Final Project Data Analysis

Jeyaraman Ramalingam

5/5/2021

Contents

Data Exploration

Child Mortality Data

Sample

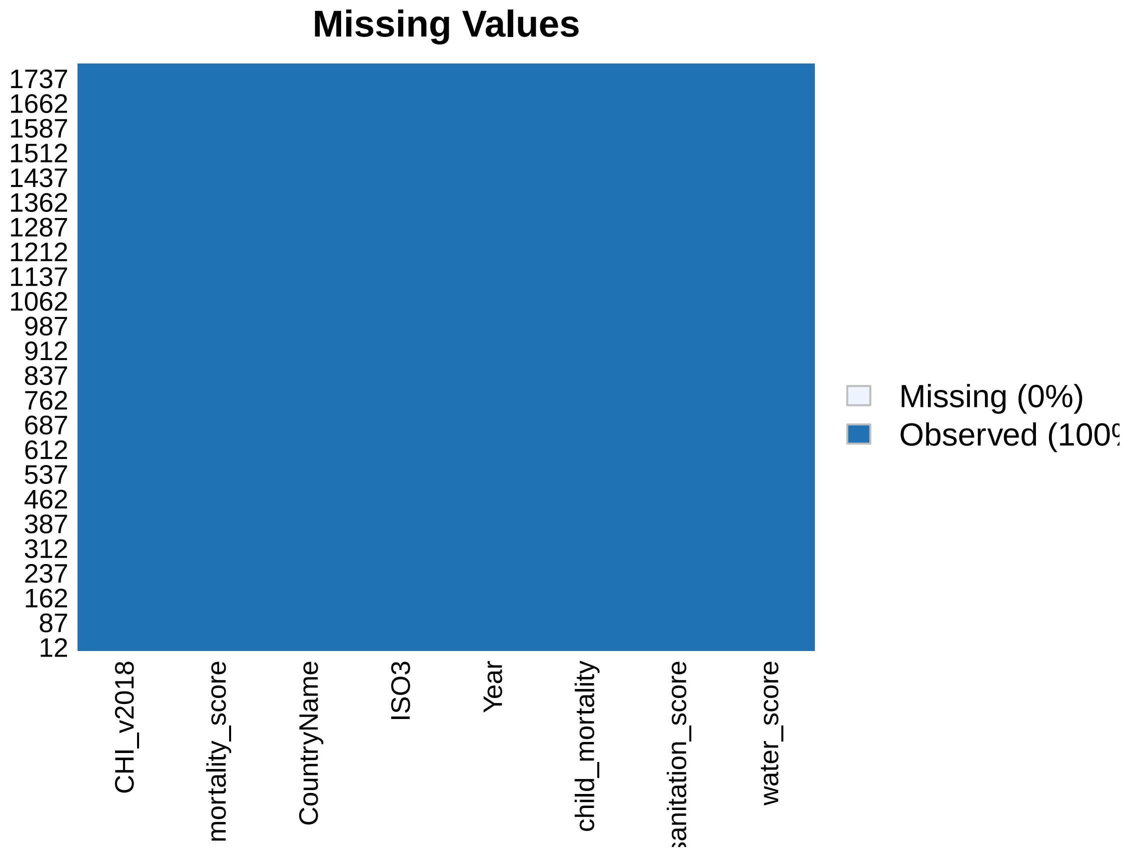
water_score	sanitation_score	child_mortality	Year	ISO3	CountryName	mortality_score	CHI_v2018
50.05	33.56	16.65	10	AFG	Afghanistan	75.80	53.14
52.62	34.67	15.39	11	AFG	Afghanistan	76.83	54.71
55.20	35.79	14.65	12	AFG	Afghanistan	77.10	56.03
57.79	36.93	14.42	13	AFG	Afghanistan	76.58	57.10
60.38	38.07	14.64	14	AFG	Afghanistan	75.31	57.92
62.98	39.22	15.13	15	AFG	Afghanistan	73.80	58.67

