Gapminder_dataset

Saahil Shroff

2022-12-10

Data exploration and multiple regression

What is life expectancy?

```
## Life expectancy is a statistical estimate of how long someone is predicted to live based on their birth year, present ag
e, and other demographic parameters such as gender. It is used to evaluate and determine a variety of critical policies that
have an influence on everyday living, such as setting the State Pension age and focusing health policy activities. The curre
nt life expectancy for U.S. in 2022 is 79.05 years, a 0.08% increase from 2021.
```

References -

Period and cohort life expectancy explained: December 2019 - Office for National Statistics. (n.d.). https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/methodologies/periodand
(https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/methodologies/periodan

U.S. Life Expectancy 1950-2022. (n.d.). MacroTrends. https://www.macrotrends.net/countries/USA/united-states/life-expectancy (https://www.macrotrends.net/countries/USA/united-states/life-expectancy)

Loading and cleaning the data to remove all cases with missing life expectancy, year and country name.

```
## Dimension of the loaded dataframe: 13055 25
```

```
## Any NULLs present in the dataframe? TRUE
```

```
## No. of total NULLs present in the dataframe: 103406
```

```
## No. of NULLs present in each column of the dataframe:
```

```
##                iso3                name                iso2              region 
##                   0                   0                   0                   0 
##          sub.region intermediate.region                time     totalPopulation 
##                   0                   0                  36                  76 
##       fertilityRate      lifeExpectancy      childMortality youthFemaleLiteracy 
##                1307                1325                2600               12134 
##   youthMaleLiteracy       adultLiteracy              GDP_PC   accessElectricity 
##               12134               12118                3585                7608 
##    agriculturalLand agricultureTractors     cerealProduction        fertilizerHa 
##                1910                6947                3606                4929 
##                 co2     greenhouseGases              co2_PC            pm2.5_35 
##                2658                4994                2661               10727 
##         battleDeaths 
##               12051
```

```
## Dimensions of the cleaned dataframe: 11618 25
```

Reference -

Delete rows with blank values in one particular column - https://stackoverflow.com/questions/9126840/delete-rows-with-blank-values-in-one-particular-column (https://stackoverflow.com/questions/9126840/delete-rows-with-blank-values-in-one-particular-column)

