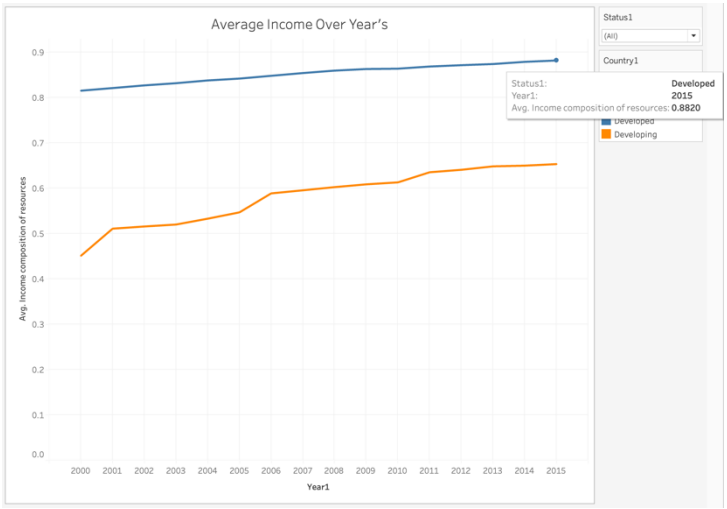


- **Income Composition of Resources:**

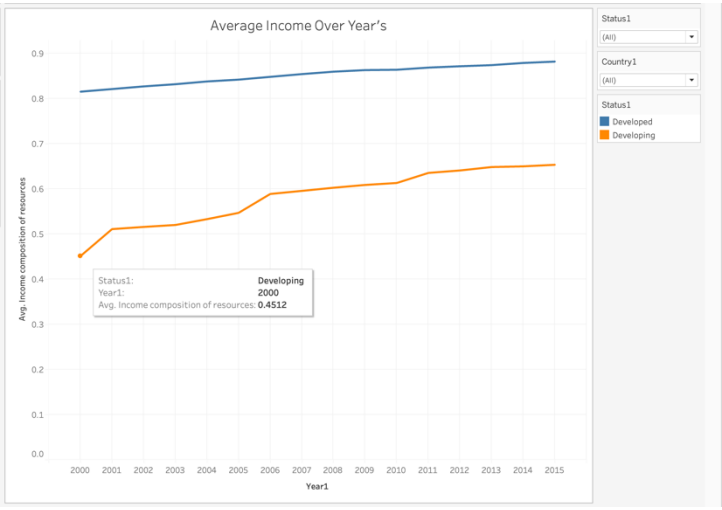
Mean Min.: 0.815 Year of Record: 2000
Country Status: Developed

