Final Project Mai Ngo

Mai Ngo

2023-08-04

Please read: Homework 3 - Part 1 starts from Line 203 / Page 15. Previous part is what I did to clean up the data.

The data set includes 2938 observations and 22 variables. Variables information:

Country - Names of the countries.
Year - Year of observations.
Status - whether developed or developing.
Life Expectancy - Average time a citizen of any country is expected to live (in years).
Adult Mortality - Probability of dying between 15 and 60 years per 1000 population.
Infant deaths - Number of Infant Deaths per 1000 population.
Alcohol - Alcohol, recorded per capita (15+) consumption (in litres).
Percentage expenditure - Expenditure on health as a percentage of GDP per capita (%).
Hepatitis B - Immunization coverage among 1-year old (%).
Measles - Number of reported cases per 1000 population.
BMI - Average Body Mass Index of entire population.
Under-five deaths - Number of under-five deaths per 1000 population.
Polio - Immunization coverage among 1-year old (%).
Total expenditure - Government expenditure on health industry as a percentage of total government expenditure (%).
Diphtheria - Immunization coverage among 1-year old (%).
HIV/AIDS - Deaths per 1000 live births HIV/AIDS (0-4 years).
GDP - Gross Domestic Product per capita (in current USD).
Population - Population of the country.
Thinness 10-19 years - Prevalence of thinness among children and adolescents for Age 10 to 19 (%).
Thinness 5-9 years - Prevalence of thinness among children for Age 5 to 9(%).
Income composition of resources - Human Development Index in terms of income composition of resources (index ranging from 0 to 1).
Schooling - Number of years of Schooling.

Read explore data.

Set Working Directory.

```r
setwd("C:/Users/maimu/OneDrive/Documents/DePaul/DSC 424")
```

Read data sets. Source: Population - https://data.worldbank.org/indicator/SP.POP.TOTL?end=2015&start=2000
GDP per capita - https://data.worldbank.org/indicator/NY.GDP.PCAP.CD

```r
expectancyData <- read.csv(file="Life Expectancy Data.csv", header=TRUE, sep=",")
populationData <- read.csv(file="World Population.csv", header=TRUE, sep=",")
gdpData <- read.csv(file="World GDP.csv", header=TRUE, sep=",")
```

Life Expectancy Data.

```r
head(expectancyData)
```

```
##       Country Year     Status Life.expectancy Adult.Mortality infant.deaths
## 1 Afghanistan 2015 Developing            65.0             263            62
## 2 Afghanistan 2014 Developing            59.9             271            64
## 3 Afghanistan 2013 Developing            59.9             268            66
## 4 Afghanistan 2012 Developing            59.5             272            69
## 5 Afghanistan 2011 Developing            59.2             275            71
## 6 Afghanistan 2010 Developing            58.8             279            74
##   Alcohol percentage.expenditure Hepatitis.B Measles  BMI under.five.deaths
## 1    0.01              71.279624          65    1154 19.1                83
## 2    0.01              73.523582          62     492 18.6                86
## 3    0.01              73.219243          64     430 18.1                89
## 4    0.01              78.184215          67    2787 17.6                93
## 5    0.01               7.097109          68    3013 17.2                97
## 6    0.01              79.679367          66    1989 16.7               102
##   Polio Total.expenditure Diphtheria HIV.AIDS       GDP Population
## 1     6              8.16         65      0.1 584.25921   33736494
## 2    58              8.18         62      0.1 612.69651     327582
## 3    62              8.13         64      0.1 631.74498   31731688
## 4    67              8.52         67      0.1 669.95900    3696958
## 5    68              7.87         68      0.1  63.53723    2978599
## 6    66              9.20         66      0.1 553.32894    2883167
##   thinness..1.19.years thinness.5.9.years Income.composition.of.resources
## 1                 17.2               17.3                           0.479
## 2                 17.5               17.5                           0.476
## 3                 17.7               17.7                           0.470
## 4                 17.9               18.0                           0.463
## 5                 18.2               18.2                           0.454
## 6                 18.4               18.4                           0.448
##   Schooling
## 1      10.1
## 2      10.0
## 3       9.9
## 4       9.8
## 5       9.5
## 6       9.2
```

Look at original expectancyData: GDP and Population have 448 and 652 missing values, respectively. Attempted to apply Listwise deletion which led to 43% data loss, this would loose the original data characteristics. Approach: Fill in missing values.

```r
summary(expectancyData)
```

```
##    Country               Year          Status          Life.expectancy
##  Length:2938        Min.   :2000   Length:2938        Min.   :36.30
##  Class :character   1st Qu.:2004   Class :character   1st Qu.:63.10
##  Mode  :character   Median :2008   Mode  :character   Median :72.10
##                     Mean   :2008                      Mean   :69.22
##                     3rd Qu.:2012                      3rd Qu.:75.70
##                     Max.   :2015                      Max.   :89.00
##                                                       NA's   :10
##  Adult.Mortality infant.deaths       Alcohol        percentage.expenditure
##  Min.   :  1.0   Min.   :   0.0   Min.   : 0.0100   Min.   :    0.000
##  1st Qu.: 74.0   1st Qu.:   0.0   1st Qu.: 0.8775   1st Qu.:    4.685
##  Median :144.0   Median :   3.0   Median : 3.7550   Median :   64.913
##  Mean   :164.8   Mean   :  30.3   Mean   : 4.6029   Mean   :  738.251
##  3rd Qu.:228.0   3rd Qu.:  22.0   3rd Qu.: 7.7025   3rd Qu.:  441.534
##  Max.   :723.0   Max.   :1800.0   Max.   :17.8700   Max.   :19479.912
##  NA's   :10                       NA's   :194
##   Hepatitis.B       Measles              BMI        under.five.deaths
##  Min.   : 1.00   Min.   :     0.0   Min.   : 1.00   Min.   :   0.00
##  1st Qu.:77.00   1st Qu.:     0.0   1st Qu.:19.30   1st Qu.:   0.00
##  Median :92.00   Median :    17.0   Median :43.50   Median :   4.00
##  Mean   :80.94   Mean   :  2419.6   Mean   :38.32   Mean   :  42.04
##  3rd Qu.:97.00   3rd Qu.:   360.2   3rd Qu.:56.20   3rd Qu.:  28.00
##  Max.   :99.00   Max.   :212183.0   Max.   :87.30   Max.   :2500.00
##  NA's   :553                        NA's   :34
##      Polio       Total.expenditure   Diphtheria       HIV.AIDS
##  Min.   : 3.00   Min.   : 0.370    Min.   : 2.00   Min.   : 0.100
##  1st Qu.:78.00   1st Qu.: 4.260    1st Qu.:78.00   1st Qu.: 0.100
##  Median :93.00   Median : 5.755    Median :93.00   Median : 0.100
##  Mean   :82.55   Mean   : 5.938    Mean   :82.32   Mean   : 1.742
##  3rd Qu.:97.00   3rd Qu.: 7.492    3rd Qu.:97.00   3rd Qu.: 0.800
##  Max.   :99.00   Max.   :17.600    Max.   :99.00   Max.   :50.600
##  NA's   :19      NA's   :226       NA's   :19
##       GDP              Population        thinness..1.19.years
##  Min.   :     1.68   Min.   :3.400e+01   Min.   : 0.10
##  1st Qu.:   463.94   1st Qu.:1.958e+05   1st Qu.: 1.60
##  Median :  1766.95   Median :1.387e+06   Median : 3.30
##  Mean   :  7483.16   Mean   :1.275e+07   Mean   : 4.84
##  3rd Qu.:  5910.81   3rd Qu.:7.420e+06   3rd Qu.: 7.20
##  Max.   :119172.74   Max.   :1.294e+09   Max.   :27.70
##  NA's   :448         NA's   :652         NA's   :34
##  thinness.5.9.years Income.composition.of.resources   Schooling
##  Min.   : 0.10      Min.   :0.0000                  Min.   : 0.00
##  1st Qu.: 1.50      1st Qu.:0.4930                  1st Qu.:10.10
##  Median : 3.30      Median :0.6770                  Median :12.30
##  Mean   : 4.87      Mean   :0.6276                  Mean   :11.99
##  3rd Qu.: 7.20      3rd Qu.:0.7790                  3rd Qu.:14.30
##  Max.   :28.60      Max.   :0.9480                  Max.   :20.70
##  NA's   :34         NA's   :167                     NA's   :163
```

Population Data.

```r
head(populationData)
```

```
##                   Country.Name     X2000     X2001     X2002     X2003     X2004
## 1                  Afghanistan  19542982  19688632  21000256  22645130  23553551
## 2 Africa Eastern and Southern 401600588 412001885 422741118 433807484 445281555
## 3  Africa Western and Central 269611898 277160097 284952322 292977949 301265247
## 4                      Albania   3089027   3060173   3051010   3039616   3026939
## 5                      Algeria  30774621  31200985  31624696  32055883  32510186
## 6               American Samoa     58230     58324     58177     57941     57626
##       X2005     X2006     X2007     X2008     X2009     X2010     X2011
## 1  24411191  25442944  25903301  26427199  27385307  28189672  29249157
## 2 457153837 469508516 482406426 495748900 509410477 523459657 537792950
## 3 309824829 318601484 327612838 336893835 346475221 356337762 366489204
## 4   3011487   2992547   2970017   2947314   2927519   2913021   2905195
## 5  32956690  33435080  33983827  34569592  35196037  35856344  36543541
## 6     57254     56837     56383     55891     55366     54849     54310
##       X2012     X2013     X2014     X2015
## 1  30466479  31541209  32716210  33753499
## 2 552530654 567892149 583651101 600008424
## 3 376797999 387204553 397855507 408690375
## 4   2900401   2895092   2889104   2880703
## 5  37260563  38000626  38760168  39543154
## 6     53691     52995     52217     51368
```

Re-arrange population data same format as expectancy data.

```r
library(tidyr)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

```
##
## Attaching package: 'dplyr'
```

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

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

```r
#Rename year columns.
populationData_newCol_names <- c("Country.Name", paste0("Year", 2000:2015))
names(populationData) <- populationData_newCol_names
#Convert to same format of original life expectancy data.
populationData_long <- pivot_longer(populationData, cols = starts_with("Year"),
                                    names_to = "Year", values_to = "Population")
populationData_long$Year <- as.numeric(gsub("Year", "", populationData_long$Year))
populationData_long <- populationData_long %>% arrange(Country.Name, desc(Year))
head(populationData_long)
```

```
## # A tibble: 6 x 3
##   Country.Name  Year Population
##   <chr>        <dbl>      <dbl>
## 1 Afghanistan   2015   33753499
## 2 Afghanistan   2014   32716210
## 3 Afghanistan   2013   31541209
## 4 Afghanistan   2012   30466479
## 5 Afghanistan   2011   29249157
## 6 Afghanistan   2010   28189672
```

Perform left join, replace original ‘Population’ column with World Bank population data.

