Executive summary and recommendations (with cross-refs)

After performing and analysing results starting from Data reparation and exploration, k-NN classification and plotting datasets using Google maps and finally at end applying multiple regression modelling using R ,we have built the predictive model to predict the Life Expectancy at birth which is major factor to be considered to improve the quality of health. The final Predictive Model we got is as per mentioned below.

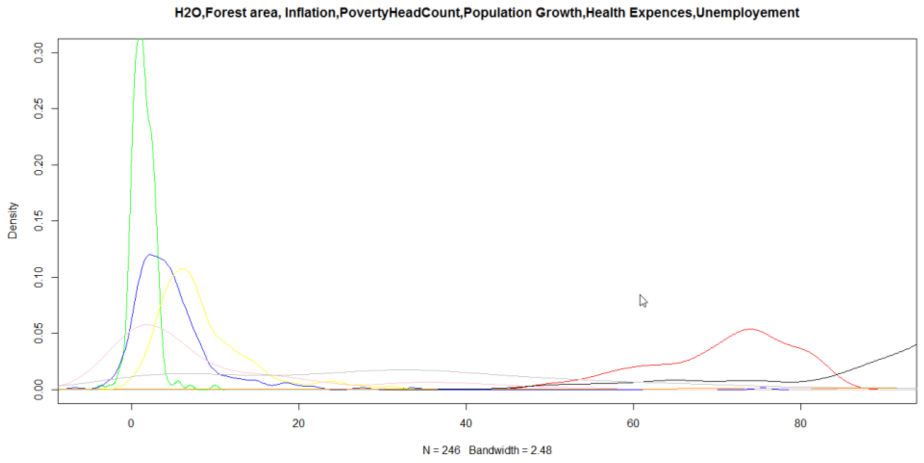
LifeExp ~ HealthExp+PovertyHcount+H2O

This Model tells us that the Life Expectancy at birth is highly depends on Health Expenditure per capita($US) , Poverty headcount ratio at $1.90 a day (2011 PPP) (% of population) and Improved water source (% of population with access). But here it is significant to know that how these factors are affecting the Life Expectancy?
The answer is discussed in the section of calculating correlation between all these where we got that increasing health expenditure is improving health which can be understood as well, but here we proved using the real statistics data and predictive model. The similar scenario with the improved water resources which go hand in hand with the life expectancy, and these technological era, water resources are greatly improved then the past which is beneficial to improve the quality of health. But if we consider third factor which is Poverty Head Count ratio then it is decreasing with year by year and which is helping people to improve the quality of their health.

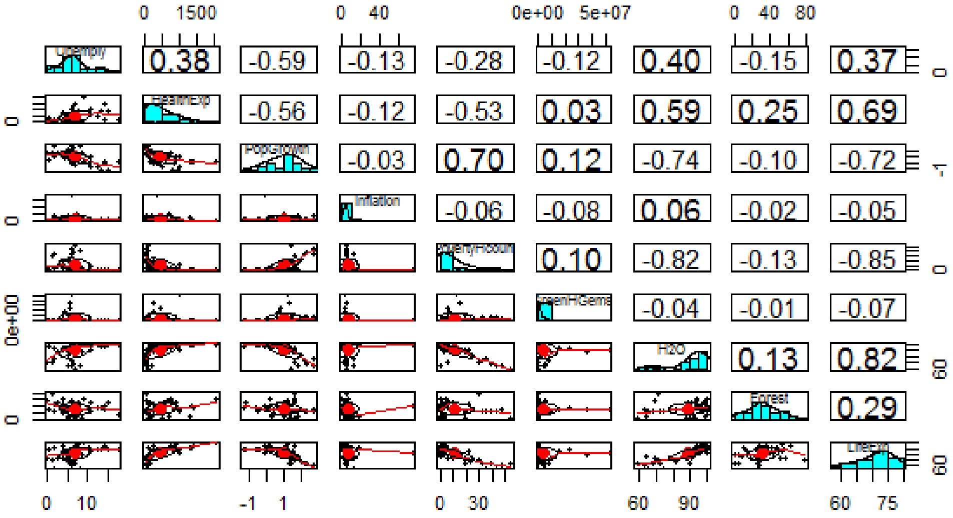
Conclusively, I would like to recommend that if we know that these are the major factors affecting the quality of health then Health organizations like WHO and UNISEF should try to focus on these factors. Even Government of poor countries can try to invest more on availing improved water resources, and increasing the financial opportunities to decrease the poverty headcount ratio. Opening more healthcare facilities, and investing more on healthcare sector would be very beneficial in the movement towards improving the quality of health.