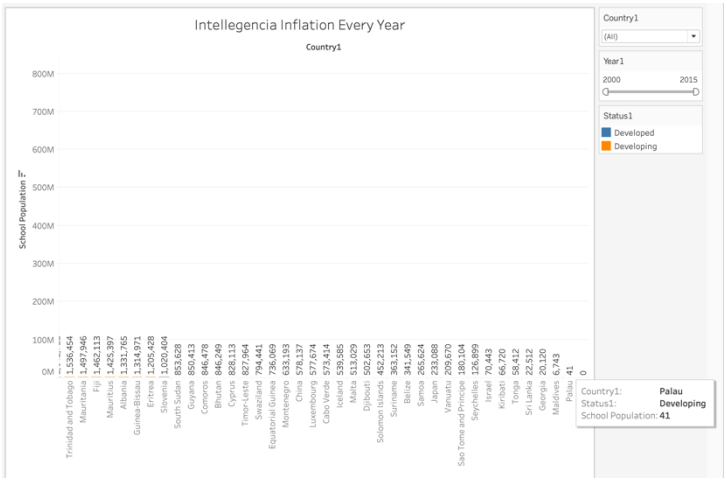


Mean Max.: 0.882 Year of Record: 2015
Country Status: Developed

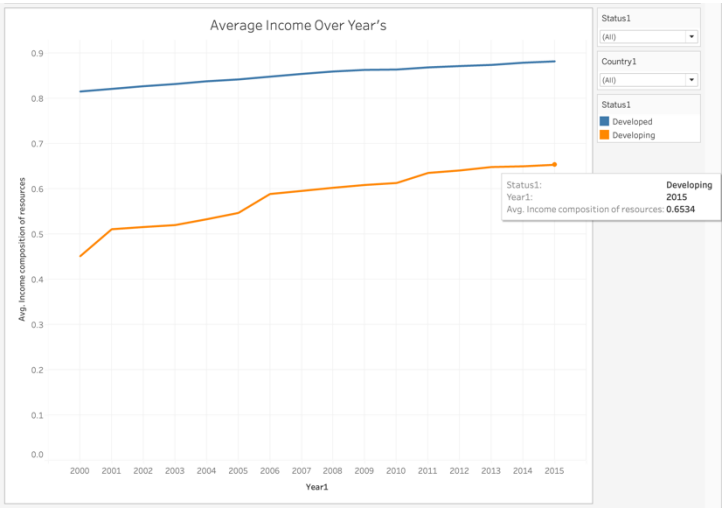


Mean Min.: 0.451 Year of Record: 2000 Count Max.: 41 Country Status: Developing Country: Palau
Country Status: Developing



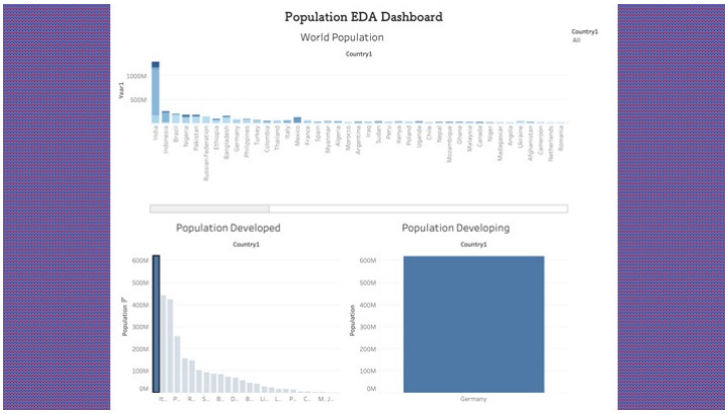


Mean Min.: 0.653 Year of Record: 2015
Country Status: Developing



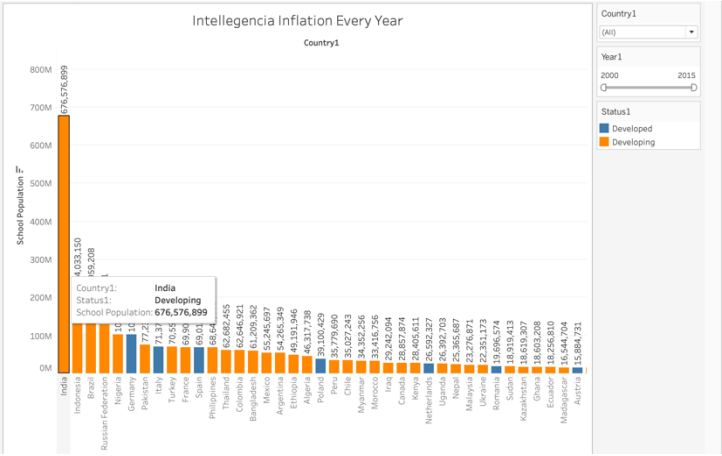
E. Dashboard Comparisons of many different aspects associated

Population Data

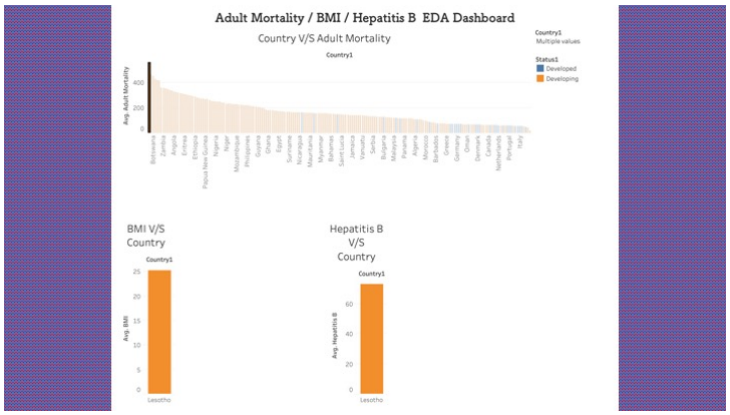


- **Schooling:**

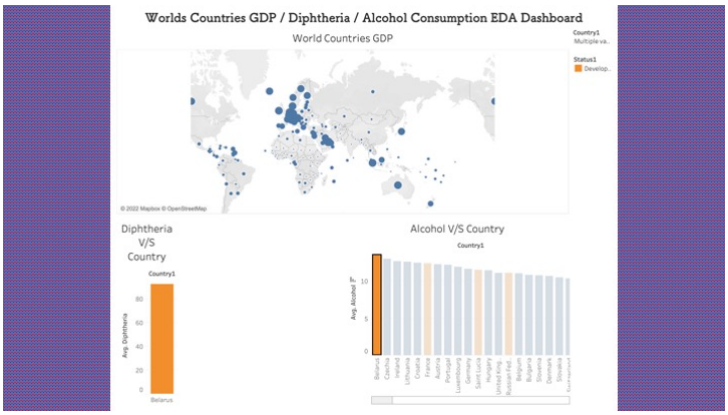
Count Max.: 676,576,899 Country Status: Developing
Country: India