```
#Rename 'Country.Name' column in populationData_long to 'Country'.
names(populationData_long)[names(populationData_long) == "Country.Name"] <- "Country"
expectancyData2 <- left_join(expectancyData, populationData_long, by = c("Country", "Year"))
expectancyData2 <- expectancyData2[, !(names(expectancyData2) %in% c('Population.x'))]
colnames(expectancyData2)[colnames(expectancyData2) == 'Population.y'] = 'Population'
head(expectancyData2)
```

```
##       Country Year     Status Life.expectancy Adult.Mortality infant.deaths
## 1 Afghanistan 2015 Developing            65.0             263            62
## 2 Afghanistan 2014 Developing            59.9             271            64
## 3 Afghanistan 2013 Developing            59.9             268            66
## 4 Afghanistan 2012 Developing            59.5             272            69
## 5 Afghanistan 2011 Developing            59.2             275            71
## 6 Afghanistan 2010 Developing            58.8             279            74
##   Alcohol percentage.expenditure Hepatitis.B Measles  BMI under.five.deaths
## 1    0.01              71.279624          65    1154 19.1                83
## 2    0.01              73.523582          62     492 18.6                86
## 3    0.01              73.219243          64     430 18.1                89
## 4    0.01              78.184215          67    2787 17.6                93
## 5    0.01               7.097109          68    3013 17.2                97
## 6    0.01              79.679367          66    1989 16.7               102
##   Polio Total.expenditure Diphtheria HIV.AIDS       GDP thinness..1.19.years
## 1     6              8.16         65      0.1 584.25921                 17.2
## 2    58              8.18         62      0.1 612.69651                 17.5
## 3    62              8.13         64      0.1 631.74498                 17.7
## 4    67              8.52         67      0.1 669.95900                 17.9
## 5    68              7.87         68      0.1  63.53723                 18.2
## 6    66              9.20         66      0.1 553.32894                 18.4
##   thinness.5.9.years Income.composition.of.resources Schooling Population
## 1               17.3                           0.479      10.1   33753499
## 2               17.5                           0.476      10.0   32716210
## 3               17.7                           0.470       9.9   31541209
## 4               18.0                           0.463       9.8   30466479
## 5               18.2                           0.454       9.5   29249157
## 6               18.4                           0.448       9.2   28189672
```

Check missing values with new Population column. 2 observations with missing values. Cook Islands and Niue don’t have data range 2000-2015.

```r
expectancyData2[is.na(expectancyData2$Population), ]
```

```
##           Country Year     Status Life.expectancy Adult.Mortality infant.deaths
## 625  Cook Islands 2013 Developing              NA              NA             0
## 1910         Niue 2013 Developing              NA              NA             0
##      Alcohol percentage.expenditure Hepatitis.B Measles  BMI under.five.deaths
## 625     0.01                      0          98       0 82.8                 0
## 1910    0.01                      0          99       0 77.3                 0
##      Polio Total.expenditure Diphtheria HIV.AIDS GDP thinness..1.19.years
## 625     98              3.58         98      0.1  NA                  0.1
## 1910    99              7.20         99      0.1  NA                  0.1
##      thinness.5.9.years Income.composition.of.resources Schooling Population
## 625                 0.1                              NA        NA         NA
## 1910                0.1                              NA        NA         NA
```

Hence, exclude these two observations. Get dimension. Using expectancyData2 forward.

```r
expectancyData2 <- expectancyData2 %>% filter(Country != 'Cook Islands' & Country != 'Niue')
dim(expectancyData2)
```

```
## [1] 2936   22
```

GDP per capita Data.

```r
head(gdpData)
```

```
##                  Country.Name      X2000      X2001      X2002      X2003
## 1                       Aruba 21023.1575 20913.2995 21377.0952 22050.8309
## 2 Africa Eastern and Southern   709.0610   630.1989   630.4791   816.4377
## 3                 Afghanistan         NA         NA   183.5328   200.4624
## 4  Africa Western and Central   522.7728   535.8979   621.8625   700.4434
## 5                      Angola   556.8842   527.4641   872.6576   982.8056
## 6                     Albania  1126.6833  1281.6598  1425.1242  1846.1201
##        X2004      X2005      X2006      X2007      X2008      X2009      X2010
## 1 24104.6462 24975.6733 25833.4456 27665.4265 29011.5592 25739.1372 24452.9284
## 2   989.2208  1124.2203  1230.1948  1374.0862  1433.2583  1417.1306  1649.6391
## 3   221.6577   255.0551   274.0007   375.0783   387.8493   443.8452   554.5947
## 4   843.9898  1003.4366  1245.8229  1420.8403  1685.3712  1467.2412  1679.6467
## 5  1254.6961  1900.7238  2597.9636  3121.3487  4081.7175  3123.6989  3496.7848
## 6  2373.5813  2673.7878  2972.7436  3595.0383  4370.5397  4114.1340  4094.3497
##        X2011      X2012      X2013      X2014      X2015
## 1 26044.4359 25609.9557 26515.6781 26942.3080 28421.3865
## 2  1799.6230  1765.2501  1736.2225  1724.5344  1545.5591
## 3   621.9124   663.1411   651.9879   628.1468   592.4762
## 4  1860.9439  1957.5196  2153.7661  2247.8575  1880.7508
## 5  4511.1532  4962.5521  5101.9839  5059.0804  3100.8307
## 6  4437.1411  4247.6314  4413.0634  4578.6332  3952.8036
```

Re-arrange population data same format as expectancy data.

```r
#Rename year columns.
gdpData_newCol_names <- c("Country.Name", paste0("Year", 2000:2015))
names(gdpData) <- gdpData_newCol_names
#Convert to same format of original life expectancy data.
gdpData_long <- pivot_longer(gdpData, cols = starts_with("Year"),
                             names_to = "Year", values_to = "GDP")
gdpData_long$Year <- as.numeric(gsub("Year", "", gdpData_long$Year))
gdpData_long <- gdpData_long %>% arrange(Country.Name, desc(Year))
head(gdpData_long)
```

```
## # A tibble: 6 x 3
##   Country.Name  Year   GDP
##   <chr>        <dbl> <dbl>
## 1 Afghanistan   2015  592.
## 2 Afghanistan   2014  628.
## 3 Afghanistan   2013  652.
## 4 Afghanistan   2012  663.
## 5 Afghanistan   2011  622.
## 6 Afghanistan   2010  555.
```

Perform left join, replace original ‘GDP’ column with World Bank GDP per capita data.

```r
#Rename 'Country.Name' column in gdpData_long to 'Country'.
names(gdpData_long)[names(gdpData_long) == "Country.Name"] <- "Country"
expectancyData3 <- left_join(expectancyData2, gdpData_long, by = c("Country", "Year"))
expectancyData3 <- expectancyData3[, !(names(expectancyData3) %in% c('GDP.x'))]
colnames(expectancyData3)[colnames(expectancyData3) == 'GDP.y'] = 'GDP'
head(expectancyData3)
```

```
##       Country Year     Status Life.expectancy Adult.Mortality infant.deaths
## 1 Afghanistan 2015 Developing            65.0             263            62
## 2 Afghanistan 2014 Developing            59.9             271            64
## 3 Afghanistan 2013 Developing            59.9             268            66
## 4 Afghanistan 2012 Developing            59.5             272            69
## 5 Afghanistan 2011 Developing            59.2             275            71
## 6 Afghanistan 2010 Developing            58.8             279            74
##   Alcohol percentage.expenditure Hepatitis.B Measles  BMI under.five.deaths
## 1    0.01              71.279624          65    1154 19.1                83
## 2    0.01              73.523582          62     492 18.6                86
## 3    0.01              73.219243          64     430 18.1                89
## 4    0.01              78.184215          67    2787 17.6                93
## 5    0.01               7.097109          68    3013 17.2                97
## 6    0.01              79.679367          66    1989 16.7               102
##   Polio Total.expenditure Diphtheria HIV.AIDS thinness..1.19.years
## 1     6              8.16         65      0.1                 17.2
## 2    58              8.18         62      0.1                 17.5
## 3    62              8.13         64      0.1                 17.7
## 4    67              8.52         67      0.1                 17.9
## 5    68              7.87         68      0.1                 18.2
## 6    66              9.20         66      0.1                 18.4
##   thinness.5.9.years Income.composition.of.resources Schooling Population
## 1               17.3                           0.479      10.1   33753499
## 2               17.5                           0.476      10.0   32716210
## 3               17.7                           0.470       9.9   31541209
## 4               18.0                           0.463       9.8   30466479
## 5               18.2                           0.454       9.5   29249157
## 6               18.4                           0.448       9.2   28189672
##        GDP
## 1 592.4762
## 2 628.1468
## 3 651.9879
## 4 663.1411
## 5 621.9124
## 6 554.5947
```

Check missing values with new GDP column. Several observations within a country have missing values.

```
head(expectancyData3[is.na(expectancyData3$GDP), ], 10)
```

```
##                                   Country Year     Status Life.expectancy
## 15                            Afghanistan 2001 Developing            55.3
## 16                            Afghanistan 2000 Developing            54.8
## 705 Democratic People’s Republic of Korea 2015 Developing            76.0
## 706 Democratic People’s Republic of Korea 2014 Developing            73.0
## 707 Democratic People’s Republic of Korea 2013 Developing            71.0
## 708 Democratic People’s Republic of Korea 2012 Developing            69.8
## 709 Democratic People’s Republic of Korea 2011 Developing            69.4
## 710 Democratic People’s Republic of Korea 2010 Developing            69.0
## 711 Democratic People’s Republic of Korea 2009 Developing            68.7
## 712 Democratic People’s Republic of Korea 2008 Developing            68.6
##     Adult.Mortality infant.deaths Alcohol percentage.expenditure Hepatitis.B
## 15              316            88    0.01               10.57473          63
## 16              321            88    0.01               10.42496          62
## 705             139             6      NA                0.00000          96
## 706             142             6    0.01                0.00000          93
## 707             146             6    3.35                0.00000          93
## 708             149             7    3.61                0.00000          96
## 709             153             8    3.39                0.00000          94
## 710             157             8    3.12                0.00000          93
## 711             161             9    3.35                0.00000          93
## 712             164             9    3.16                0.00000          92
##     Measles  BMI under.five.deaths Polio Total.expenditure Diphtheria HIV.AIDS
## 15     8762 12.6               122    35               7.8         33      0.1
## 16     6532 12.2               122    24               8.2         24      0.1
## 705       0 32.9                 7    99                NA         96      0.1
## 706       3 32.4                 8    99                NA         93      0.1
## 707       0 31.8                 8    99                NA         93      0.1
## 708       0 31.3                 9    99                NA         96      0.1
## 709       0  3.8                10    99                NA         94      0.1
## 710       0  3.3                10    99                NA         93      0.1
## 711       0 29.7                11    98                NA         93      0.1
## 712       8 29.2                12    98                NA         92      0.1
##     thinness..1.19.years thinness.5.9.years Income.composition.of.resources
## 15                   2.1                2.4                           0.340
## 16                   2.3                2.5                           0.338
## 705                  4.9                 4.9                               NA
## 706                  4.9                 4.9                               NA
## 707                  5.0                 5.0                               NA
## 708                  5.1                 5.1                               NA
## 709                  5.1                 5.2                               NA
## 710                  5.2                 5.2                               NA
## 711                  5.3                 5.3                               NA
## 712                  5.4                 5.4                               NA
##     Schooling Population GDP
## 15        5.9   19688632  NA
## 16        5.5   19542982  NA
## 705        NA   25258015  NA
## 706        NA   25126131  NA
## 707        NA   25001819  NA
## 708        NA   24887770  NA
## 709        NA   24783789  NA
## 710        NA   24686435  NA
## 711        NA   24581509  NA
## 712        NA   24469047  NA
```

Removing Democratic People's Republic of Korea due to complete missing values in both Population and GDP.
This is a good political POV since North Korea tends not to share country stats/statement globally.
Get dimension.

