To: The members of WHO

From: Chinmayi Suryakant Mahadik
Subject: Report on the factors affecting life expectancy of the world population.
Date: 5/10/2020
A report to discuss the factors like immunization factor, mortality factor, social and economic factor that are associated with the life expectancy of the world population.

EXECUTIVE SUMMARY

Major Findings:

- The life expectancy of the world population is highly associated with few factors like development status of the country, adult mortality rate of both sexes, number of infant death, number of population taken immunization of polio and diphtheria, death rate due to HIV/Aids, GDP of the country, percentage prevalence of thinness among children and adolescents for Age 10 to 19, Human Development Index in terms of income composition of resources and number of years of schooling.

- The countries with higher GDP which are generally developed countries or on the path to become developed countries have better life expectancy in age.

- In my analysis I conclude that the life expectancy of the world depends on below equation:

Life expectancy = 66.33 - 0.934 (Status) - 0.0119(Adult Mortality) - 0.7401(infant deaths^0.25) + 0.0134(Polio) + 0.029 (Diphtheria) -9.324(HIV/Aids^0.25) + 0.000044(GDP) - 0.0871(Thinness 1-19 years) +7.778 (Income Composition of resources) + 0.5622(Schooling)

- This states that if the country is a developing country, the life expectancy will decrease by 0.934 unit. If adult mortality and the fourth root of infant death increase by 1 unit per 1000 population, the life expectancy decreases by 0.0119 and 0.7401 times, respectively. The immunization coverage among 1-year-olds of Polio and Diphtheria increase by 1%, the age of life expectancy will increase by 0.0134 and 0.029 times, respectively. If the fourth root of death rate due to HIV is increased by 1, life expectancy decreases by 9.324 times and if prevalence of thinness among children and adolescents for Age 10 to 19 increases by 1%, the life expectancy will decrease by 0.0871. Lastly 1 unit increase in GDP, income composition of resource and number of years of schooling increases by 1 unit, the life expectancy will increase by 0.00044, 7.778 and 0.5622 times, respectively.

Recommendation for Action:

- Vaccinations have eliminated conditions normally resulting in certain death and increased health conditions among children and adults. Hence immunization increases life expectancy.

- Better **education** drives longer **life**. It also tends to lead to more wealth, which is why wealth and longevity associated to each other.

- Children's nutrition is very important for economic growth of country. Health is wealth.

- Life expectancy better in developed country. Developed countries have higher income which is also implies better access to housing, education, health services and other items which tend to lead to improved health, lower rates of mortality and higher life expectancy.

- According to following situation of COVID-19,

- People older than 50 years age and with severe disorders are at higher risk. This may affect few countries life expectancy but not majorly.

- Since there is no vaccination available for COVID -19, the mortality rate is high. Immunization would have reduced mortality and increase life expectancy.

- Education and awareness have helped to prevent from the virus.

Analytical Overview:

- Exploratory data analysis was first used on all the variables to determine their correlation with the life expectancy and to check normal distribution of all variables.

- To conduct a predictive model, I

- I tried a stepwise regression, best subset approach and robust regression approach to determine the life expectancy of the world population and estimate the best possible models using all possible combinations.

- All the assumptions were verified by plotting different graphs and summary results.

- To validate the models, I tried k fold cross validation method to determine accuracy in prediction in both models.

APPENDIX

Process used in Data Analysis

- Data Summarizing
- Data Checking
- Handling Outliers
- Inferences from univariant charts
- Inferences from bi-variant charts
- Inferences from multivariate charts
- Model recognition from Stepwise Regression Process
- Model recognition from Robust Regression Process
- Model recognition from Subset Regression Process
- Model recognition from advance Machine Learning approach.
- Model recognition and comparison (Regularisation)

DATA SUMMARIZING:

```
library(readr)
LifeExp <- read_csv("C:/Users/vedan/OneDrive/Desktop/Statistics/Homework/Final
Project/LifeExp.csv")
library(tidyverse)
library(funModeling)
library(Hmisc)

glimpse(LifeExp)
```

● **glimpse** function is revealed the dimensions (Observations) and names of the variables in the dataset.

```
Console ~/
> glimpse(LifeExp)
Observations: 2,496
Variables: 22
$ Country                        <chr> "Afghanistan", "Afghanistan", "Afghanistan", "Afghanistan", "Afghanistan", "Afghanistan", "Afgh
$ Year                           <dbl> 2015, 2014, 2013, 2012, 2011, 2010, 2009, 2008, 2007, 2006, 2005, 2004, 2003, 2002, 2001, 2000
$ Status                         <chr> "Developing", "Developing", "Developing", "Developing", "Developing", "Developing", "Developing
$ `Life expectancy`              <dbl> 65.0, 59.9, 59.9, 59.5, 59.2, 58.8, 58.6, 58.1, 57.5, 57.3, 57.3, 57.0, 56.7, 56.2, 55.3, 54.8
$ `Adult Mortality`              <dbl> 263, 271, 268, 272, 275, 279, 281, 287, 295, 295, 291, 293, 295, 3, 316, 321, 74, 8, 84, 86, 88
$ `infant deaths`                <dbl> 62, 64, 66, 69, 71, 74, 77, 80, 82, 84, 85, 87, 87, 88, 88, 88, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1
$ Alcohol                        <dbl> 0.01, 0.01, 0.01, 0.01, 0.01, 0.01, 0.01, 0.03, 0.02, 0.03, 0.02, 0.02, 0.01, 0.01, 0.01, 0.01
$ `percentage expenditure`       <dbl> 71.279624, 73.523582, 73.219243, 78.184215, 7.097109, 79.679367, 56.762217, 25.873925, 10.9101
$ `Hepatitis B`                  <dbl> 65, 62, 64, 67, 68, 66, 63, 64, 63, 64, 66, 67, 65, 64, 63, 62, 99, 98, 99, 99, 99, 99, 98, 99
$ Measles                        <dbl> 1154, 492, 430, 2787, 3013, 1989, 2861, 1599, 1141, 1990, 1296, 466, 798, 2486, 8762, 6532, 0,
$ BMI                            <dbl> 19.1, 18.6, 18.1, 17.6, 17.2, 16.7, 16.2, 15.7, 15.2, 14.7, 14.2, 13.8, 13.4, 13.0, 12.6, 12.2
$ `under-five deaths`            <dbl> 83, 86, 89, 93, 97, 102, 106, 110, 113, 116, 118, 120, 122, 122, 122, 122, 0, 1, 1, 1, 1, 1, 1
$ Polio                          <dbl> 6, 58, 62, 67, 68, 66, 63, 64, 63, 58, 58, 5, 41, 36, 35, 24, 99, 98, 99, 99, 99, 99, 98, 99, 9
$ `Total expenditure`            <dbl> 8.16, 8.18, 8.13, 8.52, 7.87, 9.20, 9.42, 8.33, 6.73, 7.43, 8.70, 8.79, 8.82, 7.76, 7.80, 8.20
$ Diphtheria                     <dbl> 65, 62, 64, 67, 68, 66, 63, 64, 63, 58, 58, 5, 41, 36, 33, 24, 99, 98, 99, 99, 99, 99, 98, 99,
$ `HIV/AIDS`                     <dbl> 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1,
$ GDP                            <dbl> 584.25921, 612.69651, 631.74498, 669.95900, 63.53723, 553.32894, 445.89330, 373.36112, 369.835
$ Population                     <dbl> 33736494, 327582, 31731688, 3696958, 2978599, 2883167, 284331, 2729431, 26616792, 2589345, 257
$ `thinness  1-19 years`         <dbl> 17.2, 17.5, 17.7, 17.9, 18.2, 18.4, 18.6, 18.8, 19.0, 19.2, 19.3, 19.5, 19.7, 19.9, 2.1, 2.3,
$ `thinness 5-9 years`           <dbl> 17.3, 17.5, 17.7, 18.0, 18.2, 18.4, 18.7, 18.9, 19.1, 19.3, 19.5, 19.7, 19.9, 2.2, 2.4, 2.5, 1
$ `Income composition of resources` <dbl> 0.479, 0.476, 0.470, 0.463, 0.454, 0.448, 0.434, 0.433, 0.415, 0.405, 0.396, 0.381, 0.373, 0.3
$ Schooling                      <dbl> 10.1, 10.0, 9.9, 9.8, 9.5, 9.2, 8.9, 8.7, 8.4, 8.1, 7.9, 6.8, 6.5, 6.2, 5.9, 5.5, 14.2, 14.2,
>
```

● **df_status** function identified the type of values in the attributes and the unique numbers present in it.

```
df_status(LifeExp)
```

Console ~/

```
> df_status(LifeExp)
                          variable q_zeros p_zeros q_na  p_na q_inf p_inf      type unique
1                          Country       0    0.00    0  0.00     0     0 character    156
2                             Year       0    0.00    0  0.00     0     0   numeric     16
3                           Status       0    0.00    0  0.00     0     0 character      2
4                  Life expectancy       0    0.00    0  0.00     0     0   numeric    362
5                  Adult Mortality       0    0.00    0  0.00     0     0   numeric    412
6                    infant deaths     726   29.09    0  0.00     0     0   numeric    194
7                          Alcohol       0    0.00  152  6.09     0     0   numeric   1003
8          percentage expenditure       0    0.00  177  7.09     0     0   numeric   2319
9                      Hepatitis B       0    0.00  470 18.83     0     0   numeric     84
10                         Measles     828   33.17    0  0.00     0     0   numeric    838
11                             BMI       0    0.00   16  0.64     0     0   numeric    585
12               under-five deaths     665   26.64    0  0.00     0     0   numeric    235
13                           Polio       0    0.00    8  0.32     0     0   numeric     72
14               Total expenditure       0    0.00  157  6.29     0     0   numeric    773
15                       Diphtheria      0    0.00    8  0.32     0     0   numeric     80
16                         HIV/AIDS      0    0.00    0  0.00     0     0   numeric    195
17                              GDP      0    0.00   14  0.56     0     0   numeric   2476
18                       Population      0    0.00   36  1.44     0     0   numeric   2453
19           thinness  1-19 years       0    0.00   16  0.64     0     0   numeric    196
20             thinness 5-9 years       0    0.00   16  0.64     0     0   numeric    203
21 Income composition of resources     109    4.37    0  0.00     0     0   numeric    620
22                       Schooling      14    0.56    0  0.00     0     0   numeric    173
> 
```

● **freq** function is used to give the number of times each value is repeated in the dataset. But unfortunately, it did not work as I had no categorical value recognized by software in the data set.

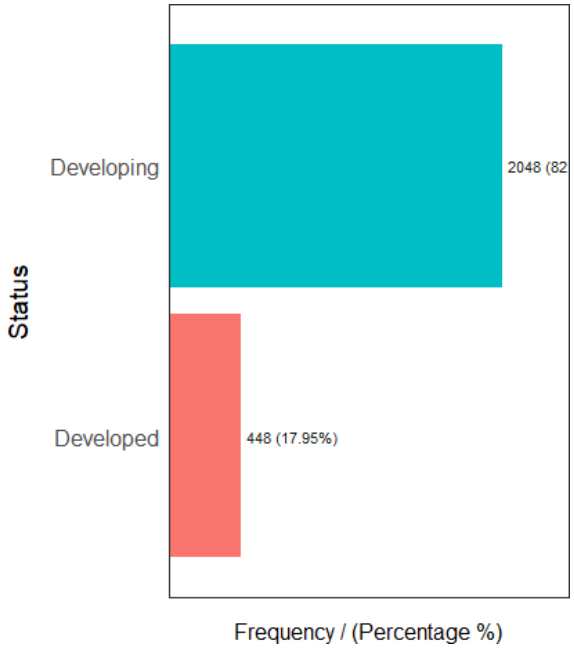
freq(LifeExp)

Console ~/

```
> freq(LifeExp)
                      Country frequency percentage cumulative_perc
1                 Afghanistan        16       0.64             0.64
2                     Albania        16       0.64             1.28
3                     Algeria        16       0.64             1.92
4                      Angola        16       0.64             2.56
5         Antigua and Barbuda        16       0.64             3.20
6                   Argentina        16       0.64             3.84
7                     Armenia        16       0.64             4.48
8                   Australia        16       0.64             5.12
9                     Austria        16       0.64             5.76
10                 Azerbaijan        16       0.64             6.40
11                    Bahrain        16       0.64             7.04
12                 Bangladesh        16       0.64             7.68
13                   Barbados        16       0.64             8.32
14                    Belarus        16       0.64             8.96
15                    Belgium        16       0.64             9.60
16                     Belize        16       0.64            10.24
17                      Benin        16       0.64            10.88
18                     Bhutan        16       0.64            11.52
19    Bosnia and Herzegovina        16       0.64            12.16
20                   Botswana        16       0.64            12.80
21                     Brazil        16       0.64            13.44
22          Brunei Darussalam        16       0.64            14.08
23                   Bulgaria        16       0.64            14.72
24               Burkina Faso        16       0.64            15.36
25                    Burundi        16       0.64            16.00
26                 Cabo Verde        16       0.64            16.64
27                   Cambodia        16       0.64            17.28
28                   Cameroon        16       0.64            17.92
29                     Canada        16       0.64            18.56
30  Central African Republic        16       0.64            19.20
31                       Chad        16       0.64            19.84
32                      Chile        16       0.64            20.48
33                      China        16       0.64            21.12
34                   Colombia        16       0.64            21.76
35                    Comoros        16       0.64            22.40
36                 Costa Rica        16       0.64            23.04
37                    Croatia        16       0.64            23.68
38                       Cuba        16       0.64            24.32
39                     Cyprus        16       0.64            24.96
40                    Denmark        16       0.64            25.60
41                   Djibouti        16       0.64            26.24
42         Dominican Republic        16       0.64            26.88
43                    Ecuador        16       0.64            27.52
44                El Salvador        16       0.64            28.16
45          Equatorial Guinea        16       0.64            28.80
46                    Eritrea        16       0.64            29.44
47                    Estonia        16       0.64            30.08
48                   Ethiopia        16       0.64            30.72
49                       Fiji        16       0.64            31.36
50                    Finland        16       0.64            32.00
51                     France        16       0.64            32.64
52                      Gabon        16       0.64            33.28
53                    Georgia        16       0.64            33.92
54                    Germany        16       0.64            34.56
55                      Ghana        16       0.64            35.20
56                     Greece        16       0.64            35.84
```

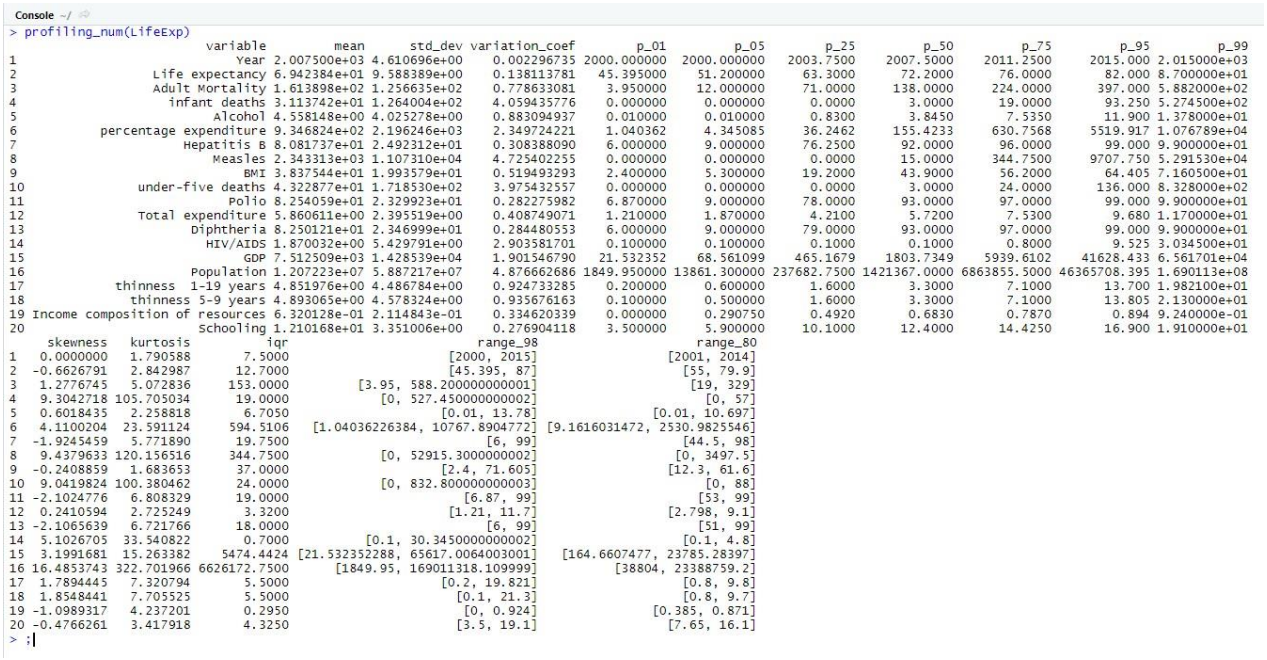
55	Ghana	16	0.64	35.20
56	Greece	16	0.64	35.84
57	Grenada	16	0.64	36.48
58	Guatemala	16	0.64	37.12
59	Guinea	16	0.64	37.76
60	Guinea-Bissau	16	0.64	38.40
61	Guyana	16	0.64	39.04
62	Haiti	16	0.64	39.68
63	Honduras	16	0.64	40.32
64	Hungary	16	0.64	40.96
65	Iceland	16	0.64	41.60
66	India	16	0.64	42.24
67	Indonesia	16	0.64	42.88
68	Iraq	16	0.64	43.52
69	Ireland	16	0.64	44.16
70	Israel	16	0.64	44.80
71	Italy	16	0.64	45.44
72	Jamaica	16	0.64	46.08
73	Japan	16	0.64	46.72
74	Jordan	16	0.64	47.36
75	Kazakhstan	16	0.64	48.00
76	Kenya	16	0.64	48.64
77	Kiribati	16	0.64	49.28
78	Kuwait	16	0.64	49.92
79	Latvia	16	0.64	50.56
80	Lebanon	16	0.64	51.20
81	Lesotho	16	0.64	51.84
82	Liberia	16	0.64	52.48
83	Libya	16	0.64	53.12
84	Lithuania	16	0.64	53.76
85	Luxembourg	16	0.64	54.40
86	Madagascar	16	0.64	55.04
87	Malawi	16	0.64	55.68
88	Malaysia	16	0.64	56.32
89	Maldives	16	0.64	56.96
90	Mali	16	0.64	57.60
91	Malta	16	0.64	58.24
92	Mauritania	16	0.64	58.88
93	Mauritius	16	0.64	59.52
94	Mexico	16	0.64	60.16
95	Mongolia	16	0.64	60.80
96	Montenegro	16	0.64	61.44
97	Morocco	16	0.64	62.08
98	Mozambique	16	0.64	62.72
99	Myanmar	16	0.64	63.36
100	Namibia	16	0.64	64.00
101	Nepal	16	0.64	64.64
102	Netherlands	16	0.64	65.28
103	New Zealand	16	0.64	65.92
104	Nicaragua	16	0.64	66.56
105	Niger	16	0.64	67.20
106	Nigeria	16	0.64	67.84
107	Norway	16	0.64	68.48
108	Oman	16	0.64	69.12
109	Pakistan	16	0.64	69.76
110	Panama	16	0.64	70.40
111	Papua New Guinea	16	0.64	71.04
112	Paraguay	16	0.64	71.68

109	Pakistan	16	0.64	69.76
110	Panama	16	0.64	70.40
111	Papua New Guinea	16	0.64	71.04
112	Paraguay	16	0.64	71.68
113	Peru	16	0.64	72.32
114	Philippines	16	0.64	72.96
115	Poland	16	0.64	73.60
116	Portugal	16	0.64	74.24
117	Qatar	16	0.64	74.88
118	Romania	16	0.64	75.52
119	Russian Federation	16	0.64	76.16
120	Rwanda	16	0.64	76.80
121	Samoa	16	0.64	77.44
122	Sao Tome and Principe	16	0.64	78.08
123	Saudi Arabia	16	0.64	78.72
124	Senegal	16	0.64	79.36
125	Serbia	16	0.64	80.00
126	Seychelles	16	0.64	80.64
127	Sierra Leone	16	0.64	81.28
128	Singapore	16	0.64	81.92
129	Slovenia	16	0.64	82.56
130	Solomon Islands	16	0.64	83.20
131	South Africa	16	0.64	83.84
132	Spain	16	0.64	84.48
133	Sri Lanka	16	0.64	85.12
134	Sudan	16	0.64	85.76
135	Suriname	16	0.64	86.40
136	Swaziland	16	0.64	87.04
137	Sweden	16	0.64	87.68
138	Switzerland	16	0.64	88.32
139	Syrian Arab Republic	16	0.64	88.96
140	Tajikistan	16	0.64	89.60
141	Thailand	16	0.64	90.24
142	Timor-Leste	16	0.64	90.88
143	Togo	16	0.64	91.52
144	Tonga	16	0.64	92.16
145	Trinidad and Tobago	16	0.64	92.80
146	Tunisia	16	0.64	93.44
147	Turkey	16	0.64	94.08
148	Turkmenistan	16	0.64	94.72
149	Uganda	16	0.64	95.36
150	Ukraine	16	0.64	96.00
151	United Arab Emirates	16	0.64	96.64
152	Uruguay	16	0.64	97.28
153	Uzbekistan	16	0.64	97.92
154	Vanuatu	16	0.64	98.56
155	Zambia	16	0.64	99.20
156	Zimbabwe	16	0.64	100.00



● **profiling_num** function is used to get more detailed information on statistical summary of the dataset. Through this I obtained the skewness, kurtosis, and some characteristic values of each attribute, which helped in understanding behavior of data set.

profiling_num(LifeExp)



```
Console ~/
> profiling_num(LifeExp)
                          variable         mean      std_dev variation_coef        p_01        p_05        p_25        p_50        p_75        p_95        p_99
1                             Year 2.007500e+03 4.610696e+00    0.002296735 2000.000000 2000.000000 2003.7500 2007.5000 2011.2500 2015.000 2.015000e+03
2                  Life expectancy 6.942384e+01 9.588389e+00    0.138113781   45.395000   51.200000   63.3000   72.2000   76.0000   82.000 8.700000e+01
3                  Adult Mortality 1.613898e+02 1.256635e+02    0.778633081    3.950000   12.000000   71.0000  138.0000  224.0000  397.000 5.882000e+02
4                    infant deaths 3.113742e+01 1.264004e+02    4.059435776    0.000000    0.000000    0.0000    3.0000   19.0000   93.250 5.274500e+02
5                          Alcohol 4.558148e+00 4.025278e+00    0.883094937    0.010000    0.010000    0.8300    3.8450    7.5350   11.900 1.378000e+01
6          percentage expenditure 9.346824e+02 2.196246e+03    2.349724221    1.040362    4.345085   36.2462  155.4233  630.7568 5519.917 1.076789e+04
7                     Hepatitis B 8.081737e+01 2.492312e+01    0.308388090    6.000000    9.000000   76.2500   92.0000   96.0000   99.000 9.900000e+01
8                         Measles 2.343313e+03 1.107310e+04    4.725402255    0.000000    0.000000    0.0000   15.0000  344.7500 9707.750 5.291530e+04
9                             BMI 3.837544e+01 1.993579e+01    0.519493293    2.400000    5.300000   19.2000   43.9000   56.2000   64.405 7.160500e+01
10               under-five deaths 4.322877e+01 1.718530e+02    3.975432557    0.000000    0.000000    0.0000    3.0000   24.0000  136.000 8.328000e+02
11                           Polio 8.254059e+01 2.329923e+01    0.282275982    6.870000    9.000000   78.0000   93.0000   97.0000   99.000 9.900000e+01
12               Total expenditure 5.860611e+00 2.395519e+00    0.408749071    1.210000    1.870000    4.2100    5.7200    7.5300    9.680 1.170000e+01
13                     Diphtheria 8.250121e+01 2.346999e+01    0.284480553    6.000000    9.000000   79.0000   93.0000   97.0000   99.000 9.900000e+01
14                       HIV/AIDS 1.870032e+00 5.429791e+00    2.903581701    0.100000    0.100000    0.1000    0.1000    0.8000    9.525 3.034500e+01
15                            GDP 7.512509e+03 1.428539e+04    1.901546790   21.532352   68.561099  465.1679 1803.7349 5939.6102 41628.433 6.561701e+04
16                     Population 1.207223e+07 5.887217e+07    4.876662686 1849.950000 13861.300000 237682.7500 1421367.0000 6863855.5000 46365708.395 1.690113e+08
17           thinness  1-19 years 4.851976e+00 4.486784e+00    0.924733285    0.200000    0.600000    1.6000    3.3000    7.1000   13.700 1.982100e+01
18             thinness 5-9 years 4.893065e+00 4.578324e+00    0.935676163    0.100000    0.500000    1.6000    3.3000    7.1000   13.805 2.130000e+01
19 Income composition of resources 6.320128e-01 2.114843e-01    0.334620339    0.000000    0.290750    0.4920    0.6830    0.7870    0.894 9.240000e-01
20                       Schooling 1.210168e+01 3.351006e+00    0.276904118    3.500000    5.900000   10.1000   12.4000   14.4250   16.900 1.910000e+01
     skewness   kurtosis          iqr                    range_98                    range_80
1   0.0000000   1.790588       7.5000                 [2000, 2015]                [2001, 2014]
2  -0.6626791   2.842987      12.7000                 [45.395, 87]                 [55, 79.9]
3   1.2776745   5.072836     153.0000         [3.95, 588.200000000001]                 [19, 329]
4   9.3042718 105.705034      19.0000             [0, 527.450000000002]                   [0, 57]
5   0.6018435   2.258818       6.7050                 [0.01, 13.78]               [0.01, 10.697]
6   4.1100204  23.591124     594.5106   [1.04036226384, 10767.8904772] [9.1616031472, 2530.9825546]
7  -1.9245459   5.771890      19.7500                      [6, 99]                  [44.5, 98]
8   9.4379633 120.156516     344.7500             [0, 52915.3000000002]                [0, 3497.5]
9  -0.2408859   1.683653      37.0000                 [2.4, 71.605]                [12.3, 61.6]
10  9.0419824 100.380462      24.0000             [0, 832.800000000003]                   [0, 88]
11 -2.1024776   6.808329      19.0000                  [6.87, 99]                   [53, 99]
12  0.2410594   2.725249       3.3200                 [1.21, 11.7]                [2.798, 9.1]
13 -2.1065639   6.721766      18.0000                      [6, 99]                  [51, 99]
14  5.1026705  33.540822       0.7000         [0.1, 30.3450000000002]                 [0.1, 4.8]
15  3.1991681  15.263382    5474.4424 [21.532352288, 65617.0064003001]  [164.6607477, 23785.28397]
16 16.4853743 322.701966 6626172.7500      [1849.95, 169011318.109999]          [38804, 23388759.2]
17  1.7894445   7.320794       5.5000                 [0.2, 19.821]                 [0.8, 9.8]
18  1.8548441   7.705525       5.5000                  [0.1, 21.3]                  [0.8, 9.7]
19 -1.0989317   4.237201       0.2950                   [0, 0.924]               [0.385, 0.871]
20 -0.4766261   3.417918       4.3250                  [3.5, 19.1]                [7.65, 16.1]
> ;
```