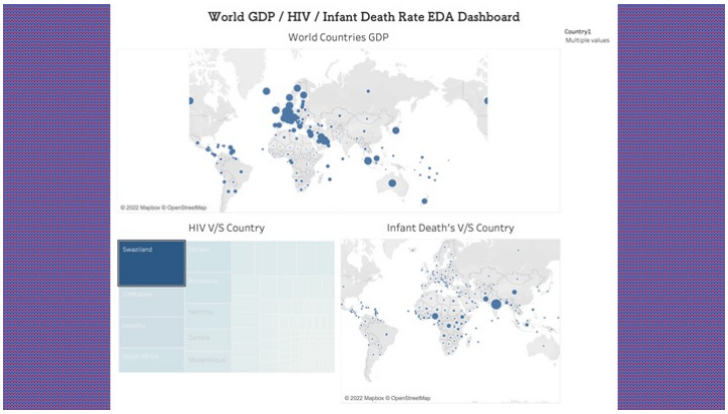
Adult Mortality affected by Hepatitis B & BMI



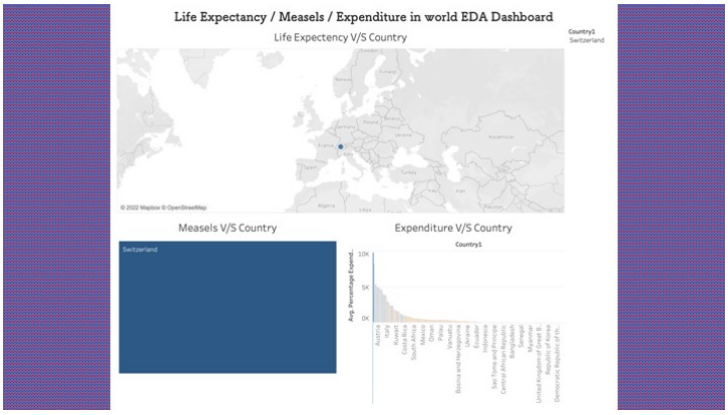
A country's GDP is affected by Alcohol consumption and Diphtheria