```r
expectancyData3 <- expectancyData3 %>% filter(Country != "Democratic People's Republic of Korea")
dim(expectancyData3)
```

```
## [1] 2920   22
```

Since there are still GDP missing values within a country. Also out of available time range (2000-2015). Hence, apply extrapolation to fill in missing values. Use the known 'Year' values and perform linear extrapolation to estimate the missing GDP values based on the given 'GDP' values.

```r
library(Hmisc)
```

```
## Warning: package 'Hmisc' was built under R version 4.2.3
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
expectancyData4 = expectancyData3
expectancyData4 <- expectancyData4 %>% group_by(Country) %>%
  mutate(GDP = if (any(is.na(GDP))) approxExtrap(x = Year[!is.na(GDP)], y = GDP[!is.na(GDP)], xout = Yea
head(expectancyData4)
```

```
## # A tibble: 6 x 22
##   Country      Year Status Life.expectancy Adult.Mortality infant.deaths Alcohol
##   <chr>       <dbl> <chr>            <dbl>           <int>         <int>   <dbl>
## 1 Afghanistan  2015 Devel~            65               263            62    0.01
## 2 Afghanistan  2014 Devel~            59.9             271            64    0.01
## 3 Afghanistan  2013 Devel~            59.9             268            66    0.01
## 4 Afghanistan  2012 Devel~            59.5             272            69    0.01
## 5 Afghanistan  2011 Devel~            59.2             275            71    0.01
## 6 Afghanistan  2010 Devel~            58.8             279            74    0.01
## # i 15 more variables: percentage.expenditure <dbl>, Hepatitis.B <int>,
## #   Measles <int>, BMI <dbl>, under.five.deaths <int>, Polio <int>,
## #   Total.expenditure <dbl>, Diphtheria <int>, HIV.AIDS <dbl>,
## #   thinness..1.19.years <dbl>, thinness.5.9.years <dbl>,
## #   Income.composition.of.resources <dbl>, Schooling <dbl>, Population <dbl>,
## #   GDP <dbl>
```

Check if there is any missing value within GDP col still.
No missing value.

```
expectancyData4[is.na(expectancyData4$GDP), ]
```

```
## # A tibble: 0 x 22
## # i 22 variables: Country <chr>, Year <dbl>, Status <chr>,
## #   Life.expectancy <dbl>, Adult.Mortality <int>, infant.deaths <int>,
## #   Alcohol <dbl>, percentage.expenditure <dbl>, Hepatitis.B <int>,
## #   Measles <int>, BMI <dbl>, under.five.deaths <int>, Polio <int>,
## #   Total.expenditure <dbl>, Diphtheria <int>, HIV.AIDS <dbl>,
## #   thinness..1.19.years <dbl>, thinness.5.9.years <dbl>,
## #   Income.composition.of.resources <dbl>, Schooling <dbl>, ...
```

Get a country sample to see how extrapolation works. Sample country Somalia.
GDP value extended nicely, given available data from 2013-2015.

```
expectancyData4 %>% filter(Country == 'Somalia')
```

```
## # A tibble: 16 x 22
##    Country  Year Status    Life.expectancy Adult.Mortality infant.deaths Alcohol
##    <chr>   <dbl> <chr>               <dbl>           <int>         <int>   <dbl>
##  1 Somalia  2015 Developi~            55               312            50   NA
##  2 Somalia  2014 Developi~            54.3             321            51    0.01
##  3 Somalia  2013 Developi~            54.2             318            51    0.01
##  4 Somalia  2012 Developi~            53.1             336            51    0.01
##  5 Somalia  2011 Developi~            53.1             329            51    0.01
##  6 Somalia  2010 Developi~            52.4             336            52    0.01
##  7 Somalia  2009 Developi~            52.2             335            52    0.01
##  8 Somalia  2008 Developi~            51.9             336            52    0.01
##  9 Somalia  2007 Developi~            51.5              34            52    0.01
## 10 Somalia  2006 Developi~            51.5             337            51    0.01
## 11 Somalia  2005 Developi~            51.6             334            50    0.01
## 12 Somalia  2004 Developi~            51.2             341            49    0.01
## 13 Somalia  2003 Developi~            51.1             344            48    0.01
## 14 Somalia  2002 Developi~            58               348            47    0.01
## 15 Somalia  2001 Developi~            57               352            46    0.01
## 16 Somalia  2000 Developi~            55               355            45    0.01
## # i 15 more variables: percentage.expenditure <dbl>, Hepatitis.B <int>,
## #   Measles <int>, BMI <dbl>, under.five.deaths <int>, Polio <int>,
## #   Total.expenditure <dbl>, Diphtheria <int>, HIV.AIDS <dbl>,
## #   thinness..1.19.years <dbl>, thinness.5.9.years <dbl>,
## #   Income.composition.of.resources <dbl>, Schooling <dbl>, Population <dbl>,
## #   GDP <dbl>
```

Using expectancyData4 moving forward. Still have 1403 missing values.

```
sum(is.na(expectancyData4))
```

```
## [1] 1403
```

Take a look at summary, other numerical attributes have missing values (but not significant count).

```
summary(expectancyData4)
```

```
##   Country               Year         Status          Life.expectancy
##  Length:2920        Min.   :2000   Length:2920        Min.   :36.30
##  Class :character   1st Qu.:2004   Class :character   1st Qu.:63.00
##  Mode  :character   Median :2008   Mode  :character   Median :72.10
##                     Mean   :2008                      Mean   :69.23
##                     3rd Qu.:2012                      3rd Qu.:75.70
##                     Max.   :2015                      Max.   :89.00
##                                                       NA's   :8
##  Adult.Mortality infant.deaths        Alcohol       percentage.expenditure
##  Min.   :  1.0   Min.   :   0.00   Min.   : 0.010   Min.   :    0.000
##  1st Qu.: 73.0   1st Qu.:   0.00   1st Qu.: 0.870   1st Qu.:    5.348
##  Median :143.5   Median :   3.00   Median : 3.790   Median :   67.338
##  Mean   :164.8   Mean   :  30.44   Mean   : 4.615   Mean   :  742.802
##  3rd Qu.:228.0   3rd Qu.:  22.00   3rd Qu.: 7.745   3rd Qu.:  445.924
##  Max.   :723.0   Max.   :1800.00   Max.   :17.870   Max.   :19479.912
##  NA's   :8                         NA's   :193
##   Hepatitis.B       Measles              BMI        under.five.deaths
##  Min.   : 1.00   Min.   :     0.0   Min.   : 1.00   Min.   :   0.00
##  1st Qu.:77.00   1st Qu.:     0.0   1st Qu.:19.30   1st Qu.:   0.00
##  Median :92.00   Median :    17.0   Median :43.75   Median :   4.00
##  Mean   :80.88   Mean   :  2433.3   Mean   :38.36   Mean   :  42.22
##  3rd Qu.:97.00   3rd Qu.:   364.8   3rd Qu.:56.20   3rd Qu.:  28.00
##  Max.   :99.00   Max.   :212183.0   Max.   :87.30   Max.   :2500.00
##  NA's   :550                        NA's   :34
##      Polio       Total.expenditure   Diphtheria       HIV.AIDS
##  Min.   : 3.00   Min.   : 0.370    Min.   : 2.00   Min.   : 0.100
##  1st Qu.:78.00   1st Qu.: 4.260    1st Qu.:78.00   1st Qu.: 0.100
##  Median :93.00   Median : 5.755    Median :93.00   Median : 0.100
##  Mean   :82.45   Mean   : 5.939    Mean   :82.31   Mean   : 1.752
##  3rd Qu.:97.00   3rd Qu.: 7.497    3rd Qu.:97.00   3rd Qu.: 0.800
##  Max.   :99.00   Max.   :17.600    Max.   :99.00   Max.   :50.600
##  NA's   :19      NA's   :210       NA's   :19
##  thinness..1.19.years thinness.5.9.years Income.composition.of.resources
##  Min.   : 0.10        Min.   : 0.10      Min.   :0.0000
##  1st Qu.: 1.60        1st Qu.: 1.50      1st Qu.:0.4930
##  Median : 3.30        Median : 3.30      Median :0.6770
##  Mean   : 4.84        Mean   : 4.87      Mean   :0.6276
##  3rd Qu.: 7.20        3rd Qu.: 7.20      3rd Qu.:0.7790
##  Max.   :27.70        Max.   :28.60      Max.   :0.9480
##  NA's   :34           NA's   :34         NA's   :149
##    Schooling       Population             GDP
##  Min.   : 0.00   Min.   :1.069e+04   Min.   :    77.4
##  1st Qu.:10.10   1st Qu.:2.177e+06   1st Qu.:  1069.8
##  Median :12.30   Median :8.086e+06   Median :  3655.0
##  Mean   :11.99   Mean   :3.671e+07   Mean   : 10873.2
##  3rd Qu.:14.30   3rd Qu.:2.496e+07   3rd Qu.: 12257.4
##  Max.   :20.70   Max.   :1.380e+09   Max.   :185055.5
##  NA's   :145
```

Get column names.

```
colnames(expectancyData4)
```

```
##  [1] "Country"                         "Year"
##  [3] "Status"                          "Life.expectancy"
##  [5] "Adult.Mortality"                 "infant.deaths"
##  [7] "Alcohol"                         "percentage.expenditure"
##  [9] "Hepatitis.B"                     "Measles"
## [11] "BMI"                             "under.five.deaths"
## [13] "Polio"                           "Total.expenditure"
## [15] "Diphtheria"                      "HIV.AIDS"
## [17] "thinness..1.19.years"            "thinness.5.9.years"
## [19] "Income.composition.of.resources" "Schooling"
## [21] "Population"                      "GDP"
```

For each country, replace missing values within each attribute by its median (calculated from available data).

```
expectancyData5 <- expectancyData4

median_replaceCols <- c('Life.expectancy', 'Adult.Mortality', 'Alcohol', 'Hepatitis.B', 'BMI',
                        'Polio', 'Total.expenditure', 'Diphtheria', 'thinness..1.19.years',
                        'thinness.5.9.years', 'Income.composition.of.resources', 'Schooling')

expectancyData5 <- expectancyData5 %>%
  group_by(Country) %>%
  mutate_at(vars(all_of(median_replaceCols)), ~ ifelse(is.na(.), median(., na.rm = TRUE), .)) %>%
  ungroup()
head(expectancyData5)
```

```
## # A tibble: 6 x 22
##   Country      Year Status Life.expectancy Adult.Mortality infant.deaths Alcohol
##   <chr>       <dbl> <chr>            <dbl>           <int>         <int>   <dbl>
## 1 Afghanistan  2015 Devel~            65               263            62    0.01
## 2 Afghanistan  2014 Devel~            59.9             271            64    0.01
## 3 Afghanistan  2013 Devel~            59.9             268            66    0.01
## 4 Afghanistan  2012 Devel~            59.5             272            69    0.01
## 5 Afghanistan  2011 Devel~            59.2             275            71    0.01
## 6 Afghanistan  2010 Devel~            58.8             279            74    0.01
## # i 15 more variables: percentage.expenditure <dbl>, Hepatitis.B <dbl>,
## #   Measles <int>, BMI <dbl>, under.five.deaths <int>, Polio <dbl>,
## #   Total.expenditure <dbl>, Diphtheria <dbl>, HIV.AIDS <dbl>,
## #   thinness..1.19.years <dbl>, thinness.5.9.years <dbl>,
## #   Income.composition.of.resources <dbl>, Schooling <dbl>, Population <dbl>,
## #   GDP <dbl>
```

Using expectancyData5 moving forward. Still have 589 missing values.