Input Dataset Summaries

```
##   water_score     sanitation_score child_mortality      Year          
##  Min.   : 18.14   Min.   :  5.69   Min.   : 0.390   Length:1782       
##  1st Qu.: 77.80   1st Qu.: 49.29   1st Qu.: 1.310   Class :character  
##  Median : 94.75   Median : 88.36   Median : 3.285   Mode  :character  
##  Mean   : 86.05   Mean   : 73.80   Mean   : 9.902                     
##  3rd Qu.: 99.30   3rd Qu.: 97.61   3rd Qu.:13.000                     
##  Max.   :100.00   Max.   :100.00   Max.   :68.810                     
##      ISO3            CountryName        mortality_score   CHI_v2018    
##  Length:1782        Length:1782        Min.   : 0.00   Min.   :17.28  
##  Class :character   Class :character   1st Qu.:78.96   1st Qu.:66.38  
##  Mode  :character   Mode  :character   Median :94.43   Median :92.81  
##                                        Mean   :83.68   Mean   :81.18  
##                                        3rd Qu.:97.82   3rd Qu.:97.78  
##                                        Max.   :99.30   Max.   :99.74  
```

Missing Data Check This test is to ensure there are no missing values ('NA') in the dataset and if there are missing values exist in the data then it has to imputed in the later stage.

```
##      water_score sanitation_score  child_mortality             Year 
##                0                0                0                0 
##             ISO3      CountryName  mortality_score        CHI_v2018 
##                0                0                0                0 
```



The missmap clearly shows that there are no missing values in the dataset.

Findings

The findings from Data Exploration on Training and Evaluation dataset are below.

1. There are no missing data
2. Year column needs to be fixed
3. Year Column need to be converted to Numeric Data type

We will perform all of these exercises in the Data Preparation step.

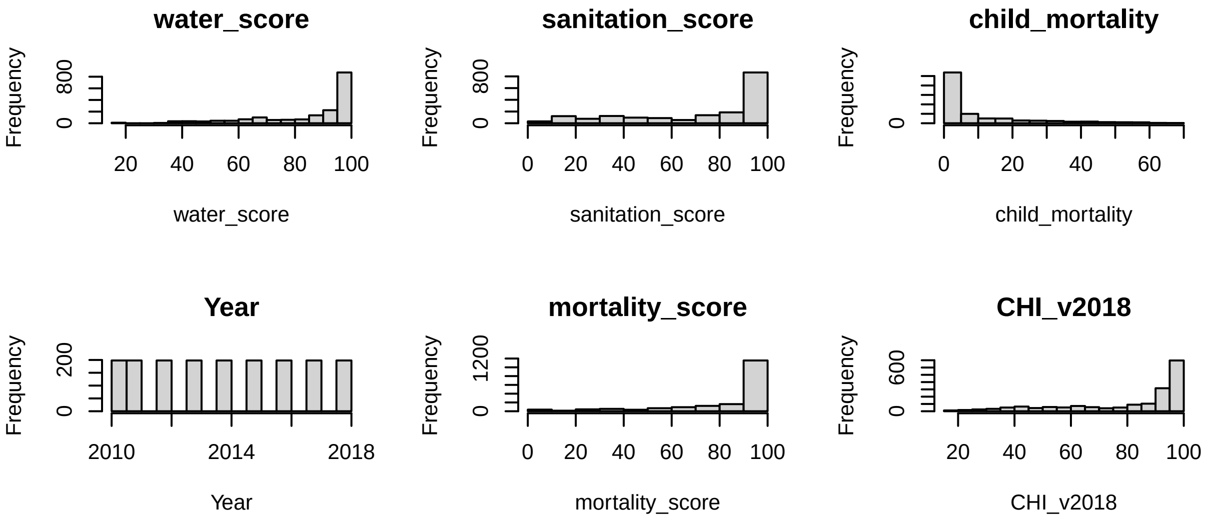
Data Preparation

Child Mortality Data - Fix Formatting

As per the data dictionary provided by the data source the Year field consists of two digit values which is equivalent to four digit Year values. The values like ‘10’ are mapped to ‘2010’. Hence it is necessary to have cleaner format before we proceed with the analysis and modeling of the data. Also the Year column was in character format at datasource and it is converted to Numeric format.