● **plot_num** function is used to find frequency count of observations for a specific category/range of each variables through graphical representation of bar chart/histogram.

plot_num(LifeExp)

● **describe** function is used to give tabular information in missing/distinct values in the dataset with its proportion percentage.

describe(LifeExp)

```
> describe(LifeExp)
LifeExp

 22  Variables      2496  Observations
---------------------------------------------------------------------------------------------------------------------------------------------------------------
Country
       n  missing distinct
    2496        0      156

lowest : Afghanistan         Albania             Algeria             Angola              Antigua and Barbuda
highest: Uruguay             Uzbekistan          Vanuatu             Zambia              Zimbabwe
---------------------------------------------------------------------------------------------------------------------------------------------------------------
Year
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    2496        0       16    0.996     2008    5.315     2000     2001     2004     2008     2011     2014     2015

lowest : 2000 2001 2002 2003 2004, highest: 2011 2012 2013 2014 2015

Value       2000  2001  2002  2003  2004  2005  2006  2007  2008  2009  2010  2011  2012  2013  2014  2015
Frequency    156   156   156   156   156   156   156   156   156   156   156   156   156   156   156   156
Proportion 0.062 0.062 0.062 0.062 0.062 0.062 0.062 0.062 0.062 0.062 0.062 0.062 0.062 0.062 0.062 0.062
---------------------------------------------------------------------------------------------------------------------------------------------------------------
Status
       n  missing distinct
    2496        0        2

value       Developed Developing
Frequency         448       2048
Proportion      0.179      0.821
---------------------------------------------------------------------------------------------------------------------------------------------------------------
Life expectancy
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    2496        0      362        1    69.42    10.68     51.2     55.0     63.3     72.2     76.0     79.9     82.0

lowest : 36.3 39.0 41.0 41.5 42.3, highest: 85.0 86.0 87.0 88.0 89.0
---------------------------------------------------------------------------------------------------------------------------------------------------------------
Adult Mortality
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    2496        0      412        1    161.4    134.4       12       19       71      138      224      329      397

lowest :   1   2   3   4   5, highest: 693 699 715 717 723
---------------------------------------------------------------------------------------------------------------------------------------------------------------
infant deaths
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    2496        0      194    0.973    31.14     53.5     0.00     0.00     0.00     3.00    19.00    57.00    93.25

lowest :    0    1    2    3    4, highest: 1400 1500 1600 1700 1800
---------------------------------------------------------------------------------------------------------------------------------------------------------------
Alcohol
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    2344      152     1003    0.999    4.558    4.519    0.010    0.010    0.830    3.845    7.535   10.697   11.900

lowest :  0.01  0.02  0.03  0.04  0.05, highest: 16.35 16.58 16.99 17.31 17.87
```

```
---------------------------------------------------------------------------------------------------------------------------------------------------------------
percentage expenditure
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    2319      177     2319        1    934.7     1488    4.345    9.162   36.246  155.423  630.757 2530.983 5519.917

lowest : 9.987219e-02 1.080560e-01 2.756483e-01 3.284181e-01 3.586514e-01, highest: 1.837933e+04 1.882287e+04 1.896135e+04 1.909905e+04 1.947991e+04
---------------------------------------------------------------------------------------------------------------------------------------------------------------
Hepatitis B
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    2026      470       84    0.997    80.82    23.19     9.00    44.50    76.25    92.00    96.00    98.00    99.00

lowest :  2  4  5  6  7, highest: 95 96 97 98 99
---------------------------------------------------------------------------------------------------------------------------------------------------------------
Measles
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    2496        0      838    0.963     2343     4354      0.0      0.0      0.0     15.0    344.8   3497.5   9707.8

lowest :      0      1      2      3      4, highest: 124219 131441 141258 168107 212183
---------------------------------------------------------------------------------------------------------------------------------------------------------------
BMI
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    2480       16      585        1    38.38    22.72      5.3     12.3     19.2     43.9     56.2     61.6     64.4

lowest :  1.4  1.8  2.0  2.1  2.2, highest: 75.7 76.2 76.7 77.1 77.6
---------------------------------------------------------------------------------------------------------------------------------------------------------------
under-five deaths
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    2496        0      235    0.978    43.23    74.44        0        0        0        3       24       88      136

lowest :    0    1    2    3    4, highest: 2100 2200 2300 2400 2500
---------------------------------------------------------------------------------------------------------------------------------------------------------------
Polio
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    2488        8       72    0.996    82.54    21.35        9       53       78       93       97       99       99

lowest :  3  4  5  6  7, highest: 95 96 97 98 99
---------------------------------------------------------------------------------------------------------------------------------------------------------------
Total expenditure
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    2339      157      773        1    5.861    2.718    1.870    2.798    4.210    5.720    7.530    9.100    9.680

lowest :  0.37  0.65  0.74  0.76  0.92, highest: 13.13 13.63 13.66 13.73 14.39
---------------------------------------------------------------------------------------------------------------------------------------------------------------
Diphtheria
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    2488        8       80    0.996     82.5    21.38        9       51       79       93       97       99       99

lowest :  2  3  4  5  6, highest: 95 96 97 98 99
---------------------------------------------------------------------------------------------------------------------------------------------------------------
HIV/AIDS
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    2496        0      195    0.783     1.87    3.124    0.100    0.100    0.100    0.100    0.800    4.800    9.525

lowest :  0.1  0.2  0.3  0.4  0.5, highest: 48.8 49.1 49.9 50.3 50.6
```

```
Population
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    2460       36     2453        1 12072226 20370353    13861    38804   237683  1421367  6863856 23388759 46365708

lowest :         43        123        135        146        419, highest: 1126135777 1144118674 1161977719 1179681239 1293859294
------------------------------------------------------------------------------------------------------------------------
thinness  1-19 years
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    2480       16      196        1    4.852    4.559      0.6      0.8      1.6      3.3      7.1      9.8     13.7

lowest :  0.1  0.2  0.3  0.4  0.5, highest: 27.2 27.3 27.4 27.5 27.7
------------------------------------------------------------------------------------------------------------------------
thinness 5-9 years
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    2480       16      203        1    4.893    4.629      0.5      0.8      1.6      3.3      7.1      9.7     13.8

lowest :  0.1  0.2  0.3  0.4  0.5, highest: 28.2 28.3 28.4 28.5 28.6
------------------------------------------------------------------------------------------------------------------------
Income composition of resources
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    2496        0      620        1    0.632   0.2289   0.2908   0.3850   0.4920   0.6830   0.7870   0.8710   0.8940

lowest : 0.000 0.253 0.255 0.261 0.266, highest: 0.939 0.941 0.942 0.945 0.948
------------------------------------------------------------------------------------------------------------------------
Schooling
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    2496        0      173        1    12.1    3.742     5.90     7.65    10.10    12.40    14.43    16.10    16.90

lowest :  0.0  2.8  2.9  3.0  3.1, highest: 20.3 20.4 20.5 20.6 20.7
------------------------------------------------------------------------------------------------------------------------
> |
```

library(DataExplorer)

- **create_report** is used to develop a html auto EDA report for the dataset

create_report(LifeExp)

Data Profiling Report.pdf

```
Console ~/
  ordinary text without R code

  |.....................................................................................
label: plot_response_scatterplot
  |.....................................................................................
  ordinary text without R code

  |.....................................................................................
label: plot_by_scatterplot

output file: C:/Users/vedan/OneDrive/Documents/report.knit.md

"C:/Program Files/RStudio/bin/pandoc/pandoc" +RTS -K512m -RTS "C:/Users/vedan/OneDrive/Dc
art --output pandoc483c46257c0f.html --email-obfuscation none --self-contained --standalc
\R\win-library\3.6\rmarkdown\rmd\h\default.html" --no-highlight --variable highlightjs=1
n-str483c1b9158b8.html" --mathjax --variable "mathjax-url:https://mathjax.rstudio.com/lat
-library/3.6/rmarkdown/rmd/lua/pagebreak.lua" --lua-filter "C:/Users/vedan/OneDrive/Docum

Output created: report.html
> |
```

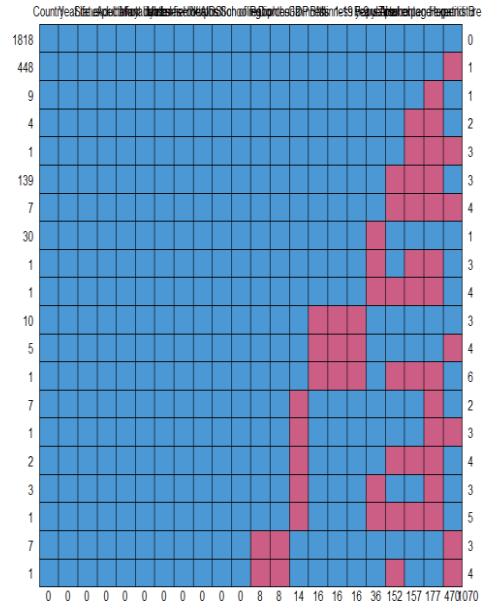
DATA CHECKING:

install.packages("mice")
library(mice)

- **md.pattern** is used to find a missing value patten in the dataset

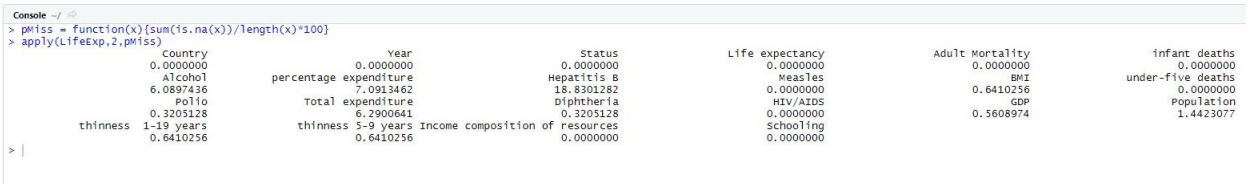
md.pattern(LifeExp)

```
> md.pattern(LifeExp)
     Country Year Status Life expectancy Adult Mortality infant deaths Measles under-five deaths HIV/AIDS Income composition of resources Schooling Polio Diphtheria GDP BMI
1818       1    1      1               1               1             1       1                 1        1                              1         1     1          1   1   1
448        1    1      1               1               1             1       1                 1        1                              1         1     1          1   1   1
9          1    1      1               1               1             1       1                 1        1                              1         1     1          1   1   1
4          1    1      1               1               1             1       1                 1        1                              1         1     1          1   1   1
1          1    1      1               1               1             1       1                 1        1                              1         1     1          1   1   1
139        1    1      1               1               1             1       1                 1        1                              1         1     1          1   1   1
7          1    1      1               1               1             1       1                 1        1                              1         1     1          1   1   1
30         1    1      1               1               1             1       1                 1        1                              1         1     1          1   1   1
1          1    1      1               1               1             1       1                 1        1                              1         1     1          1   1   1
1          1    1      1               1               1             1       1                 1        1                              1         1     1          1   1   1
10         1    1      1               1               1             1       1                 1        1                              1         1     1          1   1   0
5          1    1      1               1               1             1       1                 1        1                              1         1     1          1   1   0
1          1    1      1               1               1             1       1                 1        1                              1         1     1          1   1   0
7          1    1      1               1               1             1       1                 1        1                              1         1     1          1   0   1
1          1    1      1               1               1             1       1                 1        1                              1         1     1          1   0   1
2          1    1      1               1               1             1       1                 1        1                              1         1     1          1   0   1
3          1    1      1               1               1             1       1                 1        1                              1         1     1          1   0   1
1          1    1      1               1               1             1       1                 1        1                              1         1     1          1   0   1
7          1    1      1               1               1             1       1                 1        1                              1         1     1          0   1   1
1          1    1      1               1               1             1       1                 1        1                              1         1     0          0   1   1
           0    0      0               0               0             0       0                 0        0                              0         0     8          8  14  16
     thinness  1-19 years thinness 5-9 years Population Alcohol Total expenditure percentage expenditure Hepatitis B
1818                    1                 1          1       1                 1                          1          1    0
448                     1                 1          1       1                 1                          1          0    1
9                       1                 1          1       1                 1                          0          1    1
4                       1                 1          1       1                 0                          0          1    2
1                       1                 1          1       1                 0                          0          0    3
139                     1                 1          1       0                 0                          0          1    3
7                       1                 1          1       0                 0                          0          0    4
30                      1                 1          0       1                 1                          1          1    1
1                       1                 1          0       1                 0                          0          1    3
1                       1                 1          0       0                 0                          0          1    4
10                      0                 0          1       1                 1                          1          1    3
5                       0                 0          1       1                 1                          1          0    4
1                       0                 0          1       0                 0                          0          1    6
7                       1                 1          1       1                 1                          0          1    2
1                       1                 1          1       1                 1                          0          0    3
2                       1                 1          1       0                 0                          0          1    4
3                       1                 1          0       1                 1                          0          1    3
1                       1                 1          0       0                 0                          0          1    5
7                       1                 1          1       1                 1                          1          0    3
1                       1                 1          1       0                 1                          1          0    4
                       16                16         36     152               157                        177        470 1070
> |
```



- To find out which column has how much percentage of missing values.

```
pMiss = function(x){sum(is.na(x))/length(x)*100}
apply(LifeExp,2,pMiss)
```

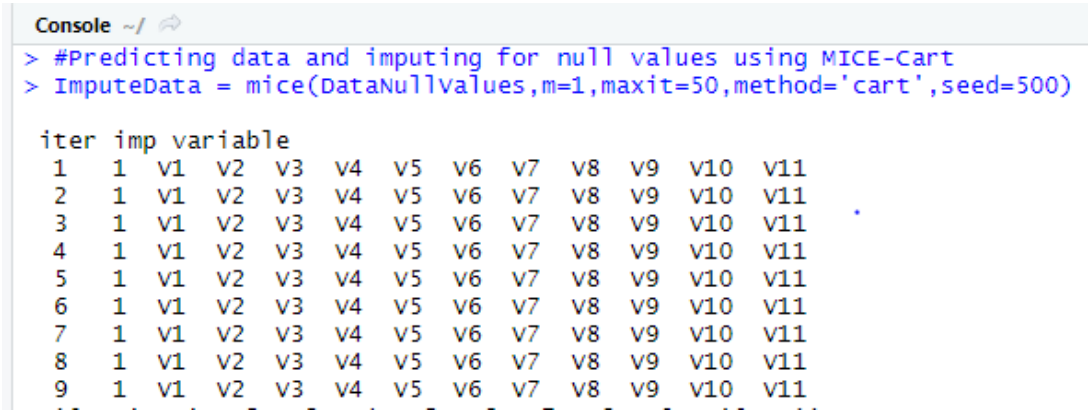


- Considering data variables having Null values

```
DataNullValues = cbind(LifeExp$Alcohol, LifeExp$`percentage expenditure`,
LifeExp$`Hepatitis B`,
                       LifeExp$BMI, LifeExp$Polio,
                       LifeExp$`Total expenditure`,LifeExp$Diphtheria,
                       LifeExp$GDP, LifeExp$Population, LifeExp$`thinness 1-19 years`,
                       LifeExp$`thinness 5-9 years`)
```

- Predicting data and imputing for null values using MICE-Cart

```
ImputeData = mice( DataNullValues, m=1, maxit=50, method='cart', seed=500)
```



1. m — Refers to 1 imputed data sets
2. maxit — Refers to no. of iterations taken to impute missing values
3. method — Refers to method used in imputation. I used CART.

#Saving Imputed Data Set Variables

Missing values I handled by imputing data of population and GDP from “The world bank data” website.
Link: https://data.worldbank.org/indicator/SP.POP.TOTL

completeData <- complete(ImputeData,1); completeData

```
Console ~/
> #Saving Imputed Data Set Varibales
> completeData <- complete(ImputeData,1); completeData
      V1          V2 V3   V4 V5   V6 V7          V8       V9  V10  V11
1   0.01   71.279624 65 19.1  6 8.16 65   584.25921 33736494 17.2 17.3
2   0.01   73.523582 62 18.6 58 8.18 62   612.69651   327582 17.5 17.5
3   0.01   73.219243 64 18.1 62 8.13 64   631.74498 31731688 17.7 17.7
4   0.01   78.184215 67 17.6 67 8.52 67   669.95900  3696958 17.9 18.0
5   0.01    7.097109 68 17.2 68 7.87 68    63.53723  2978599 18.2 18.2
6   0.01   79.679367 66 16.7 66 9.20 66   553.32894  2883167 18.4 18.4
7   0.01   56.762217 63 16.2 63 9.42 63   445.89330   284331 18.6 18.7
8   0.03   25.873925 64 15.7 64 8.33 64   373.36112  2729431 18.8 18.9
9   0.02   10.910156 63 15.2 63 6.73 63   369.83580 26616792 19.0 19.1
10  0.03   17.171518 64 14.7 58 7.43 58   272.56377  2589345 19.2 19.3
11  0.02    1.388648 66 14.2 58 8.70 58    25.29413   257798 19.3 19.5
12  0.02   15.296066 67 13.8  5 8.79  5   219.14135 24118979 19.5 19.7
13  0.01   11.089053 65 13.4 41 8.82 41   198.72854  2364851 19.7 19.9
14  0.01   16.887351 64 13.0 36 7.76 36   187.84595 21979923 19.9  2.2
15  0.01   10.574728 63 12.6 35 7.80 33   117.49698  2966463  2.1  2.4
```

#Verifying Imputed Variables- Null Values availability - should be 0

apply(completeData,2,pMiss)

```
> apply(completeData,2,pMiss)
 V1  V2  V3  V4  V5  V6  V7  V8  V9 V10 V11
  0   0   0   0   0   0   0   0   0   0   0
>
```

- Exporting Imputed Variables

write.csv(completeData, "C:/Users/vedan/OneDrive/Desktop/Statistics/Homework/Final Project/ToBeImputed.csv")

- Updated Data set

library(readr)
LifeExpImputed = read_csv("C:/Users/vedan/OneDrive/Desktop/Statistics/Homework/Final Project/LifeExpImputed.csv")

- Verifying Missing values count - should be 0

apply(LifeExpImputed,2,pMiss)

```
Console ~/
> #Verifying Missing values count - should be 0
> apply(LifeExpImputed,2,pMiss)
                       Country                           Year                         Status                Life expectancy                Adult Mortality
                             0                              0                              0                              0                              0
               infant deaths                        Alcohol      percentage expenditure                    Hepatitis B                       Measles
                             0                              0                              0                              0                              0
                           BMI             under-five deaths                         Polio            Total expenditure                    Diphtheria
                             0                              0                              0                              0                              0
                     HIV/AIDS                            GDP                     Population     thinness  1-19 years     thinness 5-9 years
                             0                              0                              0                              0                              0
Income composition of resources                     Schooling
                             0                              0
> |
```

HANDLING OUTLIERS:

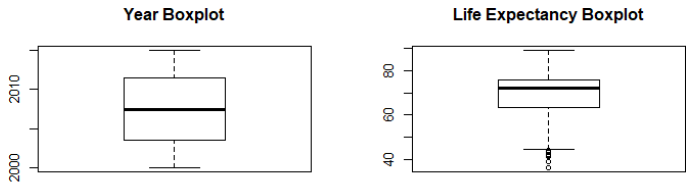
The data pulled from Kaggle link: https://www.kaggle.com/kumarajarshi/life-expectancy-who#Life%20Expectancy%20Data.csv had many extreme values or outliers. I tried to handle them using “winsorizing” method. But gave rise to new outliers in the data. Elimination of these outliers would have made dataset smaller in size and biased to few factors. Hence, I decided to proceed with my analysis with outliers.

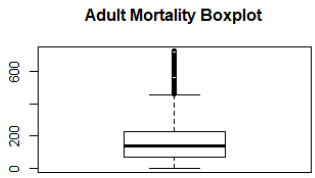
INFERNECE FROM UNIVARIATE PLOTS

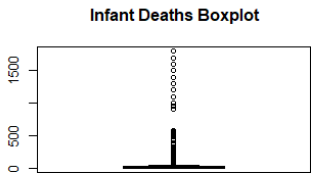
LifeExpImputed = read.csv("E:/LifeExpImputed.csv");LifeExpImputed

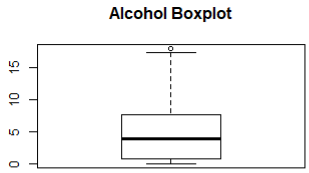
```
#.....Boxplot part 1.....
par(mfrow=c(2,2))
boxplot(LifeExpImputed$Year, main = "Year Boxplot")
boxplot(LifeExpImputed$`Life.expectancy` , main = "Life Expectancy Boxplot")
boxplot(LifeExpImputed$Adult.Mortality, main = "Adult Mortality Boxplot")

#.....Boxplot part 2.....
par(mfrow=c(2,2))
boxplot(LifeExpImputed$Alcohol, main = "Alcohol Boxplot")
boxplot(LifeExpImputed$percentage.expenditure, main = "Percentage expenditure Boxplot")
boxplot(LifeExpImputed$Hepatitis.B, main = "Hepatitis Boxplot" )
boxplot(LifeExpImputed$Measles, main = "Measles Boxplot")

#.....Boxplot part 3.....
par(mfrow=c(2,2))
boxplot(LifeExpImputed$BMI, main = "BMI Boxplot" )
boxplot(LifeExpImputed$under.five.deaths, main = "Under-five Deaths Boxplot" )
boxplot(LifeExpImputed$Polio, main = "Polio Boxplot" )
boxplot(LifeExpImputed$Total.expenditure, main = "Total expenditure Boxplot" )

#.....Boxplot part 4.....
par(mfrow=c(2,2))
boxplot(LifeExpImputed$Diphtheria, main = "Diphtheria Boxplot" )
```

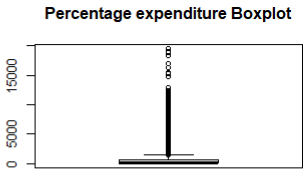
```
boxplot(LifeExpImputed$HIV.AIDS, main = "HIV/Aids Boxplot" )
boxplot(LifeExpImputed$GDP, main = "GDP Boxplot" )
boxplot(LifeExpImputed$Population, main = "Population Boxplot" )

#.....................................Boxplot part 5.....................................
par(mfrow=c(2,2))
boxplot(LifeExpImputed$thinness..1.19.years, main = "Thinness 1 - 19 years Boxplot" )
boxplot(LifeExpImputed$thinness.5.9.years, main = "Thinness 5 - 9 Years Boxplot" )
boxplot(LifeExpImputed$Income.composition.of.resources, main = "Income Composition of
resources Boxplot" )
boxplot(LifeExpImputed$Schooling, main = "Schooling Boxplot" )
```

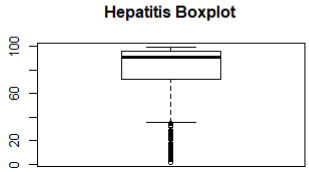


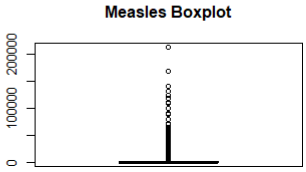


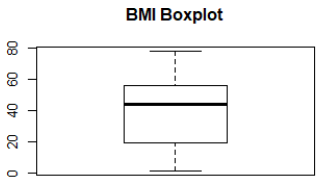


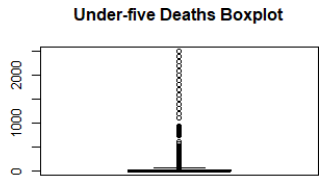


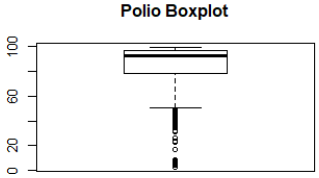


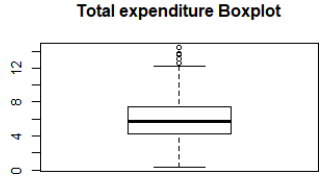


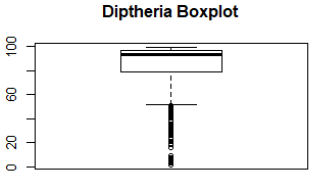


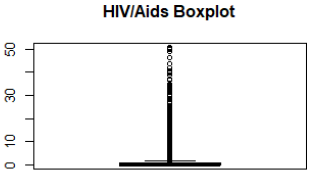

BMI Boxplot
0
20
40
60
80

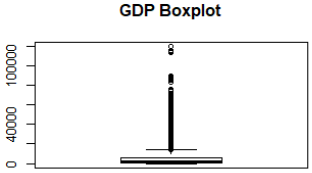

Under-five Deaths Boxplot
0
1000
2000

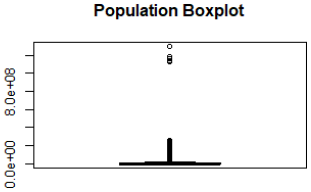

Polio Boxplot
0
20
60
100

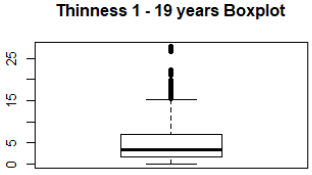

Total expenditure Boxplot
0
4
8
12

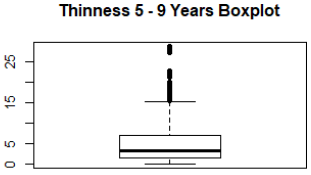

Diptheria Boxplot
0
20
60
100

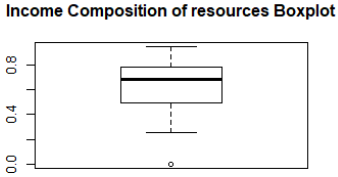

HIV/Aids Boxplot
0
10
30
50

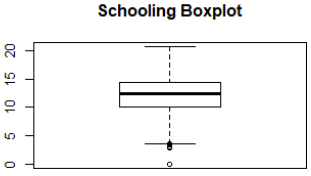

GDP Boxplot
0
40000
100000


Population Boxplot
0.0e+00
8.0e+08

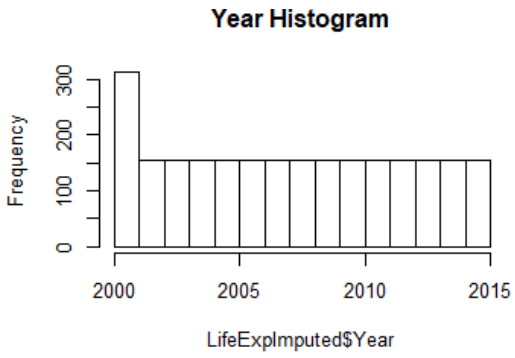

Thinness 1 - 19 years Boxplot
0
5
15
25

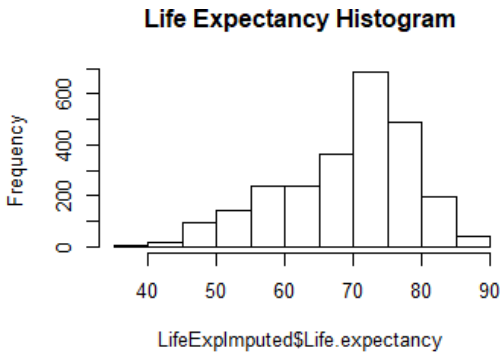

Thinness 5 - 9 Years Boxplot
0
5
15
25


Income Composition of resources Boxplot
0.0
0.4
0.8

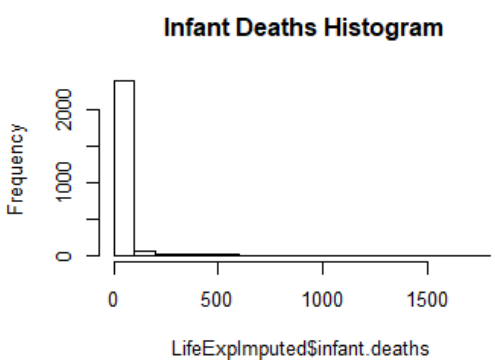

Schooling Boxplot
0
5
10
15
20

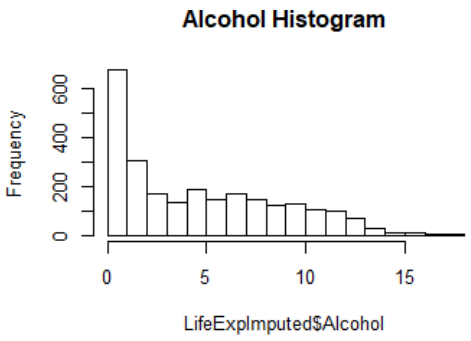
```
#.....................................Histograms.....................................
par(mfrow=c(2,2))
hist(LifeExpImputed$Year, main = "Year Histogram")
hist(LifeExpImputed$`Life.expectancy` , main = "Life Expectancy Histogram")
hist(LifeExpImputed$Adult.Mortality, main = "Adult Mortality Histogram")
hist(LifeExpImputed$infant.deaths, main = " Infant Deaths Histogram")

par(mfrow=c(2,2))
hist(LifeExpImputed$Alcohol, main = "Alcohol Histogram")
hist(LifeExpImputed$percentage.expenditure, main = "Percentage expenditure Histogram")
hist(LifeExpImputed$Hepatitis.B, main = "Hepatitis Histogram" )
hist(LifeExpImputed$Measles, main = "Measles Histogram")

par(mfrow=c(2,2))
hist(LifeExpImputed$BMI, main = "BMI Histogram" )
hist(LifeExpImputed$under.five.deaths, main = "Under-five Deaths Histogram" )
hist(LifeExpImputed$Polio, main = "Polio Histogram" )
hist(LifeExpImputed$Total.expenditure, main = "Total expenditure Histogram" )

par(mfrow=c(2,2))
hist(LifeExpImputed$Diphtheria, main = "Diphtheria Histogram" )
hist(LifeExpImputed$HIV.AIDS, main = "HIV/Aids Histogram" )
hist(LifeExpImputed$GDP, main = "GDP Histogram" )
hist(LifeExpImputed$Population, main = "Population Histogram" )

par(mfrow=c(2,2))
hist(LifeExpImputed$thinness..1.19.years, main = "Thinness 1 - 19 years Histogram" )
hist(LifeExpImputed$thinness.5.9.years, main = "Thinness 5 - 9 Years Histogram" )
hist(LifeExpImputed$Income.composition.of.resources, main = "Income Composition of
resources Histogram" )
hist(LifeExpImputed$Schooling, main = "Schooling Histogram" )
```