```
sum(is.na(expectancyData5))
```

```
## [1] 589
```

Get summary to see which column still have missing values.

```
summary(expectancyData5)
```

```
##    Country               Year         Status          Life.expectancy
##  Length:2920        Min.   :2000   Length:2920        Min.   :36.30  
##  Class :character   1st Qu.:2004   Class :character   1st Qu.:63.00  
##  Mode  :character   Median :2008   Mode  :character   Median :72.10  
##                     Mean   :2008                      Mean   :69.23  
##                     3rd Qu.:2012                      3rd Qu.:75.70  
##                     Max.   :2015                      Max.   :89.00  
##                                                       NA’s   :8      
##  Adult.Mortality infant.deaths        Alcohol       percentage.expenditure
##  Min.   :  1.0   Min.   :   0.00   Min.   : 0.010   Min.   :    0.000     
##  1st Qu.: 73.0   1st Qu.:   0.00   1st Qu.: 0.900   1st Qu.:    5.348     
##  Median :143.5   Median :   3.00   Median : 3.810   Median :   67.338     
##  Mean   :164.8   Mean   :  30.44   Mean   : 4.622   Mean   :  742.802     
##  3rd Qu.:228.0   3rd Qu.:  22.00   3rd Qu.: 7.745   3rd Qu.:  445.924     
##  Max.   :723.0   Max.   :1800.00   Max.   :17.870   Max.   :19479.912     
##  NA’s   :8                         NA’s   :17                             
##   Hepatitis.B       Measles              BMI        under.five.deaths
##  Min.   : 1.00   Min.   :     0.0   Min.   : 1.00   Min.   :   0.00  
##  1st Qu.:73.00   1st Qu.:     0.0   1st Qu.:19.30   1st Qu.:   0.00  
##  Median :91.00   Median :    17.0   Median :43.75   Median :   4.00  
##  Mean   :79.47   Mean   :  2433.3   Mean   :38.36   Mean   :  42.22  
##  3rd Qu.:96.00   3rd Qu.:   364.8   3rd Qu.:56.20   3rd Qu.:  28.00  
##  Max.   :99.00   Max.   :212183.0   Max.   :87.30   Max.   :2500.00  
##  NA’s   :144                        NA’s   :34                        
##      Polio       Total.expenditure   Diphtheria       HIV.AIDS     
##  Min.   : 3.00   Min.   : 0.370    Min.   : 2.00   Min.   : 0.100  
##  1st Qu.:77.00   1st Qu.: 4.245    1st Qu.:78.00   1st Qu.: 0.100  
##  Median :93.00   Median : 5.730    Median :93.00   Median : 0.100  
##  Mean   :82.32   Mean   : 5.920    Mean   :82.18   Mean   : 1.752  
##  3rd Qu.:97.00   3rd Qu.: 7.470    3rd Qu.:97.00   3rd Qu.: 0.800  
##  Max.   :99.00   Max.   :17.600    Max.   :99.00   Max.   :50.600  
##                  NA's   :16                                        
##  thinness..1.19.years thinness.5.9.years Income.composition.of.resources
##  Min.   : 0.10        Min.   : 0.10      Min.   :0.0000                 
##  1st Qu.: 1.60        1st Qu.: 1.50      1st Qu.:0.4930                 
##  Median : 3.30        Median : 3.30      Median :0.6770                 
##  Mean   : 4.84        Mean   : 4.87      Mean   :0.6276                 
##  3rd Qu.: 7.20        3rd Qu.: 7.20      3rd Qu.:0.7790                 
##  Max.   :27.70        Max.   :28.60      Max.   :0.9480                 
##  NA's   :34           NA's   :34         NA's   :149                    
##    Schooling       Population             GDP          
##  Min.   : 0.00   Min.   :1.069e+04   Min.   :    77.4  
##  1st Qu.:10.10   1st Qu.:2.177e+06   1st Qu.:  1069.8  
##  Median :12.30   Median :8.086e+06   Median :  3655.0  
##  Mean   :11.99   Mean   :3.671e+07   Mean   : 10873.2  
##  3rd Qu.:14.30   3rd Qu.:2.496e+07   3rd Qu.: 12257.4  
##  Max.   :20.70   Max.   :1.380e+09   Max.   :185055.5  
##  NA's   :145
```

Take a look at missing values of 'Hepatitis.B'.

```
expectancyData5[is.na(expectancyData5$Hepatitis.B), ]
```

```
## # A tibble: 144 x 22
##    Country  Year Status    Life.expectancy Adult.Mortality infant.deaths Alcohol
##    <chr>   <dbl> <chr>               <dbl>           <int>         <int>   <dbl>
##  1 Denmark  2015 Developed            86                71             0   11.0 
##  2 Denmark  2014 Developed            84                73             0    9.64
##  3 Denmark  2013 Developed            81                75             0    9.5 
##  4 Denmark  2012 Developed            80                76             0    9.26
##  5 Denmark  2011 Developed            79.7              79             0   10.5 
##  6 Denmark  2010 Developed            79.2              84             0   10.3 
##  7 Denmark  2009 Developed            78.9              86             0   10.1 
##  8 Denmark  2008 Developed            78.8              88             0   10.7 
##  9 Denmark  2007 Developed            78.4              93             0   11.0 
## 10 Denmark  2006 Developed            78.1              93             0   11.0 
## # i 134 more rows
## # i 15 more variables: percentage.expenditure <dbl>, Hepatitis.B <dbl>,
## #   Measles <int>, BMI <dbl>, under.five.deaths <int>, Polio <dbl>,
## #   Total.expenditure <dbl>, Diphtheria <dbl>, HIV.AIDS <dbl>,
## #   thinness..1.19.years <dbl>, thinness.5.9.years <dbl>,
## #   Income.composition.of.resources <dbl>, Schooling <dbl>, Population <dbl>,
## #   GDP <dbl>
```

In this case we have some country doesn't have complete data for certain attribute. Apply Listwise deletion, assign to expectancyData6. Final data have 2608 observations, 22 columns. Remove 11.23% data from the original data set.

```
expectancyData6 <- na.omit(expectancyData5)
dim(expectancyData6)
```

```
## [1] 2608   22
```

Double confirmation missing values: None.

```
sum(is.na(expectancyData6))
```

```
## [1] 0
```

Export expectancyData6 to a csv file.

```
#write.csv(expectancyData6, file = "Final_Life ExpectancyData.csv", row.names = FALSE)
```

Homework 3 - Part 1 - Logistic regression.

Research question: Identify strong determinants that can distinguish between developing and developed countries? Apply logistic regression model to find these determinants, given developed country coded as '1' and developing country coded as '0'.

Get data to work with: 2608 observations, 22 attributes.

```
head(expectancyData6)
```

```
## # A tibble: 6 x 22
##   Country      Year Status Life.expectancy Adult.Mortality infant.deaths Alcohol
##   <chr>       <dbl> <chr>            <dbl>           <int>         <int>   <dbl>
## 1 Afghanistan  2015 Devel~            65               263            62    0.01
## 2 Afghanistan  2014 Devel~            59.9             271            64    0.01
## 3 Afghanistan  2013 Devel~            59.9             268            66    0.01
## 4 Afghanistan  2012 Devel~            59.5             272            69    0.01
## 5 Afghanistan  2011 Devel~            59.2             275            71    0.01
## 6 Afghanistan  2010 Devel~            58.8             279            74    0.01
## # i 15 more variables: percentage.expenditure <dbl>, Hepatitis.B <dbl>,
## #   Measles <int>, BMI <dbl>, under.five.deaths <int>, Polio <dbl>,
## #   Total.expenditure <dbl>, Diphtheria <dbl>, HIV.AIDS <dbl>,
## #   thinness..1.19.years <dbl>, thinness.5.9.years <dbl>,
## #   Income.composition.of.resources <dbl>, Schooling <dbl>, Population <dbl>,
## #   GDP <dbl>
```

Remove 'Country' and 'Year' attribute. Not relevent to the analysis.

```
expectancyData7 <- subset(expectancyData6, select = -c(Country, Year))
head(expectancyData7)
```

```
## # A tibble: 6 x 20
##   Status     Life.expectancy Adult.Mortality infant.deaths Alcohol
##   <chr>                <dbl>           <int>         <int>   <dbl>
## 1 Developing            65               263            62    0.01
## 2 Developing            59.9             271            64    0.01
## 3 Developing            59.9             268            66    0.01
## 4 Developing            59.5             272            69    0.01
## 5 Developing            59.2             275            71    0.01
## 6 Developing            58.8             279            74    0.01
## # i 15 more variables: percentage.expenditure <dbl>, Hepatitis.B <dbl>,
## #   Measles <int>, BMI <dbl>, under.five.deaths <int>, Polio <dbl>,
## #   Total.expenditure <dbl>, Diphtheria <dbl>, HIV.AIDS <dbl>,
## #   thinness..1.19.years <dbl>, thinness.5.9.years <dbl>,
## #   Income.composition.of.resources <dbl>, Schooling <dbl>, Population <dbl>,
## #   GDP <dbl>
```

Correlation.

Get pairs of high correlation, greater than |0.7|. There are 14 pairs of high correlation, we will use this to exclude attributes from the model later.

```
numData = expectancyData7 %>% select_if(is.numeric)
numData_corr <- cor(numData, method = "spearman")
high_corr_pairs <- which(abs(numData_corr) > 0.7, arr.ind = TRUE)
high_corr_pairs <- high_corr_pairs[high_corr_pairs[, 1] < high_corr_pairs[, 2], ]
# Extract the column names from the matrix index
high_corr_names <- data.frame(attr1 = rownames(numData_corr)[high_corr_pairs[, 1]],
                              attr2 = colnames(numData_corr)[high_corr_pairs[, 2]],
                              correlation = numData_corr[high_corr_pairs])

high_corr_names
```

```
##                              attr1                           attr2 correlation
## 1                    infant.deaths               under.five.deaths   0.9931559
## 2                      Hepatitis.B                           Polio   0.7579417
## 3                      Hepatitis.B                      Diphtheria   0.7811522
## 4                            Polio                      Diphtheria   0.9313965
## 5                  Life.expectancy                        HIV.AIDS  -0.7413106
## 6             thinness..1.19.years              thinness.5.9.years   0.9405629
## 7                  Life.expectancy Income.composition.of.resources   0.8490504
## 8                  Life.expectancy                       Schooling   0.7930403
## 9  Income.composition.of.resources                       Schooling   0.8882958
## 10                   infant.deaths                      Population   0.7612081
## 11               under.five.deaths                      Population   0.7513259
## 12                 Life.expectancy                             GDP   0.8057295
## 13 Income.composition.of.resources                             GDP   0.8825881
## 14                       Schooling                             GDP   0.8149533
```

Eigenvectors of Correlation Matrix.
We are also using eigenvectors to detect multi-collinearity. Distinct difference between the largest and smallest eigenvalues. Hence, at this point we can further assume there is multi-collinearity.