How many countries do we have in these data?

```
## There are 204 in total. They are:
```

```
##   [1] "Afghanistan"
##   [2] "Albania"
##   [3] "Algeria"
##   [4] "Angola"
##   [5] "Antigua and Barbuda"
##   [6] "Argentina"
##   [7] "Armenia"
##   [8] "Aruba"
##   [9] "Australia"
##  [10] "Austria"
##  [11] "Azerbaijan"
##  [12] "Bahamas"
##  [13] "Bahrain"
##  [14] "Bangladesh"
##  [15] "Barbados"
##  [16] "Belarus"
##  [17] "Belgium"
##  [18] "Belize"
##  [19] "Benin"
##  [20] "Bermuda"
##  [21] "Bhutan"
##  [22] "Bolivia (Plurinational State of)"
##  [23] "Bosnia and Herzegovina"
##  [24] "Botswana"
##  [25] "Brazil"
##  [26] "Brunei Darussalam"
##  [27] "Bulgaria"
##  [28] "Burkina Faso"
##  [29] "Burundi"
##  [30] "Cabo Verde"
##  [31] "Cambodia"
##  [32] "Cameroon"
##  [33] "Canada"
##  [34] "Cayman Islands"
##  [35] "Central African Republic"
##  [36] "Chad"
##  [37] "Chile"
##  [38] "China"
##  [39] "Colombia"
##  [40] "Comoros"
##  [41] "Congo"
##  [42] "Congo, Democratic Republic of the"
##  [43] "Costa Rica"
##  [44] "Côte d'Ivoire"
##  [45] "Croatia"
##  [46] "Cuba"
##  [47] "Cyprus"
##  [48] "Czechia"
##  [49] "Denmark"
##  [50] "Djibouti"
##  [51] "Dominica"
##  [52] "Dominican Republic"
##  [53] "Ecuador"
##  [54] "Egypt"
##  [55] "El Salvador"
##  [56] "Equatorial Guinea"
##  [57] "Eritrea"
##  [58] "Estonia"
##  [59] "Eswatini"
##  [60] "Ethiopia"
##  [61] "Faroe Islands"
##  [62] "Fiji"
##  [63] "Finland"
##  [64] "France"
##  [65] "French Polynesia"
##  [66] "Gabon"
##  [67] "Gambia"
##  [68] "Georgia"
##  [69] "Germany"
##  [70] "Ghana"
##  [71] "Greece"
##  [72] "Greenland"
##  [73] "Grenada"
##  [74] "Guam"
##  [75] "Guatemala"
##  [76] "Guinea"
##  [77] "Guinea-Bissau"
##  [78] "Guyana"
##  [79] "Haiti"
##  [80] "Honduras"
##  [81] "Hong Kong"
##  [82] "Hungary"
##  [83] "Iceland"
##  [84] "India"
##  [85] "Indonesia"
##  [86] "Iran (Islamic Republic of)"
##  [87] "Iraq"
##  [88] "Ireland"
##  [89] "Israel"
##  [90] "Italy"
##  [91] "Jamaica"
##  [92] "Japan"
##  [93] "Jordan"
##  [94] "Kazakhstan"
##  [95] "Kenya"
##  [96] "Kiribati"
##  [97] "Korea (Democratic People's Republic of)"
##  [98] "Korea, Republic of"
##  [99] "Kuwait"
## [100] "Kyrgyzstan"
## [101] "Lao People's Democratic Republic"
## [102] "Latvia"
## [103] "Lebanon"
## [104] "Lesotho"
## [105] "Liberia"
## [106] "Libya"
## [107] "Liechtenstein"
## [108] "Lithuania"
## [109] "Luxembourg"
## [110] "Macao"
## [111] "Madagascar"
## [112] "Malawi"
## [113] "Malaysia"
## [114] "Maldives"
## [115] "Mali"
## [116] "Malta"
## [117] "Marshall Islands"
## [118] "Mauritania"
## [119] "Mauritius"
## [120] "Mexico"
## [121] "Micronesia (Federated States of)"
## [122] "Moldova, Republic of"
## [123] "Mongolia"
## [124] "Montenegro"
## [125] "Morocco"
## [126] "Mozambique"
## [127] "Myanmar"
## [128] "Namibia"
## [129] "Nepal"
## [130] "Netherlands"
## [131] "New Caledonia"
## [132] "New Zealand"
## [133] "Nicaragua"
## [134] "Niger"
## [135] "Nigeria"
## [136] "North Macedonia"
## [137] "Norway"
## [138] "Oman"
## [139] "Pakistan"
## [140] "Palau"
## [141] "Palestine, State of"
## [142] "Panama"
## [143] "Papua New Guinea"
## [144] "Paraguay"
## [145] "Peru"
## [146] "Philippines"
## [147] "Poland"
## [148] "Portugal"
## [149] "Puerto Rico"
## [150] "Qatar"
## [151] "Romania"
## [152] "Russian Federation"
## [153] "Rwanda"
## [154] "Saint Kitts and Nevis"
## [155] "Saint Lucia"
## [156] "Saint Martin (French part)"
## [157] "Saint Vincent and the Grenadines"
## [158] "Samoa"
## [159] "San Marino"
## [160] "Sao Tome and Principe"
## [161] "Saudi Arabia"
## [162] "Senegal"
## [163] "Serbia"
## [164] "Seychelles"
## [165] "Sierra Leone"
## [166] "Singapore"
## [167] "Sint Maarten (Dutch part)"
## [168] "Slovakia"
## [169] "Slovenia"
## [170] "Solomon Islands"
## [171] "Somalia"
## [172] "South Africa"
## [173] "South Sudan"
## [174] "Spain"
## [175] "Sri Lanka"
## [176] "Sudan"
## [177] "Suriname"
## [178] "Sweden"
## [179] "Switzerland"
## [180] "Syrian Arab Republic"
## [181] "Tajikistan"
## [182] "Tanzania, United Republic of"
## [183] "Thailand"
## [184] "Timor-Leste"
## [185] "Togo"
## [186] "Tonga"
## [187] "Trinidad and Tobago"
## [188] "Tunisia"
## [189] "Turkey"
## [190] "Turkmenistan"
## [191] "Uganda"
## [192] "Ukraine"
## [193] "United Arab Emirates"
## [194] "United Kingdom of Great Britain and Northern Ireland"
## [195] "United States of America"
## [196] "Uruguay"
## [197] "Uzbekistan"
## [198] "Vanuatu"
## [199] "Venezuela (Bolivarian Republic of)"
## [200] "Viet Nam"
## [201] "Virgin Islands (U.S.)"
## [202] "Yemen"
## [203] "Zambia"
## [204] "Zimbabwe"
```