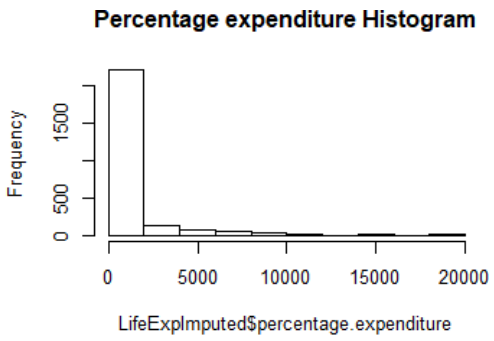

Year Histogram
Frequency
0
100
200
300
2000
2005
2010
2015
LifeExpImputed$Year

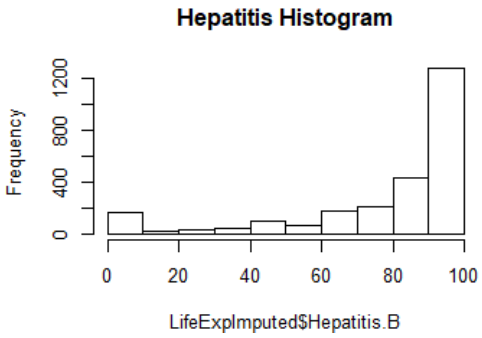

Life Expectancy Histogram
Frequency
0
200
400
600
40
50
60
70
80
90
LifeExpImputed$Life.expectancy

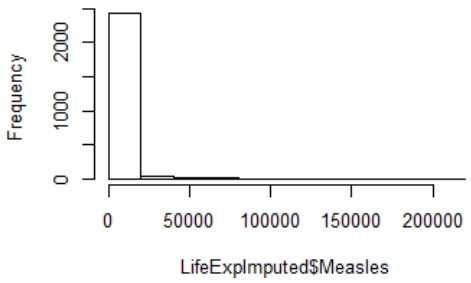

Adult Mortality Histogram
Frequency
0
200
400
0
200
400
600
LifeExpImputed$Adult.Mortality

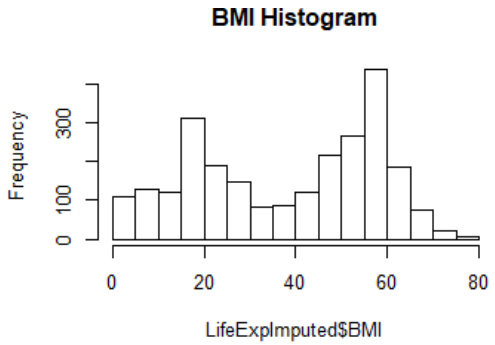

Infant Deaths Histogram
Frequency
0
1000
2000
0
500
1000
1500
LifeExpImputed$infant.deaths

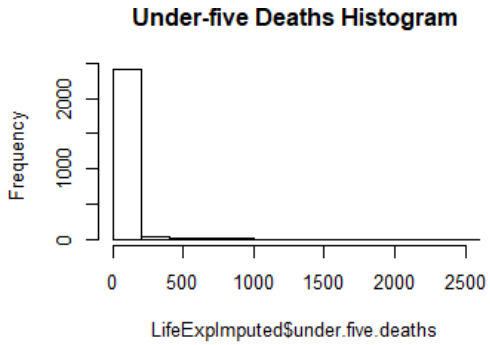

Alcohol Histogram
Frequency
0
200
400
600
0
5
10
15
LifeExpImputed$Alcohol

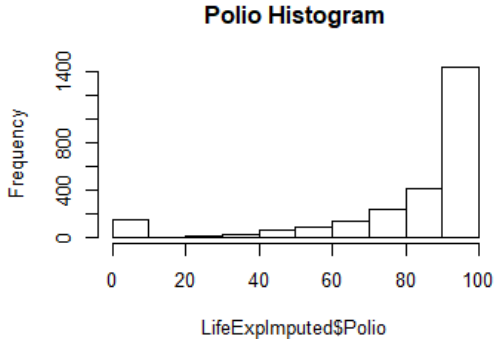

Percentage expenditure Histogram
Frequency
0
500
1500
0
5000
10000
15000
20000
LifeExpImputed$percentage.expenditure

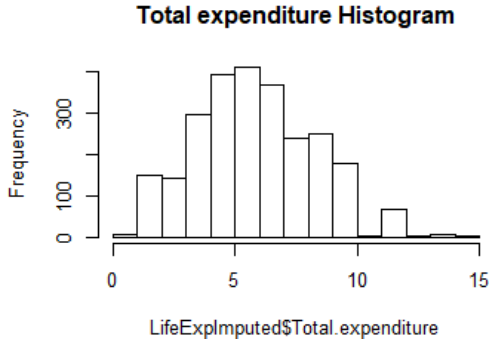

Hepatitis Histogram
Frequency
0
400
800
1200
0
20
40
60
80
100
LifeExpImputed$Hepatitis.B

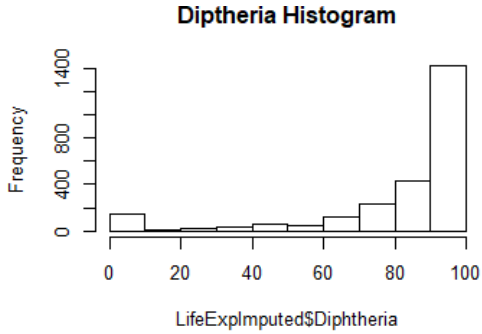

Measles Histogram
Frequency
0
1000
2000
0
50000
100000
150000
200000
LifeExpImputed$Measles

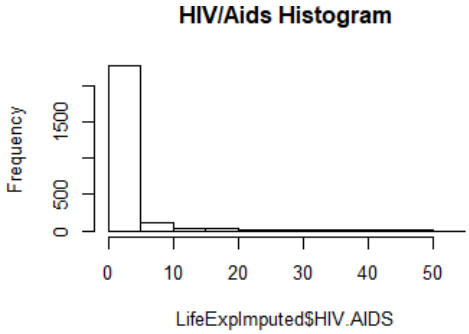

BMI Histogram
Frequency
0
100
300
0
20
40
60
80
LifeExpImputed$BMI

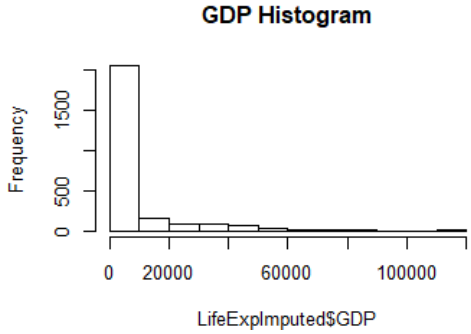

Under-five Deaths Histogram
Frequency
0
1000
2000
0
500
1000
1500
2000
2500
LifeExpImputed$under.five.deaths

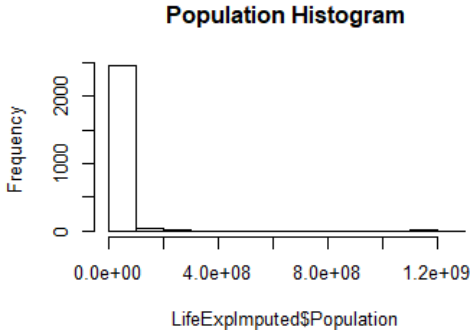

Polio Histogram
Frequency
0
400
800
1400
0
20
40
60
80
100
LifeExpImputed$Polio


Total expenditure Histogram
Frequency
0
100
300
0
5
10
15
LifeExpImputed$Total.expenditure


Diptheria Histogram
Frequency
0
400
800
1400
0
20
40
60
80
100
LifeExpImputed$Diphtheria

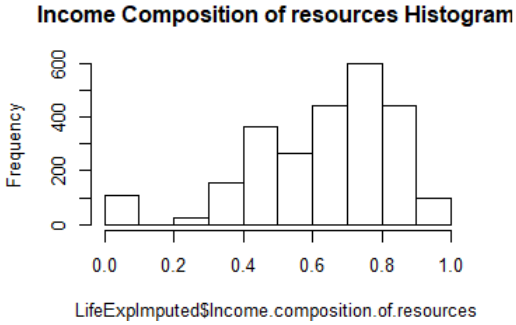

HIV/Aids Histogram
Frequency
0
500
1500
0
10
20
30
40
50
LifeExpImputed$HIV.AIDS

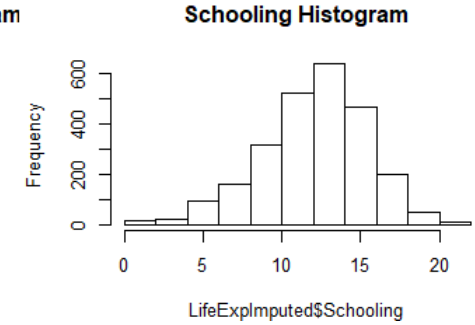

GDP Histogram
Frequency
0
500
1500
0
20000
60000
100000
LifeExpImputed$GDP


Population Histogram
Frequency
0
1000
2000
0.0e+00
4.0e+08
8.0e+08
1.2e+09
LifeExpImputed$Population


Thinness 1 - 19 years Histogram
Frequency
0
200
600
0
5
10
15
20
25
LifeExpImputed$thinness..1.19.years

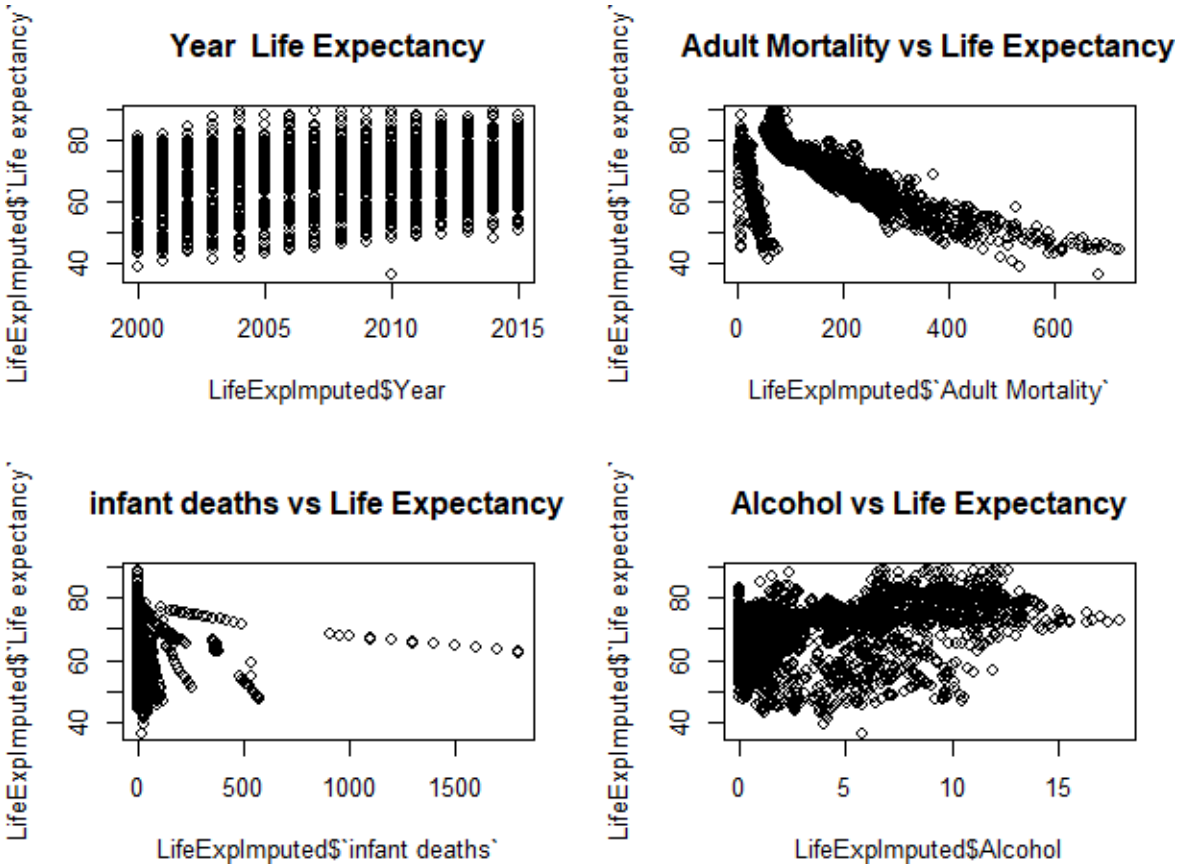

Thinness 5 - 9 Years Histogram
Frequency
0
200
600
0
5
10
15
20
25
30
LifeExpImputed$thinness.5.9.years


Income Composition of resources Histogram
Frequency
0
200
400
600
0.0
0.2
0.4
0.6
0.8
1.0
LifeExpImputed$Income.composition.of.resources

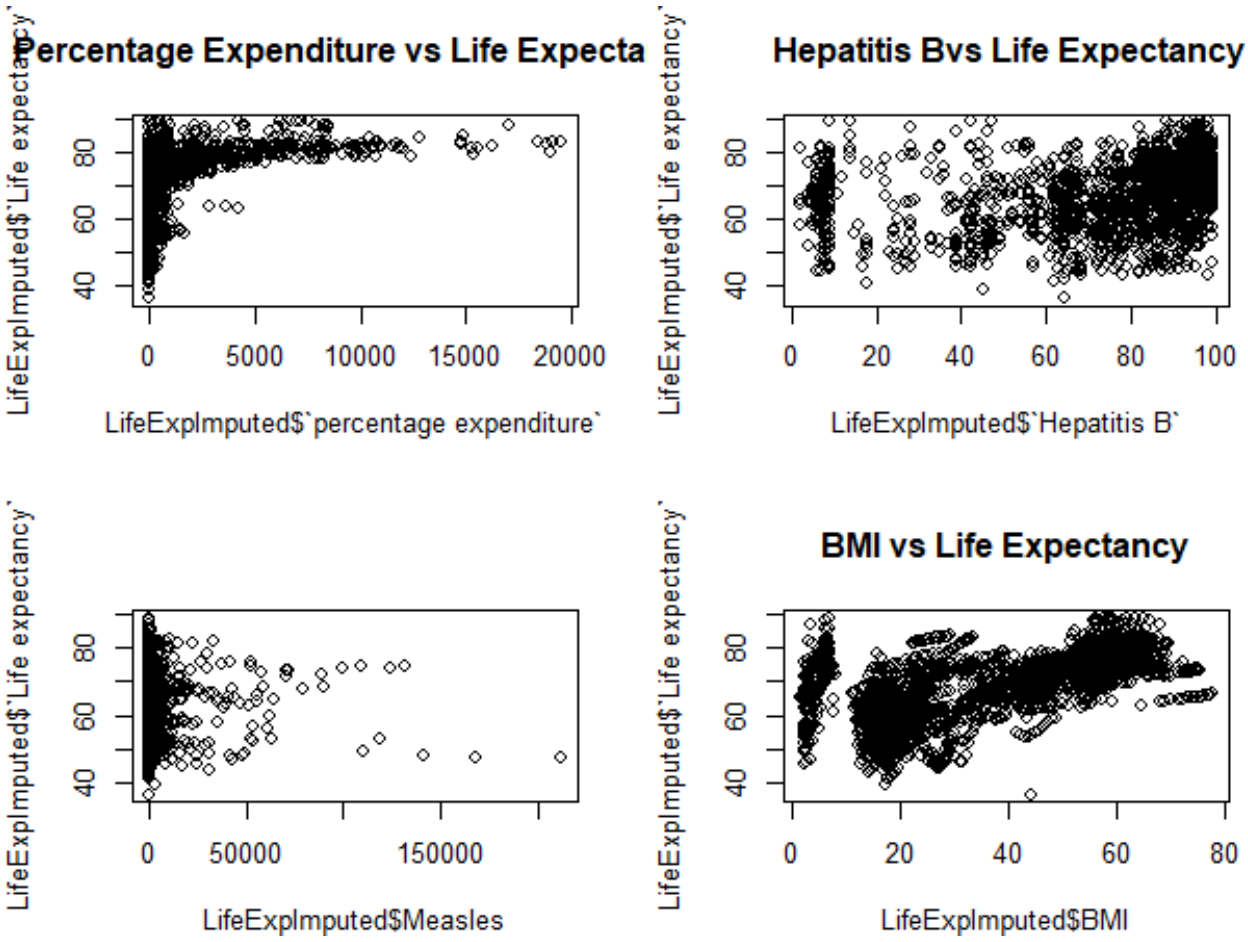

Schooling Histogram
Frequency
0
200
400
600
0
5
10
15
20
LifeExpImputed$Schooling

INFERNECE FROM BIVARIATE PLOTS

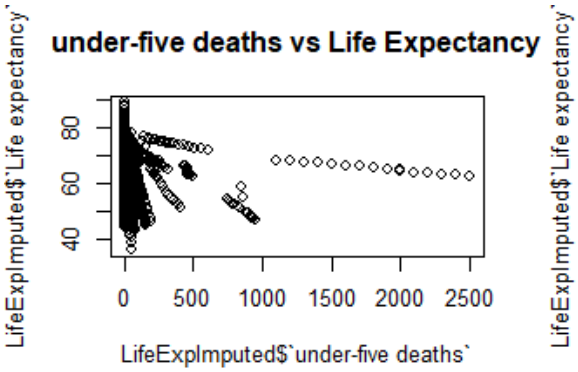
```
par(mfrow=c(2,2))
plot(LifeExpImputed$Year,LifeExpImputed$`Life expectancy`,main = "Year Life Expectancy")
plot(LifeExpImputed$`Adult Mortality`, LifeExpImputed$`Life expectancy`,main = "Adult
Mortality vs Life Expectancy")
plot(LifeExpImputed$`infant deaths`, LifeExpImputed$`Life expectancy`,main = "infant deaths
vs Life Expectancy")
plot(LifeExpImputed$Alcohol,LifeExpImputed$`Life expectancy`,main = "Alcohol vs Life
Expectancy")
```

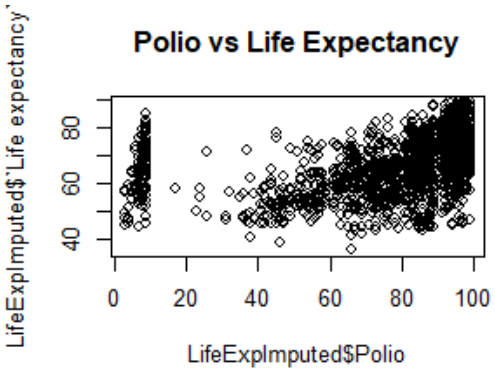


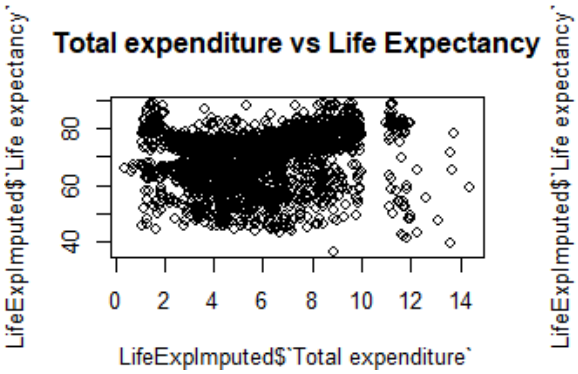

```
par(mfrow=c(2,2))
plot(LifeExpImputed$`percentage expenditure`, LifeExpImputed$`Life expectancy`,main = "
Percentage Expenditure vs Life Expectancy")
plot(LifeExpImputed$`Hepatitis B`, LifeExpImputed$`Life expectancy`,main = " Hepatitis Bvs
Life Expectancy")
plot(LifeExpImputed$Measles, LifeExpImputed$`Life expectancy`)
plot(LifeExpImputed$BMI, LifeExpImputed$`Life expectancy`,main = "BMI vs Life
Expectancy")
```

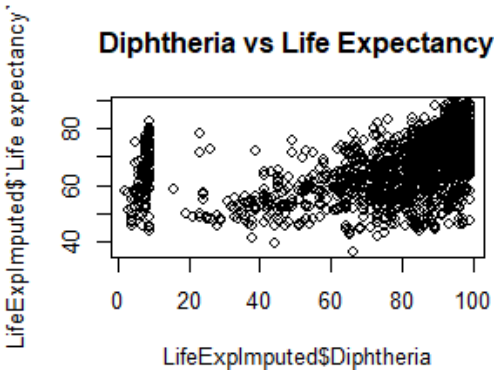



```
par(mfrow=c(2,2))
plot(LifeExpImputed$`under-five deaths`, LifeExpImputed$`Life expectancy`,main = "under-five deaths vs Life Expectancy")
plot(LifeExpImputed$Polio, LifeExpImputed$`Life expectancy`,main = "Polio vs Life Expectancy")
plot(LifeExpImputed$`Total expenditure`, LifeExpImputed$`Life expectancy`,main = "Total expenditure vs Life Expectancy")
plot(LifeExpImputed$Diphtheria, LifeExpImputed$`Life expectancy`,main = "Diphtheria vs Life Expectancy")
```

