Life Expectancy

Debatosh Chakraborty

24/02/2022

1 Life Expectancy dataset

```
life = read.csv('life-expectancy.csv')
summary(life)
```

```
##     Entity              Code                Year      Life.expectancy
##  Length:19028       Length:19028       Min.   :1543   Min.   :17.76
##  Class :character   Class :character   1st Qu.:1961   1st Qu.:52.31
##  Mode  :character   Mode  :character   Median :1980   Median :64.71
##                                        Mean   :1975   Mean   :61.75
##                                        3rd Qu.:2000   3rd Qu.:71.98
##                                        Max.   :2019   Max.   :86.75
```

2 Introduction

2.1 RStudio Version 1.3.1093

2.2 R Version 1.0.2

```
R.version
```

```
##                _
## platform       x86_64-w64-mingw32
## arch           x86_64
## os             mingw32
## crt            ucrt
## system         x86_64, mingw32
## status
## major          4
## minor          2.1
## year           2022
## month          06
## day            23
## svn rev        82513
## language       R
## version.string R version 4.2.1 (2022-06-23 ucrt)
## nickname       Funny-Looking Kid
```

```
library(dplyr)
library(ggplot2)
library(magrittr)
library(summarytools)
```

```
library(tibble)
library(skimr)
```

3 Data Analysis Task

Let's start by checking for null or missing values

```
glimpse(life)
```

```
## Rows: 19,028
## Columns: 4
## $ Entity          <chr> "Afghanistan", "Afghanistan", "Afghanistan", "Afghanis~
## $ Code            <chr> "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", "AFG"~
## $ Year            <int> 1950, 1951, 1952, 1953, 1954, 1955, 1956, 1957, 1958, ~
## $ Life.expectancy <dbl> 27.638, 27.878, 28.361, 28.852, 29.350, 29.854, 30.365~
```

```
sum(is.na(life))
```

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

We see that there are no missing values. So let's go through the entire dataset.

We guarantee that all the columns in the dataset are in their correct format. But there are some 0 character in Code. So we want some investigation regarding their existence

```
is.null(life)
```

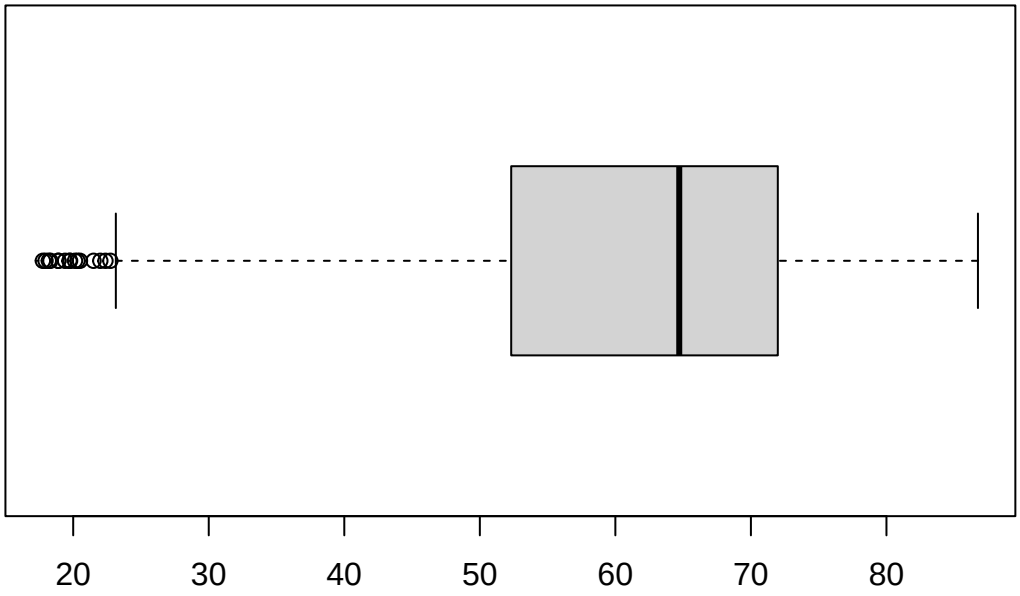
```
## [1] FALSE
```

```
dist = life %>%
  filter(nchar(Code) == 0) %>%
  distinct(Entity)
dist
```