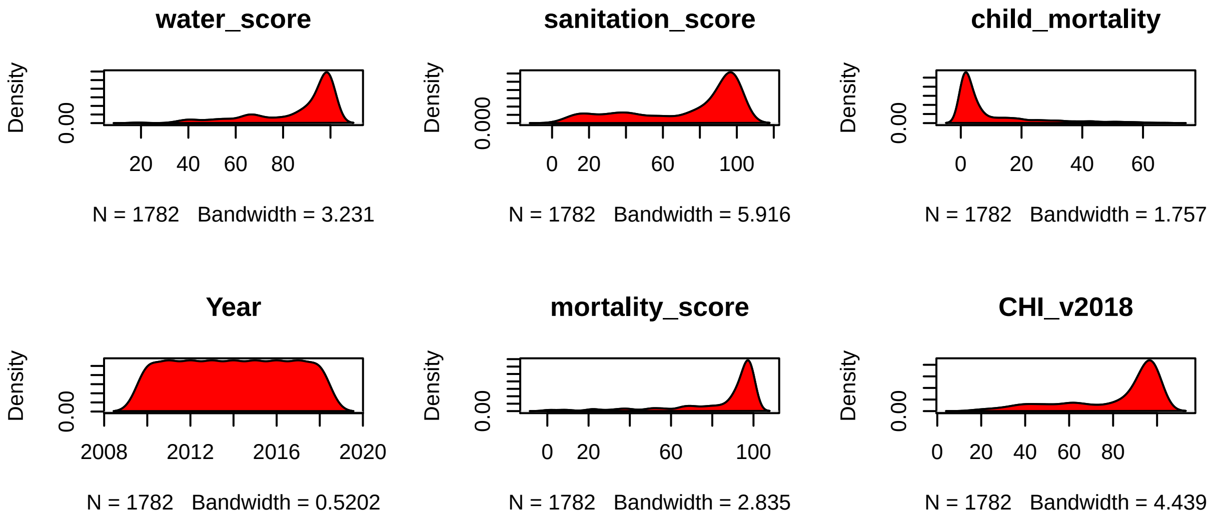
Child Mortality Data - Histograms

The histograms shows that the water, sanitation, mortality scores are left skewed and the child mortality is right skewed. Usually , the skewed datasets need to be fixed before proceeding with the models but the data consists of Yearly snapshot of scores and there wont be lot of derivation/data transformation with the source data. Hence we can proceed with the source dataset without any transformation / fix.



Child Mortality Data - Density Plots

The Density Plots clearly illustrates the conclusion we arrived by looking at histograms.



Training Data - Skewness Report

This is a programmatic skewness test used to justify the conclusion we made earlier. The values of skewness and function are below.