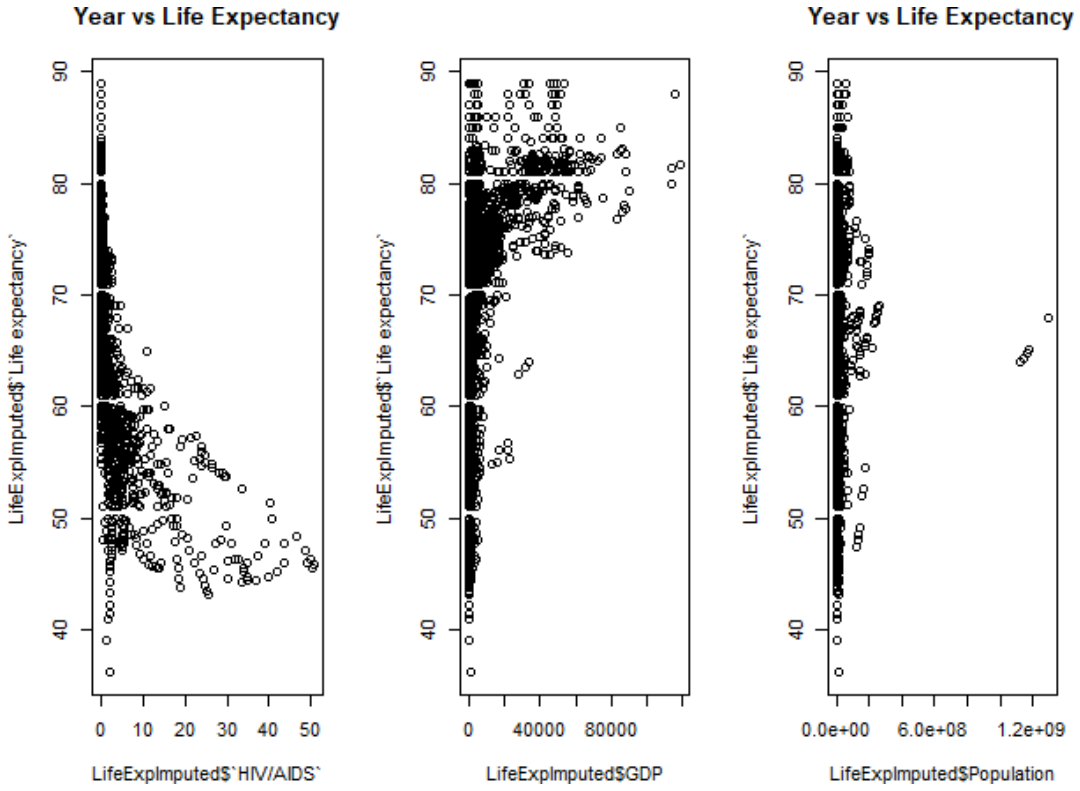




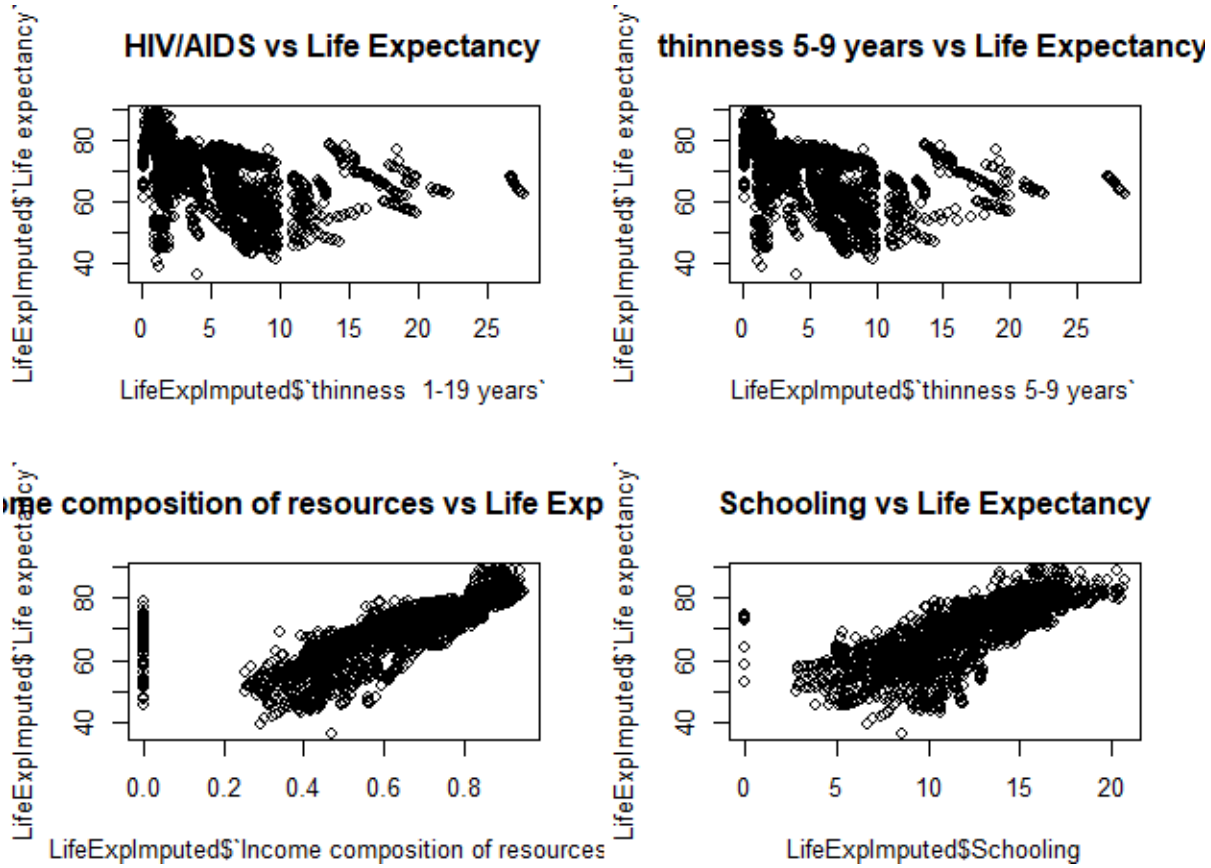





```
par(mfrow=c(1,3))
plot(LifeExpImputed$`HIV/AIDS`, LifeExpImputed$`Life expectancy`,main = "Year vs Life Expectancy")
plot(LifeExpImputed$GDP, LifeExpImputed$`Life expectancy`)
plot(LifeExpImputed$Population, LifeExpImputed$`Life expectancy`,main = "Year vs Life Expectancy")
```




```
par(mfrow=c(2,2))
plot(LifeExpImputed$`thinness 1-19 years`, LifeExpImputed$`Life expectancy`,main =
"HIV/AIDS vs Life Expectancy")
plot(LifeExpImputed$`thinness 5-9 years`, LifeExpImputed$`Life expectancy`,main = "thinness
5-9 years vs Life Expectancy")
plot(LifeExpImputed$`Income composition of resources`, LifeExpImputed$`Life
expectancy`,main = "Income composition of resources vs Life Expectancy")
plot(LifeExpImputed$Schooling, LifeExpImputed$`Life expectancy`,main = "Schooling vs Life
Expectancy")
```



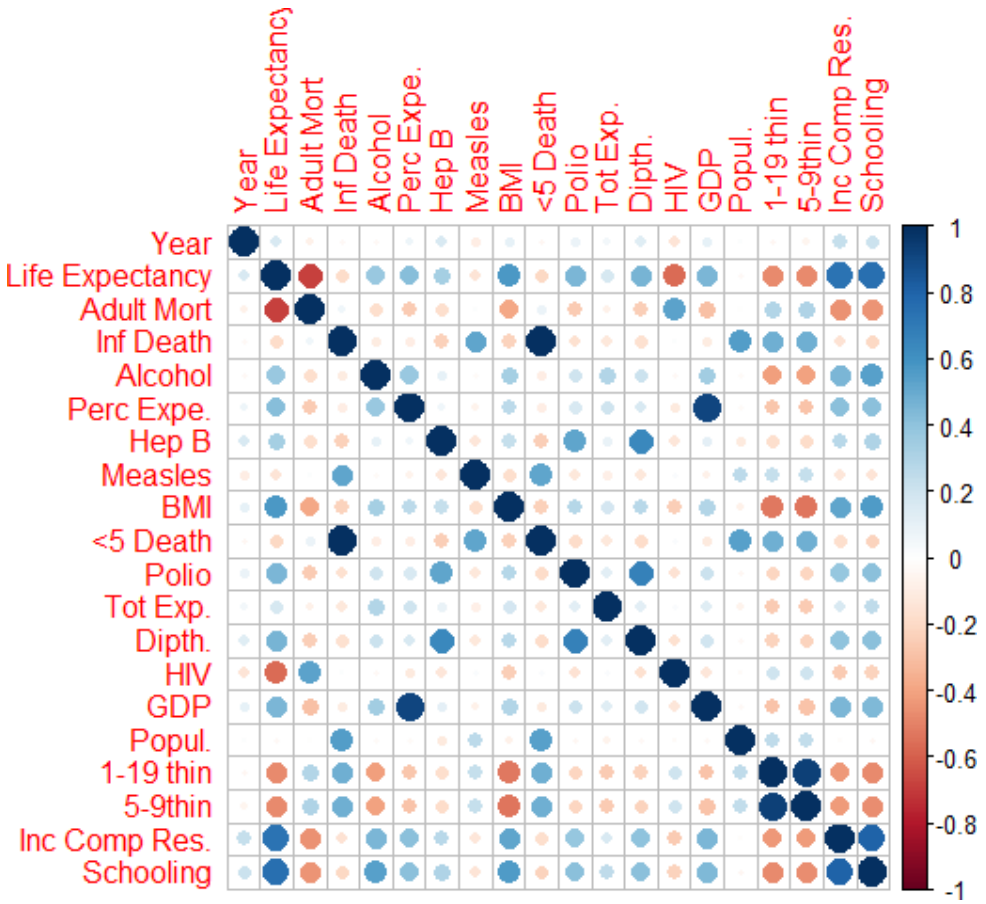
- Bivariate Analysis:

From the graphs, I can see that most of the predictors have no linear relationship with the dependent variable – Life Expectancy. Since all variables are associated with every country present in the data, there are outliers present on all linear relationship with dependent variables. And columns like HIV/AIDS and thinness between 1- 19 years vary drastically based on the population of each country, they do not how any patterns on the graph. Whereas, Income composition of resources and schooling had a roper linear relationship with life expectancy making them crucial variables when modelling the dataset.

Correlation plot:

```
par(mfrow=c(1,1))


library(corrplot)
Lifecor = data.frame(LifeExpImputed$Year, LifeExpImputed$`Life expectancy`,
LifeExpImputed$`Adult Mortality`,
                     LifeExpImputed$`infant deaths`, LifeExpImputed$Alcohol,
LifeExpImputed$`percentage expenditure`,
                     LifeExpImputed$`Hepatitis B`,LifeExpImputed$Measles, LifeExpImputed$BMI,
```

```
            LifeExpImputed$`under-five deaths`, LifeExpImputed$Polio,
LifeExpImputed$`Total expenditure`,
            LifeExpImputed$Diphtheria, LifeExpImputed$`HIV/AIDS`,
LifeExpImputed$GDP, LifeExpImputed$Population,
            LifeExpImputed$`thinness 1-19 years`, LifeExpImputed$`thinness 5-9 years`,
            LifeExpImputed$`Income composition of resources`, LifeExpImputed$Schooling)
names(Lifecor) = c("Year","Life Expectancy","Adult Mort","Inf Death",
            "Alcohol","Perc Expe.","Hep B","Measles","BMI",
            "<5 Death", "Polio", "Tot Exp.", "Dipth.",
            "HIV", "GDP", "Popul.", "1-19 thin", "5-9thin", "Inc Comp Res.", "Schooling" )

lifecorrelation = cor(Lifecor)
corrplot(lifecorrelation)
```



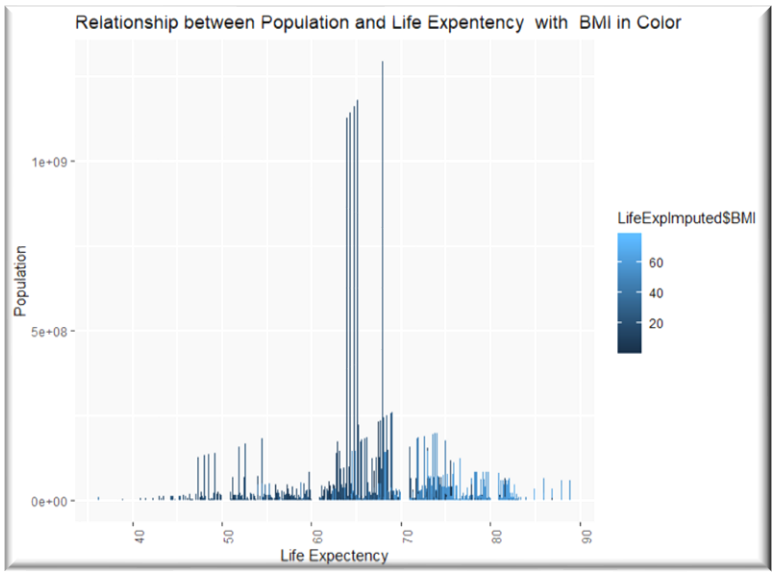
From the correlation plot, I can see that there is high autocorrelation between the following variables:

- 1-19 thinness and 5-9 thinness
- Income composition of resources and Schooling
- Under 5 deaths and infant deaths
- GDP and percentage expenditure

Predictors which have auto correlation do not giveme proper significance values while performing modeling on them and all these predictors should be eliminated or undergo some transformations before data models.

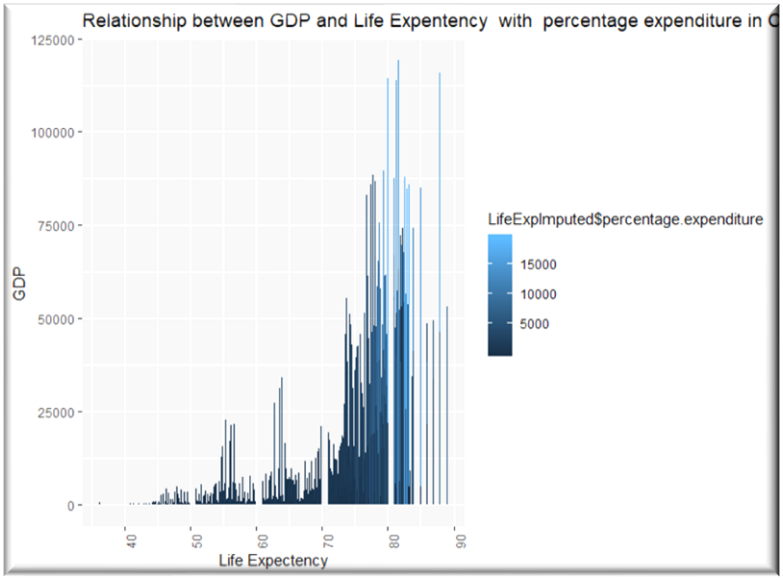
INFERNECES FROM MULTIVARIATE PLOTS

```
library(ggplot2)
#Relationship between population and Life Expectency with BMI in Color
ggplot(LifeExpImputed, aes(x=LifeExpImputed$Life.expectancy ,
y=LifeExpImputed$Population)) +
  geom_bar(aes(fill =LifeExpImputed$BMI ), stat="identity",position=position_dodge()) +
  labs(title = "Relationship between Population and Life Expentency with BMI in Color",
       x= " Life Expectency ",
       y = "Population",
       colour="BMI") +
  theme(axis.text.x = element_text(angle = 90))
```



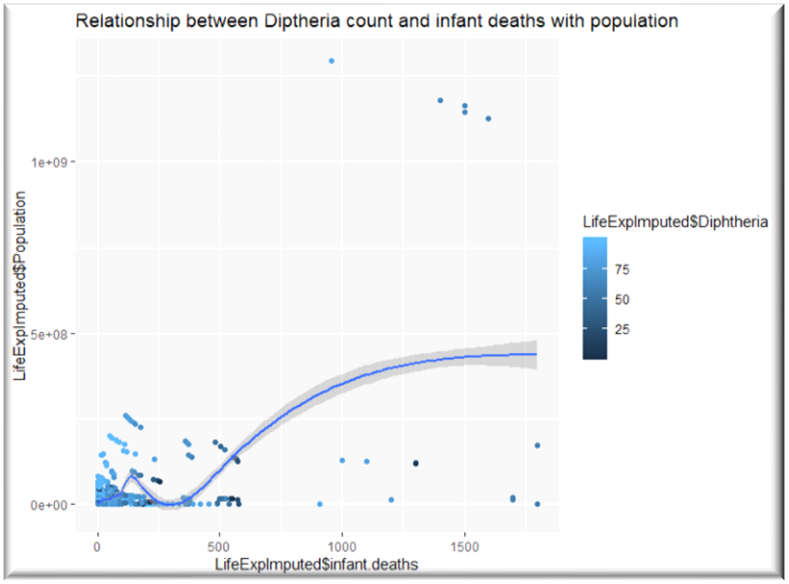
- From graph 1, I found out that people with high BMI have better life expectancies based on population.

```
#Relationship between GDP and Life Expentency with percentage expenditure in Color
ggplot(LifeExpImputed, aes(x=LifeExpImputed$Life.expectancy , y=LifeExpImputed$GDP)) +
geom_bar(aes(fill =LifeExpImputed$percentage.expenditure ),
stat="identity",position=position_dodge()) +
  labs(title = "Relationship between GDP and Life Expentency with percentage expenditure in
Color",
       x= "Life Expectency",
       y = "GDP",
       colour="Percentage expenditure") +
  theme(axis.text.x = element_text(angle = 90))
```



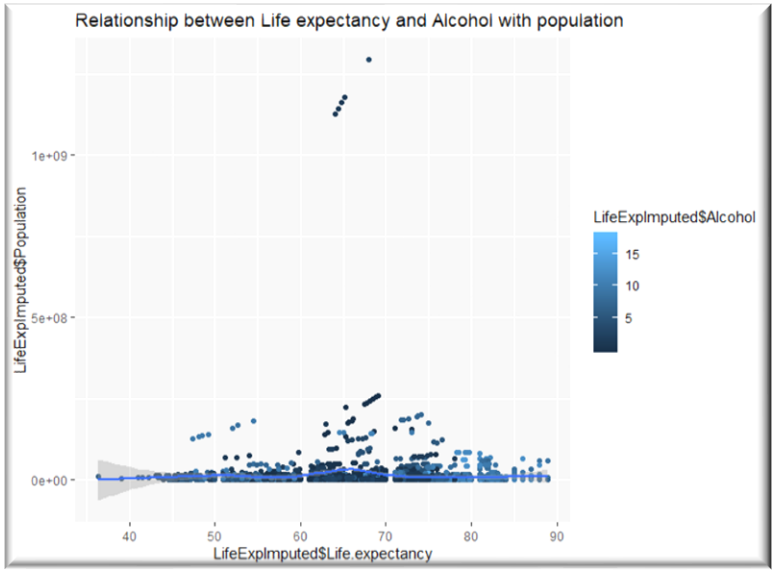
- From graph 2, I found out countries with higher GDP have good Life expectancies than the others.

#Relationship between Diptheria count and infant deaths with population
qplot(LifeExpImputed$infant.deaths,LifeExpImputed$Population,
color =LifeExpImputed$Diphtheria ,geom = c("point","smooth"),
main = "Relationship between Diptheria count and infant deaths with population ")



- From graph 4, Infant deaths is related with Diphtheria rate in every country and outliers are analyzed.

```
#Relationship between Life expectancy and Alcohol with populationr
qplot(LifeExpImputed$Life.expectancy,LifeExpImputed$Population,
      color =LifeExpImputed$Alcohol ,geom = c("point","smooth"),
       main = "Relationship between Life expectancy and Alcohol with population ")
```



- From graph 3, Relationship between alcoholism rate and life expectancies is found alongside with population.

MODEL RECOGNITION FROM STEPWISE REGRESSION PROCESS

#..... Linear regression with outliers ……..

- **Full linear regression**

model1 = lm(lifeEx$Life.expectancy ~ lifeEx$Adult.Mortality + lifeEx$infant.deaths + lifeEx$Alcohol + lifeEx$percentage.expenditure + lifeEx$Hepatitis.B + lifeEx$Measles + lifeEx$BMI + lifeEx$under.five.deaths +lifeEx$Polio + lifeEx$Total.expenditure + lifeEx$Diphtheria + lifeEx$HIV.AIDS + lifeEx$GDP + lifeEx$Population + lifeEx$thinness..1.19.years + lifeEx$thinness.5.9.years + lifeEx$Income.composition.of.resources + lifeEx$Schooling)
summary(model1)

```
Residuals:
     Min       1Q   Median       3Q      Max
-21.1454  -2.1556  -0.0846   2.1853  19.0000

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      5.315e+01  5.652e-01  94.036  < 2e-16 ***
lifeEx$Adult.Mortality          -1.698e-02  8.283e-04 -20.494  < 2e-16 ***
lifeEx$infant.deaths             8.969e-02  8.252e-03  10.870  < 2e-16 ***
lifeEx$Alcohol                  -7.832e-03  2.532e-02  -0.309   0.7571
lifeEx$percentage.expenditure    1.997e-04  9.476e-05   2.107   0.0352 *
lifeEx$Hepatitis.B               1.957e-03  4.107e-03   0.476   0.6338
lifeEx$Measles                  -1.121e-05  8.437e-06  -1.328   0.1842
lifeEx$BMI                       3.581e-02  5.254e-03   6.814 1.18e-11 ***
lifeEx$under.five.deaths        -6.606e-02  6.054e-03 -10.911  < 2e-16 ***
lifeEx$Polio                     2.327e-02  4.728e-03   4.922 9.15e-07 ***
lifeEx$Total.expenditure         3.399e-02  3.572e-02   0.952   0.3414
lifeEx$Diphtheria                2.548e-02  5.230e-03   4.872 1.18e-06 ***
lifeEx$HIV.AIDS                 -4.761e-01  1.737e-02 -27.413  < 2e-16 ***
lifeEx$GDP                       3.149e-05  1.454e-05   2.166   0.0304 *
lifeEx$Population               -1.150e-09  1.638e-09  -0.702   0.4826
lifeEx$thinness..1.19.years     -8.769e-02  4.926e-02  -1.780   0.0752 .
lifeEx$thinness.5.9.years       -2.699e-02  4.865e-02  -0.555   0.5792
lifeEx$Income.composition.of.resources  7.325e+00  6.580e-01  11.133  < 2e-16 ***
lifeEx$Schooling                 8.095e-01  4.496e-02  18.007  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.911 on 2477 degrees of freedom
Multiple R-squared:  0.8349,    Adjusted R-squared:  0.8337
F-statistic: 695.7 on 18 and 2477 DF,  p-value: < 2.2e-16
```

- Few independent variables in full linear model are not significant. To select best combination of independent variables I tried stepwise AIC.

model1_Step = stepAIC(model1)

- The result of stepwise AIC gave me a combination of independent variable with least AIC value

```
Step:  AIC=6818.34
lifeEx$Life.expectancy ~ lifeEx$Adult.Mortality + lifeEx$infant.deaths +
    lifeEx$percentage.expenditure + lifeEx$BMI + lifeEx$under.five.deaths +
    lifeEx$Polio + lifeEx$Diphtheria + lifeEx$HIV.AIDS + lifeEx$GDP +
    lifeEx$thinness..1.19.years + lifeEx$Income.composition.of.resources +
    lifeEx$Schooling

                                     Df Sum of Sq   RSS    AIC
<none>                                             37939 6818.3
- lifeEx$GDP                         1      66.7 38006 6820.7
- lifeEx$percentage.expenditure      1      81.0 38020 6821.7
- lifeEx$thinness..1.19.years        1     327.0 38266 6837.8
- lifeEx$Polio                       1     397.8 38337 6842.4
- lifeEx$Diphtheria                  1     491.7 38431 6848.5
- lifeEx$BMI                         1     770.4 38709 6866.5
- lifeEx$Income.composition.of.resources 1    1901.2 39840 6938.4
- lifeEx$infant.deaths               1    1905.0 39844 6938.6
- lifeEx$under.five.deaths           1    1947.9 39887 6941.3
- lifeEx$Schooling                   1    5546.1 43485 7156.9
- lifeEx$Adult.Mortality             1    6484.8 44424 7210.2
- lifeEx$HIV.AIDS                    1   11674.1 49613 7486.0
```

- Using the result of stepwise AIC, created model2

```
model2 =lm(lifeEx$Life.expectancy ~ lifeEx$Adult.Mortality + lifeEx$infant.deaths +
        lifeEx$percentage.expenditure + lifeEx$BMI + lifeEx$under.five.deaths +
        lifeEx$Polio + lifeEx$Diphtheria + lifeEx$HIV.AIDS + lifeEx$GDP +
        lifeEx$thinness..1.19.years + lifeEx$Income.composition.of.resources +
        lifeEx$Schooling)
summary(model2)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-20.9303  -2.1606  -0.0927   2.1916  18.9846

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                          5.326e+01  5.303e-01 100.443  < 2e-16 ***
lifeEx$Adult.Mortality              -1.692e-02  8.215e-04 -20.601  < 2e-16 ***
lifeEx$infant.deaths                 8.910e-02  7.980e-03  11.166  < 2e-16 ***
lifeEx$percentage.expenditure        2.116e-04  9.191e-05   2.302   0.0214 *
lifeEx$BMI                           3.685e-02  5.189e-03   7.101 1.61e-12 ***
lifeEx$under.five.deaths            -6.633e-02  5.874e-03 -11.291  < 2e-16 ***
lifeEx$Polio                         2.381e-02  4.666e-03   5.102 3.61e-07 ***
lifeEx$Diphtheria                    2.656e-02  4.683e-03   5.673 1.57e-08 ***
lifeEx$HIV.AIDS                     -4.763e-01  1.723e-02 -27.641  < 2e-16 ***
lifeEx$GDP                           2.985e-05  1.428e-05   2.090   0.0367 *
lifeEx$thinness..1.19.years         -1.103e-01  2.384e-02  -4.626 3.91e-06 ***
lifeEx$Income.composition.of.resources  7.301e+00  6.546e-01  11.155  < 2e-16 ***
lifeEx$Schooling                     8.103e-01  4.253e-02  19.052  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.909 on 2483 degrees of freedom
Multiple R-squared:  0.8346,    Adjusted R-squared:  0.8338
F-statistic:  1044 on 12 and 2483 DF,  p-value: < 2.2e-16
```

- In the model 2, all independent variable are significant values (since p-value < 0.05) but when I check VIF values some of them were highly correlated. This gives rise to multicollinearity.

```
> vif(model2)
                lifeEx$Adult.Mortality                    lifeEx$infant.deaths
                              1.739952                              166.119387
         lifeEx$percentage.expenditure                             lifeEx$BMI
                              6.480950                                1.746374
              lifeEx$under.five.deaths                           lifeEx$Polio
                            166.422948                                1.925679
                     lifeEx$Diphtheria                         lifeEx$HIV.AIDS
                              1.967909                                1.429184
                             lifeEx$GDP           lifeEx$thinness..1.19.years
                              6.784722                                1.868024
lifeEx$Income.composition.of.resources                    lifeEx$Schooling
                              3.129120                                3.316563
```

- Infant death rate, percentage expenditure, number of deaths under age five and GDP have VIF > 5.

- This leads to multicollinearity and violations of assumption of linear regression. Hence creating next model without these independent variables.