Data preparation and exploration in R

- Research of social, economic, environmental factors that affects the quality of health leads to the selection of the Life Expectancy as the most significant variable to predict the quality of health (Robine, J.M., Romieu, I. and Cambois, E., 1999) and then the independent factors that significantly affects Life Expectancy are mentioned below.
- The Selected Data Indicators from World bank Data :
  1. Life expectancy at birth, total (years)
  2. Health expenditure per capita (current US$)
  3. Population growth (annual %)
  4. Inflation, GDP deflator (annual %)
  5. Poverty headcount ratio at $1.90 a day (2011 PPP) (% of population)
  6. Total greenhouse gas emissions (kt of CO2 equivalent)
  7. Improved water source (% of population with access)
  8. Forest area (% of land area)
- Firstly, reading all csv datasets and analyse them to be familiar with chosen dataset then exploring basic data representations like boxplots, histograms, density plots and summery measures for the year 2012. Density plots of all the datasets are shown below.

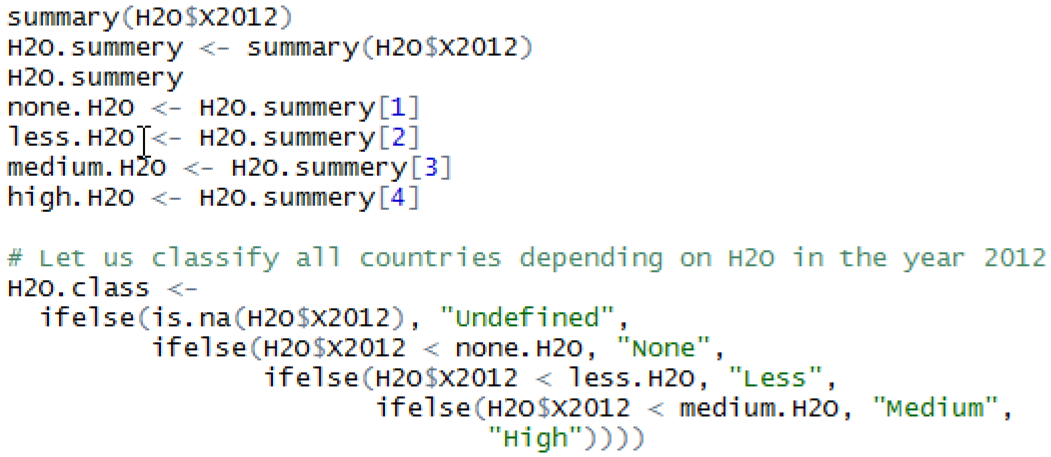


- Here, all the seven datasets are in numeric, so firstly I have converted all of them into categorical by using classification and categorized them in low, medium and high based on the dataset summery.
- I have created new dataframe (MasterData) and included all the original dataset and classified datasets for the year 20012.Now, there are many NA values inside the dataframe which we can remove those observations to clean our dataframe. Here, I have cleaned the masterData dataframe and make new dataset (MasterData.clean). to explore them further, even in Google map ,we will need cleaned dataset.
- Then I have plotted Correlation between all cleaned datasets using "psych" library functions and I got some interesting graphical plots.