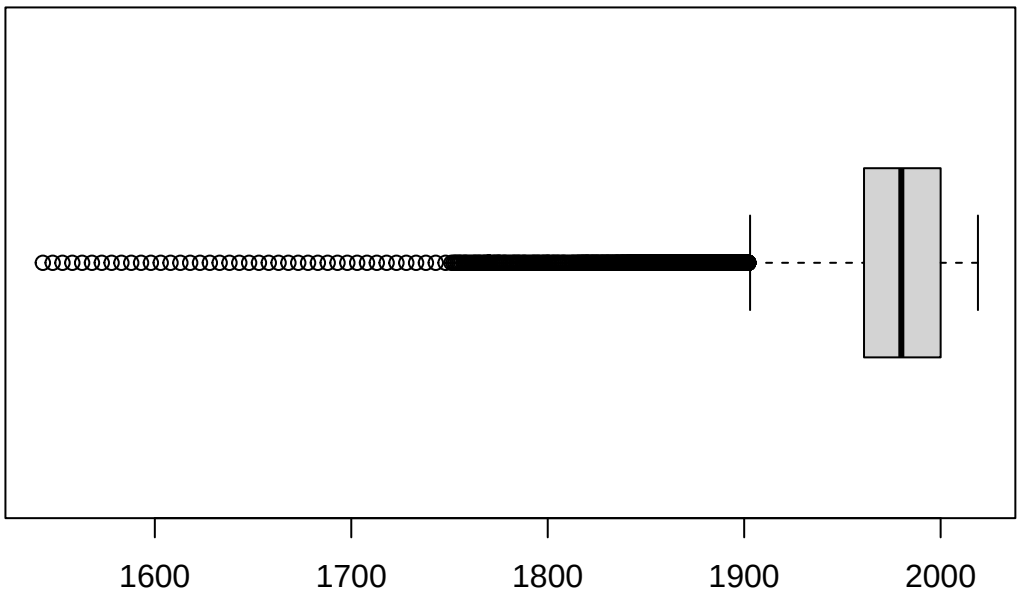
```
##                            Entity
## 1                          Africa
## 2                        Americas
## 3                            Asia
## 4                          Europe
## 5 Latin America and the Caribbean
## 6                 Northern America
## 7                         Oceania
## 8                  Saint Barthlemy
```

So, mostly the country groups has no code.

```
life %$%
  boxplot(Life.expectancy, horizontal = T)
```



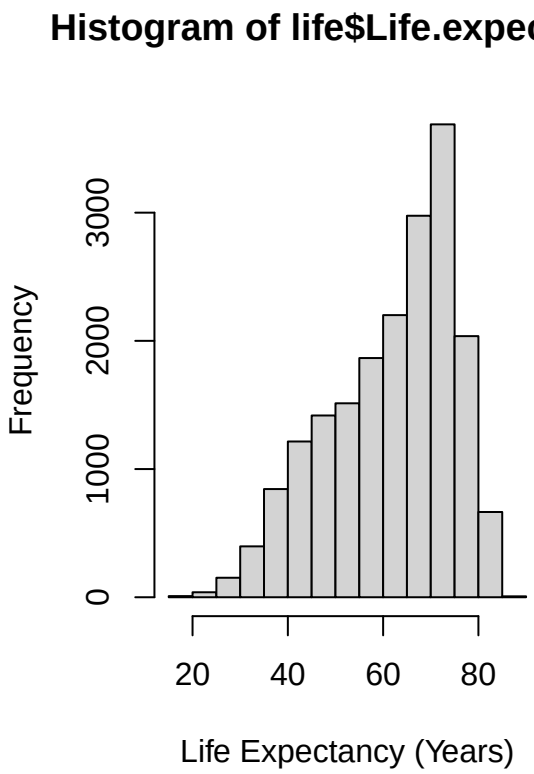

```
life %$%
  boxplot(Year, horizontal = T)
```