World GDP affected by HIV & Infant Death

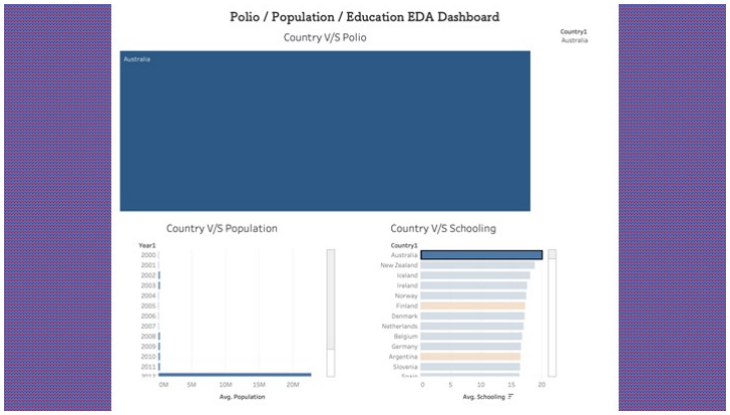


Main Factors That are responsible for good life expectancy

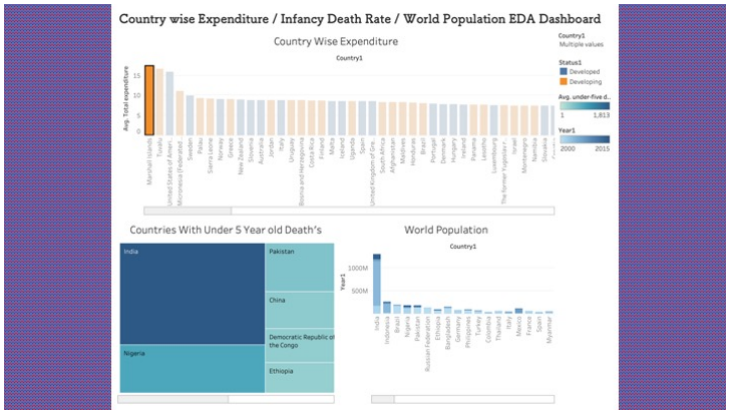


<# IF A COUNTRY IS SPENDING ITS FINANCES ON ITS OWN COUNTRY, THEREBY IT ENABLES INCREASE IN LIFE OF ITS RESIDING CITIZEN, ALSO A FIRM CONTROL ON MEASLES WILL FURTHER AID THE CAUSE OF KEEPING GOOD HEALTH OF PEOPLE #>

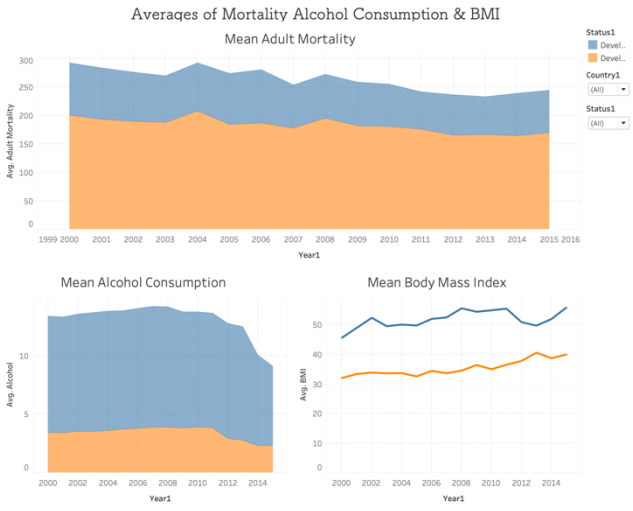
Effects of Polio and Educated Population



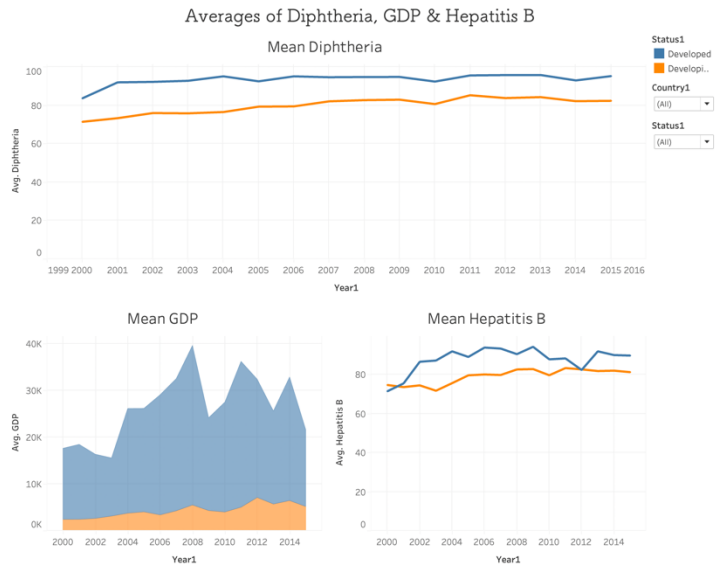
World Population with Infancy Death and Country's Expenditure



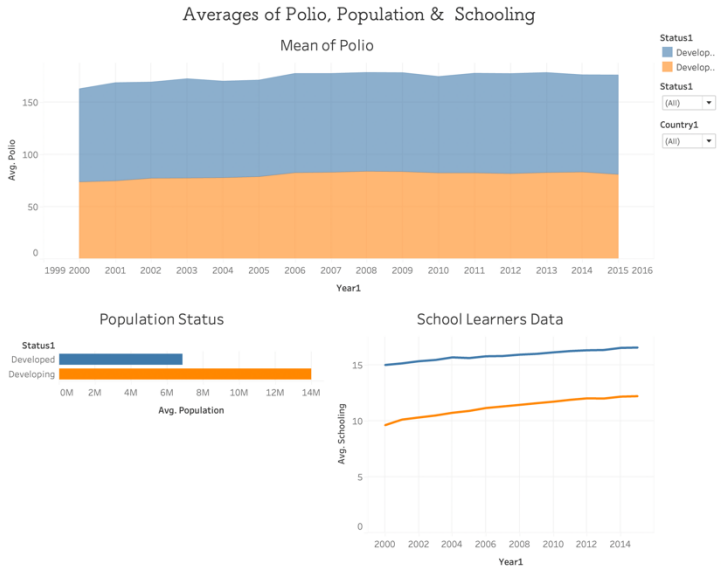
Averages of Mortality Alcohol Consumption & BMI