```r
eigenvaluesCorr <- eigen(numData_corr)$values
eigenvaluesCorr
```

```
##  [1] 8.47960609 2.17358023 2.11540474 1.17728720 0.99967273 0.75658865
##  [7] 0.65886782 0.56249007 0.49760353 0.43608461 0.30534053 0.25977653
## [13] 0.17564061 0.12709280 0.07956803 0.06789888 0.06411138 0.05716517
## [19] 0.00622043
```

```r
smallest_eigenvalue <- min(eigenvaluesCorr)
largest_eigenvalue <- max(eigenvaluesCorr)

print(paste("Smallest Eigenvalue:", smallest_eigenvalue))
```

```
## [1] "Smallest Eigenvalue: 0.00622043016611957"
```

```r
print(paste("Largest Eigenvalue:", largest_eigenvalue))
```

```
## [1] "Largest Eigenvalue: 8.47960608806205"
```

Predicted value.

Predicted value: Convert ‘Status’ column to dummy variable of 0 and 1.
‘1’ coded as developed and ‘0’ coded as developing.

```r
expectancyData8 <- expectancyData7 %>% mutate(Developed.Country = ifelse(Status == "Developing", 0, 1))
expectancyData8 <- expectancyData8 %>% select(-Status)
head(expectancyData8)
```

```
## # A tibble: 6 x 20
##   Life.expectancy Adult.Mortality infant.deaths Alcohol percentage.expenditure
##             <dbl>           <int>         <int>   <dbl>                  <dbl>
## 1            65               263            62    0.01                  71.3
## 2            59.9             271            64    0.01                  73.5
## 3            59.9             268            66    0.01                  73.2
## 4            59.5             272            69    0.01                  78.2
## 5            59.2             275            71    0.01                   7.10
## 6            58.8             279            74    0.01                  79.7
## # i 15 more variables: Hepatitis.B <dbl>, Measles <int>, BMI <dbl>,
## #   under.five.deaths <int>, Polio <dbl>, Total.expenditure <dbl>,
## #   Diphtheria <dbl>, HIV.AIDS <dbl>, thinness..1.19.years <dbl>,
## #   thinness.5.9.years <dbl>, Income.composition.of.resources <dbl>,
## #   Schooling <dbl>, Population <dbl>, GDP <dbl>, Developed.Country <dbl>
```

Distribution visualization.

Life expectancy distribution between developed and developing countries. On average, developed country has higher life expectancy of 78.94 years compared to developing country of 67.4 years. Correspondingly, majority of people in developing country have life expectancy between 62 and 74 years, while higher for developed countries from 76 to 81.5 years. This perhaps associated to better social-economic factors in developed countries.

```r
lifeExpect_stat <- expectancyData8 %>% group_by(Developed.Country) %>%
  summarise(Min_lifeExpectancy = min(Life.expectancy),
            Avg_lifeExpectancy = mean(Life.expectancy),
            Max_lifeExpectancy = max(Life.expectancy),
            Q1_lifeExpectancy = quantile(Life.expectancy, 0.25),
            Q3_lifeExpectancy = quantile(Life.expectancy, 0.75))
lifeExpect_stat
```

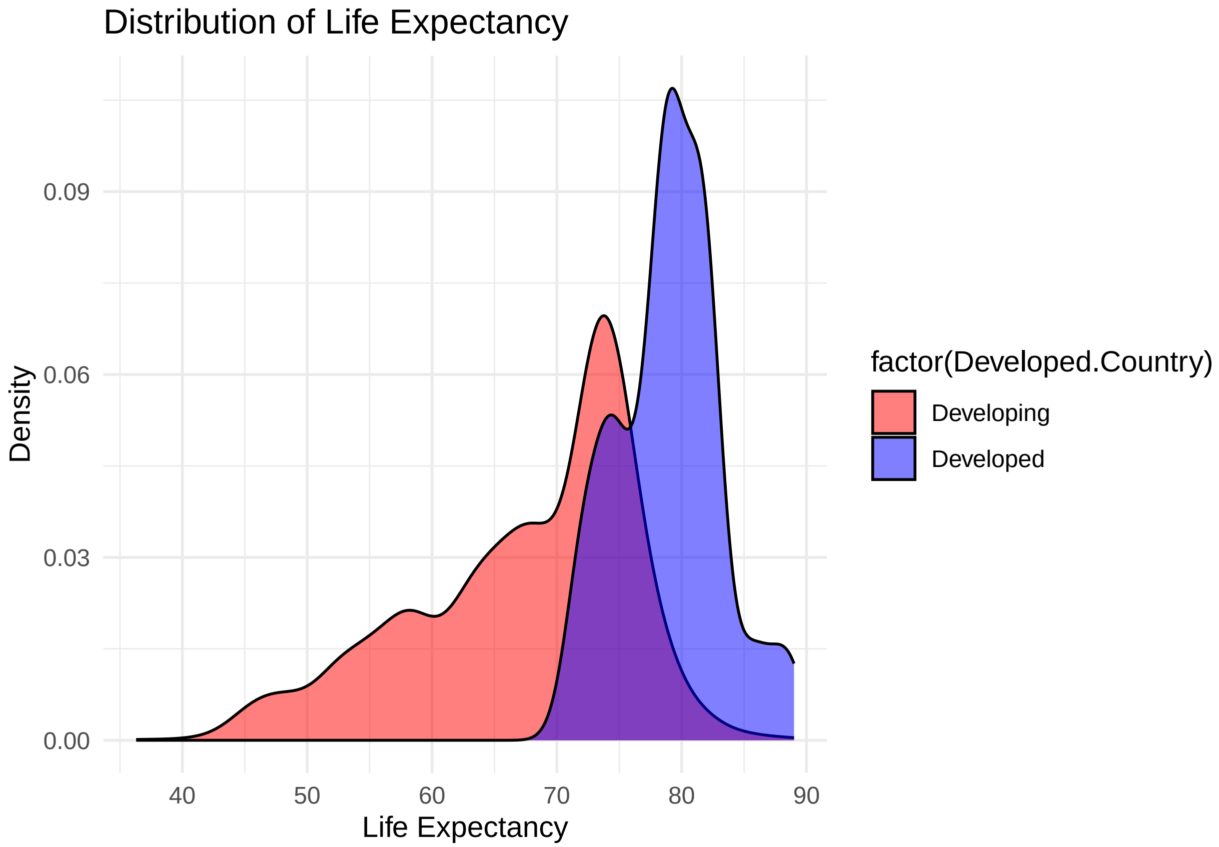
```
## # A tibble: 2 x 6
##   Developed.Country Min_lifeExpectancy Avg_lifeExpectancy Max_lifeExpectancy
##               <dbl>              <dbl>              <dbl>              <dbl>
## 1                 0               36.3               67.4                 89
## 2                 1               69.9               78.9                 89
## # i 2 more variables: Q1_lifeExpectancy <dbl>, Q3_lifeExpectancy <dbl>
```

Developed country has higher life expectancy than developing country. Especially, minimum life expectancy in developing country is significantly small compared to developed countries: 36.3 and 69.9 years, respectively. This is a good topic to explore further, perhaps breakdown data to continents to explore further.

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

```r
ggplot(expectancyData8, aes(x = Life.expectancy, fill = factor(Developed.Country))) +
  geom_density(alpha = 0.5) +
  labs(title = "Distribution of Life Expectancy", x = "Life Expectancy", y = "Density") +
  scale_fill_manual(values = c("0" = "red", "1" = "blue"), labels = c("Developing", "Developed")) +
  theme_minimal()
```



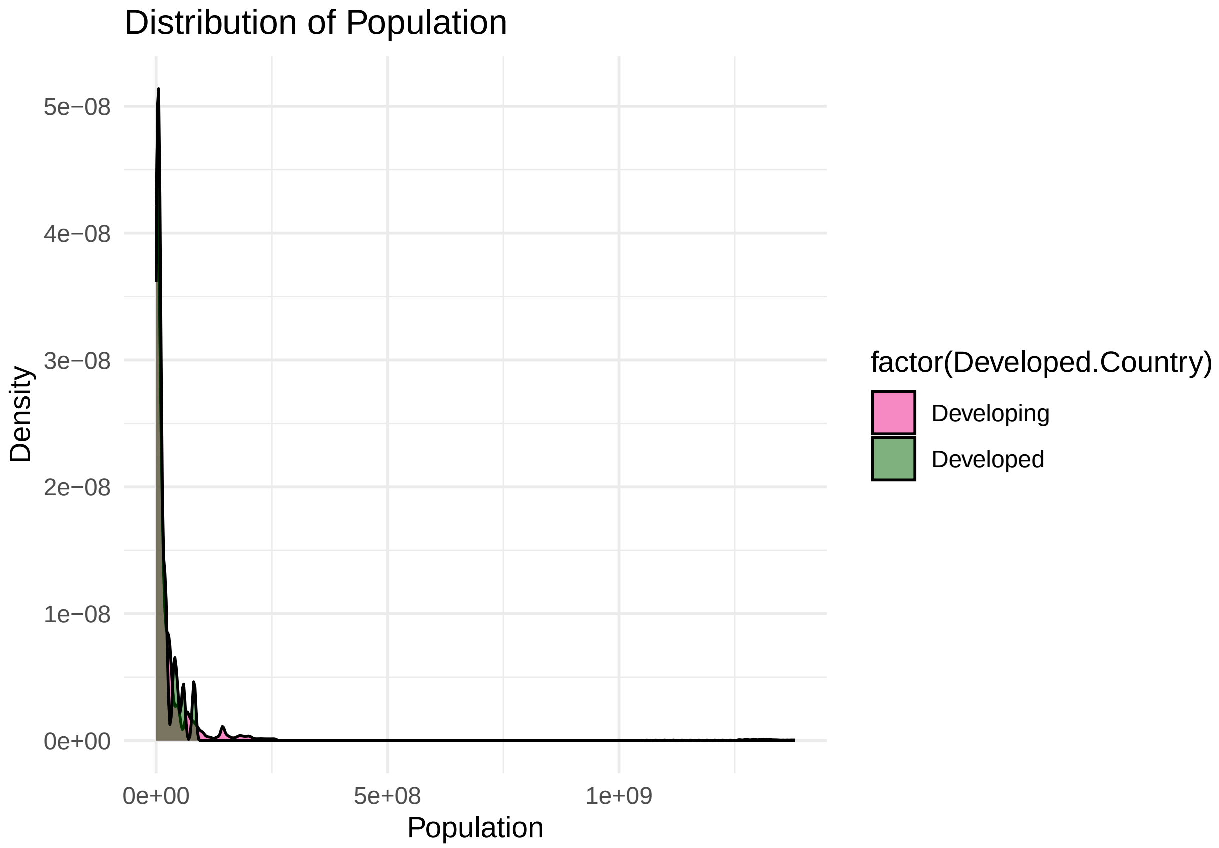
Population distribution between developed and developing countries. Developing country China has significantly highest maximum population of 1,379,860,000, compared to developed country Germany with 82,534,176 people. The average population of developing country also 2.5% higher than developed country. Noticeably, majority of developing country has population between 1.8 to 25 millions, while it is more controlled for developed country of 4 to 20 millions.