Insights from k-NN modelling and visualisation with Google Maps in R

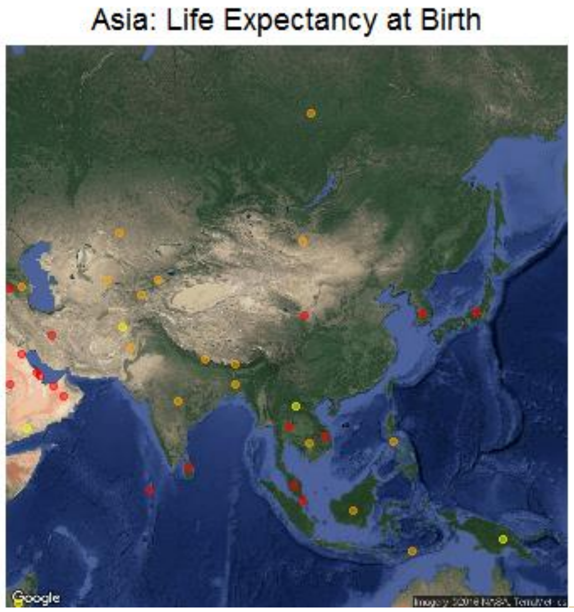
- k-NN classification is done on all the required datasets. As described in above section, I have created new dataframe which includes all the original data as well as the datasets after classification. The classification is considered based on summery measures. As Google map also need more of classified data and so we will use this classified datasets to plot Improved water resources ,population growth and Life Expectancy on google map and highlight them with various colours ,sizes and shapes.
- Below Screenshot describes the k-NN classification technique which is applied to all needed dataset to classify them. In this case, Improved Water Resources is classified.

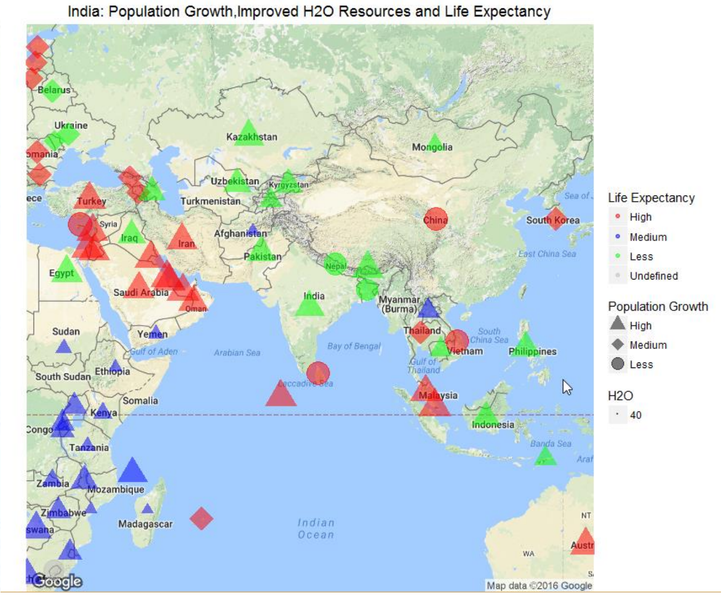


```
summary(H2O$X2012)
H2O.summery <- summary(H2O$X2012)
H2O.summery
none.H2O <- H2O.summery[1]
less.H2O <- H2O.summery[2]
medium.H2O <- H2O.summery[3]
high.H2O <- H2O.summery[4]

# Let us classify all countries depending on H2O in the year 2012
H2O.class <-
  ifelse(is.na(H2O$X2012), "Undefined",
         ifelse(H2O$X2012 < none.H2O, "None",
                ifelse(H2O$X2012 < less.H2O, "Less",
                       ifelse(H2O$X2012 < medium.H2O, "Medium",
                              "High"))))
```

- Now, for locations we have geolocations(latitude, longitude) in csv file(Countries_with_GeoLocs.csv),so we will use that and add that datasets into our final dataframe(popVSinflation.clean) and the we have put all in new dataframe(MasterData.coords) which is our master dataframe for Google maps.
- Below Satellite Map titled “Asia :Life Expectancy at Birth” represents the different levels of Life Expectancy in different colours in which red coloured datapoints are for High life expectancy, orange is for Medium and yellow is for Low Life expectancy.
- Below Map titled “India : Population Growth, Improved H2O and Life Expectancy” represents the Population Growth as different shapes, Life Expectancy as colours and Improved Water Resources as size of the data-points in the geolocations near around Asia.