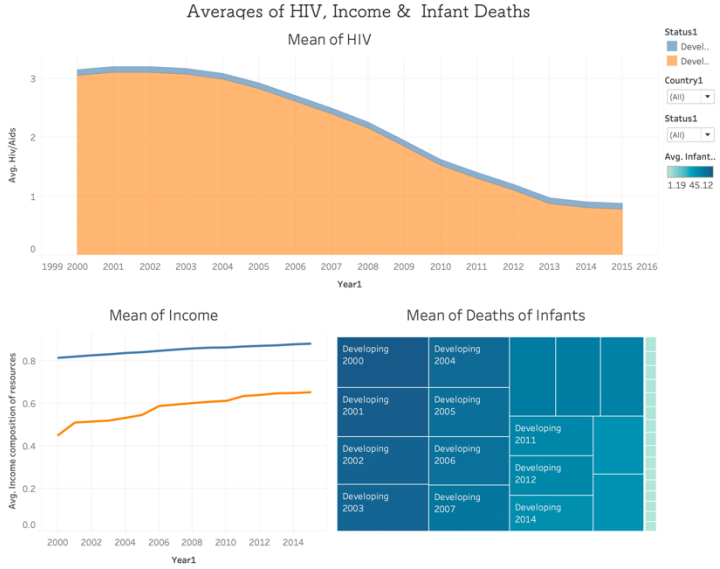
Averages of Diphtheria, GDP & Hepatitis B



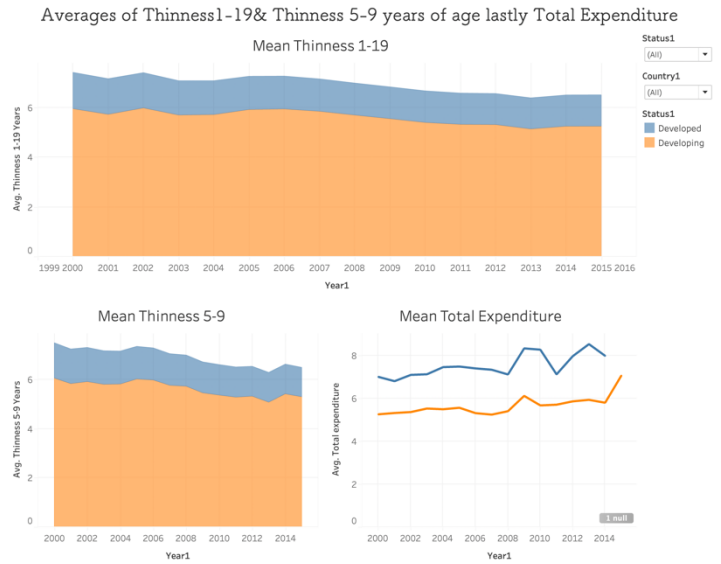
Averages of Polio, Population & Schooling