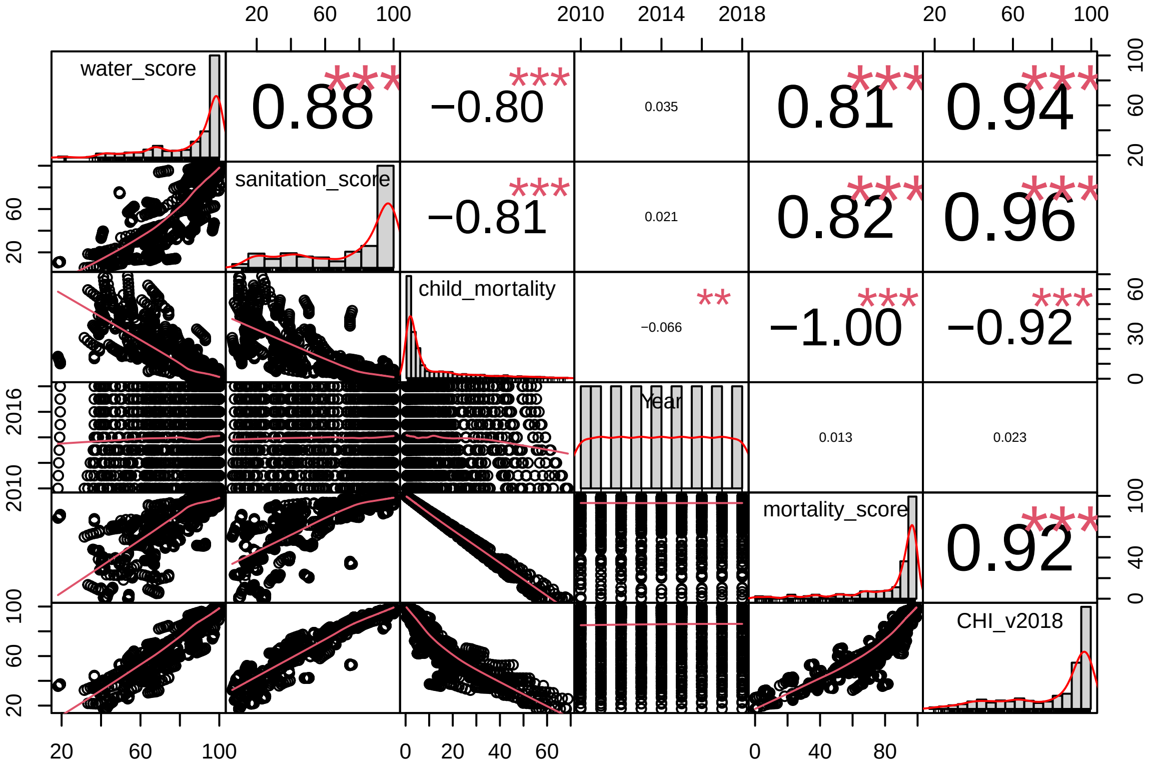
1. water_score < 0 (Left Skewed)
2. sanitation_score < 0 (Left Skewed)
3. child mortality < 0 (Right Skewed)
4. mortality score < 0 (Left Skewed)
5. CHI v2018 < 0 (Left Skewed)

Training Data - Correlation Report

The correlation matrix is used to investigate the dependence between multiple variables at the same time. The result is a table containing the correlation coefficients between each variable and the others.

```
##                  water_score sanitation_score child_mortality        Year
## water_score       1.00000000       0.87614019     -0.80464073  0.03488539
## sanitation_score  0.87614019       1.00000000     -0.81263941  0.02095698
## child_mortality  -0.80464073      -0.81263941      1.00000000 -0.06604295
## Year              0.03488539       0.02095698     -0.06604295  1.00000000
## mortality_score   0.80544486       0.81557616     -0.99559988  0.01302549
## CHI_v2018         0.93740836       0.96242438     -0.92189410  0.02323020
##                    mortality_score  CHI_v2018
## water_score             0.80544486  0.9374084
## sanitation_score        0.81557616  0.9624244
## child_mortality        -0.99559988 -0.9218941
## Year                    0.01302549  0.0232302
## mortality_score         1.00000000  0.9249347
## CHI_v2018               0.92493470  1.0000000
```

```
##                    water_score sanitation_score child_mortality        Year
## water_score                 NA        0.0000000     0.000000000 0.141005329
## sanitation_score     0.0000000               NA     0.000000000 0.376615901
## child_mortality      0.0000000        0.0000000              NA 0.005286959
## Year                 0.1410053        0.3766159     0.005286959          NA
## mortality_score      0.0000000        0.0000000     0.000000000 0.582667817
## CHI_v2018            0.0000000        0.0000000     0.000000000 0.327047725
##                    mortality_score CHI_v2018
## water_score              0.0000000 0.0000000
## sanitation_score         0.0000000 0.0000000
## child_mortality          0.0000000 0.0000000
## Year                     0.5826678 0.3270477
## mortality_score                 NA 0.0000000
## CHI_v2018                0.0000000        NA
```



As per the Correlation matrix the columns which are significant to child_mortality are water_score ,mortality_score, sanitation_score and CHI_v2018. This concludes that the source dataset is good for consideration of modeling.