Model 3

```
model3 =lm(lifeEx$Life.expectancy ~ lifeEx$Adult.Mortality + lifeEx$BMI +
        lifeEx$Polio + lifeEx$Diphtheria + lifeEx$HIV.AIDS +
        lifeEx$thinness..1.19.years + lifeEx$Income.composition.of.resources +
        lifeEx$Schooling)
summary(model3)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-20.9369  -2.2058  -0.0415   2.1739  20.1799

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                      51.4711549  0.5349068  96.225  < 2e-16 ***
lifeEx$Adult.Mortality           -0.0179061  0.0008498 -21.071  < 2e-16 ***
lifeEx$BMI                        0.0370862  0.0053909   6.879 7.58e-12 ***
lifeEx$Polio                      0.0269280  0.0048359   5.568 2.85e-08 ***
lifeEx$Diphtheria                 0.0317192  0.0048307   6.566 6.26e-11 ***
lifeEx$HIV.AIDS                  -0.4805646  0.0178561 -26.913  < 2e-16 ***
lifeEx$thinness..1.19.years      -0.1148021  0.0221250  -5.189 2.29e-07 ***
lifeEx$Income.composition.of.resources  8.7971481  0.6659270  13.210  < 2e-16 ***
lifeEx$Schooling                  0.8655730  0.0438120  19.757  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.066 on 2487 degrees of freedom
Multiple R-squared:  0.8208,    Adjusted R-squared:  0.8202
F-statistic:  1423 on 8 and 2487 DF,  p-value: < 2.2e-16
```

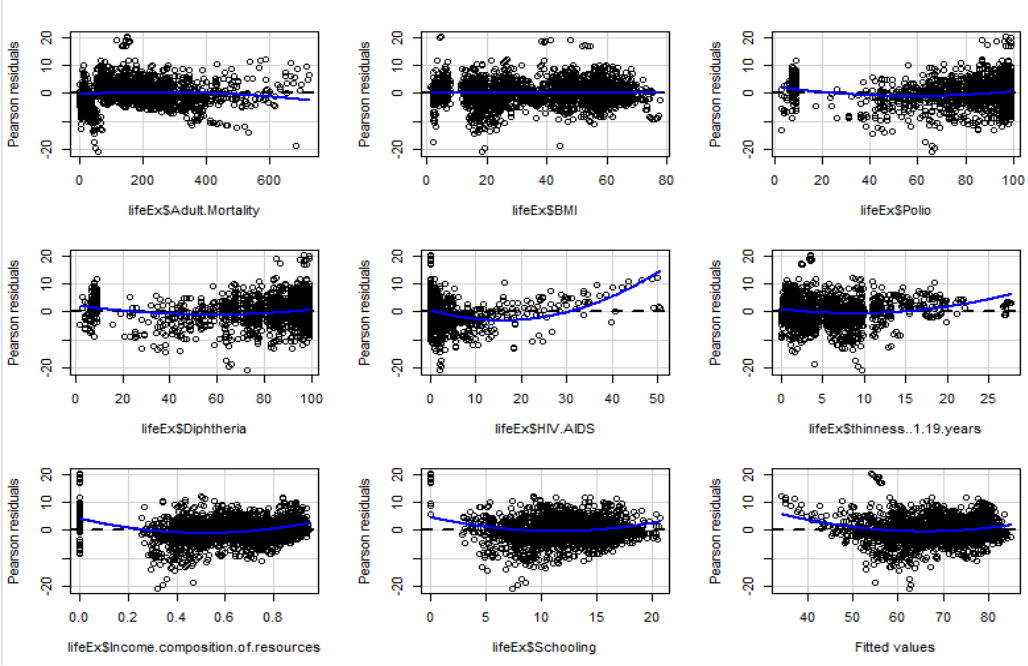
```
> AIC(model3)
[1] 14096.39
> BIC(model3)
[1] 14154.61
> vif(model3)
                lifeEx$Adult.Mortality                            lifeEx$BMI
                              1.721069                              1.741776
                        lifeEx$Polio                     lifeEx$Diphtheria
                              1.911546                              1.935539
                     lifeEx$HIV.AIDS           lifeEx$thinness..1.19.years
                              1.418658                              1.487494
lifeEx$Income.composition.of.resources                  lifeEx$Schooling
                              2.993286                              3.252939
```

- The independent variables are significant values. P-value is < 0.001, which means the variable has 99.9% of confidence according to T-test.
- Also, AIC and BIC value is less than previous value.
- The independent variables have VIF values less than 5, then no multicollinearity.
- Adjust R- square = 0.8202, which determines 82.02% variance in Life expectancy.

```
> durbinWatsonTest(model3)
 lag Autocorrelation D-W Statistic p-value
   1       0.6821965     0.6323139       0
 Alternative hypothesis: rho != 0
```

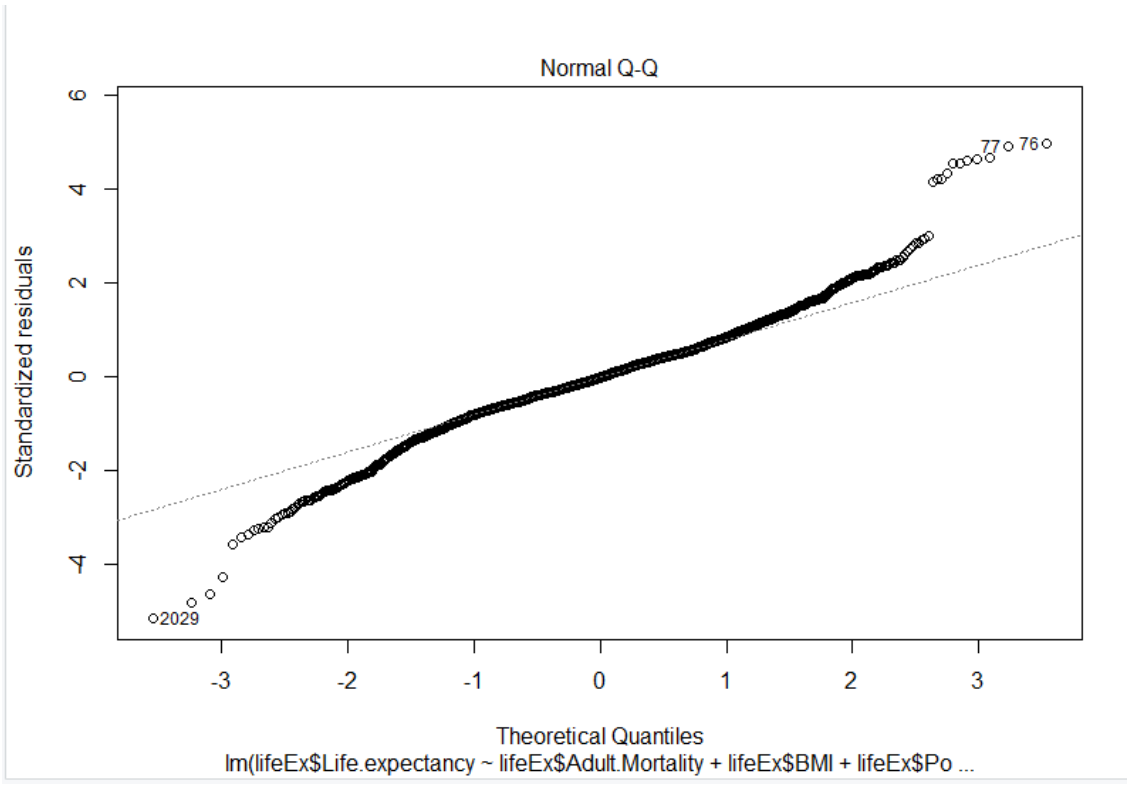
- In Durbin Watson Test, DW statistics < 2 , indicating that the residuals are correlated, and violation of the assumption of randomness and independence.

Residual plots

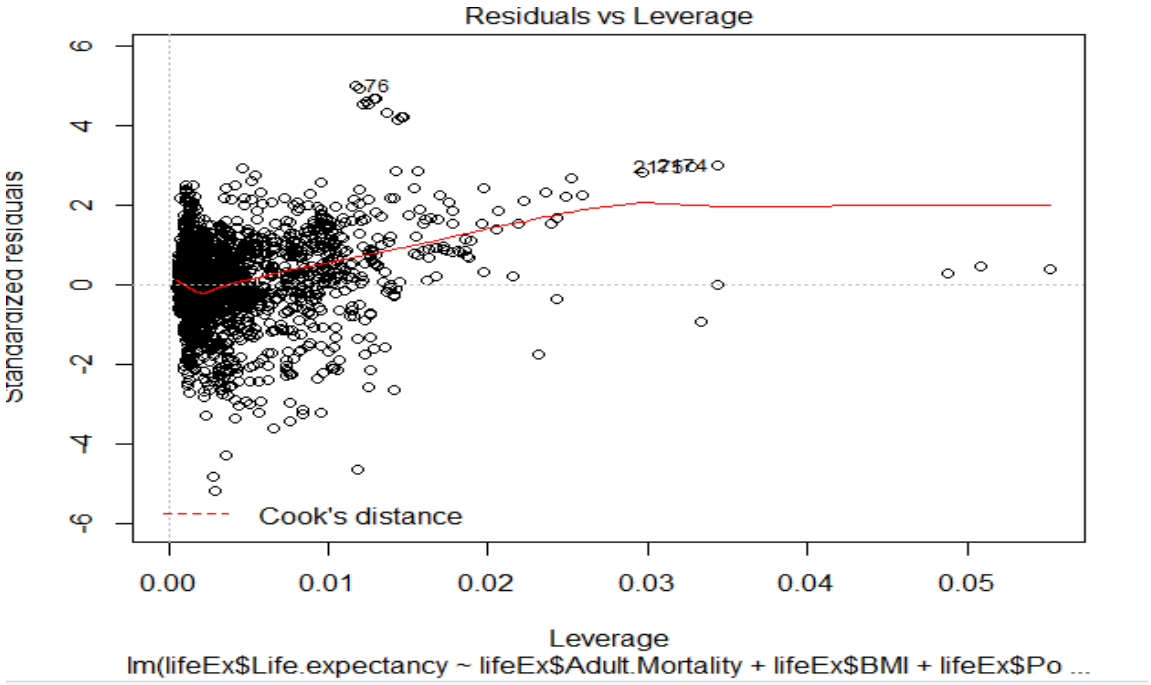


- Linear regression model is highly sensitive to outliers, hence even when the model gives high R-square value it violates linear regression assumption.

- The variance in residuals is not constant, hence assumption of homoscedasticity has not met.



- According to Q-Q graph, the residuals doesn't follow normal pattern. The main reason is outliers.



According to cook's distance graph, linear regression model is highly influenced by outliers.

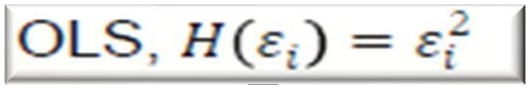
```
> print(cv_model)
Linear Regression

2496 samples
   8 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 2248, 2247, 2245, 2246, 2248, 2246, ...
Resampling results:

  RMSE      Rsquared  MAE
  4.081173  0.81894   3.004081

Tuning parameter 'intercept' was held constant at a value of TRUE
> #Rsquare value of every fold
> cv_model$resample
       RMSE  Rsquared      MAE Resample
1  4.693738 0.7852761 3.438392   Fold01
2  3.829296 0.8555441 2.778281   Fold02
3  3.579516 0.8579499 2.665677   Fold03
4  4.067288 0.8376782 2.945911   Fold04
5  4.108711 0.8059327 2.955726   Fold05
6  4.481222 0.7803930 3.324886   Fold06
7  3.893934 0.8067905 2.896623   Fold07
8  4.030631 0.8056577 2.999324   Fold08
9  4.109714 0.8212279 3.032830   Fold09
10 4.017683 0.8329501 3.003158   Fold10
```

K-folded cross validation gave good value of RMSE but this model doesn't obeys linear regression assumption.

MODEL RECOGNITION FROM ROBUST REGRESSION PROCESS

- In ordinary least regression (generally known as linear regression) is not robust to outliers since the least sum of the squared of residuals is influenced by extreme values.

- The summation of square of residuals:



- For outlier, the deviation is large value hence it's square will also be a large value. When I have lots out outliers in my data I need to use "ROBUST REGRESSION"

I tried different robust regression methods

A) <u>Least Absolute Deviations (LAD) Regression:</u>

- Instead of minimizing the sum of squares of the residuals in this method I minimized the sum of the absolute values of residuals.

- The outliers I weighted linearly and not quadratically.

$$\text{Minimize } S = \sum |y_i - \hat{y}_i|$$

```
Call: rq(formula = lifeEx$Life.expectancy ~ lifeEx$Adult.Mortality +
    lifeEx$BMI + lifeEx$Polio + lifeEx$Diphtheria + lifeEx$HIV.AIDS +
    lifeEx$thinness..1.19.years + lifeEx$Income.composition.of.resources +
    lifeEx$Schooling, tau = 0.5)

tau: [1] 0.5

Coefficients:
                                        Value     Std. Error t value   Pr(>|t|)
(Intercept)                            53.42435   0.78115   68.39164   0.00000
lifeEx$Adult.Mortality                 -0.02207   0.00109  -20.23687   0.00000
lifeEx$BMI                              0.01560   0.00259    6.01680   0.00000
lifeEx$Polio                            0.02104   0.00355    5.92260   0.00000
lifeEx$Diphtheria                       0.01898   0.00350    5.42227   0.00000
lifeEx$HIV.AIDS                        -0.44669   0.02041  -21.88907   0.00000
lifeEx$thinness..1.19.years            -0.11760   0.02235   -5.26253   0.00000
lifeEx$Income.composition.of.resources 14.54449   1.57776    9.21843   0.00000
lifeEx$Schooling                        0.63875   0.05815   10.98446   0.00000
```

The variables are significant variables since p-value < 0.05

B) Huber M- Estimator loss:

- The Huber M-estimator attempts to get the best of both the least-square estimator (easy to find the minimum) and the absolute deviation estimator (more robust)

$$H(\varepsilon) = \begin{cases} \varepsilon^2/2 & for\ |\varepsilon| \le k \\ k|\varepsilon| - k^2/2 & for\ |\varepsilon| > k \end{cases}$$

- Generally, value of k = 1.345 * standard deviation. This approach gives 95% efficiency than OLS method even without outliers.

```
Call: rlm(formula = lifeEx$Life.expectancy ~ lifeEx$Adult.Mortality +
    lifeEx$BMI + lifeEx$Polio + lifeEx$Diphtheria + lifeEx$HIV.AIDS +
    lifeEx$thinness..1.19.years + lifeEx$Income.composition.of.resources +
    lifeEx$Schooling, k2 = 1.345)
Residuals:
      Min        1Q    Median        3Q       Max
-21.21784  -2.13832  -0.08786   2.14565  20.95304

Coefficients:
                                    Value    Std. Error t value
(Intercept)                          52.2695   0.4633   112.8269
lifeEx$Adult.Mortality               -0.0199   0.0007   -27.1006
lifeEx$BMI                            0.0235   0.0047     5.0269
lifeEx$Polio                          0.0206   0.0042     4.9250
lifeEx$Diphtheria                     0.0249   0.0042     5.9494
lifeEx$HIV.AIDS                      -0.4606   0.0155   -29.7851
lifeEx$thinness..1.19.years          -0.0950   0.0192    -4.9558
lifeEx$Income.composition.of.resources 10.2460   0.5767    17.7652
lifeEx$Schooling                      0.8746   0.0379    23.0497

Residual standard error: 3.179 on 2487 degrees of freedom
>
```

- The variables are significant variables since p-value < 0.0001 and RSE (Residual standard error) is less than linear regression.

C) Least Mean Square (LMS) Regression

```
lqs.formula(formula = lifeEx$Life.expectancy ~ lifeEx$Adult.Mortality +
    lifeEx$BMI + lifeEx$Polio + lifeEx$Diphtheria + lifeEx$HIV.AIDS +
    lifeEx$thinness..1.19.years + lifeEx$Income.composition.of.resources +
    lifeEx$Schooling, method = "lms")

Coefficients:
                            (Intercept)                  lifeEx$Adult.Mortality
                              78.423749                               -0.073820
                           lifeEx$BMI                           lifeEx$Polio
                              -0.004365                                0.006473
                     lifeEx$Diphtheria                        lifeEx$HIV.AIDS
                               0.002900                                0.447154
           lifeEx$thinness..1.19.years  lifeEx$Income.composition.of.resources
                              -0.217508                               -0.719364
                     lifeEx$Schooling
                               0.449737

Scale estimates 2.527 2.284
```

D) Least Trimmed squares (LTS) Regression

```
lqs.formula(formula = lifeEx$Life.expectancy ~ lifeEx$Adult.Mortality +
    lifeEx$BMI + lifeEx$Polio + lifeEx$Diphtheria + lifeEx$HIV.AIDS +
    lifeEx$thinness..1.19.years + lifeEx$Income.composition.of.resources +
    lifeEx$Schooling, method = "lts")

Coefficients:
                          (Intercept)                 lifeEx$Adult.Mortality
                            66.751989                             -0.047035
                        lifeEx$BMI                          lifeEx$Polio
                             0.002275                              0.034752
                    lifeEx$Diphtheria                       lifeEx$HIV.AIDS
                            -0.002142                             -0.151442
         lifeEx$thinness..1.19.years  lifeEx$Income.composition.of.resources
                            -0.236577                             22.043796
                    lifeEx$Schooling
                            -0.362802

Scale estimates 2.497 2.399
```

E) S- estimator regression

```
lqs.formula(formula = lifeEx$Life.expectancy ~ lifeEx$Adult.Mortality +
    lifeEx$BMI + lifeEx$Polio + lifeEx$Diphtheria + lifeEx$HIV.AIDS +
    lifeEx$thinness..1.19.years + lifeEx$Income.composition.of.resources +
    lifeEx$Schooling, method = "S")

Coefficients:
                          (Intercept)                 lifeEx$Adult.Mortality
                             66.97430                              -0.05915
                        lifeEx$BMI                          lifeEx$Polio
                              0.10317                               0.01931
                    lifeEx$Diphtheria                       lifeEx$HIV.AIDS
                              0.02045                               0.13506
         lifeEx$thinness..1.19.years  lifeEx$Income.composition.of.resources
                              0.05779                              14.40842
                    lifeEx$Schooling
                             -0.28351

Scale estimates 2.409
```

F) MM-estimator regression

```
Call: rlm(formula = lifeEx$Life.expectancy ~ lifeEx$Adult.Mortality +
    lifeEx$BMI + lifeEx$Polio + lifeEx$Diphtheria + lifeEx$HIV.AIDS +
    lifeEx$thinness..1.19.years + lifeEx$Income.composition.of.resources +
    lifeEx$Schooling, method = "MM")
Residuals:
     Min       1Q   Median       3Q      Max
-25.8657  -2.1515  -0.2138   1.7645  16.8254

Coefficients:
                                    Value    Std. Error t value
(Intercept)                          61.2347   0.4041   151.5500
lifeEx$Adult.Mortality               -0.0386   0.0006   -60.1726
lifeEx$BMI                            0.0091   0.0041     2.2408
lifeEx$Polio                          0.0131   0.0037     3.5922
lifeEx$Diphtheria                     0.0127   0.0036     3.4869
lifeEx$HIV.AIDS                      -0.1262   0.0135    -9.3532
lifeEx$thinness..1.19.years          -0.1266   0.0167    -7.5728
lifeEx$Income.composition.of.resources  6.4177   0.5030    12.7582
lifeEx$Schooling                      0.7798   0.0331    23.5642

Residual standard error: 2.411 on 2487 degrees of freedom
```

- The variables are significant variables since p-value < 0.05 and residual standard error is less than linear regression model.

Models/ Coeficients	OLS	Least mean square	Least Trimmed squares	Least Absolute Deviations(LAD)	S-estimator	Huber Loss	MM-estimator
Intercept	51.4711549	58.593957	68.88349	53.42435	60.59928	52.2695	61.2453
Adult.Mortality	-0.0179061	-0.030748	-0.0554	-0.02207	-0.04291	-0.0199	-0.0386
BMI	0.0370862	-0.005296	0.02729	0.0156	-0.02499	0.0235	0.0091
Polio	0.026928	-0.014092	0.008576	0.02104	-0.03915	0.0206	0.0131
Diphtheria	0.0317192	-0.009322	-0.02106	0.01898	0.02994	0.0249	0.0127
HIV AIDS	-0.4805646	-0.234046	0.179071	-0.4467	-0.08377	-0.4606	-0.1259
Thinness 1 to 19 years	-0.1148021	-0.123844	0.09597	-0.1176	-0.04622	-0.095	-0.1266
Income composition of resources	8.7971481	35.697632	16.3582	14.5445	26.4642	10.246	6.4165
Schooling	0.865573	-0.314554	0.00487	0.63875	0.07994	0.8746	0.7795

- Comparison of Linear regression with robust regression

	OLS- Linear regression	Robust rgression
R-square	0.82	0.8
Min_Max accuracy	0.872	0.873
RSE (Residual Standard error)	4.066	2.409
RMSE	4.074196	4.093033
MAE	3.005417	2.996708

- The robust regression didn't improved the model significantly but it did handled outliers in data.