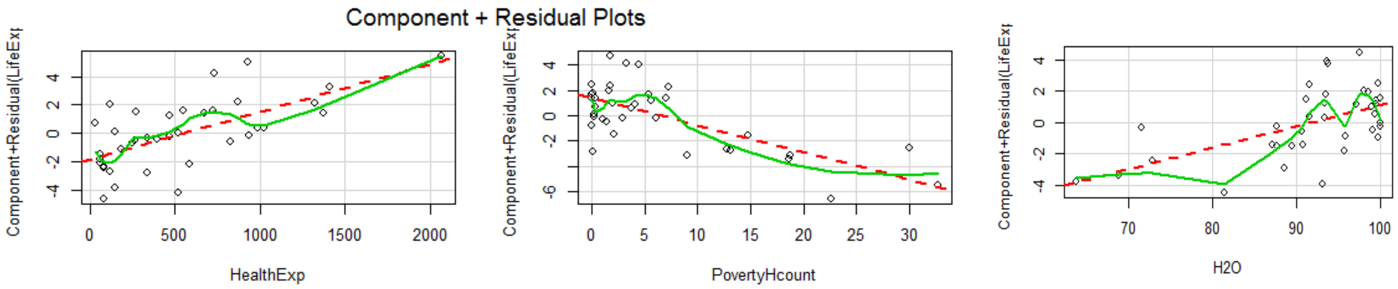
Insights from Multiple Regression modelling in R

To build the predictive model to predict the Life Expectancy at birth, which is dependent on major seven factors considered above. So firstly, creating a dataframe(Data2012) of all the dataset together for the year 2012. Then Plotting correlation chart to see the relationships between all the original data. By analysing their correlation coefficients it is clearly seen that Life Expectancy is highly correlated with Improved Water Resources, Poverty headcount ratio and Health expenditure per capita. Moreover, if any correlation coefficient greater than “0.8” in between all dependent variables show multicollinearity between those variables. So we can remove that here or at the time of model building.
Removing NA values is the significant step in modelling. So there are methods like imputing mean or ignoring NA values. But here I have used “mice” method of the “mice” package to check the missing values and removed NA values and cleaned the dataframe. Then some important plots like ggplot is plotted to see the linier relationship.
Now, we have cleaned dataframe to use for better outcome of the predictive model. So going towards building Predictive model which to life expectancy at birth, training and validation datasets are created by using the splitting ratio of 0.7 which means 70% values will be TRUE and 30% values will be FALSE. Then start building model from training dataset and refining it based on p-value(p>0.05).

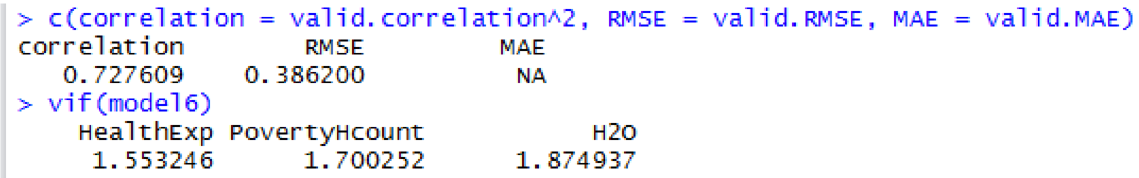
Model1 : LifeExp ~ Unemply+HealthExp+PopGrowth+Inflation+PovertyHcount+GreenHGems+H2O+Forest

Removing insignificant dataset and refining model recursively will take to the final strong model which is mentioned below.

Model6 : LifeExp ~ HealthExp+PovertyHcount+H2O



From the upper cr-plots, it is evident to say that there are linier relationships exist in the datasets present in the final model.
Even we can see that there are some outliers present into the dataset which can be removed to improve the power of the prediction model. So here I have used the cooks distance = 4/(n-k-1) to deal with the present outliers. Every time after removing outliers ,checking summery of the model and checked F-statistics which is increasing in every iteration. Here, I have iterate this four times and finally got the effective outcome with higher F-test values and lower P-value.
The next step to be followed is to validating the final model using validate dataset and find the correlation, root mean square values. Then vif is calculated for the final model and the outcome is shown in below.