```
population_stat <- expectancyData8 %>% group_by(Developed.Country) %>%
  summarise(Min_population = min(Population),
            Avg_population = mean(Population),
            Max_population = max(Population),
            Q1_population = quantile(Population, 0.25),
            Q3_population = quantile(Population, 0.75))
population_stat
```

```
## # A tibble: 2 x 6
##   Developed.Country Min_population Avg_population Max_population Q1_population
##               <dbl>          <dbl>          <dbl>          <dbl>         <dbl>
## 1                 0          75055      39955286.     1379860000      1820192.
## 2                 1         390087      16287028.       82534176      4027715.
## # i 1 more variable: Q3_population <dbl>
```

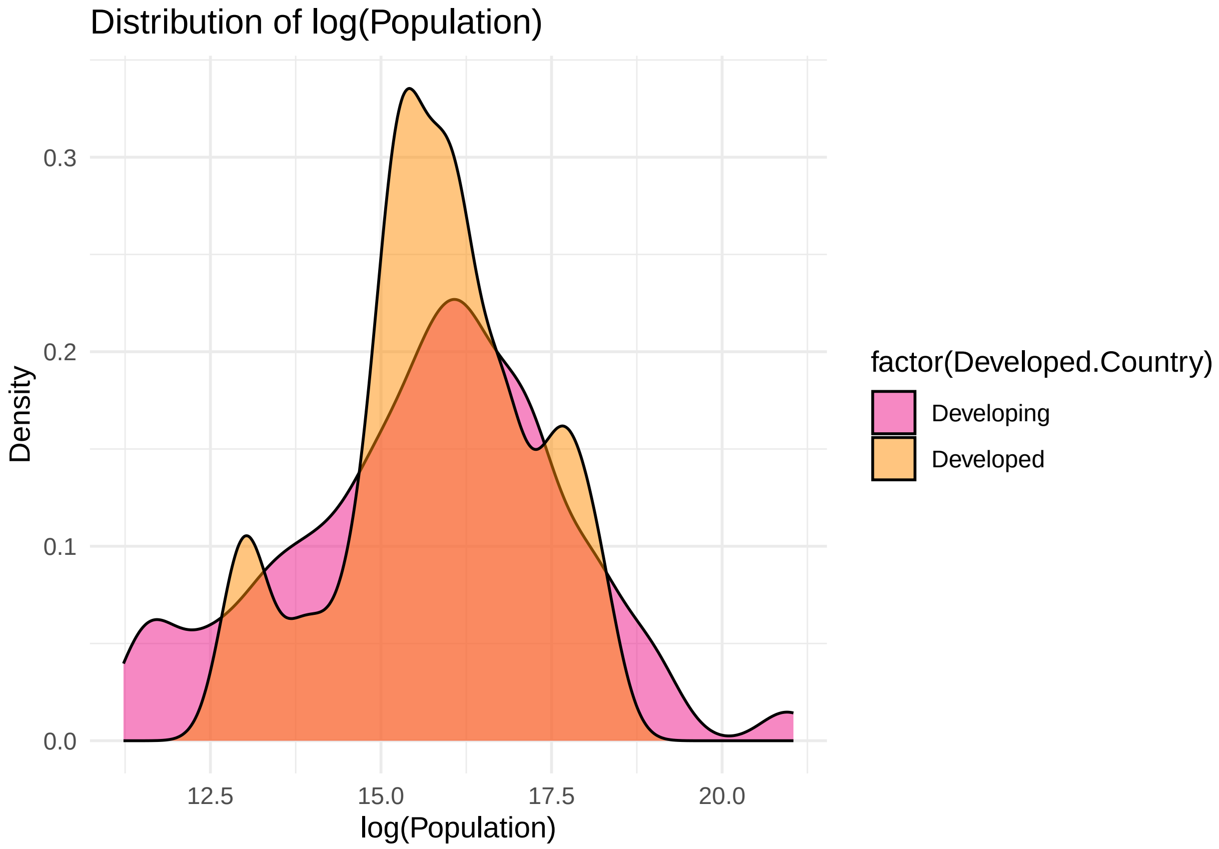
Population distribution is very right skewed. We will need to apply transformation.

```
ggplot(expectancyData8, aes(x = Population, fill = factor(Developed.Country))) +
  geom_density(alpha = 0.5) +
  labs(title = "Distribution of Population", x = "Population", y = "Density") +
  scale_fill_manual(values = c("0" = "deeppink2", "1" = "darkgreen"), labels = c("Developing", "Develope
  theme_minimal()
```



Apply log transformation on population. The population distribution looks much better now, we will using log(Population) for the model.

```
expectancyData8$logPopulation <- log(expectancyData8$Population)
ggplot(expectancyData8, aes(x = logPopulation, fill = factor(Developed.Country))) +
  geom_density(alpha = 0.5) +
  labs(title = "Distribution of log(Population)", x = "log(Population)", y = "Density") +
  scale_fill_manual(values = c("0" = "deeppink2", "1" = "darkorange"), labels = c("Developing", "Develop
  theme_minimal()
```



HIV.AIDS distribution between developed and developing countries. All developed country has consistent low - same number of HIV.AIDS death: 1 person per 1,000 live births. While on average in developing country is 2 person, maximum is 50 person. This distinguish stats will cause perfect separation.

```
HIV.AIDS_stat <- expectancyData8 %>% group_by(Developed.Country) %>%
  summarise(Min_HIV.AIDS = min(HIV.AIDS),
            Avg_HIV.AIDS = mean(HIV.AIDS),
            Max_HIV.AIDS = max(HIV.AIDS),
            Q1_HIV.AIDS = quantile(HIV.AIDS, 0.25),
            Q3_HIV.AIDS = quantile(HIV.AIDS, 0.75))
HIV.AIDS_stat
```

```
## # A tibble: 2 x 6
##   Developed.Country Min_HIV.AIDS Avg_HIV.AIDS Max_HIV.AIDS Q1_HIV.AIDS
##               <dbl>        <dbl>        <dbl>        <dbl>       <dbl>
## 1                 0          0.1         2.11         50.6         0.1
## 2                 1          0.1         0.1           0.1         0.1
## # i 1 more variable: Q3_HIV.AIDS <dbl>
```

Infant deaths distribution between developed and developing countries. Developed countries has significantly low number of infant deaths compared to developing countries, given maximum values of 4 and 1,800 deaths, respectively. This woud cause perfect seperation.

```
infantDeaths_stat <- expectancyData8 %>% group_by(Developed.Country) %>%
  summarise(Min_infantDeaths = min(infant.deaths),
            Avg_infantDeaths = mean(infant.deaths),
            Max_infantDeaths = max(infant.deaths),
            Q1_infantDeaths = quantile(infant.deaths, 0.25),
            Q3_infantDeaths = quantile(infant.deaths, 0.75))
infantDeaths_stat
```

```
## # A tibble: 2 x 6
##   Developed.Country Min_infantDeaths Avg_infantDeaths Max_infantDeaths
##               <dbl>            <int>            <dbl>            <int>
## 1                 0                0           35.2               1800
## 2                 1                0            0.662                4
## # i 2 more variables: Q1_infantDeaths <dbl>, Q3_infantDeaths <dbl>
```

Train-Test split.

Data imbalance: 352 values for developed countries / 2256 values for developing countries.

Thus, balanced class ratio of 13.5% for developed countries and 86.5% for developing countries in both the testing and training data.

```
#Get predicted value data count.
table(expectancyData8$Developed.Country)
```

```
##
##    0    1
## 2256  352
```

Split data into train and test sets.

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.2.3
```

```
## Loading required package: lattice
```

```
x <- expectancyData8[, -which(names(expectancyData8) == "Developed.Country")]
y <- expectancyData8$Developed.Country
set.seed(2023)
trainIndex <- createDataPartition(y, p = 0.7, list = FALSE)
x_train <- x[trainIndex, ]
y_train <- y[trainIndex]
x_test <- x[-trainIndex, ]
y_test <- y[-trainIndex]
```

Count of classes in training and testing data. Qualified given original data ratio.
Train data: Developing (0): 86.5% | Developed (1): 13.5%
Test data: Developing (0): 86.5% | Developed (1): 13.5%

```r
train_classCounts <- table(y_train)
print(train_classCounts)
```

```
## y_train
##    0    1
## 1580  246
```

```r
test_classCounts <- table(y_test)
print(test_classCounts)
```

```
## y_test
##   0   1
## 676 106
```

First full logistic model to obtain VIF values.

Check for multi-collinearity.

```r
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 4.2.3
```

```
## Loading required package: Matrix
```

```
## Warning: package 'Matrix' was built under R version 4.2.3
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Loaded glmnet 4.1-7
```

```r
library(car)
```

```
## Warning: package 'car' was built under R version 4.2.3
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
logisticReg <- glm(y_train ~ . - Population, family = binomial(link = 'logit'), data = cbind(y_train, x
```

```
## Warning: glm.fit: algorithm did not converge
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
VIF <- vif(logisticReg)
VIF
```

```
##                 Life.expectancy                 Adult.Mortality
##                        3.738205                        1.375978
##                   infant.deaths                         Alcohol
##                       12.270711                        1.733750
##          percentage.expenditure                     Hepatitis.B
##                        1.647851                        1.401412
##                         Measles                             BMI
##                        1.071966                        1.181998
##               under.five.deaths                           Polio
##                       12.014094                        1.224763
##               Total.expenditure                      Diphtheria
##                        1.392734                        1.252556
##                        HIV.AIDS            thinness..1.19.years
##                        1.000004                        5.278964
##              thinness.5.9.years Income.composition.of.resources
##                        6.469432                        8.430720
##                       Schooling                             GDP
##                        3.149106                        4.254528
##                   logPopulation
##                        3.808583
```

From the first model we receive 2 warnings: glm.fit: algorithm did not converge, and Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred. 1st warning indicates that output parameters are not optimal (combination of the input parameters are not stable). 2nd warning indicates that one of the attributes can perfectly separate the predicted value which makes the model unreliable.
Given above VIF values and correlation scores, there is multi-collinearity issue. We also have attributes with high p-values greater than 0.05. Thus, we will start to minimize the model in terms of number of attributes.

```
summary(logisticReg)
```

```
##
## Call:
## glm(formula = y_train ~ . - Population, family = binomial(link = "logit"),
##     data = cbind(y_train, x_train))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.80182  -0.02413   0.00000   0.00000   3.02358
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -1.453e+01  1.475e+02  -0.099 0.921488
## Life.expectancy                 -4.312e-02  5.990e-02  -0.720 0.471622    
## Adult.Mortality                 -3.934e-03  2.727e-03  -1.443 0.149127    
## infant.deaths                   -4.960e-01  4.656e-01  -1.065 0.286782    
## Alcohol                          2.978e-01  4.484e-02   6.643 3.07e-11 ***
## percentage.expenditure           4.839e-05  6.747e-05   0.717 0.473227    
## Hepatitis.B                      2.775e-02  6.284e-03   4.416 1.01e-05 ***
## Measles                          1.360e-05  6.939e-05   0.196 0.844572    
## BMI                             -1.838e-02  8.536e-03  -2.153 0.031309 *  
## under.five.deaths               -6.541e-01  4.014e-01  -1.629 0.103241    
## Polio                            1.293e-02  1.103e-02   1.172 0.241296    
## Total.expenditure               -4.379e-02  5.961e-02  -0.735 0.462526    
## Diphtheria                      -7.432e-03  1.209e-02  -0.615 0.538801    
## HIV.AIDS                        -1.159e+02  1.474e+03  -0.079 0.937298    
## thinness..1.19.years            -1.134e+00  2.279e-01  -4.974 6.54e-07 ***
## thinness.5.9.years               7.754e-01  2.227e-01   3.482 0.000497 ***
## Income.composition.of.resources  2.644e+01  6.237e+00   4.239 2.24e-05 ***
## Schooling                       -2.385e-01  1.407e-01  -1.695 0.090052 .  
## GDP                             -2.727e-05  1.384e-05  -1.970 0.048805 *  
## logPopulation                    6.458e-01  1.552e-01   4.160 3.18e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 1443.50  on 1825  degrees of freedom
## Residual deviance:  402.16  on 1806  degrees of freedom
## AIC: 442.16
## 
## Number of Fisher Scoring iterations: 25
```

Smaller logistic models.