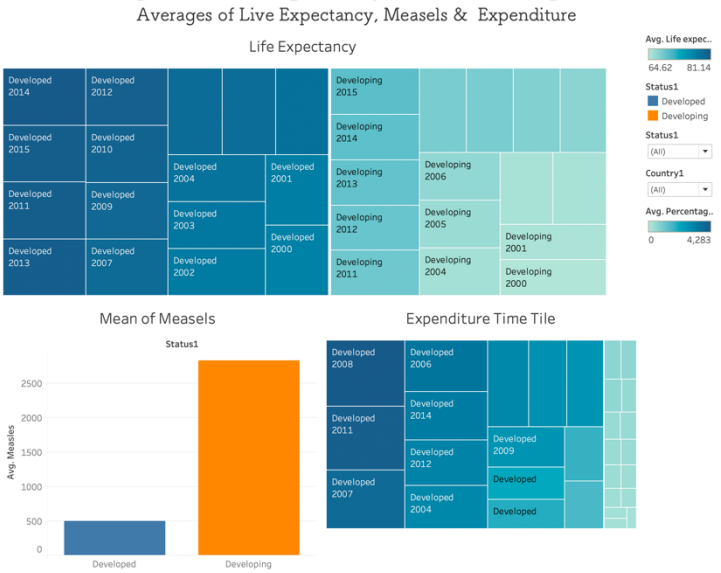
Averages of HIV, Income & Infant Deaths



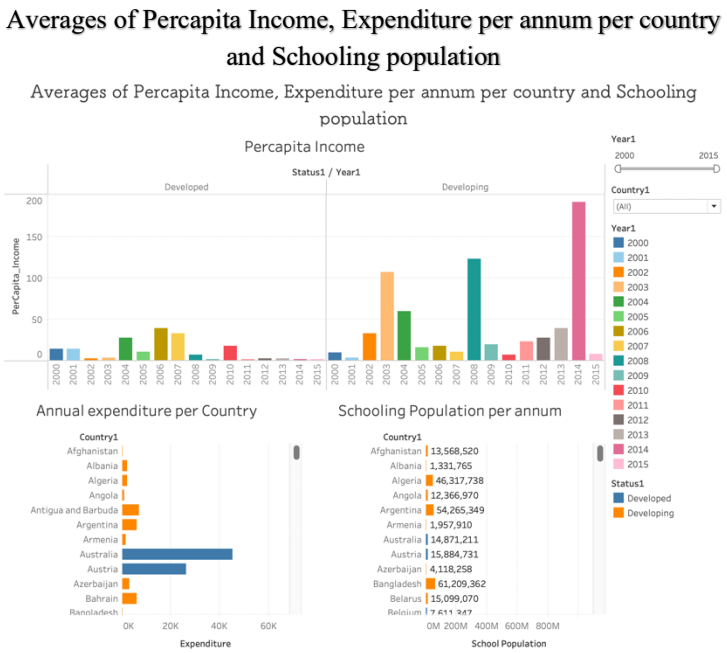
Averages of Thinness1-19& Thinness 5-9 years of age lastly Total Expenditure

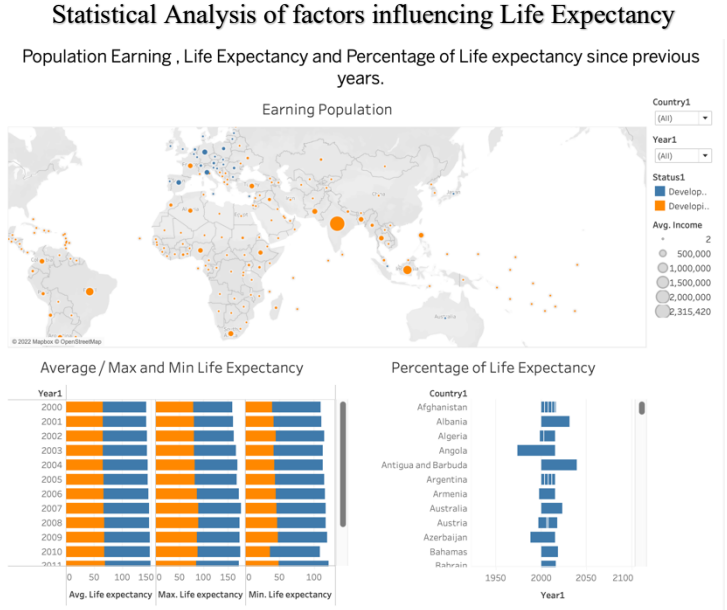


Averages of Live Expectancy, Measles & Expenditure



{Please Scroll Down}





F. Predictive Analysis

Loading libraries and reading file

```
library(dplyr)

## Warning: package 'dplyr' was built under R version 4.0.5

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(purrr)
library(tidyr)

## Warning: package 'tidyr' was built under R version 4.0.5

library(data.table)

## Warning: package 'data.table' was built under R version 4.
0.5

##
## Attaching package: 'data.table'

## The following object is masked from 'package:purrr':
##
##     transpose

## The following objects are masked from 'package:dplyr':
##
##     between, first, last

library(caret)

## Warning: package 'caret' was built under R version 4.0.5

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.0.
5

## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift

library(tidyverse)

## -- Attaching packages --------------------------------------- tid
yverse 1.3.1 --

##  v tibble  3.1.6      v stringr 1.4.0
## v readr   2.1.2      v forcats 0.5.1

## Warning: package 'tibble' was built under R version 4.0.5

## Warning: package 'readr' was built under R version 4.0.5

## Warning: package 'forcats' was built under R version 4.0.5

## -- Conflicts ------------------------------------------ tidyverse_
conflicts() --
##  x data.table::between()  masks dplyr::between()
##  x dplyr::filter()        masks stats::filter()
##  x data.table::first()    masks dplyr::first()
##  x dplyr::lag()           masks stats::lag()
```

```
##   x   data.table::last()                masks   dplyr::last()
##   x   caret::lift()                    masks   purrr::lift()
## x data.table::transpose() masks purrr::transpose()

dat<- read.csv("Life Expectancy Data.csv")
```

Checking for NA values and removing them

```
sum(is.na(dat))

## [1] 2563

#There   are   2563   NA   values.Removing   them.

dat1<-                                  na.omit(dat)
#Now no NA values in the dataset.
```

Lets consider Life expectancy to be our target variable.