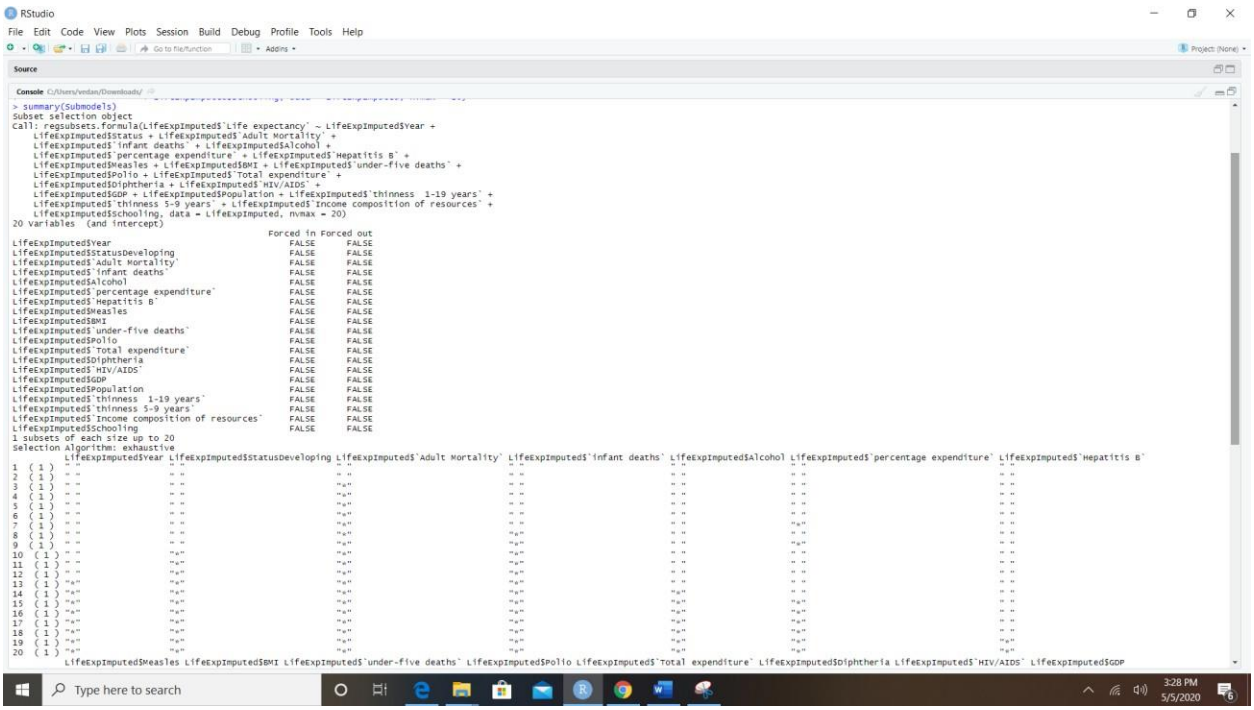
MODEL RECOGNITION FROM SUBSET REGRESSION PROCESS

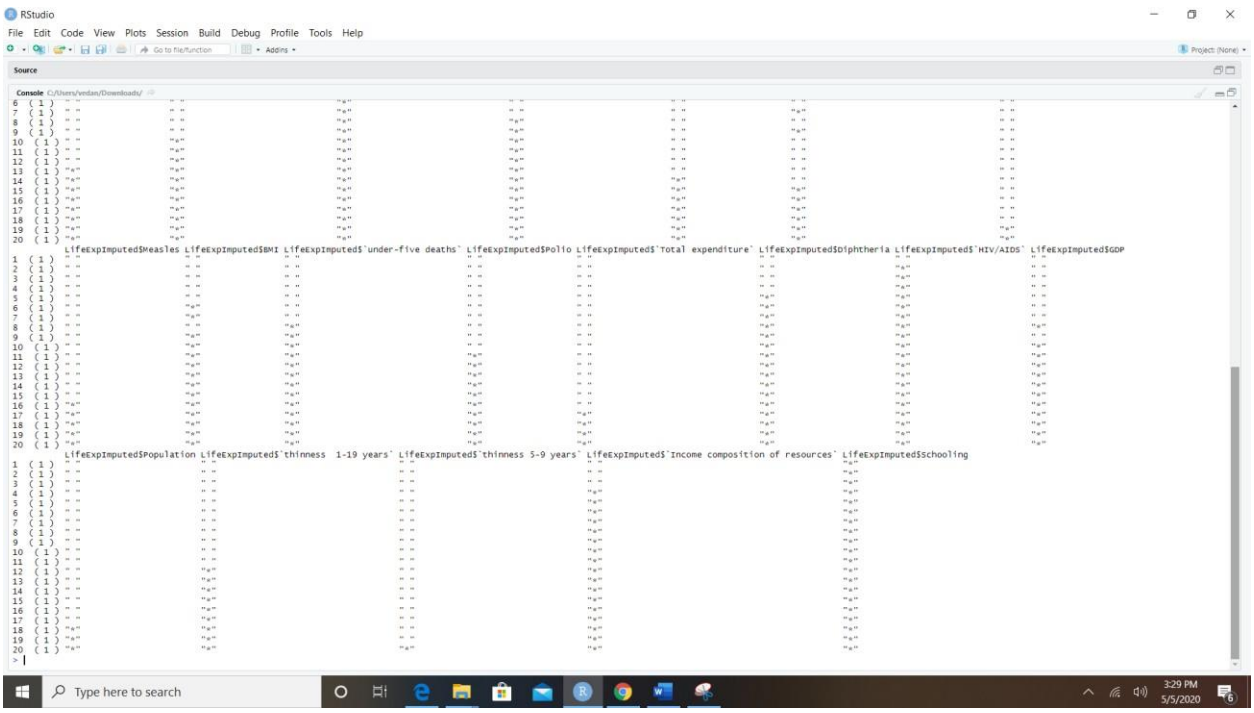
```
#.....Model Finding from Subset Algorithm.....

library(tidyverse)
library(caret)
library(leaps)

#Best Model finding by not considering Countries

modelSub = lm(LifeExpImputed$`Life expectancy` ~ LifeExpImputed$Year +
LifeExpImputed$Status
        + LifeExpImputed$`Adult Mortality` + LifeExpImputed$`infant deaths` +
LifeExpImputed$Alcohol
        + LifeExpImputed$`percentage expenditure` + LifeExpImputed$`Hepatitis B` +
LifeExpImputed$Measles
        + LifeExpImputed$BMI + LifeExpImputed$`under-five deaths` +
LifeExpImputed$Polio
        + LifeExpImputed$`Total expenditure` +LifeExpImputed$Diphtheria +
LifeExpImputed$`HIV/AIDS`
        +LifeExpImputed$GDP + LifeExpImputed$Population + LifeExpImputed$`thinness 1-
19 years`
        + LifeExpImputed$`thinness 5-9 years` + LifeExpImputed$`Income composition of
resources` + LifeExpImputed$Schooling, data = LifeExpImputed)

Summary(modelSub)
anova(modelSub)
```

```
Console C:/Users/vedan/Downloads/
> summary(modelSub)

Call:
lm(formula = LifeExpImputed$`Life expectancy` ~ LifeExpImputed$Year +
    LifeExpImputed$Status + LifeExpImputed$`Adult Mortality` +
    LifeExpImputed$`infant deaths` + LifeExpImputed$Alcohol +
    LifeExpImputed$`percentage expenditure` + LifeExpImputed$`Hepatitis B` +
    LifeExpImputed$Measles + LifeExpImputed$BMI + LifeExpImputed$`under-five deaths` +
    LifeExpImputed$Polio + LifeExpImputed$`Total expenditure` +
    LifeExpImputed$Diphtheria + LifeExpImputed$`HIV/AIDS` + LifeExpImputed$GDP +
    LifeExpImputed$Population + LifeExpImputed$`thinness  1-19 years` +
    LifeExpImputed$`thinness 5-9 years` + LifeExpImputed$`Income composition of resources` +
    LifeExpImputed$Schooling, data = LifeExpImputed)

Residuals:
     Min       1Q   Median       3Q      Max
-21.1653  -2.1930  -0.0361   2.2078  19.1780

Coefficients:
                                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                                     1.650e+02  3.635e+01   4.540 5.89e-06 ***
LifeExpImputed$Year                            -5.505e-02  1.818e-02  -3.028  0.00249 **
LifeExpImputed$StatusDeveloping                -1.416e+00  2.821e-01  -5.019 5.55e-07 ***
LifeExpImputed$`Adult Mortality`               -1.638e-02  8.291e-04 -19.754  < 2e-16 ***
LifeExpImputed$`infant deaths`                  8.923e-02  8.200e-03  10.881  < 2e-16 ***
LifeExpImputed$Alcohol                         -6.958e-02  2.711e-02  -2.566  0.01034 *
LifeExpImputed$`percentage expenditure`         1.337e-04  9.477e-05   1.411  0.15843
LifeExpImputed$`Hepatitis B`                    1.934e-03  4.104e-03   0.471  0.63745
LifeExpImputed$Measles                         -1.145e-05  8.396e-06  -1.363  0.17292
LifeExpImputed$BMI                              3.728e-02  5.237e-03   7.117 1.44e-12 ***
LifeExpImputed$`under-five deaths`             -6.580e-02  6.017e-03 -10.935  < 2e-16 ***
LifeExpImputed$Polio                            2.189e-02  4.705e-03   4.652 3.46e-06 ***
LifeExpImputed$`Total expenditure`              2.766e-02  3.563e-02   0.776  0.43760
LifeExpImputed$Diphtheria                       2.712e-02  5.201e-03   5.215 1.99e-07 ***
LifeExpImputed$`HIV/AIDS`                      -4.799e-01  1.738e-02 -27.608  < 2e-16 ***
LifeExpImputed$GDP                              3.257e-05  1.445e-05   2.254  0.02431 *
LifeExpImputed$Population                      -8.739e-10  1.627e-09  -0.537  0.59116
LifeExpImputed$`thinness  1-19 years`          -8.379e-02  4.892e-02  -1.713  0.08687 .
LifeExpImputed$`thinness 5-9 years`            -2.161e-02  4.833e-02  -0.447  0.65478
LifeExpImputed$`Income composition of resources`  7.282e+00  6.614e-01  11.010  < 2e-16 ***
LifeExpImputed$Schooling                        8.079e-01  4.513e-02  17.902  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.883 on 2475 degrees of freedom
Multiple R-squared:  0.8373,    Adjusted R-squared:  0.836
F-statistic:   637 on 20 and 2475 DF,  p-value: < 2.2e-16
```

```
Console C:/Users/vedan/Downloads/
> anova(modelSub)
Analysis of Variance Table

Response: LifeExpImputed$`Life expectancy`
                                                Df  Sum Sq Mean Sq   F value    Pr(>F)
LifeExpImputed$Year                              1    6406    6406  424.8940 < 2.2e-16 ***
LifeExpImputed$Status                            1   54635   54635 3624.0601 < 2.2e-16 ***
LifeExpImputed$`Adult Mortality`                 1   69949   69949 4639.8711 < 2.2e-16 ***
LifeExpImputed$`infant deaths`                   1    2617    2617  173.5791 < 2.2e-16 ***
LifeExpImputed$Alcohol                           1    3322    3322  220.3767 < 2.2e-16 ***
LifeExpImputed$`percentage expenditure`          1    2592    2592  171.9553 < 2.2e-16 ***
LifeExpImputed$`Hepatitis B`                     1    5112    5112  339.0910 < 2.2e-16 ***
LifeExpImputed$Measles                           1     870     870   57.7210 4.256e-14 ***
LifeExpImputed$BMI                               1   10111   10111  670.6728 < 2.2e-16 ***
LifeExpImputed$`under-five deaths`               1    6593    6593  437.3024 < 2.2e-16 ***
LifeExpImputed$Polio                             1    2915    2915  193.3817 < 2.2e-16 ***
LifeExpImputed$`Total expenditure`               1      43      43    2.8559   0.09116 .
LifeExpImputed$Diphtheria                        1    1287    1287   85.4023 < 2.2e-16 ***
LifeExpImputed$`HIV/AIDS`                        1   10597   10597  702.9056 < 2.2e-16 ***
LifeExpImputed$GDP                               1     486     486   32.2575 1.508e-08 ***
LifeExpImputed$Population                        1       1       1    0.0472   0.82804
LifeExpImputed$`thinness  1-19 years`            1     666     666   44.2103 3.616e-11 ***
LifeExpImputed$`thinness 5-9 years`              1       5       5    0.3519   0.55308
LifeExpImputed$`Income composition of resources`  1    9033    9033  599.1902 < 2.2e-16 ***
LifeExpImputed$Schooling                         1    4831    4831  320.4780 < 2.2e-16 ***
Residuals                                     2475   37312      15
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

#Using regsubsets to find out the top best subset model to predict Life Exp - Not Considering Countries
Submodels = regsubsets(LifeExpImputed$`Life expectancy` ~ LifeExpImputed$Year + LifeExpImputed$Status
+ LifeExpImputed$`Adult Mortality` + LifeExpImputed$`infant deaths` + LifeExpImputed$Alcohol
+ LifeExpImputed$`percentage expenditure` + LifeExpImputed$`Hepatitis B` + LifeExpImputed$Measles
+ LifeExpImputed$BMI + LifeExpImputed$`under-five deaths` + LifeExpImputed$Polio
+ LifeExpImputed$`Total expenditure` + LifeExpImputed$Diphtheria + LifeExpImputed$`HIV/AIDS`
+LifeExpImputed$GDP + LifeExpImputed$Population + LifeExpImputed$`thinness 1-19 years`
+ LifeExpImputed$`thinness 5-9 years` + LifeExpImputed$`Income composition of resources`
+ LifeExpImputed$Schooling, data = LifeExpImputed, nvmax = 20)
summary(Submodels)






```
#considering parameters and finding best possible subset
Var_sum = summary(Submodels)
data.frame(
  Adj_R_Sq = which.max(Var_sum$adjr2),
  CP_Value = which.min(Var_sum$cp),
  BIC_Value = which.min(Var_sum$bic)
)
```

```
Console C:/Users/vedan/Downloads/
> #considering parameters and finding best possible subset
> Var_sum = summary(Submodels)
> data.frame(
+   Adj_R_Sq = which.max(Var_sum$adjr2),
+   CP_Value = which.min(Var_sum$cp),
+   BIC_Value = which.min(Var_sum$bic)
+ )
  Adj_R_Sq CP_Value BIC_Value
1       16       15        12
>
```

Considering 12 Variables – as per the Principle of Parsimony

- #Considering 12 Variables I get (Principle of Parsimonious)

```
SubModel1 = lm(LifeExpImputed$`Life expectancy` ~ LifeExpImputed$Status +
LifeExpImputed$`Adult Mortality` + LifeExpImputed$`infant deaths`
        + LifeExpImputed$BMI + LifeExpImputed$`under-five deaths` +
LifeExpImputed$Polio
        + LifeExpImputed$Diphtheria + LifeExpImputed$`HIV/AIDS` +
LifeExpImputed$GDP
        + LifeExpImputed$`thinness 1-19 years`+ LifeExpImputed$`Income composition of
resources` + LifeExpImputed$Schooling,
        data = LifeExpImputed)

summary(SubModel1)
anova(SubModel1)
library(car)
vif(SubModel1)
```

```
Console C:/Users/vedan/Downloads/
> summary(SubModel1)

Call:
lm(formula = LifeExpImputed$`Life expectancy` ~ LifeExpImputed$Status +
    LifeExpImputed$`Adult Mortality` + LifeExpImputed$`infant deaths` +
    LifeExpImputed$BMI + LifeExpImputed$`under-five deaths` +
    LifeExpImputed$Polio + LifeExpImputed$Diphtheria + LifeExpImputed$`HIV/AIDS` +
    LifeExpImputed$GDP + LifeExpImputed$`thinness  1-19 years` +
    LifeExpImputed$`Income composition of resources` + LifeExpImputed$Schooling,
    data = LifeExpImputed)

Residuals:
     Min       1Q   Median       3Q      Max
-21.0426  -2.2473  -0.0602   2.2370  18.5368

Coefficients:
                                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                                     5.503e+01  6.209e-01  88.629  < 2e-16 ***
LifeExpImputed$StatusDeveloping                -1.373e+00  2.574e-01  -5.336 1.04e-07 ***
LifeExpImputed$`Adult Mortality`               -1.671e-02  8.183e-04 -20.423  < 2e-16 ***
LifeExpImputed$`infant deaths`                  9.224e-02  7.963e-03  11.584  < 2e-16 ***
LifeExpImputed$BMI                              3.845e-02  5.175e-03   7.429 1.50e-13 ***
LifeExpImputed$`under-five deaths`             -6.877e-02  5.865e-03 -11.726  < 2e-16 ***
LifeExpImputed$Polio                            2.280e-02  4.642e-03   4.911 9.63e-07 ***
LifeExpImputed$Diphtheria                       2.698e-02  4.662e-03   5.788 8.04e-09 ***
LifeExpImputed$`HIV/AIDS`                      -4.762e-01  1.715e-02 -27.771  < 2e-16 ***
LifeExpImputed$GDP                              4.942e-05  6.535e-06   7.562 5.53e-14 ***
LifeExpImputed$`thinness  1-19 years`          -9.534e-02  2.393e-02  -3.985 6.95e-05 ***
LifeExpImputed$`Income composition of resources` 6.954e+00  6.522e-01  10.663  < 2e-16 ***
LifeExpImputed$Schooling                        7.704e-01  4.313e-02  17.861  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.891 on 2483 degrees of freedom
Multiple R-squared:  0.8361,    Adjusted R-squared:  0.8353
F-statistic:  1056 on 12 and 2483 DF,  p-value: < 2.2e-16
```

```
Console C:/Users/vedan/Downloads/
> anova(SubModel1)
Analysis of Variance Table

Response: LifeExpImputed$`Life expectancy`
                                        Df Sum Sq Mean Sq  F value    Pr(>F)
LifeExpImputed$Status                    1  54635   54635 3608.979 < 2.2e-16 ***
LifeExpImputed$`Adult Mortality`         1  72795   72795 4808.577 < 2.2e-16 ***
LifeExpImputed$`infant deaths`           1   2838    2838  187.501 < 2.2e-16 ***
LifeExpImputed$BMI                       1  16158   16158 1067.366 < 2.2e-16 ***
LifeExpImputed$`under-five deaths`       1   7734    7734  510.897 < 2.2e-16 ***
LifeExpImputed$Polio                     1   5905    5905  390.090 < 2.2e-16 ***
LifeExpImputed$Diphtheria                1   2182    2182  144.118 < 2.2e-16 ***
LifeExpImputed$`HIV/AIDS`                1  10577   10577  698.705 < 2.2e-16 ***
LifeExpImputed$GDP                       1   3073    3073  203.011 < 2.2e-16 ***
LifeExpImputed$`thinness  1-19 years`    1   1068    1068   70.562 < 2.2e-16 ***
LifeExpImputed$`Income composition of resources`  1   9998    9998  660.449 < 2.2e-16 ***
LifeExpImputed$Schooling                 1   4830    4830  319.027 < 2.2e-16 ***
Residuals                             2483  37589      15
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

```
Console C:/Users/vedan/Downloads/
> library(car)
> vif(SubModel1)
                          LifeExpImputed$Status               LifeExpImputed$`Adult Mortality`               LifeExpImputed$`infant deaths`
                                       1.608972                                       1.742935                                     166.960474
                             LifeExpImputed$BMI             LifeExpImputed$`under-five deaths`                         LifeExpImputed$Polio
                                       1.753146                                     167.413195                                       1.923687
                     LifeExpImputed$Diphtheria                    LifeExpImputed$`HIV/AIDS`                            LifeExpImputed$GDP
                                       1.968465                                       1.429019                                       1.434053
          LifeExpImputed$`thinness  1-19 years` LifeExpImputed$`Income composition of resources`                     LifeExpImputed$Schooling
                                       1.899853                                       3.135259                                       3.443190
>
```

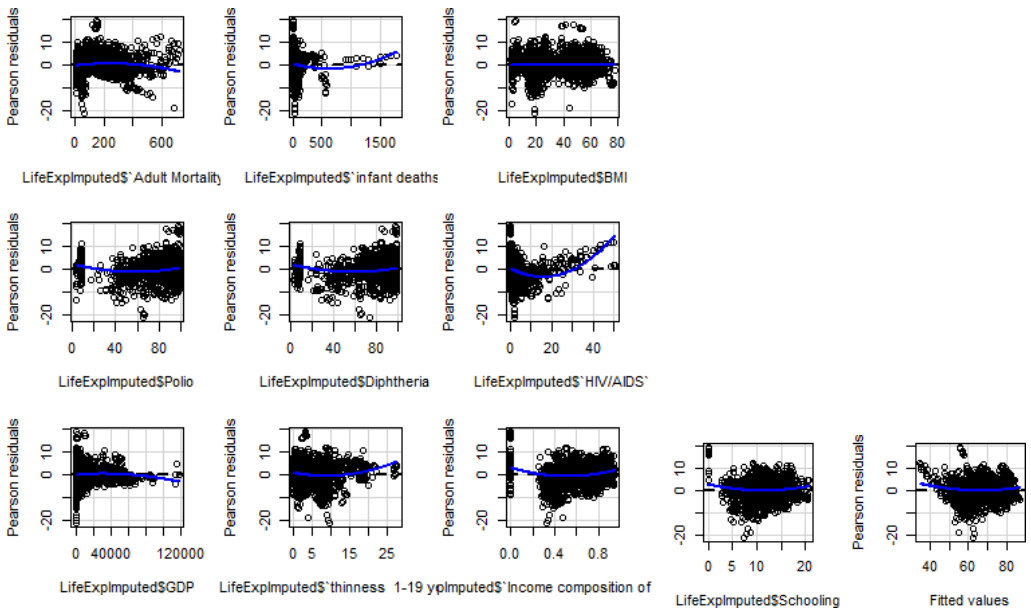
- Removing LifeExpImputed$`under-five deaths` having 167.41 vif value

SubModel2 = lm(LifeExpImputed$`Life expectancy` ~ LifeExpImputed$Status + LifeExpImputed$`Adult Mortality` + LifeExpImputed$`infant deaths`
+ LifeExpImputed$BMI + LifeExpImputed$Polio + LifeExpImputed$Diphtheria + LifeExpImputed$`HIV/AIDS` + LifeExpImputed$GDP
+ LifeExpImputed$`thinness 1-19 years`+ LifeExpImputed$`Income composition of resources` + LifeExpImputed$Schooling,
data = LifeExpImputed)
summary(SubModel2)

- #got decreased significance of independent variable infant deaths

- #Poor residual plots

```
> summary(SubModel2)

Call:
lm(formula = LifeExpImputed$`Life expectancy` ~ LifeExpImputed$Status +
    LifeExpImputed$`Adult Mortality` + LifeExpImputed$`infant deaths` +
    LifeExpImputed$BMI + LifeExpImputed$Polio + LifeExpImputed$Diphtheria +
    LifeExpImputed$`HIV/AIDS` + LifeExpImputed$GDP + LifeExpImputed$`thinness  1-19 years` +
    LifeExpImputed$`Income composition of resources` + LifeExpImputed$Schooling,
    data = LifeExpImputed)

Residuals:
    Min      1Q  Median      3Q     Max
-21.196  -2.239  -0.074   2.304  19.093

Coefficients:
                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                      5.341e+01  6.217e-01  85.903  < 2e-16 ***
LifeExpImputed$StatusDeveloping                 -1.141e+00  2.636e-01  -4.329 1.56e-05 ***
LifeExpImputed$`Adult Mortality`                -1.722e-02  8.394e-04 -20.511  < 2e-16 ***
LifeExpImputed$`infant deaths`                  -7.571e-04  7.336e-04  -1.032  0.30215
LifeExpImputed$BMI                               3.879e-02  5.316e-03   7.296 3.96e-13 ***
LifeExpImputed$Polio                             2.544e-02  4.762e-03   5.342 1.00e-07 ***
LifeExpImputed$Diphtheria                        3.312e-02  4.758e-03   6.960 4.33e-12 ***
LifeExpImputed$`HIV/AIDS`                       -4.867e-01  1.759e-02 -27.666  < 2e-16 ***
LifeExpImputed$GDP                               4.574e-05  6.704e-06   6.823 1.12e-11 ***
LifeExpImputed$`thinness  1-19 years`           -7.834e-02  2.453e-02  -3.194  0.00142 **
LifeExpImputed$`Income composition of resources` 7.717e+00  6.665e-01  11.578  < 2e-16 ***
LifeExpImputed$Schooling                         7.852e-01  4.428e-02  17.731  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.996 on 2484 degrees of freedom
Multiple R-squared:  0.8271,    Adjusted R-squared:  0.8263
F-statistic:  1080 on 11 and 2484 DF,  p-value: < 2.2e-16

>
```



- Now, I have 10 independent variables, and performed some transformation to refine residual plots

```
SubModel3 = lm(LifeExpImputed$`Life expectancy` ~ LifeExpImputed$Status +
LifeExpImputed$`Adult Mortality` + sqrt(sqrt(LifeExpImputed$`infant deaths`) )
        + LifeExpImputed$Polio + LifeExpImputed$Diphtheria +
sqrt(sqrt(LifeExpImputed$`HIV/AIDS`)) + LifeExpImputed$GDP
        + LifeExpImputed$`thinness 1-19 years`+ LifeExpImputed$`Income composition of
resources` + LifeExpImputed$Schooling,
        data = LifeExpImputed)

summary(SubModel3)
anova(SubModel3)
library(car)
vif(SubModel3)
qqnorm(SubModel3$residuals)

AIC(SubModel3)
BIC(SubModel3)
residualPlots(SubModel3)
```

- Significant Variables and good r square value

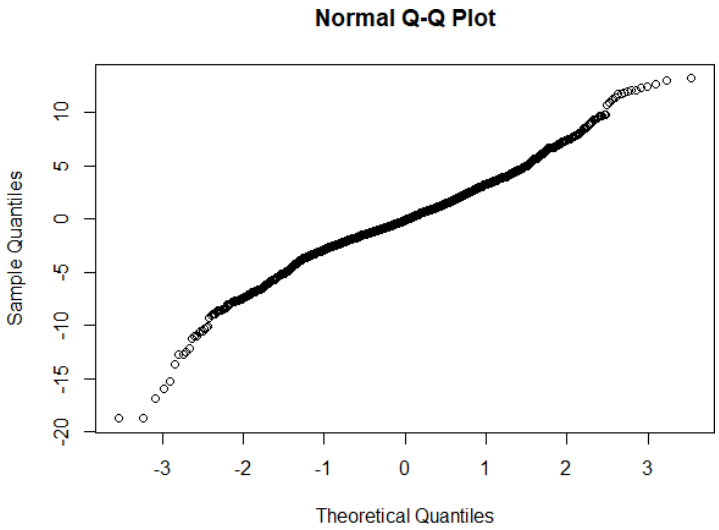
```
> summary(SubModel3)

Call:
lm(formula = LifeExpImputed$`Life expectancy` ~ LifeExpImputed$Status +
    LifeExpImputed$`Adult Mortality` + sqrt(sqrt(LifeExpImputed$`infant deaths`)) +
    LifeExpImputed$Polio + LifeExpImputed$Diphtheria + sqrt(sqrt(LifeExpImputed$`HIV/AIDS`)) +
    LifeExpImputed$GDP + LifeExpImputed$`thinness  1-19 years` +
    LifeExpImputed$`Income composition of resources` + LifeExpImputed$Schooling,
    data = LifeExpImputed)

Residuals:
     Min       1Q   Median       3Q      Max
-18.7504  -1.9363  -0.1152   2.0301  13.1669

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                               6.633e+01  6.137e-01 108.079  < 2e-16 ***
LifeExpImputed$StatusDeveloping          -9.342e-01  2.314e-01  -4.037 5.59e-05 ***
LifeExpImputed$`Adult Mortality`         -1.199e-02  7.601e-04 -15.771  < 2e-16 ***
sqrt(sqrt(LifeExpImputed$`infant deaths`))  -7.401e-01  7.994e-02  -9.258  < 2e-16 ***
LifeExpImputed$Polio                      1.344e-02  4.212e-03   3.190  0.00144 **
LifeExpImputed$Diphtheria                 2.029e-02  4.200e-03   4.832 1.43e-06 ***
sqrt(sqrt(LifeExpImputed$`HIV/AIDS`))    -9.324e+00  2.291e-01 -40.693  < 2e-16 ***
LifeExpImputed$GDP                        4.434e-05  5.925e-06   7.483 9.99e-14 ***
LifeExpImputed$`thinness  1-19 years`    -8.717e-02  1.913e-02  -4.556 5.48e-06 ***
LifeExpImputed$`Income composition of resources`  7.778e+00  5.879e-01  13.229  < 2e-16 ***
LifeExpImputed$Schooling                  5.622e-01  3.932e-02  14.300  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.517 on 2485 degrees of freedom
Multiple R-squared:  0.866,     Adjusted R-squared:  0.8655
F-statistic:  1606 on 10 and 2485 DF,  p-value: < 2.2e-16

> |
```

```
> anova(SubModel3)
Analysis of Variance Table

Response: LifeExpImputed$`Life expectancy`
                                                 Df Sum Sq Mean Sq F value    Pr(>F)
LifeExpImputed$Status                             1  54635   54635 4416.80 < 2.2e-16 ***
LifeExpImputed$`Adult Mortality`                  1  72795   72795 5884.91 < 2.2e-16 ***
sqrt(sqrt(LifeExpImputed$`infant deaths`))        1  16343   16343 1321.24 < 2.2e-16 ***
LifeExpImputed$Polio                              1   7731    7731  624.98 < 2.2e-16 ***
LifeExpImputed$Diphtheria                         1   3337    3337  269.79 < 2.2e-16 ***
sqrt(sqrt(LifeExpImputed$`HIV/AIDS`))             1  27485   27485 2221.97 < 2.2e-16 ***
LifeExpImputed$GDP                                1   2670    2670  215.85 < 2.2e-16 ***
LifeExpImputed$`thinness  1-19 years`             1   1507    1507  121.85 < 2.2e-16 ***
LifeExpImputed$`Income composition of resources`  1   9612    9612  777.04 < 2.2e-16 ***
LifeExpImputed$Schooling                          1   2529    2529  204.48 < 2.2e-16 ***
Residuals                                      2485  30739      12
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

- VIF < 5

```
> library(car)
Loading required package: carData
> vif(SubModel3)
                        LifeExpImputed$Status                LifeExpImputed$`Adult Mortality`       sqrt(sqrt(LifeExpImputed$`infant deaths`))
                                     1.591621                                        1.840038                                        1.693018
                          LifeExpImputed$Polio                       LifeExpImputed$Diphtheria             sqrt(sqrt(LifeExpImputed$`HIV/AIDS`))
                                     1.937768                                        1.955358                                        1.936518
                           LifeExpImputed$GDP LifeExpImputed$`thinness  1-19 years` LifeExpImputed$`Income composition of resources`
                                     1.442648                                        1.486949                                        3.118499
                    LifeExpImputed$Schooling
                                     3.501098
>
```

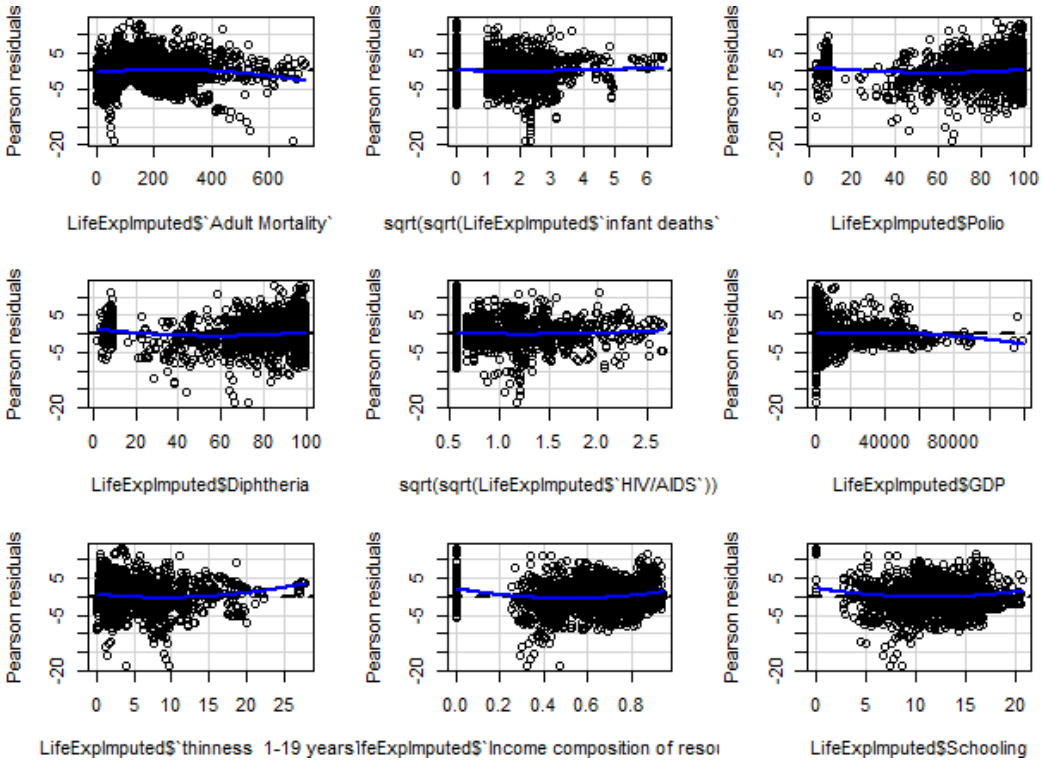
- Moderate Normally distributed qq plot

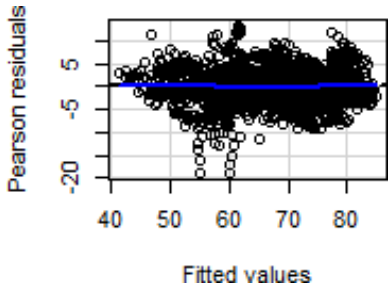


- Good AIC BIC Values

```
> AIC(SubModel3)
[1] 13374.39
> BIC(SubModel3)
[1] 13444.26
> residualPlots(SubModel3)
Hit <Return> to see next plot:
```

- Better residual curves meet assumptions of Constant variance.





- Durbin Watson Test Satisfied. No Autocorrelation found in the variables. P value < 0.05.

```
Console ~/
> durbinWatsonTest(SubModel3)
 lag Autocorrelation D-W Statistic p-value
   1       0.7133733     0.5728189       0
 Alternative hypothesis: rho != 0
>
```

- Cross validation: good RMSE value found.