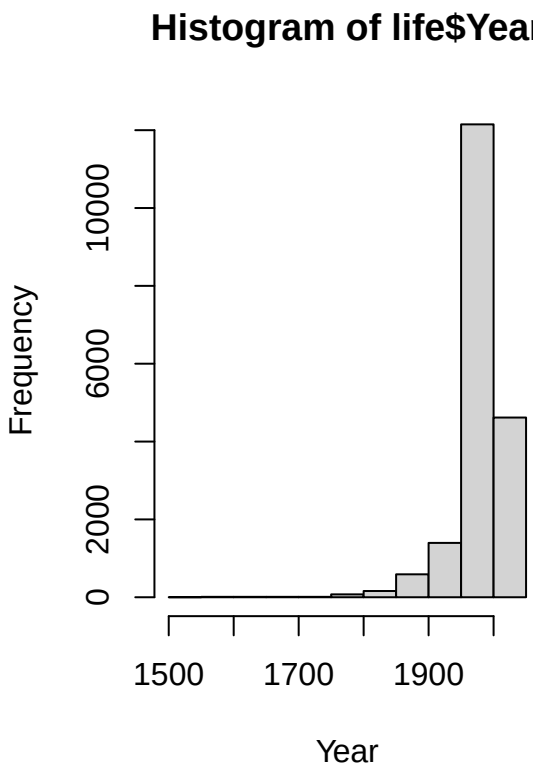


So we observe that there are outliers in Life expectancy and in terms of years, the data is hugely available after 20th century which is quite intuitive.

To check the outliers, let us first check the frequency

```
par(mfrow = c(1,2))
hist(life$Life.expectancy, xlab = "Life Expectancy (Years)")
hist(life$Year, xlab = "Year")
```





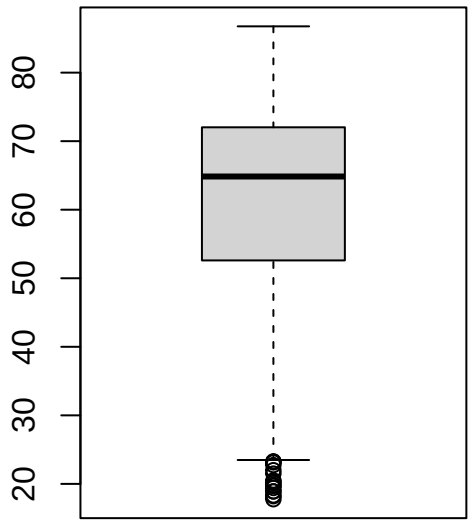
Now as the histogram says, there are fairly low amount of records less than 1800, and so we tend to discard them as they can affect the model.

```
life_over = life%>%
  filter(Year >= 1800)

life_under = life%>%
  filter(Year <= 1800)

summary(cbind(life_over$Life.expectancy, life_under$Life.expectancy))
```

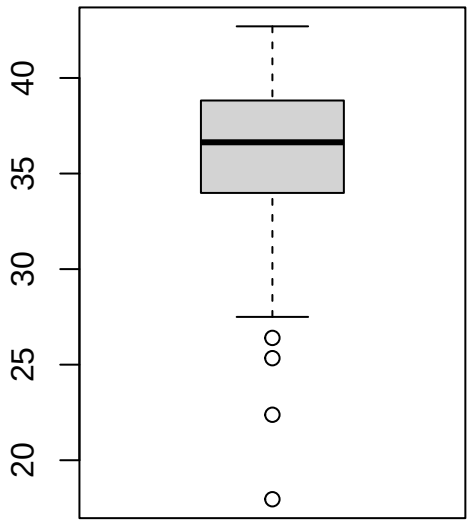
```
## Warning in cbind(life_over$Life.expectancy, life_under$Life.expectancy): number
## of rows of result is not a multiple of vector length (arg 2)
```

```
##        V1              V2
##  Min.   :17.76   Min.   :17.96
##  1st Qu.:52.60   1st Qu.:33.94
##  Median :64.85   Median :36.61
##  Mean   :61.90   Mean   :36.06
##  3rd Qu.:72.02   3rd Qu.:38.82
##  Max.   :86.75   Max.   :42.70
```

```
par(mfrow = c(1,2))
boxplot(life_over$Life.expectancy, xlab = "Life Expectancy over 1800 (Years)")
boxplot(life_under$Life.expectancy, xlab = "Life Expectancy under 1800 (Years)")
```