What is the first and last year with valid life expectancy data?

```
## First year with valid life expectancy data: 1960 and last year with valid life expectancy data: 2019
```

What is the lowest and highest life expectancy values? Which country/year do they correspond to?

Lowest life expectancy details:

```
##          name time lifeExpectancy
## 5642 Cambodia 1977         18.907
```

Highest life expectancy details:

```
##            name time lifeExpectancy
## 9561 San Marino 2012       85.41707
```

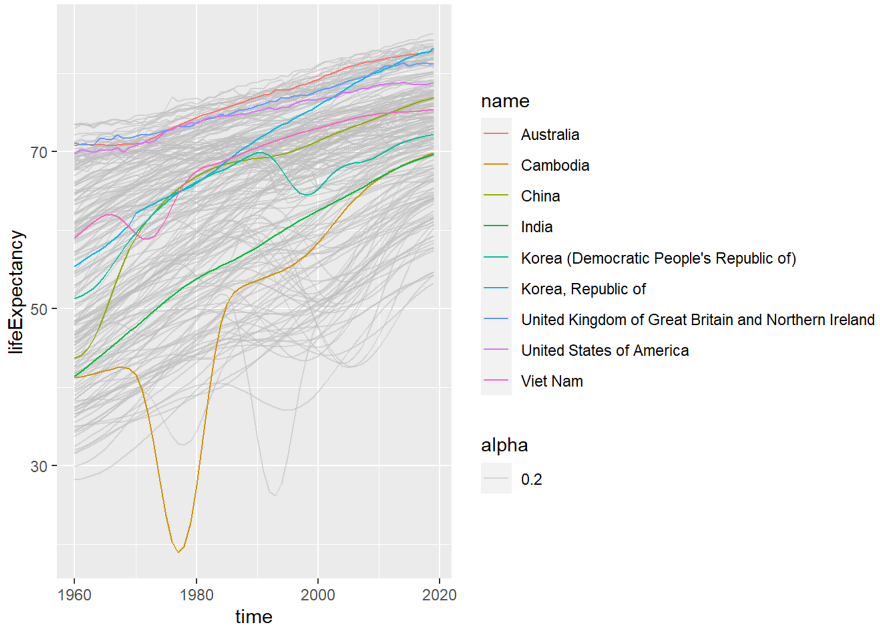
The shortest life expectancy corresponds to a well-known event. What is the event?

```
## The Cambodian Genocide, which lasted four years (from 1975 and 1979), was a wave of mass violence that murdered between
1.5 and 3 million people at the hands of the Khmer Rouge, a communist political organization. Following the Cambodian Civil
War, the Khmer Rouge seized power in the country.
## After seizing power, the Khmer Rouge initiated a dramatic overhaul of Cambodian society. This entailed the forcible reloc
ation of city people to the countryside, where they would be compelled to work as farmers, digging canals and tending crops.
Mismanagement of the country's economy resulted in food and medication shortages, and untold thousands of people died of sic
kness and famine. Families were also divided. The Khmer Rouge established work brigades, dividing them into categories based
on age and gender. Hundreds of thousands of Cambodians died as a result of this program.
```

Reference -

Cambodia. (n.d.). College of Liberal Arts. https://cla.umn.edu/chgs/holocaust-genocide-education/resource-guides/cambodia (https://cla.umn.edu/chgs/holocaust-genocide-education/resource-guides/cambodia)