Given high count of numerical attributes. We will break down these attributes into smaller groups that represent as contributors to a country's life expectancy: Socioeconomic indicators, Health Development, Mortality, and Immunization. Run logistic regression model for each group, account multi-collinearity, and retain only significant attribute(s) based on p-value. Then we will run an accumulate logistic regression with all significant attributes.

Socioeconomic indicators: Life Expectancy, Alcohol, GDP, and Schooling are significant with P-value < 0.05

Life Expectancy - Average time a citizen of any country is expected to live(in years).
Alcohol - Alcohol, recorded per capita (15+) consumption (in litres).
BMI - Average Body Mass Index of entire population.
GDP - Gross Domestic Product per capita (in USD).
Population - Population of the country.
Schooling - Number of years of Schooling.

```
logisticReg2 <- glm(y_train ~ Life.expectancy + Alcohol + BMI + GDP + logPopulation + Schooling,
                    family = binomial(link = 'logit'),
                    data = cbind(y_train, x_train))
summary(logisticReg2)
```

```
##
## Call:
## glm(formula = y_train ~ Life.expectancy + Alcohol + BMI + GDP +
##     logPopulation + Schooling, family = binomial(link = "logit"),
##     data = cbind(y_train, x_train))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.66607  -0.21574  -0.07145  -0.01290   2.90976
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -2.458e+01  2.669e+00  -9.208  < 2e-16 ***
## Life.expectancy  2.173e-01  3.783e-02   5.743 9.29e-09 ***
## Alcohol          3.512e-01  3.338e-02  10.523  < 2e-16 ***
## BMI             -7.730e-03  6.694e-03  -1.155 0.248193
## GDP              1.957e-05  8.591e-06   2.278 0.022756 *
## logPopulation   -2.323e-02  5.721e-02  -0.406 0.684715
## Schooling        3.239e-01  8.338e-02   3.885 0.000102 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1443.50  on 1825  degrees of freedom
## Residual deviance:  589.16  on 1819  degrees of freedom
## AIC: 603.16
##
## Number of Fisher Scoring iterations: 8
```

Health Development: Thinness 10-19 years and Income composition of resources are significant with p-value < 0.05

Percentage expenditure - Expenditure on health as a percentage of GDP per capita (%).
Total expenditure - Government expenditure on health industry as a percentage of total government expenditure(%).
Thinness 10-19 years - Prevalence of thinness among children and adolescents for Age 10 to 19 (%).
Thinness 5-9 years - Prevalence of thinness among children for Age 5 to 9(%).
Income composition of resources - Human Development Index in terms of income composition of resources (index ranging from 0 to 1).
Thinness 10-19 years and 5-9 years are highly positive correlated 0.94. Remove Thinness 5-9 years due to higher VIF value of 6.47.

```r
logisticReg3 <- glm(y_train ~ percentage.expenditure + Total.expenditure + thinness..1.19.years + Incom
                    family = binomial(link = 'logit'),
                    data = cbind(y_train, x_train))
summary(logisticReg3)
```

```
##
## Call:
## glm(formula = y_train ~ percentage.expenditure + Total.expenditure +
##     thinness..1.19.years + Income.composition.of.resources, family = binomial(link = "logit"),
##     data = cbind(y_train, x_train))
```

```
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.48822  -0.24554  -0.03610  -0.00144   2.68613
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -1.921e+01  1.696e+00 -11.326  < 2e-16 ***
## percentage.expenditure         -5.783e-05  5.383e-05  -1.074    0.283
## Total.expenditure               2.885e-02  4.687e-02   0.616    0.538
## thinness..1.19.years           -4.899e-01  8.204e-02  -5.972 2.35e-09 ***
## Income.composition.of.resources  2.424e+01  2.091e+00  11.596  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1443.50  on 1825  degrees of freedom
## Residual deviance:  628.11  on 1821  degrees of freedom
## AIC: 638.11
##
## Number of Fisher Scoring iterations: 8
```

Mortality: Adult.Mortality, Under-five deaths and Measles are signficant with p-value < 0.05.

Adult Mortality - Probability of dying between 15 and 60 years per 1000 population.
Infant deaths - Number of Infant Deaths per 1000 population.
Measles - Number of reported cases per 1000 population.
Under-five deaths - Number of under-five deaths per 1000 population.
HIV/AIDS - Deaths per 1000 live births HIV/AIDS (0-4 years).
All developed country has HIV.AIDS score = 0.1, extremely small number of infant deaths, under-five deaths (< 5) compared to wide range of developing countries. This could cause perfect separation, exclude from the model.
Infant deaths and Under-five deaths are extremely positive correlated at 0.99.

```r
logisticReg4 <- glm(y_train ~ Adult.Mortality + Measles,
                    family = binomial(link = 'logit'), data = cbind(y_train, x_train))
summary(logisticReg4)
```

```
##
## Call:
## glm(formula = y_train ~ Adult.Mortality + Measles, family = binomial(link = "logit"),
##     data = cbind(y_train, x_train))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0071  -0.5653  -0.3924  -0.1632   3.2073
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -4.039e-01  1.196e-01  -3.377 0.000732 ***
## Adult.Mortality -1.094e-02  9.910e-04 -11.035  < 2e-16 ***
## Measles         -1.421e-04  4.837e-05  -2.937 0.003314 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1443.5  on 1825  degrees of freedom
## Residual deviance: 1239.2  on 1823  degrees of freedom
## AIC: 1245.2
##
## Number of Fisher Scoring iterations: 8
```

Immuniuzation: Polio is significant with p-value < 0.05.

Hepatitis B - Immunization coverage among 1-year old (%).
Polio - Immunization coverage among 1-year old (%).
Diphtheria - Immunization coverage among 1-year old (%). Polio and Diphtheria are highly positive correlated at 0.931. Exclude Diphtheria due to higher VIF value of 1.26.

```r
logisticReg5 <- glm(y_train ~ Hepatitis.B + Polio, family = binomial(link = 'logit'),
                    data = cbind(y_train, x_train))
summary(logisticReg5)
```

```
##
## Call:
## glm(formula = y_train ~ Hepatitis.B + Polio, family = binomial(link = "logit"),
##     data = cbind(y_train, x_train))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7755  -0.6524  -0.4850  -0.1685   3.9397
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.224082   0.891205  -9.228  < 2e-16 ***
## Hepatitis.B -0.001999   0.003783  -0.528    0.597
## Polio        0.072633   0.010270   7.073 1.52e-12 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1443.5  on 1825  degrees of freedom
## Residual deviance: 1319.0  on 1823  degrees of freedom
## AIC: 1325
##
## Number of Fisher Scoring iterations: 7
```

Logistic regression with significant attribute from each category.

Check for correlation between significant attributes.

```
significantAttr <- c("Life.expectancy", "Alcohol", "GDP", "Schooling", "thinness..1.19.years",
                     "Income.composition.of.resources", "Measles", "Adult.Mortality", "Polio")

significantAttr_data <- expectancyData8[, significantAttr]

significantAttr_cor <- cor(significantAttr_data, method = "spearman")
significantAttr_high_corr_pairs <- which(abs(significantAttr_cor) > 0.7, arr.ind = TRUE)
significantAttr_high_corr_pairs <- significantAttr_high_corr_pairs[significantAttr_high_corr_pairs[, 1]
significantAttr_high_corr_names <- data.frame(attr1 = rownames(significantAttr_cor)[significantAttr_hig
                                  attr2 = colnames(significantAttr_cor)[significantAttr_high_corr_pairs
                                  correlation = significantAttr_cor[significantAttr_high_corr_pairs])

significantAttr_high_corr_names
```

```
##             attr1                           attr2 correlation
## 1 Life.expectancy                             GDP   0.8057295
## 2 Life.expectancy                       Schooling   0.7930403
## 3             GDP                       Schooling   0.8149533
## 4 Life.expectancy Income.composition.of.resources   0.8490504
## 5             GDP Income.composition.of.resources   0.8825881
## 6       Schooling Income.composition.of.resources   0.8882958
```

Remove Income composition of resources and Schooling due to high correlation.

```
logisticReg6 <- glm(y_train ~ Life.expectancy + Alcohol + GDP + thinness..1.19.years + Measles + Adult.
                    family = binomial(link = 'logit'),
                    data = cbind(y_train, x_train))
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(logisticReg6)
```

```
##
## Call:
## glm(formula = y_train ~ Life.expectancy + Alcohol + GDP + thinness..1.19.years +
##     Measles + Adult.Mortality + Polio + under.five.deaths, family = binomial(link = "logit"),
##     data = cbind(y_train, x_train))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.21418  -0.16944  -0.00658   0.00000   2.99811
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -1.314e+01  3.490e+00  -3.765 0.000167 ***
## Life.expectancy       1.194e-01  4.385e-02   2.722 0.006492 **
## Alcohol               3.366e-01  3.226e-02  10.435  < 2e-16 ***
## GDP                   1.818e-05  9.356e-06   1.943 0.051963 .
## thinness..1.19.years -5.122e-01  1.062e-01  -4.822 1.42e-06 ***
## Measles               2.769e-05  6.272e-05   0.442 0.658804
## Adult.Mortality      -4.272e-03  2.155e-03  -1.982 0.047428 *
## Polio                  2.420e-02  8.866e-03   2.729 0.006354 **
## under.five.deaths    -4.127e-01  9.826e-02  -4.200 2.67e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1443.50  on 1825  degrees of freedom
## Residual deviance:  530.88  on 1817  degrees of freedom
## AIC: 548.88
##
## Number of Fisher Scoring iterations: 13
```

Final model.