Life Expectancy over 1800 (Years)



Life Expectancy under 1800 (Years)

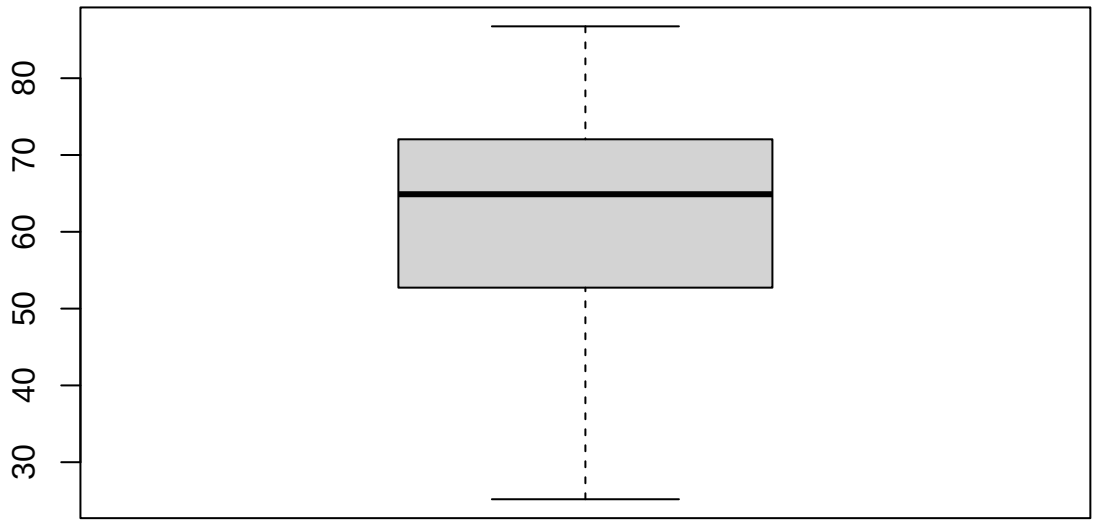
Now this is interesting case that before 1800, life expectancy was far less than that when it is in 19th century. It may be due to advancement in medical technology or other factors.

Now, there are some significant trial of outliers which can't be discarded.

```
life_over %>%
  count(cut_width(Life.expectancy, 10))
```

```
##   cut_width(Life.expectancy, 10)    n
## 1                        [15,25]   46
## 2                        (25,35]  517
## 3                        (35,45] 1980
## 4                        (45,55] 2931
## 5                        (55,65] 4067
## 6                        (65,75] 6665
## 7                        (75,85] 2702
## 8                        (85,95]    7
```