Plotting the life expectancy over time for all countries



Reference - Rapp, A. (n.d.). Albert Rapp - 4 Ways to use colors in ggplot more efficiently. https://albert-rapp.de/posts/ggplot2-tips/07_four_ways_colors_more_efficiently/07_four_ways_colors_more_efficiently.html (https://albert-rapp.de/posts/ggplot2-tips/07_four_ways_colors_more_efficiently/07_four_ways_colors_more_efficiently.html)

```
## The countries taken for my analyses are - Viet Nam, India, United Kingdom (represented as United Kingdom of Great Britain
and Northern Ireland), Australia, and North Korea (represented as Korea, Republic of), Cambodia, China, United States of Ame
rica (or U.S.A), South Korea (represented as Korea (Democratic People's Republic of)). All in all, I have taken a different
mix of countries that are not only from different continents, but also can be identified as First,Second, or Third World Cou
ntries.
```

```
## First world countries - U.S.A, U.K., and Australia.
```

```
## Second world countries - China, South Korea, and Cambodia.
```

```
## Third world countries - North Korea, India, and Viet Nam.
```

References -

Wikipedia contributors. (2022, December 7). List of civil wars. Wikipedia. https://en.wikipedia.org/wiki/List_of_civil_wars (https://en.wikipedia.org/wiki/List_of_civil_wars)

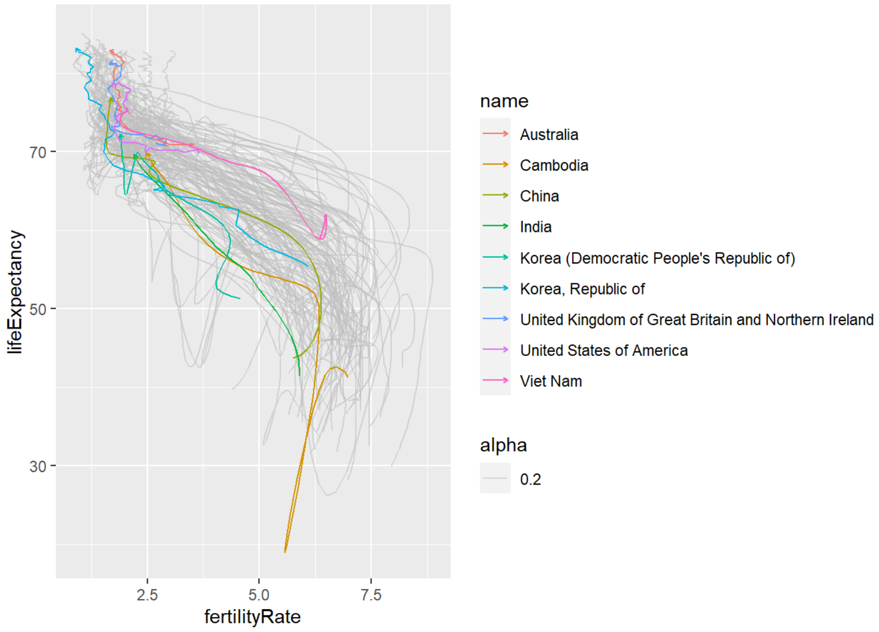
First, Second and Third World. (n.d.). http://www.hartford-hwp.com/archives/10/150.html (http://www.hartford-hwp.com/archives/10/150.html)

Inference on how selected countries behave?

```
## Looking at the plot, it seems that life expectancy of 1st world countries w.r.t time steadily increases. These countries
have had amongst the highest life expectancy from the beginning (year 1960) and have been able to keep their place in the to
p. For second world countries, we see some set backs to the life expectancy (a trough in the plot indicates this), but they
have soon been able to recover from it and have picked up pace. In case of third world countries, the growth of life expecta
ncy (seen by the steepness of the line) has been highest across the other 2 categories of countries.
```

Creating a fertility rate versus life expectancy plot of all countries with selected

countries highlighted (with arrows to mark which way the time goes)



Reference -

How to draw a nice arrow in ggplot2. (2016, June 24). Stack Overflow. https://stackoverflow.com/questions/38008863/how-to-draw-a-nice-arrow-in-ggplot2 (https://stackoverflow.com/questions/38008863/how-to-draw-a-nice-arrow-in-ggplot2)