Using Density Plot To Check If Response Variable Is Close To Normal

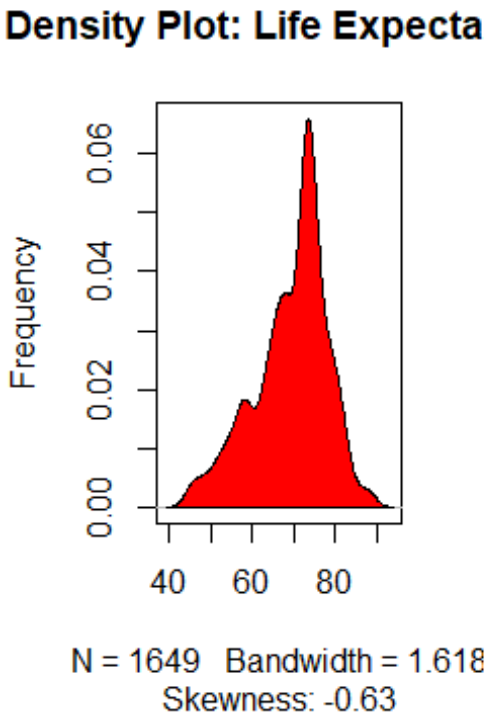
```
library(e1071)  # for skewness function

## Warning: package 'e1071' was built under R version 4.0.5

par(mfrow=c(1,  2))   #  divide  graph  area  in  2  columns

plot(density(dat1$Life.expectancy), main="Density Plot: Life
Expectancy", ylab="Frequency", sub=paste("Skewness:", rou
nd(e1071::skewness(dat1$Life.expectancy), 2)))  # density pl
ot  for  'Life  Expectancy'

polygon(density(dat1$Life.expectancy), col="red")
```



1. Removing the variables not to be used and partitioning the data into the Test and Training set.

```
dat2<-                                  dat1[c(-1,-2,-3)]

set.seed(123)
training.samples   <-   dat2$Life.expectancy   %>%
 createDataPartition(p   =   0.8,   list   =   FALSE)
train.data          <-          dat2[training.samples,          ]
test.data <- dat2[-training.samples, ]
```

80% of the data is partitioned into a training set and 20% in the Test set.

2.Fitting the model

```
model<-   lm(Life.expectancy~.,   data   =   train.data)

summary(model)

##
##                                                  Call:
## lm(formula = Life.expectancy ~ ., data = train.data)
##
##                                              Residuals:
##      Min       1Q   Median       3Q      Max
##  -16.919   -2.080    0.048    2.214   12.216
##
##                                           Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             5.283e+01  8.201e-01  64.413  < 2
e-16                                                      ***
## Adult.Mortality         -1.668e-02  1.054e-03 -15.827
<                         2e-16                           ***
## infant.deaths            9.153e-02  1.146e-02   7.988 3.0
0e-15                                                     ***
## Alcohol                 -5.669e-02  3.483e-02  -1.628   0.1
038
## percentage.expenditure        4.188e-04  1.939e-04   2.15
9                                  0.0310                 *
## Hepatitis.B              -3.392e-03  4.988e-03  -0.680   0.
4966
## Measles                 -3.756e-06  1.276e-05  -0.294   0.7
685
## BMI                      3.331e-02  6.746e-03   4.937 8.98e
-07                                                       ***
## under.five.deaths          -6.904e-02  8.291e-03  -8.328  <
2e-16                                                     ***
## Polio                    6.365e-03  5.928e-03   1.074   0.28
31
## Total.expenditure             8.030e-02  4.541e-02   1.768
0.0772                                                    .
## Diphtheria               1.347e-02  6.739e-03   1.999   0.
0458                                                      *
## HIV.AIDS                 -4.383e-01  1.969e-02 -22.266  <
2e-16                                                     ***
```

```
## GDP                              5.865e-06  3.015e-05   0.195   0.8
458
## Population                      -8.152e-10  1.888e-09  -0.432   0.
6660
## thinness..1.19.years             7.772e-03  5.620e-02   0.138
0.8900
## thinness.5.9.years              -6.494e-02  5.553e-02  -1.169
0.2424
## Income.composition.of.resources 9.442e+00 9.187e-01 1
0.277                          <          2e-16          ***
## Schooling                        9.370e-01  6.495e-02  14.426  <
2e-16                                                      ***
##                                                           ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.59 on 1301 degrees of freedom
## Multiple R-squared:  0.8369, Adjusted R-squared:  0.8346
## F-statistic: 370.8 on 18 and 1301 DF,  p-value: < 2.2e-16
```