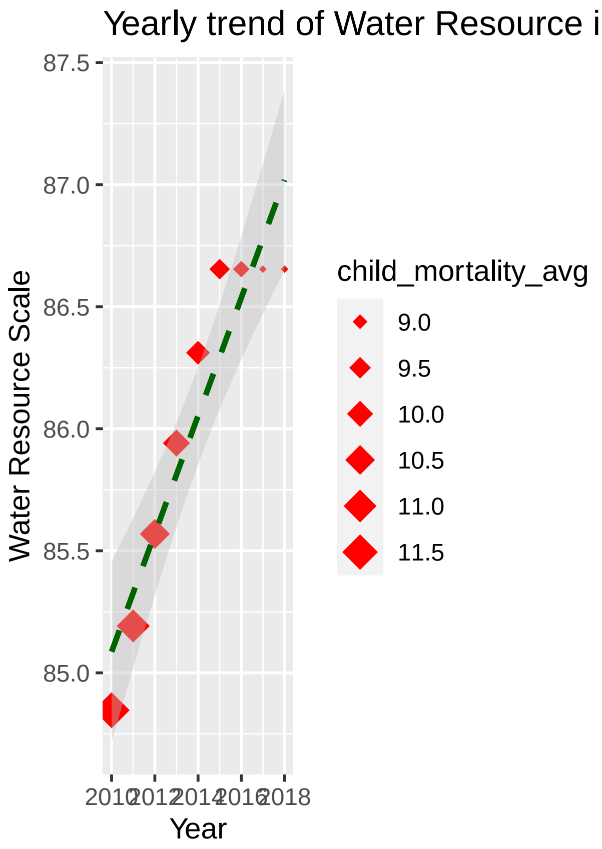
Exploratory Data Analysis

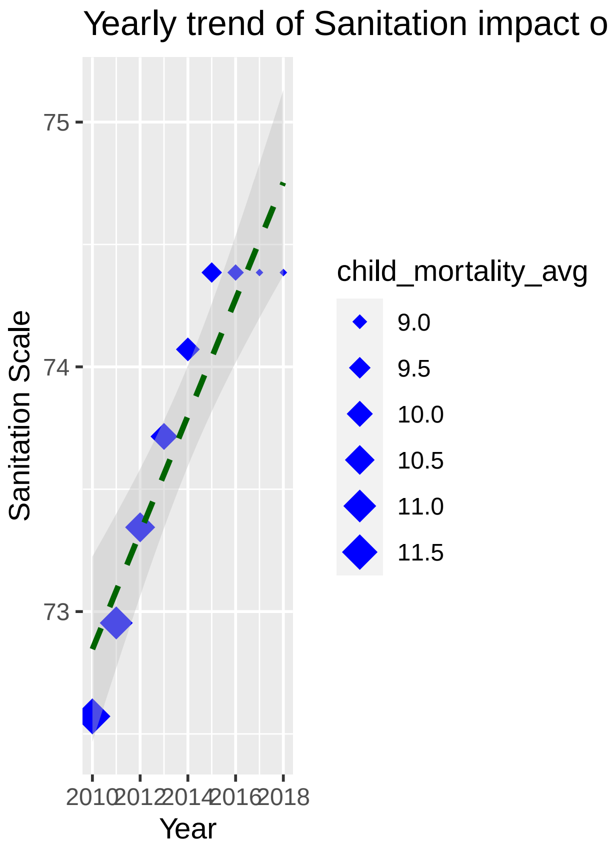
Yearly Impact on Mortality - Water Resource and Sanitation

This Analysis is to understand the dataset a little more.

The Objectives are below.

1. To get to know how much impact water resource and sanitation has on the child mortality
2. Get the Yearly trend of the impact
3. Understand how water resources and Sanitation levels are maintained globally over the years





As per the two plots the following observations are made.