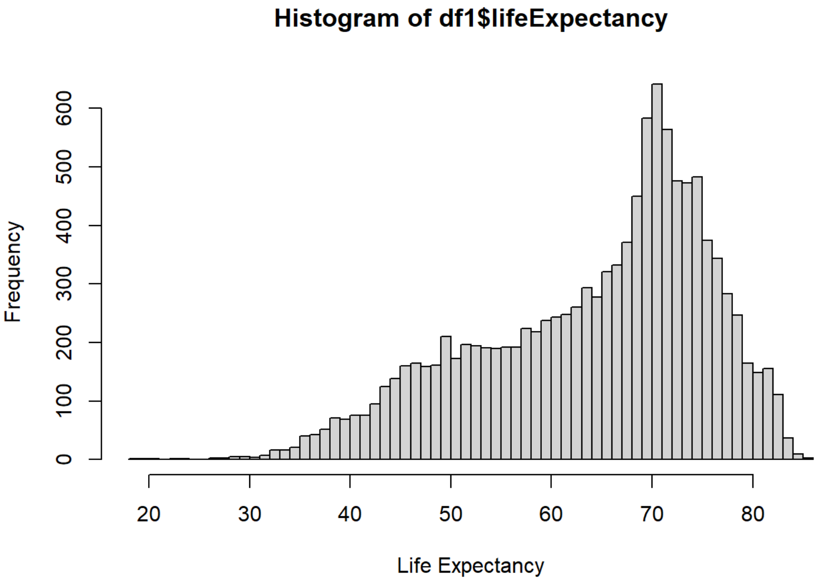
Comment on the results. Where is the world going? Where are the highlighted countries going?

```
## The above plot cements the common real-world observation. In earlier times (1960's - 2000's), the life expectancy was low
and people had larger families (indicated by higher fertility rate); one can attribute lower life expectancy to low advancem
ents in the healtcare and medical domain. However, since the 2000's, people have started having nuclear families of 3 or 4 i
ndividuals, which is shown by fertility rate going <2.5, and with the new medical equipments and better standard of living,
individuals are living for a longer time (indicated by the arrow < 70). This phenomena is observed throughout the world, thu
s in the future we can hope to see smaller families with higher life expectancy.
```

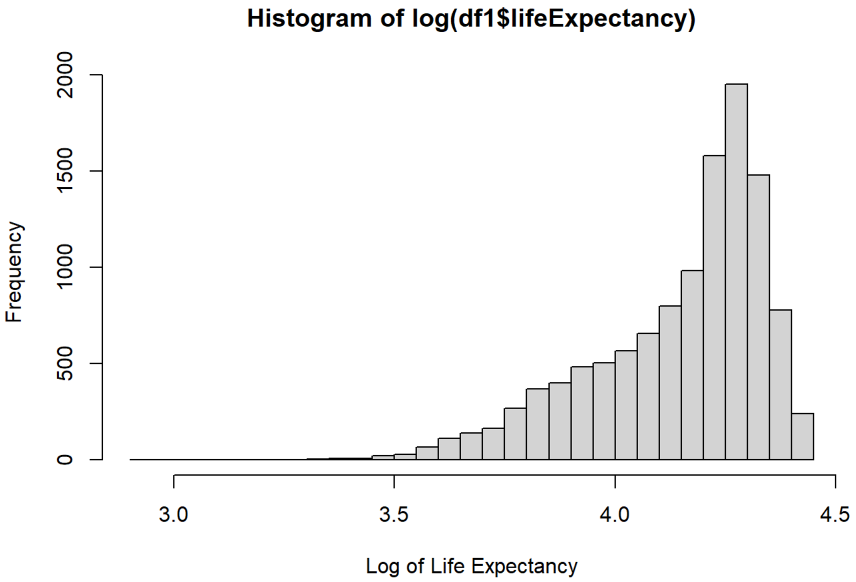
Modeling life expectancy

Distribution of life expectancy. How does it look like?