Conclusion- (a) From the model summary we can get that the R square value is 83%, which is quite good. Also, the p-value is less than significant at 0.05. (b) The output shows the variables impacting the model. The variables having a p-value less than 0.05 are highly significant ones. (c) Therefore, removing the insignificant ones.

Model Update

```
model_updated<- lm(Life.expectancy~Adult.Mortality+infan
t.deaths+percentage.expenditure+          under.five.deaths+
                Total.expenditure+ Diphtheria+ HIV.AIDS+ Inc
ome.composition.of.resources+Schooling, data = train.data)

summary(model_updated)

##
##                                                      Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + infant.
deaths                                                    +
##     percentage.expenditure + under.five.deaths + Total.exp
enditure                                                  +
##     Diphtheria + HIV.AIDS + Income.composition.of.resou
rces                                                      +
##                         Schooling,     data     =     train.data)
##
##                                                   Residuals:
##        Min         1Q     Median         3Q        Max
##  -17.9747    -2.0848     0.0807     2.3567    12.0196
##
##                                                  Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 5.278e+01  6.866e-01  76.878  < 2
e-16                                                      ***
## Adult.Mortality             -1.755e-02  1.055e-03 -16.630
<                          2e-16                          ***
## infant.deaths                8.996e-02  1.064e-02   8.455  < 2
e-16                                                      ***
## percentage.expenditure       4.336e-04  6.225e-05   6.96
6                          5.16e-12                       ***
## under.five.deaths           -6.941e-02  7.911e-03  -8.774  <
2e-16                                                     ***
## Total.expenditure            1.003e-01  4.558e-02   2.200
0.028                                                     *
## Diphtheria                   1.274e-02  5.177e-03   2.462   0.
014                                                       *
## HIV.AIDS                    -4.486e-01  1.978e-02 -22.685  <
2e-16                                                     ***
## Income.composition.of.resources  1.007e+01  9.027e-01  1
1.160                          <          2e-16          ***
## Schooling                    1.005e+00  5.967e-02  16.842  <
2e-16                                                     ***
##                                                          ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.638 on 1310 degrees of freedom
## Multiple R-squared:  0.8314, Adjusted R-squared:  0.8302
## F-statistic: 717.6 on 9 and 1310 DF,  p-value: < 2.2e-16
```

Now all variables are highly significant.

2. Predicting the with test set-

```
# Predict test data based on model
predict_reg <- model_updated %>% predict(test.data)

linear_rmse<-                                        data.frame(
  RMSE = caret::RMSE(predict_reg, test.data$Life.expectanc
y),
  Rsquare = caret::R2(predict_reg, test.data$Life.expectancy)
)

sigma(model_updated)/mean(test.data$Life.expectancy)

## [1] 0.05233031
```

1. We can see that the RMSE value is low i.e 3.689 and the R square is high i.e 81.8%, meaning the model is good.
2. Also, the average prediction error rate is 5%.

THE FINAL CONCLUSION OF THE PREDICTIVE ANALYSIS SHARED THAT OUR TEST, TRAINING AND MODEL CREATION IS DONE SUCCESSFULLY WITH 81.8%

IV. Acknowledgement

I Rohin Mehra, Student of Griffith College South Circular Road (author of the report) gratefully acknowledge the contributions of professor Dr Waseem Akhtar, For his persistent motivation and always making tough decisions so simple to take that it makes learner's knowledge and also embed students with new innovative ideas in their careers.

V. References

References are mostly taken from WHO servers and Kaggel the link for the same areas follow.

1. https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who?datasetId=12603&language=R
2. https://www.who.int/data/gho/data/themes/topics/indicator-groups/indicator-group-details/GHO/life-expectancy-and-healthy-life-expectancy
3. https://www.who.int/data/gho/info/gho-odata-api