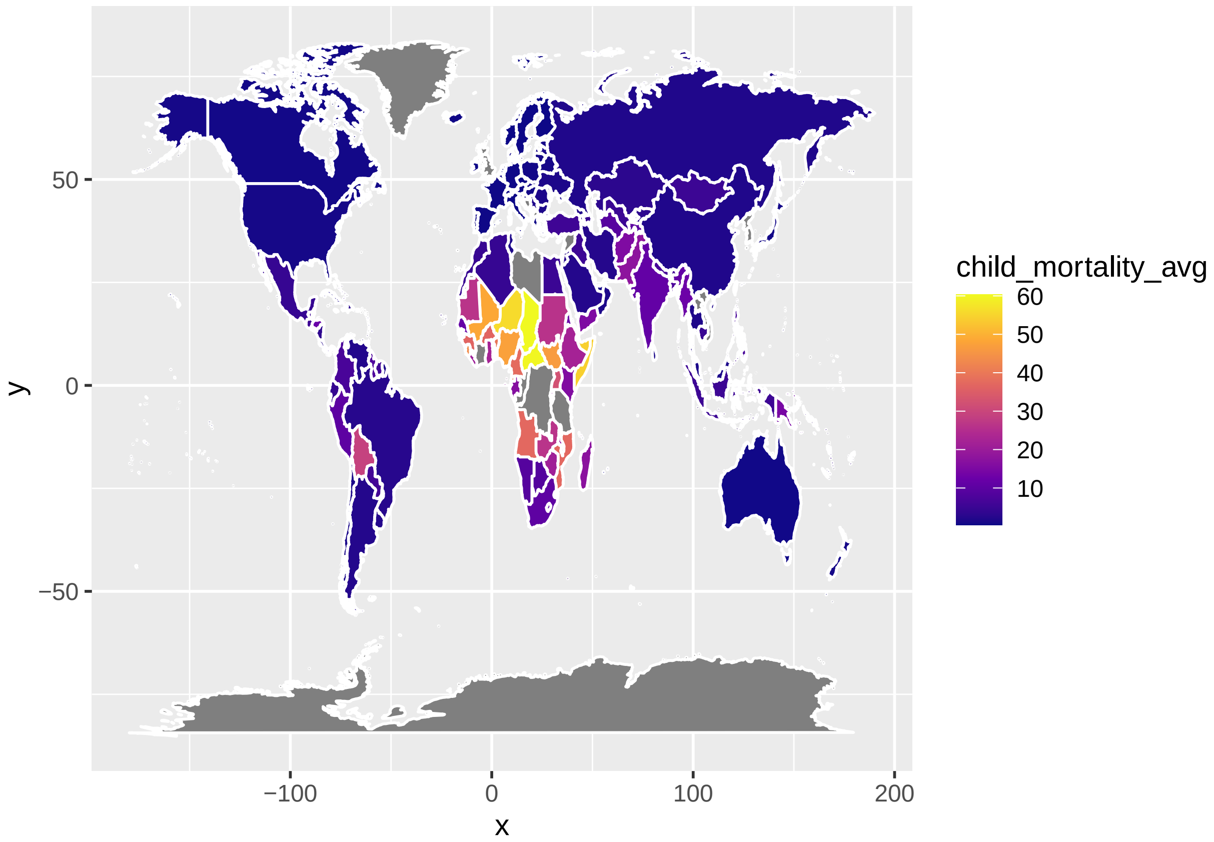
1. Globally the Water resources have been in the increasing trend which means we were able to create and leverage existing water resources in the world.

2. Sanitation levels are also in the rising trend as more awareness and investments towards sanitation has improved overall global sanitation levels.

3. As the water resources and sanitation levels improve the child mortality is in the decreasing trend which concludes the strong correlation betwee the Child Mortality and Water resources/Sanitation levels.

World Map - Child Mortality Impacted Countries

The next step of the analysis is to find out which areas are severely impacted in the world. If those areas are identified then it will be useful to invest more in the countries which has severe impact.



The above world map with impact shows clearly that the African countries as well as few south american countries are severely impacted by child mortality . This could be possibly due to the lack of water resources and poor sanitation practices in those countries.