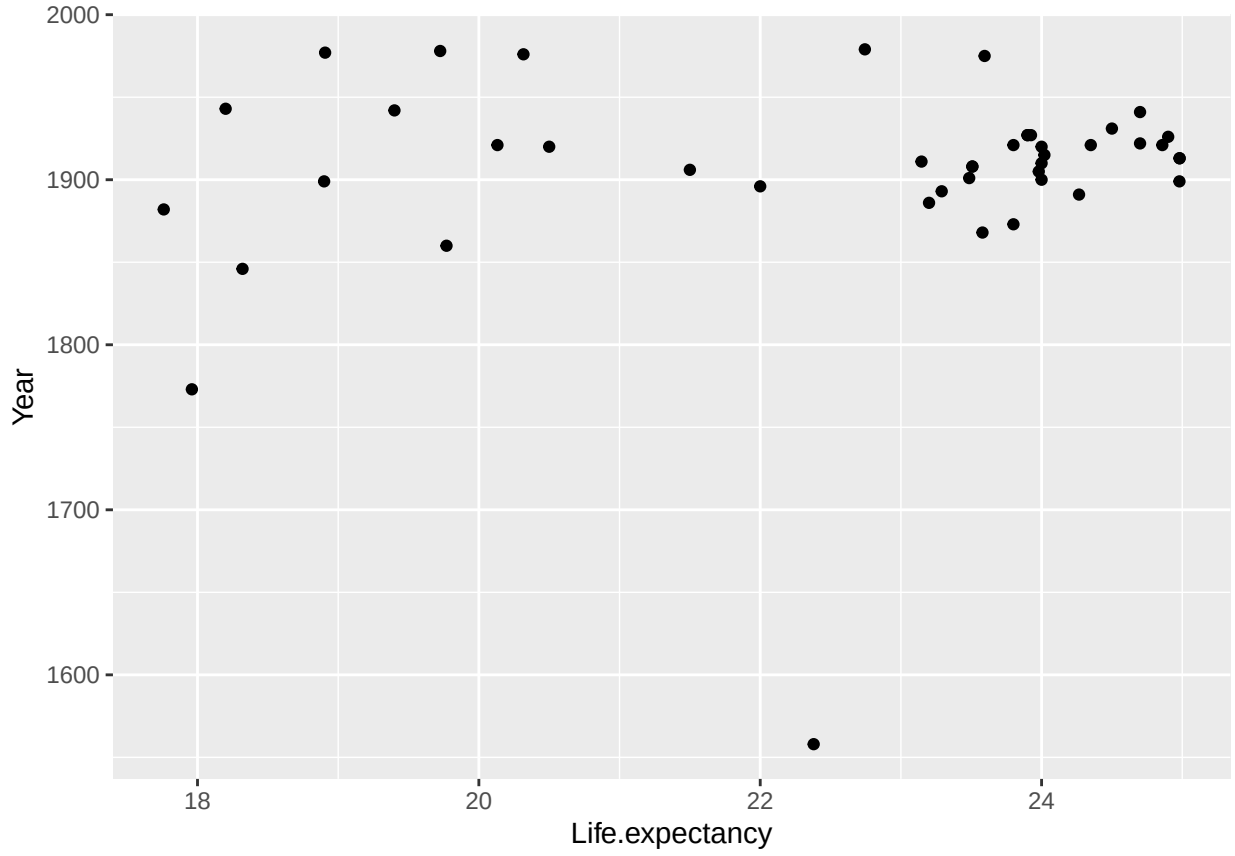
```
under_age = life %>%
  filter(Life.expectancy < 25)

clean_life = life_over %>%
  filter(Life.expectancy > 25)

boxplot(clean_life$Life.expectancy)
```



Now, we have divided the dataset into 3 clean parts, one before and another after 19th century, and the last one for underage data. The first one can be used to model regression and the others can be used for analysis of those times. So, we have the possibility to discard the records less than 25 but we don't know about those data, so we need to do some further study on that.

Doing some year based outlier analysis,

```
ggplot(data = under_age, aes(x = Life.expectancy, y = Year))+
  geom_point()
```



This plot clearly shows that main concentration of the data is mainly towards the upper part i.e more than 1950s. And this plot encourages to assume that a particular country is underaged. In that case we can't discard them without proper analysis.

```
tab = as.data.frame(table(under_age$Entity))
tab
```