Distribution of life expectancy:



Distribution of log of life expectancy:



```
## Looking at the 1st plot - Distribution of life expectancy, I thought that log transformation would be required. While the
plot is normally distributed, it was right-skewed; hence, I thought doing a log-transformation will create the plot normally
distributed in the centre. However, when I checked the distribution of the log-transformed life expectancy variable, my assu
mption was wrong; the plot was still right-skewed. All in all, there is no difference between the 2 plots in terms of skewne
ss. As a result, I am NOT taking a log-transformation and going with the original distribution.
```

Creating a linear model between life expectancy with just time. (Using year − 2000 instead of just year for time)

```
##   iso3  name iso2   region                      sub.region intermediate.region
## 1  ABW Aruba   AW Americas Latin America and the Caribbean           Caribbean
## 2  ABW Aruba   AW Americas Latin America and the Caribbean           Caribbean
## 3  ABW Aruba   AW Americas Latin America and the Caribbean           Caribbean
## 4  ABW Aruba   AW Americas Latin America and the Caribbean           Caribbean
## 5  ABW Aruba   AW Americas Latin America and the Caribbean           Caribbean
## 6  ABW Aruba   AW Americas Latin America and the Caribbean           Caribbean
##   time totalPopulation fertilityRate lifeExpectancy childMortality
## 1 1960           54211         4.820         65.662             NA
## 2 1961           55438         4.655         66.074             NA
## 3 1962           56225         4.471         66.444             NA
## 4 1963           56695         4.271         66.787             NA
## 5 1964           57032         4.059         67.113             NA
## 6 1965           57360         3.842         67.435             NA
##   youthFemaleLiteracy youthMaleLiteracy adultLiteracy GDP_PC accessElectricity
## 1                  NA                NA            NA     NA                NA
## 2                  NA                NA            NA     NA                NA
## 3                  NA                NA            NA     NA                NA
## 4                  NA                NA            NA     NA                NA
## 5                  NA                NA            NA     NA                NA
## 6                  NA                NA            NA     NA                NA
##   agriculturalLand agricultureTractors cerealProduction fertilizerHa      co2
## 1               NA                  NA               NA           NA 11092.67
## 2               20                  NA               NA           NA 11576.72
## 3               20                  NA               NA           NA 12713.49
## 4               20                  NA               NA           NA 12178.11
## 5               20                  NA               NA           NA 11840.74
## 6               20                  NA               NA           NA 10623.30
##   greenhouseGases   co2_PC pm2.5_35 battleDeaths time_ref_2000
## 1              NA 204.6204       NA           NA           -40
## 2              NA 208.8228       NA           NA           -39
## 3              NA 226.1181       NA           NA           -38
## 4              NA 214.8004       NA           NA           -37
## 5              NA 207.6158       NA           NA           -36
## 6              NA 185.2040       NA           NA           -35
```

```
##
## Call:
## lm(formula = lifeExpectancy ~ time_ref_2000, data = year2000)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.350  -7.603   2.505   8.042  18.542
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   67.358008   0.109226  616.68   <2e-16 ***
## time_ref_2000  0.308758   0.005441   56.74   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.14 on 11616 degrees of freedom
## Multiple R-squared:  0.217,  Adjusted R-squared:  0.217
## F-statistic:  3220 on 1 and 11616 DF,  p-value: < 2.2e-16
```

```
## The process of doing (year-2000) is called mean-centering. It is done so that it changes the interpretation of the interc
ept in a very helpful way. For instance, if we do not scale/mean-centre the year, the life expectancy will come out as negat
ive, which cannot be the case. Thus, it is a required step for furhter model interpretation.
```

Reference - Lohninger, H. (n.d.). Scaling of Data. http://www.statistics4u.com/fundstat_eng/cc_scaling.html (http://www.statistics4u.com/fundstat_eng/cc_scaling.html)

Interpret the results of the model (both b0 and b1).

```
##  b0 indicates the life expectancy at year 2000 (or time = 0), while b1 indicates that with as time ahead moves by 1year,
the life expectancy in the world increases by 0.3087 years
```