Remove Measles due to insignificance, p-value > 0.05. The final model has all significant attribute with p-values less than 0.05. There is no warning about perfect separation. Life expectancy, Alcohol consumption, GDP per capita, and Polio immunization have positive relationships with the odds of a country being developed. And vice versa for Pprevalence thinness of children and adult mortality. Akaike Information Criterion (AIC) value of 592.38. Null deviance of 1443.50 compared to much lower residual deviance of 578.38, indicates that the model is a good fit for the data. Coefficient interpretation example: When life expectancy increase by one year, 1.512e-01, holdings other variable constant, the odds of a country being developed increases by a factor of exp(0.1512), or 16.3%.

```r
logisticReg7 <- glm(y_train ~ Life.expectancy + Alcohol + GDP + thinness..1.19.years + Adult.Mortality +
                    family = binomial(link = 'logit'), data = cbind(y_train, x_train))
summary(logisticReg7)
```

```
##
## Call:
## glm(formula = y_train ~ Life.expectancy + Alcohol + GDP + thinness..1.19.years +
##     Adult.Mortality + Polio, family = binomial(link = "logit"),
##     data = cbind(y_train, x_train))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.62963  -0.21564  -0.06780  -0.00771   2.99820
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -1.658e+01  3.472e+00  -4.776 1.78e-06 ***
## Life.expectancy       1.512e-01  4.282e-02   3.530 0.000416 ***
## Alcohol               3.691e-01  3.347e-02  11.028  < 2e-16 ***
## GDP                   2.352e-05  8.935e-06   2.632 0.008486 **
## thinness..1.19.years -3.372e-01  9.970e-02  -3.383 0.000718 ***
## Adult.Mortality      -5.398e-03  2.034e-03  -2.654 0.007955 **
## Polio                 2.296e-02  8.634e-03   2.660 0.007825 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1443.50  on 1825  degrees of freedom
## Residual deviance:  578.38  on 1819  degrees of freedom
## AIC: 592.38
##
## Number of Fisher Scoring iterations: 8
```

Confusion matrix and ROC curve for evaluation.

Area Under the Curve (AUC) value of 0.9761, indicates that the model performs well in classifying developed/developing countries. ROC curve very close to the upper-left corner of the plot.

```
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 4.2.3
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(caret)
predictedProbs <- predict(logisticReg7, newdata = x_test, type = "response")

rocCurve <- roc(y_test, predictedProbs)
```

```
## Setting levels: control = 0, case = 1
```

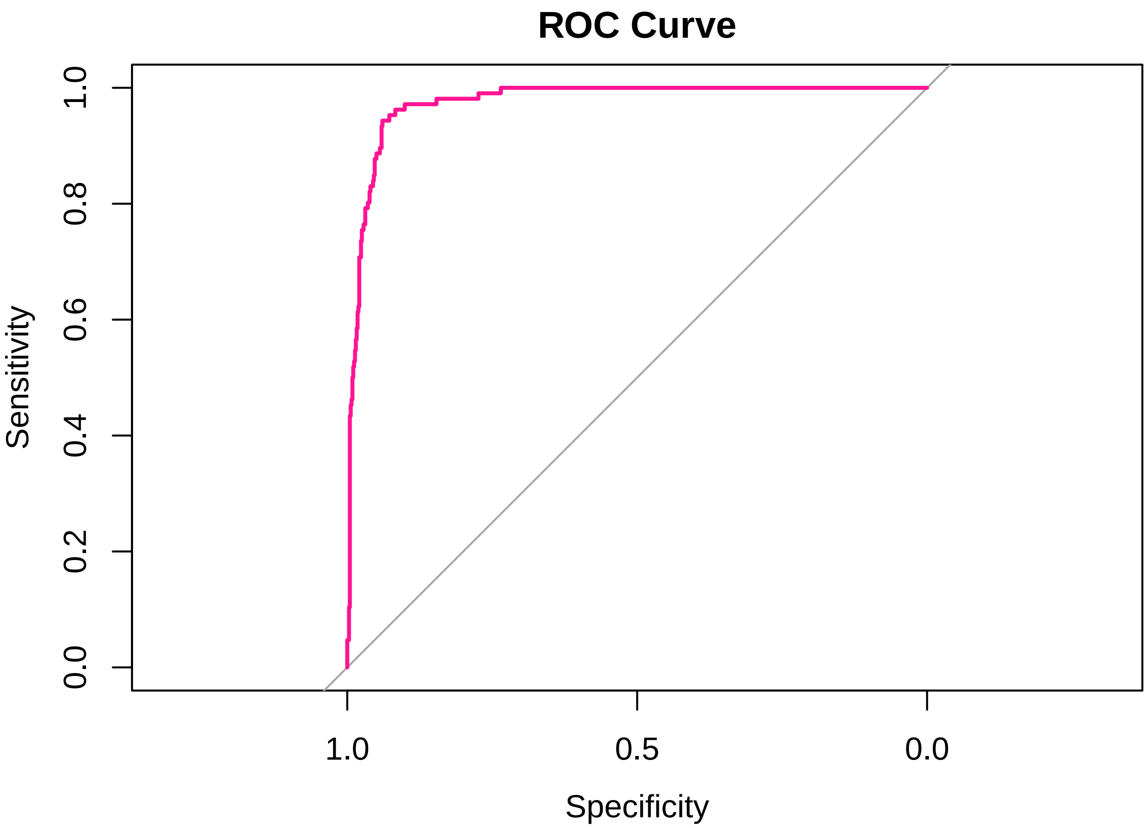
```
## Setting direction: controls < cases
```

```
aucValue <- auc(rocCurve)

print(aucValue)
```

```
## Area under the curve: 0.9761
```

```
plot(rocCurve, main = "ROC Curve", col = "deeppink1")
```



Confusion matrix: True positive - correctly predicted 652 developing countries - coded as '0'. Model accuracy is 94.25% given 95% Confidence Interval of 92.38 - 95.77. Sensitivity indicates that 96.45% of developing country were predicted as it is. Specificity indicates that 80.19% of developed countries were accurately predicted as it is. F1-score of 0.966 indicates that 96.6% of times the model makes correct predictions.

```
predictedClasses <- factor(ifelse(predictedProbs > 0.5, 1, 0), levels = c(0, 1))
y_test <- factor(y_test, levels = c(0, 1))

confusionMatrix <- confusionMatrix(data = predictedClasses, reference = y_test)
print(confusionMatrix)
```

```
## Confusion Matrix and Statistics
##
##             Reference
## Prediction   0   1
##          0 652  21
##          1  24  85
##
##                Accuracy : 0.9425
##                  95% CI : (0.9238, 0.9577)
##     No Information Rate : 0.8645
##     P-Value [Acc > NIR] : 1.224e-12
##
##                   Kappa : 0.7573
##
##  Mcnemar's Test P-Value : 0.7656
```

```
##
##             Sensitivity : 0.9645
##             Specificity : 0.8019
##          Pos Pred Value : 0.9688
##          Neg Pred Value : 0.7798
##              Prevalence : 0.8645
##          Detection Rate : 0.8338
##    Detection Prevalence : 0.8606
##       Balanced Accuracy : 0.8832
##
##        'Positive' Class : 0
##
```

```
accuracy <- confusionMatrix$overall["Accuracy"]
precision <- confusionMatrix$byClass["Pos Pred Value"]
recall <- confusionMatrix$byClass["Sensitivity"]
specificity <- confusionMatrix$byClass["Specificity"]
f1_score <- confusionMatrix$byClass["F1"]

print(paste("Precision:", precision))
```

```
## [1] "Precision: 0.968796433878158"
```

```
print(paste("F1-Score:", f1_score))
```

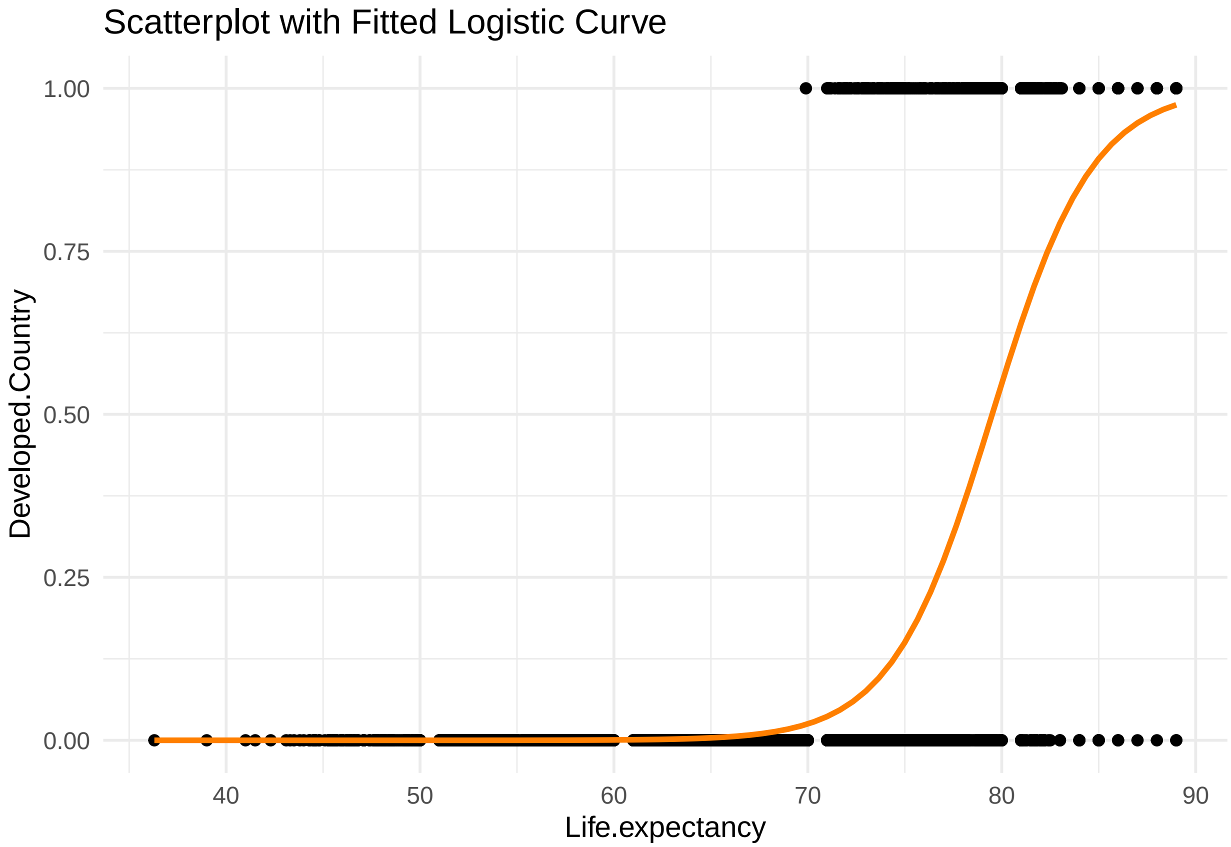
```
## [1] "F1-Score: 0.966641957005189"
```

Logistic Curve.

Given a country with life expectancy of 80 year-old, there is a 53% chance that this country is developed.

```
ggplot(expectancyData8, aes(x = Life.expectancy, y = Developed.Country)) +
  geom_point() +
  labs(title = 'Scatterplot with Fitted Logistic Curve') +
  theme_minimal() +
  geom_smooth(method = "glm", method.args = list(family = "binomial"), se = FALSE, col = "darkorange1")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Conclusion:

1) From analysis, HIV.AIDS and Infant deaths can significantly distinguish developed and developing countries.
2) From logistic model, Life expectancy, Alcohol consumption, Children Age 10 to 19 thinness, GDP per capita, Adult mortality, and Immunization coverage can determine probability of a country is developing or developed.
3) Globally, on average a person can live up to 69 year-old.

4) Developing countries has much fluctuated and wider range of population given China of over 1.3 billion people, Germany has highest population in developed country of 82.5 millions.

5) Developed country has significantly low - same number of HIV.AIDS death: 1 person per 1,000 live births. Also have low number of infant deaths of 1 to 4 with average of 0.6.
6) In terms of life expectancy, in order for a country to be classify as developed (53% chance), life expectancy needs to be at least 79 year-old.