```
library(caret)
set.seed(123)
train.control <- trainControl(method = "cv", number = 10)
cv_model <- train (`Life expectancy` ~ Status + `Adult Mortality`
                   + sqrt(sqrt(`infant deaths`) ) + Polio + Diphtheria
                   + sqrt(sqrt(`HIV/AIDS`)) + GDP +`thinness 1-19 years`
                   + `Income composition of resources` +Schooling, data = LifeExpImputed,
                    method = "lm", trControl = train.control)
# Summarize the results
print(cv_model)
#Rsquare value of every fold
cv_model$resample
```

```
Console ~/
> # Summarize the results
> print(cv_model)
Linear Regression

2496 samples
  10 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 2246, 2247, 2246, 2247, 2246, 2246, ...
Resampling results:

  RMSE     Rsquared  MAE
  3.52222  0.866096  2.634214

Tuning parameter 'intercept' was held constant at a value of TRUE
>
```

```
Console ~/
> #Rsquare value of every fold
> cv_model$resample
       RMSE  Rsquared      MAE Resample
1  3.488854 0.8585689 2.632310   Fold01
2  3.463574 0.8702248 2.626280   Fold02
3  4.044286 0.8372899 2.855633   Fold03
4  3.680712 0.8604763 2.712605   Fold04
5  3.263727 0.8786005 2.453429   Fold05
6  3.514592 0.8609573 2.721901   Fold06
7  3.336934 0.8888613 2.536085   Fold07
8  3.365911 0.8823599 2.562776   Fold08
9  3.582028 0.8533746 2.631186   Fold09
10 3.481580 0.8702466 2.609935   Fold10
>
```

#.....Considering few Significant countries.....

LifeExpLimitedCountries.csv

```
SubModel4 = lm(LifeExpLimitedCountries$`Life expectancy` ~
LifeExpLimitedCountries$Country
          + LifeExpLimitedCountries$Year
           + LifeExpLimitedCountries$`Adult Mortality`
           + sqrt(sqrt(LifeExpLimitedCountries$`infant deaths`) )
           + LifeExpLimitedCountries$Polio + LifeExpLimitedCountries$Diphtheria
           + sqrt(sqrt(LifeExpLimitedCountries$`HIV/AIDS`)) + LifeExpLimitedCountries$GDP
           + LifeExpLimitedCountries$`thinness 1-19 years`
           + LifeExpLimitedCountries$`Income composition of resources`
           + LifeExpLimitedCountries$Schooling,
           data = LifeExpLimitedCountries)

summary(SubModel4)
```

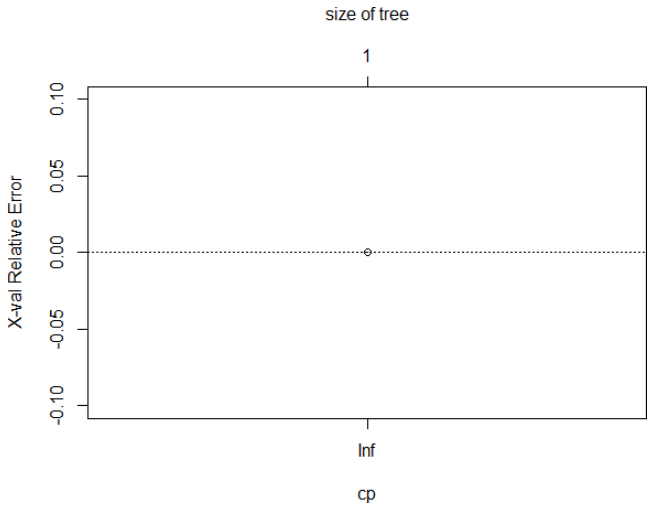
```
Residuals:
    Min      1Q  Median      3Q     Max
-1.3216 -0.3136 -0.0088  0.2453  4.0202

Coefficients:
                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                      1.951e+02  1.903e+02   1.025 0.310286
LifeExpLimitedCountries$CountryAlgeria          -8.983e-01  1.128e+00  -0.796 0.429617
LifeExpLimitedCountries$CountryAngola           -2.350e+01  7.506e+00  -3.131 0.002908 **
LifeExpLimitedCountries$CountryArgentina        -5.514e+00  1.532e+00  -3.599 0.000733 ***
LifeExpLimitedCountries$Year                    -8.180e-02  1.008e-01  -0.812 0.420774
LifeExpLimitedCountries$`Adult Mortality`       -3.017e-04  1.871e-03  -0.161 0.872571
sqrt(sqrt(LifeExpLimitedCountries$`infant deaths`))  2.101e-01  5.251e-01   0.400 0.690735
LifeExpLimitedCountries$Polio                   -1.162e-02  1.300e-02  -0.894 0.375742
LifeExpLimitedCountries$Diphtheria               1.808e-02  1.184e-02   1.527 0.133016
sqrt(sqrt(LifeExpLimitedCountries$`HIV/AIDS`))   1.497e+01  8.960e+00   1.670 0.101077
LifeExpLimitedCountries$GDP                      2.028e-05  4.730e-05   0.429 0.670003
LifeExpLimitedCountries$`thinness  1-19 years`  -9.897e-02  8.354e-02  -1.185 0.241754
LifeExpLimitedCountries$`Income composition of resources`  4.369e+01  2.464e+01   1.773 0.082252 .
LifeExpLimitedCountries$Schooling                3.472e-01  4.212e-01   0.824 0.413685
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7578 on 50 degrees of freedom
Multiple R-squared:  0.9965,    Adjusted R-squared:  0.9955
F-statistic:  1084 on 13 and 50 DF,  p-value: < 2.2e-16
```

- Good r square value but,
- When adding few countries in Level the significance of other imp. variable is going down.
- And the prediction scope is limited for fewer countries only.
- for all the year in dataset, the Life Exp of each country has no significant change.
- Hence, rejecting this Model.

Some Advance Machine Learning Models

```
library(caret)
library(rpart)

SubModelFit = rpart(LifeExpImputed$`Life expectancy` ~ LifeExpImputed$Status +
LifeExpImputed$`Adult Mortality` + LifeExpImputed$`infant deaths`
        + LifeExpImputed$Polio + LifeExpImputed$Diphtheria
+LifeExpImputed$`HIV/AIDS` + LifeExpImputed$GDP
        + LifeExpImputed$`Income composition of resources` + LifeExpImputed$Schooling,
        method="class", data = LifeExpImputed)
printcp(SubModelFit) # display the results
plotcp(SubModelFit) # visualize cross-validation results
summary(SubModelFit) # detailed summary of splits
```

```
Warning message:
package 'caret' was built under R version 3.6.3
> library(rpart)
> SubModelFit = rpart(LifeExpImputed$`Life expectancy` ~ LifeExpImputed$Status + LifeExpImputed$`Adult Mortality` + LifeExpImputed$`infant deaths`
+                 + LifeExpImputed$Polio + LifeExpImputed$Diphtheria +LifeExpImputed$`HIV/AIDS` + LifeExpImputed$GDP
+                 + LifeExpImputed$`Income composition of resources` + LifeExpImputed$Schooling,
+                 method="class", data = LifeExpImputed)
> printcp(SubModelFit) # display the results

Classification tree:
rpart(formula = LifeExpImputed$`Life expectancy` ~ LifeExpImputed$Status +
    LifeExpImputed$`Adult Mortality` + LifeExpImputed$`infant deaths` +
    LifeExpImputed$Polio + LifeExpImputed$Diphtheria + LifeExpImputed$`HIV/AIDS` +
    LifeExpImputed$GDP + LifeExpImputed$`Income composition of resources` +
    LifeExpImputed$Schooling, data = LifeExpImputed, method = "class")

Variables actually used in tree construction:
character(0)

Root node error: 2454/2496 = 0.98317

n= 2496

       CP nsplit rel error xerror xstd
1 0.00815      0         1      0    0
```



```
> summary(SubModelFit) # detailed summary of splits
Call:
rpart(formula = LifeExpImputed$`Life expectancy` ~ LifeExpImputed$Status +
    LifeExpImputed$`Adult Mortality` + LifeExpImputed$`infant deaths` +
    LifeExpImputed$Polio + LifeExpImputed$Diphtheria + LifeExpImputed$`HIV/AIDS` +
    LifeExpImputed$GDP + LifeExpImputed$`Income composition of resources` +
    LifeExpImputed$Schooling, data = LifeExpImputed, method = "class")
  n= 2496

          CP nsplit rel error xerror xstd
1 0.008149959      0         1      0    0

Node number 1: 2496 observations
  predicted class=73  expected loss=0.9831731  P(node) =1
    class counts:     1     1     1     1     1     1     1     1     1     1     1     1     2
 2     1     4     1     2     1     2     5     2     2     3     4     1     2
   2     2     1     2     1     4     1     5     6     2     4     2     3     4
     3     3     3     2     4     3     2     2     4     3     3     1     8
 10     2     3     4     4     1     2     8     3     4     2     2     5     1
    7     9     8     4     3     6     3     7     4     3     8     4     6     7
```

SubModelFit$cptable

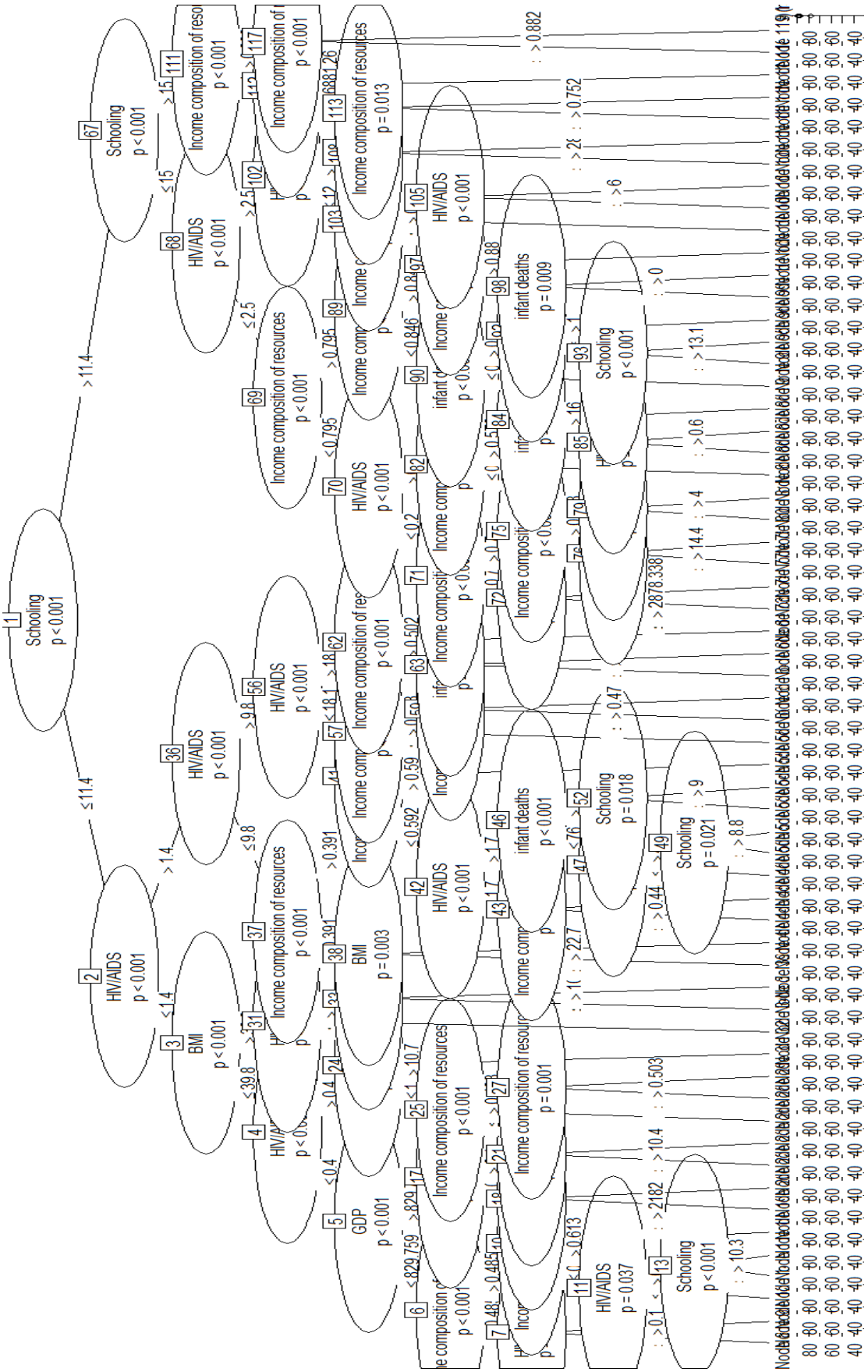
```
Console ~/
> SubModelFit$cptable
          CP nsplit rel error xerror xstd
1 0.008149959      0         1      0    0
> |
```

#.....R -Decision tree Analysis.....

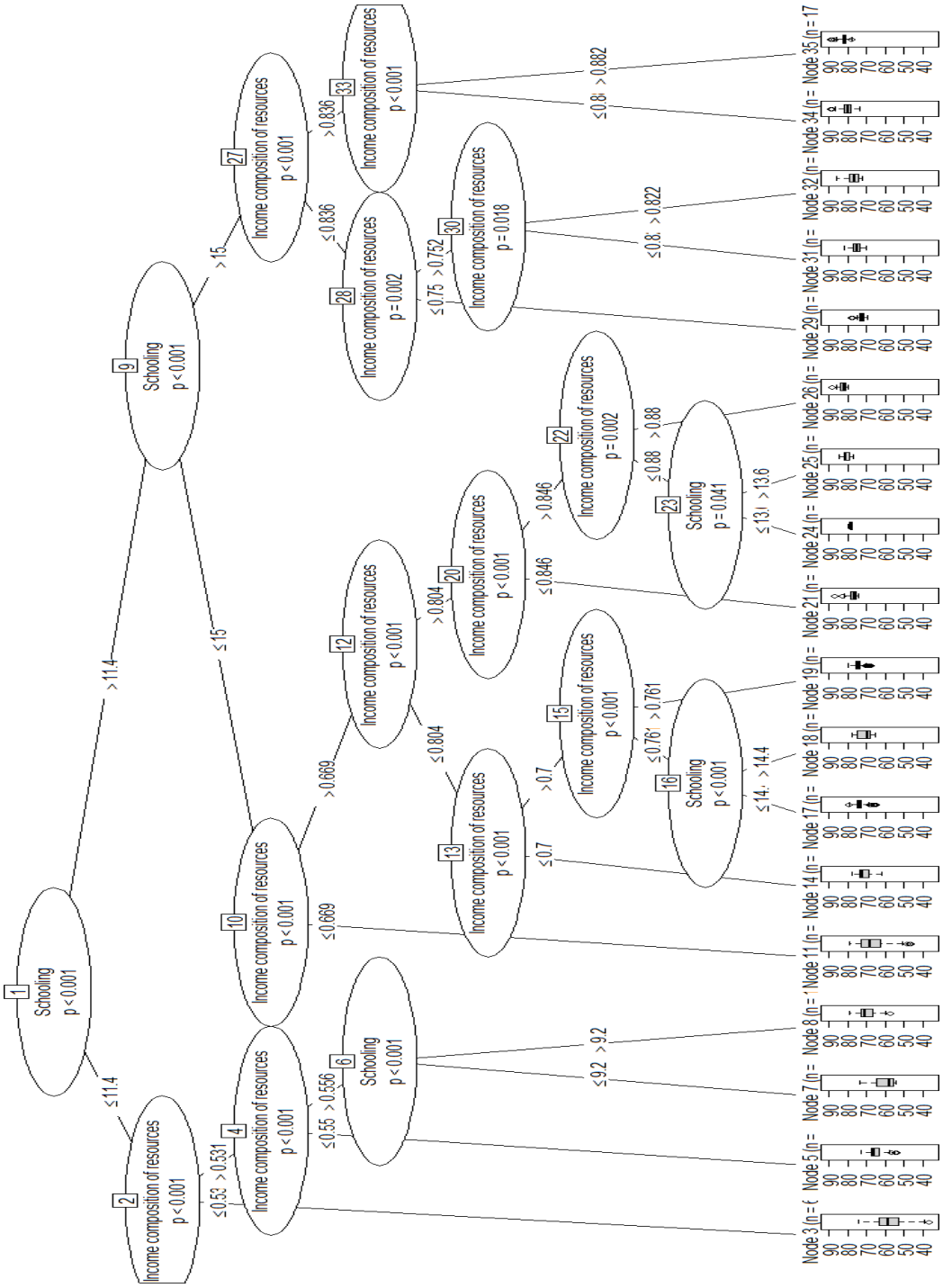
install.packages("matrixStats")
library(party)
library("matrixStats")
#considerd top most correlatd and significant independent variables to build a decision tree

Ctree.LifeExp = ctree(LifeExpImputed$`Life expectancy` ~ `infant deaths`
+ BMI + `HIV/AIDS` + GDP + `Income composition of resources`
+ Schooling, data = LifeExpImputed)
plot(Ctree.LifeExp)

dev.off()


1
Schooling
p < 0.001
≤ 11.4
> 11.4
2
HIV/AIDS
p < 0.001
≤ 1.4
> 1.4
3
BMI
p < 0.001
≤ 1.4
> 1.4
4
≤ 39.8
> 39.8
5
GDP
p < 0.001
≤ 0.4
> 0.4
6
≤ 829.759
> 829.759
7
≤ 0.485
> 0.485
36
HIV/AIDS
p < 0.001
≤ 9.8
> 9.8
37
Income composition of resources
p < 0.001
≤ 0.391
> 0.391
38
BMI
p = 0.003
≤ 0.592
> 0.592
42
HIV/AIDS
p < 0.001
≤ 1.7
> 1.7
43
Income composition of resources
p = 0.001
≤ 0.613
> 0.613
56
HIV/AIDS
p < 0.001
≤ 18.1
> 18.1
57
Income composition of resources
p < 0.001
≤ 0.502
> 0.502
62
Income composition of resources
p < 0.001
≤ 0.59
> 0.59
63
infant deaths
p < 0.001
≤ 1.7
> 1.7
67
Schooling
p < 0.001
≤ 15
> 15
68
HIV/AIDS
p < 0.001
≤ 2.5
> 2.5
69
Income composition of resources
p < 0.001
≤ 0.795
> 0.795
70
HIV/AIDS
p < 0.001
≤ 0.2
> 0.2
71
Income composition of resources
p < 0.001
≤ 0.47
> 0.47
72
Income composition of resources
p < 0.001
≤ 0.76
> 0.76
75
infant deaths
p < 0.001
≤ 0.5
> 0.5
78
≤ 0.7
> 0.7
79
≤ 0.338
> 0.338
82
Income composition of resources
p < 0.001
≤ 0.846
> 0.846
84
infant deaths
p = 0.009
≤ 1
> 1
85
≤ 16
> 16
88
Income composition of resources
p < 0.001
≤ 0.882
> 0.882
89
Income composition of resources
p < 0.001
≤ 0.795
> 0.795
90
infant deaths
p < 0.001
≤ 0.88
> 0.88
93
Schooling
p < 0.001
≤ 0.6
> 0.6
97
≤ 0.08
> 0.08
98
≤ 0.5
> 0.5
102
≤ 2.5
> 2.5
103
≤ 0.2
> 0.2
105
HIV/AIDS
p < 0.001
≤ 0.752
> 0.752
111
Income composition of resources
p < 0.001
≤ 0.882
> 0.882
112
≤ 15
> 15
113
Income composition of resources
p = 0.013
≤ 0.795
> 0.795
117
Income composition of resources
p < 0.001
≤ 0.882
> 0.882
121
≤ 0.7
> 0.7
127
≤ 0.4
> 0.4
131
≤ 0.4
> 0.4
24
≤ 1
> 1
25
≤ 10.7
> 10.7
46
infant deaths
p < 0.001
≤ 1.7
> 1.7
47
≤ 76
> 76
49
≤ 0.44
> 0.44
52
Schooling
p = 0.018
≤ 22.7
> 22.7
111
HIV/AIDS
p = 0.037
≤ 0.1
> 0.1
112
≤ 2182
> 2182
13
Schooling
p < 0.001
≤ 10.3
> 10.3
Schooling
p = 0.021
≤ 8.8
> 8.8
≤ 9
> 9
≤ 10.4
> 10.4
≤ 0.503
> 0.503
≤ 2878.338
> 2878.338
≤ 14.4
> 14.4
≤ 4
> 4
≤ 13.1
> 13.1
Schooling
p < 0.001
≤ 0
> 0
≤ 6
> 6
≤ 21
> 21
≤ 0.752
> 0.752
≤ 0.882
> 0.882
Node
80 80 80

#If I consier `Income composition of resources` + Schooling as predictors
Ctree.LifeExp = ctree(LifeExpImputed$`Life expectancy` ~ `Income composition of resources`
+ Schooling, data = LifeExpImputed)
plot(Ctree.LifeExp)

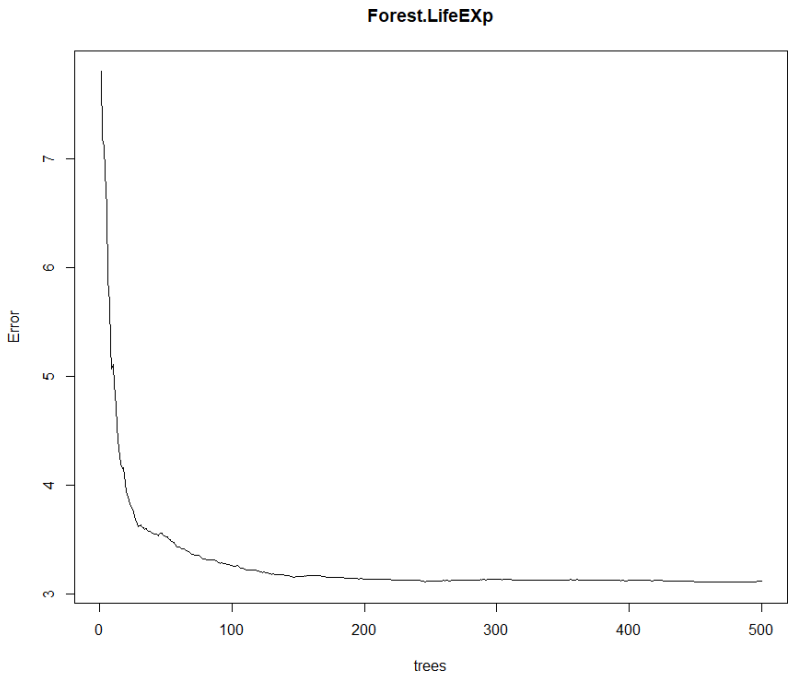


#.....Random Forest Analysis.....

```
install.packages("randomForest")

library(randomForest)

Forest.LifeEXp = randomForest(`Life expectancy` ~ LifeExpImputed$`Adult Mortality`
                    + LifeExpImputed$`infant deaths`+ Polio + Diphtheria +
LifeExpImputed$`HIV/AIDS`
                    + LifeExpImputed$GDP + LifeExpImputed$`thinness 1-19 years`
                    + LifeExpImputed$`Income composition of resources`
                    + Schooling,data = LifeExpImputed)
print(Forest.LifeEXp)

# To find Importance of each predictor.
print(importance(fit,type = 2))
```

```
> print(Forest.LifeEXp)

Call:
 randomForest(formula = `Life expectancy` ~ LifeExpImputed$`Adult Mo
 + LifeExpImputed$`HIV/AIDS` +      LifeExpImputed$GDP + LifeExpImpu
n of resources` + Schooling,      data = LifeExpImputed)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 3

          Mean of squared residuals: 3.131358
                    % Var explained: 96.59
> |
```

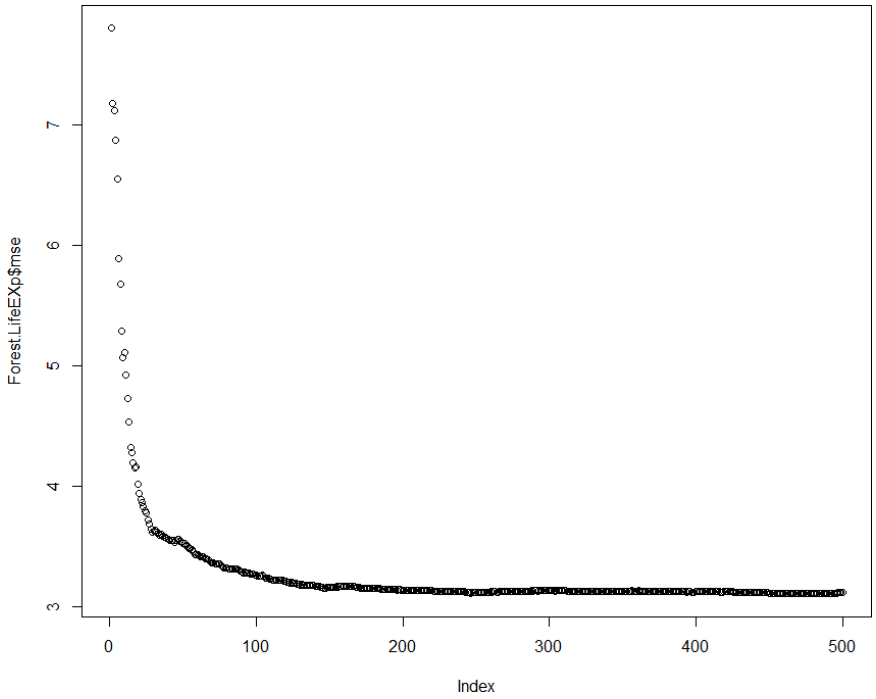
```
plot(Forest.LifeEXp)
```



predict(Forest.LifeEXp)

```
Console ~/
> plot(Forest.LifeEXp$mse)
> Forest.LifeEXp$importance
                                                IncNodePurity
LifeExpImputed$`Adult Mortality`                    38142.207
LifeExpImputed$`infant deaths`                       9605.386
Polio                                                4528.981
Diphtheria                                           4531.889
LifeExpImputed$`HIV/AIDS`                           59295.365
LifeExpImputed$GDP                                   6098.315
LifeExpImputed$`thinness  1-19 years`                8149.041
LifeExpImputed$`Income composition of resources`    65181.446
Schooling                                           32398.715
>
```

plot(Forest.LifeEXp$mse)



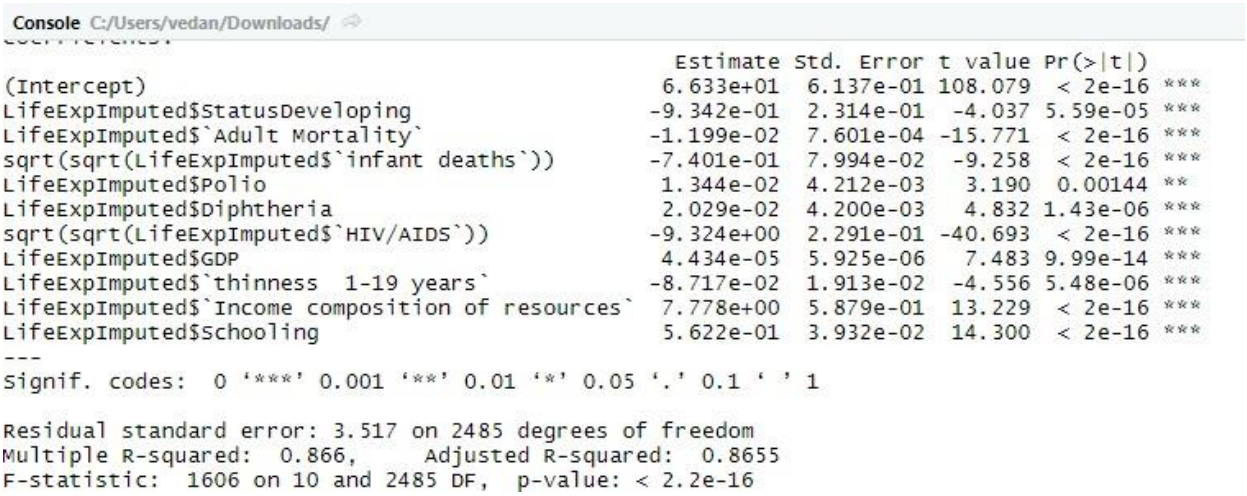

```
Console ~/
> Forest.LifeEXp$ntree
[1] 500
>
```

mean(Forest.LifeEXp$rsq)

```
>
> mean(Forest.LifeEXp$rsq)
[1] 0.9644036
>
```

#.....Regularization approach to see change in selected Model.....