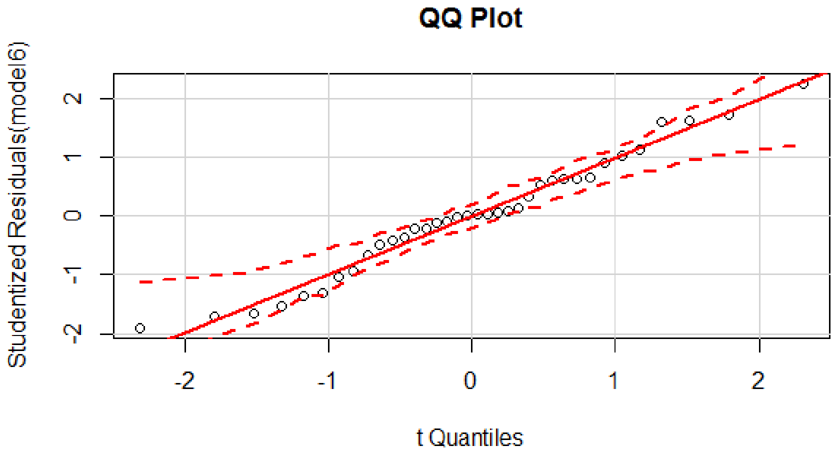


```
> c(correlation = valid.correlation^2, RMSE = valid.RMSE, MAE = valid.MAE)
correlation         RMSE          MAE
   0.727609     0.386200           NA
> vif(model6)
    HealthExp PovertyHcount           H2O
     1.553246      1.700252      1.874937
```

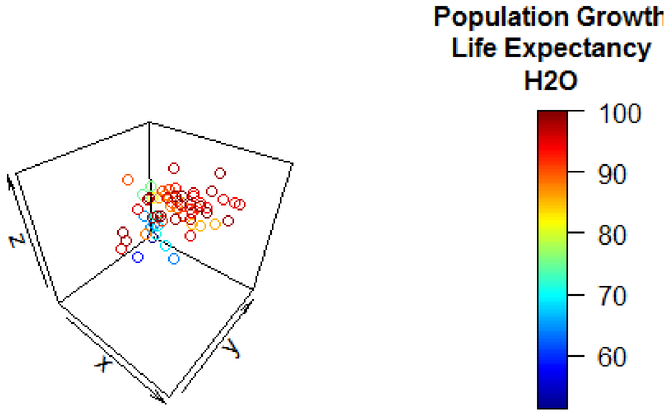
Here, vif which is Variance Inflation Factors is used to check the presence of multicollinearity of the model. Large values of the variables are considered as multicollinearity. But in our case, we got less vif for all three datasets.

Other insights from research and other explorations in R

While selecting indicators from World bank indicators website, I have read the news and highlights going on the website(http://data.worldbank.org/) and I came across the Poverty headcount ratio at $1.90 a day (2011 PPP) (% of population) and it was mentioned that “Extreme Poverty : The proportion of the world’s population living in extreme poverty has dropped significantly” and then I have start reading through and got the World Development Indicators 2016 where I got Life Expectancy as the major indicator for health of the people. Then continuous reading of Life Expectancy gave me insights of searching similar data in world bank indicators and Then after having all mentioned independent factors ,I have started exploring them using R. Below is the qqplot, plotted while removing outliers.



Here is the 3D plots plotted between Population Growth, Life Expectancy and Improved Water Resources.



Then after building and validating my life expectancy prediction model , I have researched again for justifying my results. I found some similar studies already been done titled Predicting Life Expectancy: A Cross-Country Empirical Analysis (Hendricks, A. and Graves, P.E., 2009) but this study has focussed on the data of the year 2002 while my study is on the data of the year 2012. Still the dependency of outcome is mirrored in the conclusion.

References

Robine, J.M., Romieu, I. and Cambois, E., 1999. Health expectancy indicators. Bulletin-World Health Organization, 77, pp.181-185

Hendricks, A. and Graves, P.E., 2009. Predicting Life Expectancy: A Cross-Country Empirical Analysis. Available at SSRN 1477594.