Estimating the life expectancy through multiple regression model (adding continent)

```
##
## Call:
## lm(formula = lifeExpectancy ~ time_ref_2000 + region, data = year2000)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.172  -4.057   0.565   4.041  20.037
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    55.941322   0.123110  454.40   <2e-16 ***
## time_ref_2000   0.304745   0.003574   85.27   <2e-16 ***
## regionAmericas 15.872056   0.182335   87.05   <2e-16 ***
## regionAsia     12.147162   0.169536   71.65   <2e-16 ***
## regionEurope   20.831659   0.180406  115.47   <2e-16 ***
## regionOceania  13.570858   0.264889   51.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.661 on 11612 degrees of freedom
## Multiple R-squared:  0.6624, Adjusted R-squared:  0.6623
## F-statistic:  4557 on 5 and 11612 DF,  p-value: < 2.2e-16
```

Interpreting the model and evaulating this model against the previous model.

```
## All the independent variables - time_ref_200 and region - are statistically significant to predict the life expectancy. I
t can be said since each of the factors are <2e-16, which is outside the 95% confidence interval.
## The intercept represents the African-region and the time trend is for 80 years (1960 - 2020), centered at 2000.
## This model has a better R^2 value as compared to the previous model, which suggests that this is a stronger and a more re
liable model as compare to the previous one. Additionally, the difference between R^2 and adjusted R^2 is minimal, which mea
ns each independent variables actually have an effect on the performance of the model.
```

Adding two additional variables to the model: log of GDP per capita, and fertility rate.

```
##
## Call:
## lm(formula = lifeExpectancy ~ time_ref_2000 + region + log(GDP_PC) +
##     fertilityRate, data = year2000)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.292  -2.477   0.289   2.724  12.250
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    49.93572    0.50900   98.11   <2e-16 ***
## time_ref_2000   0.13778    0.00355   38.81   <2e-16 ***
## regionAmericas  6.03430    0.15968   37.79   <2e-16 ***
## regionAsia      5.84118    0.15009   38.92   <2e-16 ***
## regionEurope    5.42126    0.20713   26.17   <2e-16 ***
## regionOceania   5.75319    0.22491   25.58   <2e-16 ***
## log(GDP_PC)     2.49027    0.04699   53.00   <2e-16 ***
## fertilityRate  -2.23512    0.04635  -48.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.345 on 8970 degrees of freedom
##   (2640 observations deleted due to missingness)
## Multiple R-squared:  0.8472, Adjusted R-squared:  0.8471
## F-statistic:  7107 on 7 and 8970 DF,  p-value: < 2.2e-16
```

```
## The intercept represents the African-region and the time trend is for 80 years (1960 - 2020), centered at 2000. The new p
arameters - GDP_PC and fertilityRate - are both statistically significant as well as their p-value is <2e-16 (<0.05), which
means that it is outside the 95% confidence interval and plays a role in prediciting the dependent variable.
## This model has a better R^2 value as compared to the all the previous models, which suggests that this is a stronger and
a more reliable model as compare to both the previous ones. Additionally, the difference between R^2 and adjusted R^2 is min
imal, which means each independent variables actually have an effect on the performance of the model.
```

Additional variables made the ranking of continents to look different than the previous models.

```
## Europe was the leading region in Q5, whereas America is the leading region now when we introduce other socio-economic var
iables such as GDP & fertility rate. The major cause of the order of regions changing as we keep adding additional variables
is because every new beta value has an influence on the cohesiveness with which all the other factors are impacting the depe
ndent variable.
## Furthermore, studies have shown that higher values of GDP per capita and lower values of infant mortality levels lead to
higher life expectancy at birth suggesting that longevity of people in these five countries is increasing, which is cemented
through the above model.
```

Reference - Miladinov, G. (2020, January 10). Socioeconomic development and life expectancy relationship: evidence from the EU accession candidate countries - Genus. SpringerOpen. https://genus.springeropen.com/articles/10.1186/s41118-019-0071-0 (https://genus.springeropen.com/articles/10.1186/s41118-019-0071-0)

Which continent has the highest and lowest life expectancy?

```
## Looking at the latest model, America has the highest life expectancy, while Europe has the lowest life expectancy.
```