```
SubModel3 = lm(LifeExpImputed$`Life expectancy` ~ LifeExpImputed$Status +
LifeExpImputed$`Adult Mortality` + sqrt(sqrt(LifeExpImputed$`infant deaths`) )
        + LifeExpImputed$Polio + LifeExpImputed$Diphtheria +
sqrt(sqrt(LifeExpImputed$`HIV/AIDS`)) + LifeExpImputed$GDP
        + LifeExpImputed$`thinness 1-19 years`+ LifeExpImputed$`Income composition of
resources` + LifeExpImputed$Schooling,
        data = LifeExpImputed)

summary(SubModel3)
```



```
Console C:/Users/vedan/Downloads/

                                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                                  6.633e+01  6.137e-01 108.079  < 2e-16 ***
LifeExpImputed$StatusDeveloping             -9.342e-01  2.314e-01  -4.037 5.59e-05 ***
LifeExpImputed$`Adult Mortality`            -1.199e-02  7.601e-04 -15.771  < 2e-16 ***
sqrt(sqrt(LifeExpImputed$`infant deaths`))  -7.401e-01  7.994e-02  -9.258  < 2e-16 ***
LifeExpImputed$Polio                         1.344e-02  4.212e-03   3.190  0.00144 **
LifeExpImputed$Diphtheria                    2.029e-02  4.200e-03   4.832 1.43e-06 ***
sqrt(sqrt(LifeExpImputed$`HIV/AIDS`))       -9.324e+00  2.291e-01 -40.693  < 2e-16 ***
LifeExpImputed$GDP                           4.434e-05  5.925e-06   7.483 9.99e-14 ***
LifeExpImputed$`thinness  1-19 years`       -8.717e-02  1.913e-02  -4.556 5.48e-06 ***
LifeExpImputed$`Income composition of resources`  7.778e+00  5.879e-01  13.229  < 2e-16 ***
LifeExpImputed$Schooling                     5.622e-01  3.932e-02  14.300  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.517 on 2485 degrees of freedom
Multiple R-squared:  0.866,     Adjusted R-squared:  0.8655
F-statistic:  1606 on 10 and 2485 DF,  p-value: < 2.2e-16
```

- Code Reference-class Notes
- For ridge and lasso regression, I will be using the `glmnet` library.
- Need to tune my hyperparameter, λ to find the 'best' ridge or lasso model to implement.

```
library(glmnet)
x = model.matrix(LifeExpImputed$`Life expectancy` ~., LifeExpImputed)[,-1]
y = LifeExpImputed$`Life expectancy`
```

- Ridge regression
- Then I would want to build in a cross-validation process to choose 'best' λ. I can do this using `cv.glmnet,`

```
cv_ridge = cv.glmnet(x, y, alpha = 0)
```

- ridge regression is performed by default using `alpha = 0`

cv_ridge$lambda.min

```
Console C:/Users/vedan/Downloads/
> #Code Reference-class Notes
> #For ridge and lasso regression, we
> #Remember, we need to tune our hyper
> library(glmnet)
> x = model.matrix(LifeExpImputed$`Lif
> y = LifeExpImputed$`Life expectancy`
> #Ridge regression
> #Then we would want to build in a cr
> cv_ridge = cv.glmnet(x, y, alpha = 0
> #ridge regression is performed by de
> cv_ridge$lambda.min
[1] 0.721332
>
```

- The cross-validated model with a $\lambda =0.721332 $ provides the optimal model in terms of minimizing MSE

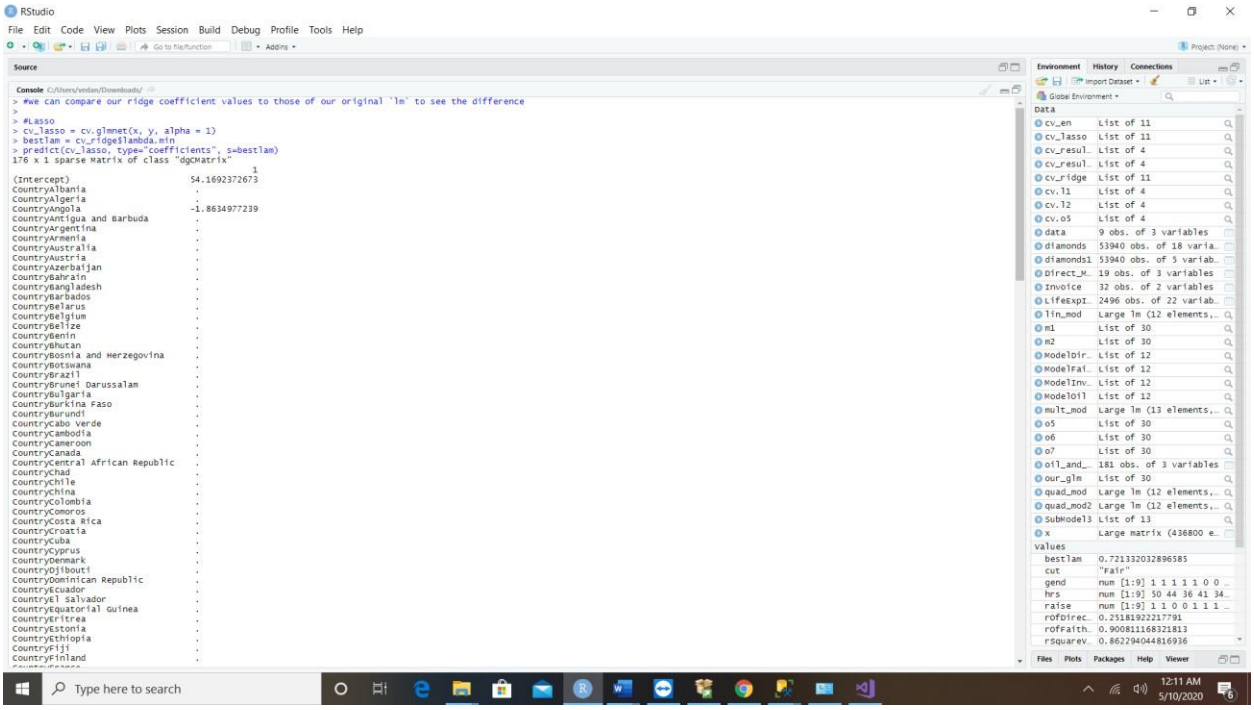
predict(cv_ridge, type="coefficients", s=4.15826)

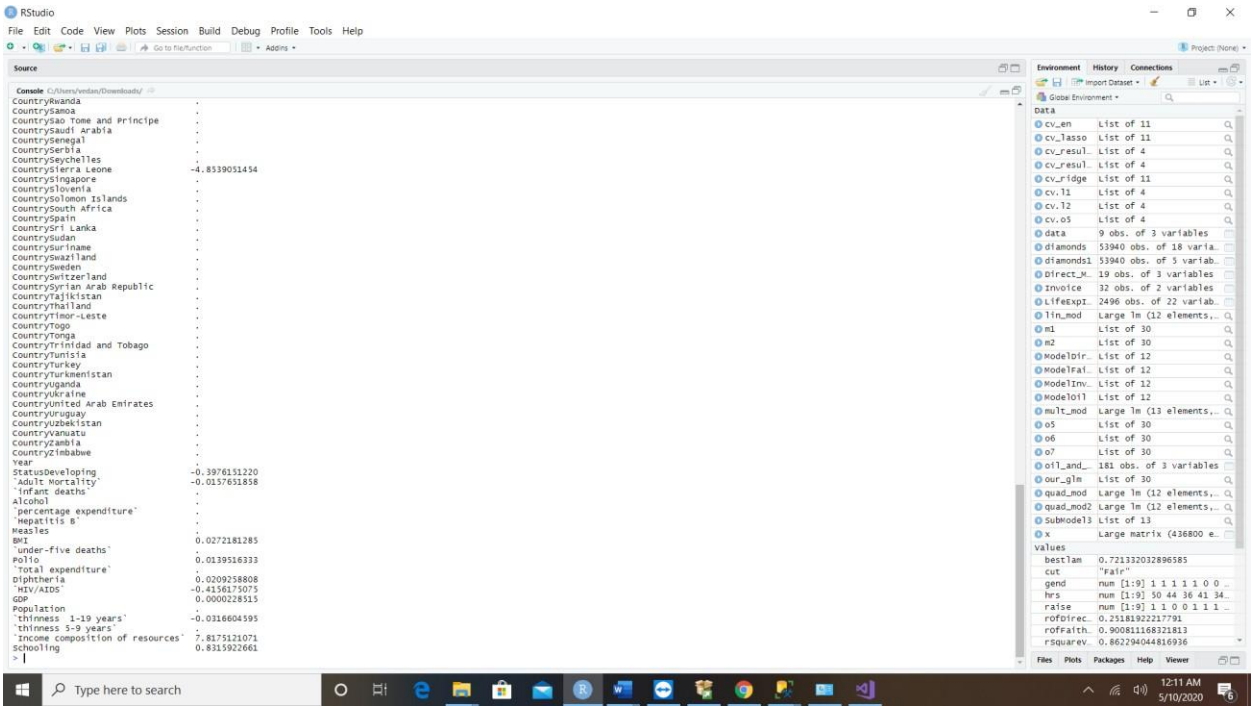
```
Console C:/Users/vedan/Downloads/
> #We see that the cross-validated model with a $\lambda = 1.975$ provides the optimal model in terms of minimizing MSE
> predict(cv_ridge, type="coefficients", s=4.15826)
176 x 1 sparse Matrix of class "dgCMatrix"
                                          1
(Intercept)                       -7.596656e+01
CountryAlbania                     1.815817e+00
CountryAlgeria                     2.073657e+00
CountryAngola                     -9.184505e+00
CountryAntigua and Barbuda         4.116705e+00
CountryArgentina                   4.390583e-01
CountryArmenia                     1.682678e+00
CountryAustralia                   1.152880e+00
CountryAustria                     2.824893e+00
CountryAzerbaijan                  5.400313e-01
CountryBahrain                     1.881562e+00
CountryBangladesh                  3.245357e+00
CountryBarbados                    9.844239e-01
CountryBelarus                    -1.787751e+00
CountryBelgium                     2.013997e+00
CountryBelize                     -1.093337e+00
CountryBenin                      -4.140820e+00
CountryBhutan                      2.379064e+00
CountryBosnia and Herzegovina      4.243001e+00
CountryBotswana                   -4.367461e+00
CountryBrazil                      7.196136e-01
CountryBrunei Darussalam           2.630095e+00
CountryBulgaria                   -1.341048e+00
CountryBurkina Faso               -4.124959e+00
CountryBurundi                    -5.051196e+00
CountryCabo Verde                  3.024597e+00
CountryCambodia                   -3.123183e-01
CountryCameroon                   -6.236949e+00
CountryCanada                      4.097604e+00
CountryCentral African Republic   -6.876743e+00
CountryChad                       -6.500865e+00
CountryChile                       3.782561e+00
CountryChina                       3.922221e+00
CountryColombia                    1.638157e+00
CountryComoros                    -1.841899e+00
CountryCosta Rica                  4.819823e+00
CountryCroatia                     2.625267e-01
CountryCuba                        3.309943e+00
CountryCyprus                      2.339931e+00
CountryDenmark                     4.028676e-01
CountryDjibouti                   -1.121767e+00
CountryDominican Republic          1.508174e+00
CountryEcuador                     2.486784e+00
CountryEl Salvador                 7.868982e-01
CountryEquatorial Guinea          -4.735702e+00
CountryEritrea                    -7.042240e-01
CountryEstonia                     7.109938e-01
CountryEthiopia                   -1.055951e+00
CountryFiji                       -2.245024e+00
CountryFinland                     2.942301e+00
CountryFrance                      4.357583e+00
CountryGabon                      -2.109008e+00
CountryGeorgia                     1.775098e+00
CountryGermany                     2.164627e+00
```

```
CountryRwanda                      -3.570961e+00
CountrySamoa                        2.117142e+00
CountrySao Tome and Principe       -1.506917e+00
CountrySaudi Arabia                 7.561115e-01
CountrySenegal                     -1.208151e+00
CountrySerbia                       1.009317e+00
CountrySeychelles                   1.331783e+00
CountrySierra Leone                -1.148649e+01
CountrySingapore                    3.990535e+00
CountrySlovenia                     1.802731e+00
CountrySolomon Islands             -1.100614e-01
CountrySouth Africa                -2.396558e+00
CountrySpain                        2.858560e+00
CountrySri Lanka                    3.550032e+00
CountrySudan                       -1.806712e+00
CountrySuriname                     8.083757e-01
CountrySwaziland                   -3.898760e+00
CountrySweden                       3.045977e+00
CountrySwitzerland                  1.877866e+00
CountrySyrian Arab Republic         2.202368e+00
CountryTajikistan                  -1.230816e+00
CountryThailand                     2.564380e+00
CountryTimor-Leste                 -2.702602e-01
CountryTogo                        -4.766630e+00
CountryTonga                        3.759508e-01
CountryTrinidad and Tobago          1.233630e+00
CountryTunisia                      1.401161e+00
CountryTurkey                       2.108352e+00
CountryTurkmenistan                -1.446899e+00
CountryUganda                      -4.770936e+00
CountryUkraine                     -5.400714e-01
CountryUnited Arab Emirates         1.828084e+00
CountryUruguay                      1.776498e+00
CountryUzbekistan                  -1.071035e+00
CountryVanuatu                      2.856270e+00
CountryZambia                      -5.316500e+00
CountryZimbabwe                    -4.666905e+00
Year                                6.782590e-02
StatusDeveloping                   -1.801338e+00
`Adult Mortality`                  -8.842770e-03
`infant deaths`                    -3.397924e-04
Alcohol                             5.248699e-02
`percentage expenditure`           1.226789e-04
`Hepatitis B`                       9.131688e-03
Measles                            -1.507514e-05
BMI                                 3.040850e-02
`under-five deaths`                -8.577046e-04
Polio                               1.884229e-02
`Total expenditure`                 2.618539e-03
Diphtheria                          1.857959e-02
`HIV/AIDS`                         -2.747793e-01
GDP                                 2.569198e-05
Population                          1.032909e-09
`thinness  1-19 years`             -1.018555e-01
`thinness 5-9 years`               -9.178407e-02
`Income composition of resources`   5.218348e+00
Schooling                           3.954481e-01
>
```

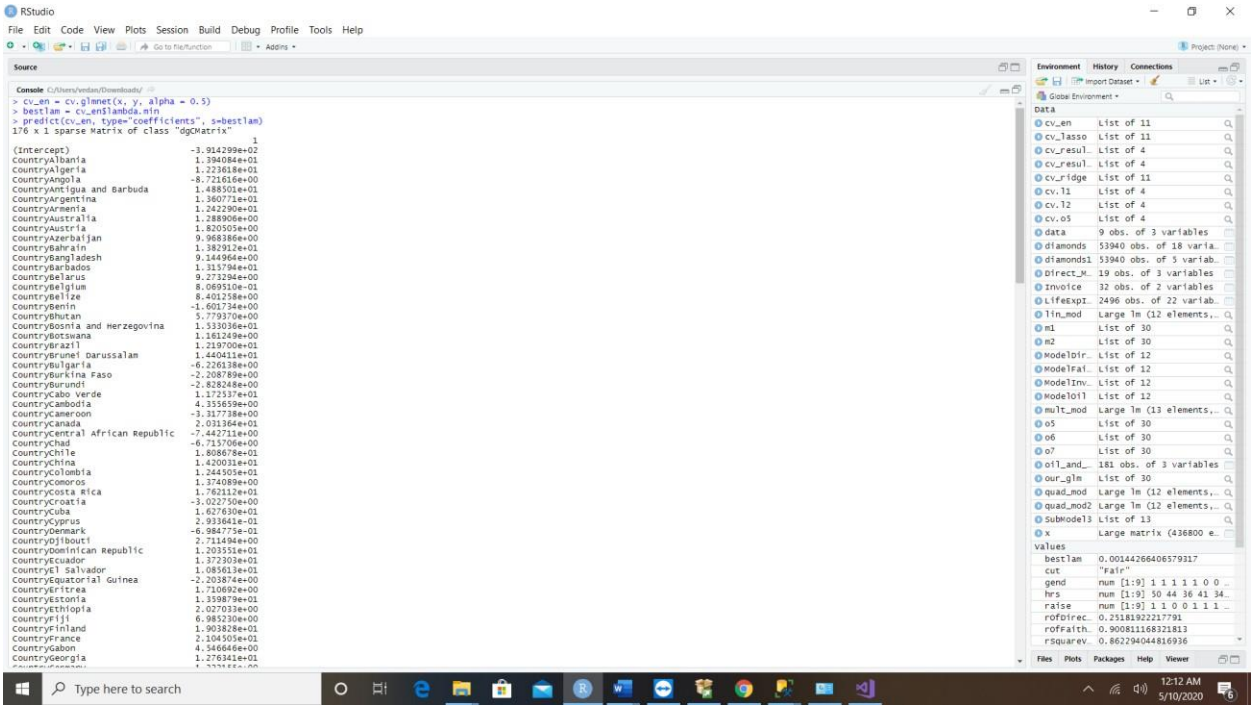
- I can compare my ridge coefficient values to those of my original `lm` to see the difference

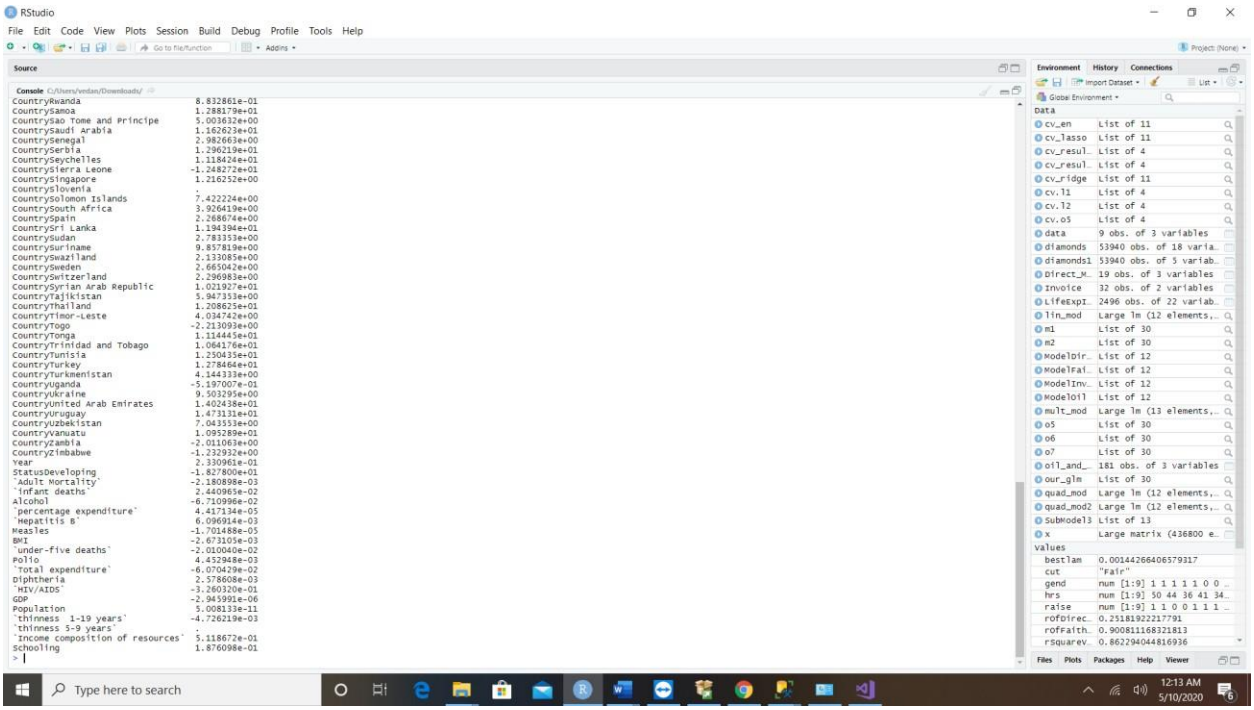
#Lasso
cv_lasso = cv.glmnet(x, y, alpha = 1)
bestlam = cv_ridge$lambda.min
predict(cv_lasso, type="coefficients", s=bestlam)


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
> #we can compare our ridge coefficient values to those of our original `lm` to see the difference
> #Lasso
> cv_lasso = cv.glmnet(x, y, alpha = 1)
> bestlam = cv_ridge$lambda.min
> predict(cv_lasso, type="coefficients", s=bestlam)
176 x 1 sparse Matrix of class "dgCMatrix"
1
(Intercept) 54.1692372673
CountryAlbania .
CountryAlgeria .
CountryAngola -1.8634977239
CountryAntigua and Barbuda .
CountryArgentina .
CountryArmenia .
CountryAustralia .
CountryAustria .
CountryAzerbaijan .
CountryBahrain .
CountryBangladesh .
CountryBarbados .
CountryBelarus .
CountryBelgium .
CountryBelize .
CountryBenin .
CountryBhutan .
CountryBosnia and Herzegovina .
CountryBotswana .
CountryBrazil .
CountryBrunei Darussalam .
CountryBulgaria .
CountryBurkina Faso .
CountryBurundi .
CountryCabo Verde .
CountryCambodia .
CountryCameroon .
CountryCanada .
CountryCentral African Republic .
CountryChad .
CountryChile .
CountryChina .
CountryColombia .
CountryComoros .
CountryCosta Rica .
CountryCroatia .
CountryCuba .
CountryCyprus .
CountryDenmark .
CountryDjibouti .
CountryDominican Republic .
CountryEcuador .
CountryEl Salvador .
CountryEquatorial Guinea .
CountryEritrea .
CountryEstonia .
CountryEthiopia .
CountryFiji .
CountryFinland .
Environment History Connections
Import Dataset
Global Environment
Data
cv_en List of 11
cv_lasso List of 11
cv_resul... List of 4
cv_resul... List of 4
cv_ridge List of 11
cv.l1 List of 4
cv.l2 List of 4
cv.o5 List of 4
data 9 obs. of 3 variables
diamonds 53940 obs. of 18 varia...
diamonds1 53940 obs. of 5 variab...
Direct_M... 19 obs. of 3 variables
Invoice 32 obs. of 2 variables
LifeExpI... 2496 obs. of 22 variab...
lin_mod Large lm (12 elements,...
m1 List of 30
m2 List of 30
Modeldir... List of 12
Modelfai... List of 12
ModelInv... List of 12
Modeloil List of 12
mult_mod Large lm (13 elements,...
o5 List of 30
o6 List of 30
o7 List of 30
oil_and_... 181 obs. of 3 variables
our_glm List of 30
quad_mod Large lm (12 elements,...
quad_mod2 Large lm (12 elements,...
SubModel3 List of 13
x Large matrix (436800 e...
values
bestlam 0.721332032896585
cut "Fair"
gend num [1:9] 1 1 1 1 1 0 0 ...
hrs num [1:9] 50 44 36 41 34...
raise num [1:9] 1 1 0 0 1 1 1 ...
rofDirec... 0.25181922217791
rofFaith... 0.900811168321813
rsquarev... 0.862294044816936
Files Plots Packages Help Viewer
Type here to search
12:11 AM
5/10/2020


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
CountryRwanda .
CountrySamoa .
CountrySao Tome and Principe .
CountrySaudi Arabia .
CountrySenegal .
CountrySerbia .
CountrySeychelles .
CountrySierra Leone -4.8539051454
CountrySingapore .
CountrySlovenia .
CountrySolomon Islands .
CountrySouth Africa .
CountrySpain .
CountrySri Lanka .
CountrySudan .
CountrySuriname .
CountrySwaziland .
CountrySweden .
CountrySwitzerland .
CountrySyrian Arab Republic .
CountryTajikistan .
CountryThailand .
CountryTimor-Leste .
CountryTogo .
CountryTonga .
CountryTrinidad and Tobago .
CountryTunisia .
CountryTurkey .
CountryTurkmenistan .
CountryUganda .
CountryUkraine .
CountryUnited Arab Emirates .
CountryUruguay .
CountryUzbekistan .
CountryVanuatu .
CountryZambia .
CountryZimbabwe .
Year .
StatusDeveloping -0.3976151220
`Adult Mortality` -0.0157651858
`infant deaths` .
Alcohol .
`percentage expenditure` .
`Hepatitis B` .
Measles .
BMI 0.0272181285
`under-five deaths` .
Polio 0.0139516333
`Total expenditure` .
Diphtheria 0.0209258808
`HIV/AIDS` -0.4156175075
GDP 0.0000228515
Population .
`thinness 1-19 years` -0.0316604595
`thinness 5-9 years` .
`Income composition of resources` 7.8175121071
Schooling 0.8315922661
Environment History Connections
Import Dataset
Global Environment
Data
cv_en List of 11
cv_lasso List of 11
cv_resul... List of 4
cv_resul... List of 4
cv_ridge List of 11
cv.l1 List of 4
cv.l2 List of 4
cv.o5 List of 4
data 9 obs. of 3 variables
diamonds 53940 obs. of 18 varia...
diamonds1 53940 obs. of 5 variab...
Direct_M... 19 obs. of 3 variables
Invoice 32 obs. of 2 variables
LifeExpI... 2496 obs. of 22 variab...
lin_mod Large lm (12 elements,...
m1 List of 30
m2 List of 30
Modeldir... List of 12
Modelfai... List of 12
ModelInv... List of 12
Modeloil List of 12
mult_mod Large lm (13 elements,...
o5 List of 30
o6 List of 30
o7 List of 30
oil_and_... 181 obs. of 3 variables
our_glm List of 30
quad_mod Large lm (12 elements,...
quad_mod2 Large lm (12 elements,...
SubModel3 List of 13
x Large matrix (436800 e...
values
bestlam 0.721332032896585
cut "Fair"
gend num [1:9] 1 1 1 1 1 0 0 ...
hrs num [1:9] 50 44 36 41 34...
raise num [1:9] 1 1 0 0 1 1 1 ...
rofDirec... 0.25181922217791
rofFaith... 0.900811168321813
rsquarev... 0.862294044816936
Files Plots Packages Help Viewer
Type here to search
12:11 AM
5/10/2020

#elastic net
cv_en = cv.glmnet(x, y, alpha = 0.5)
bestlam = cv_en$lambda.min
predict(cv_en, type="coefficients", s=bestlam)





- "lm" came out to be feasible and better model.