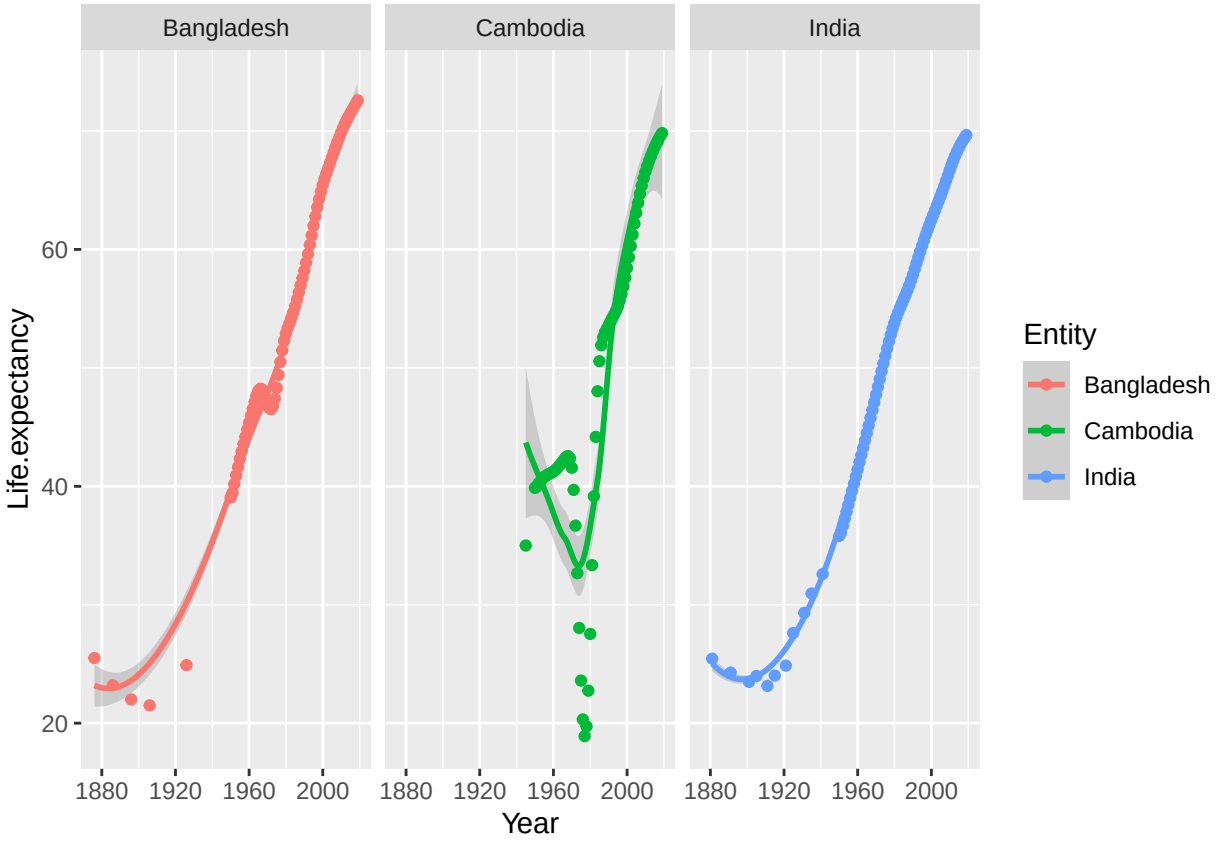
```
##             Var1 Freq
## 1     Bangladesh    4
## 2       Cambodia    5
## 3           Cuba    1
## 4      Guatemala    2
## 5        Iceland    3
## 6          India    6
## 7     Kazakhstan    2
## 8          Kenya    1
## 9         Mexico    2
## 10     Nicaragua    1
## 11         Niger    1
## 12   North Korea    2
## 13      Pakistan    1
## 14        Russia    3
## 15       Senegal    1
## 16  Sierra Leone    1
## 17   South Korea    2
## 18        Sweden    1
## 19        Uganda    1
```

```
## 20        Ukraine    3
## 21 United Kingdom    1
```

This table draws our attention towards some particular countries including our country India. So it definitely demands an analysis of those countries.

```
tab_great = tab %>%
  filter(Freq>3)

life %>%
  filter(Entity %in% tab_great$Var1)%>%
  ggplot(data = ., aes(x = Year, y = Life.expectancy, color = Entity))+
  geom_smooth()+ geom_point()+
  facet_wrap(~Entity)
```

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



This shows that for countries like Cambodia, these are mere outliers or an indication that something serious happened after 1970 that repidly boosted their develoment. But for our country it is as good as it can be being slow and steady to a good position now.

Now as there are so many countries, graphical analysis of each and every country is nearly impossible. So, making a function that would detect that for us.

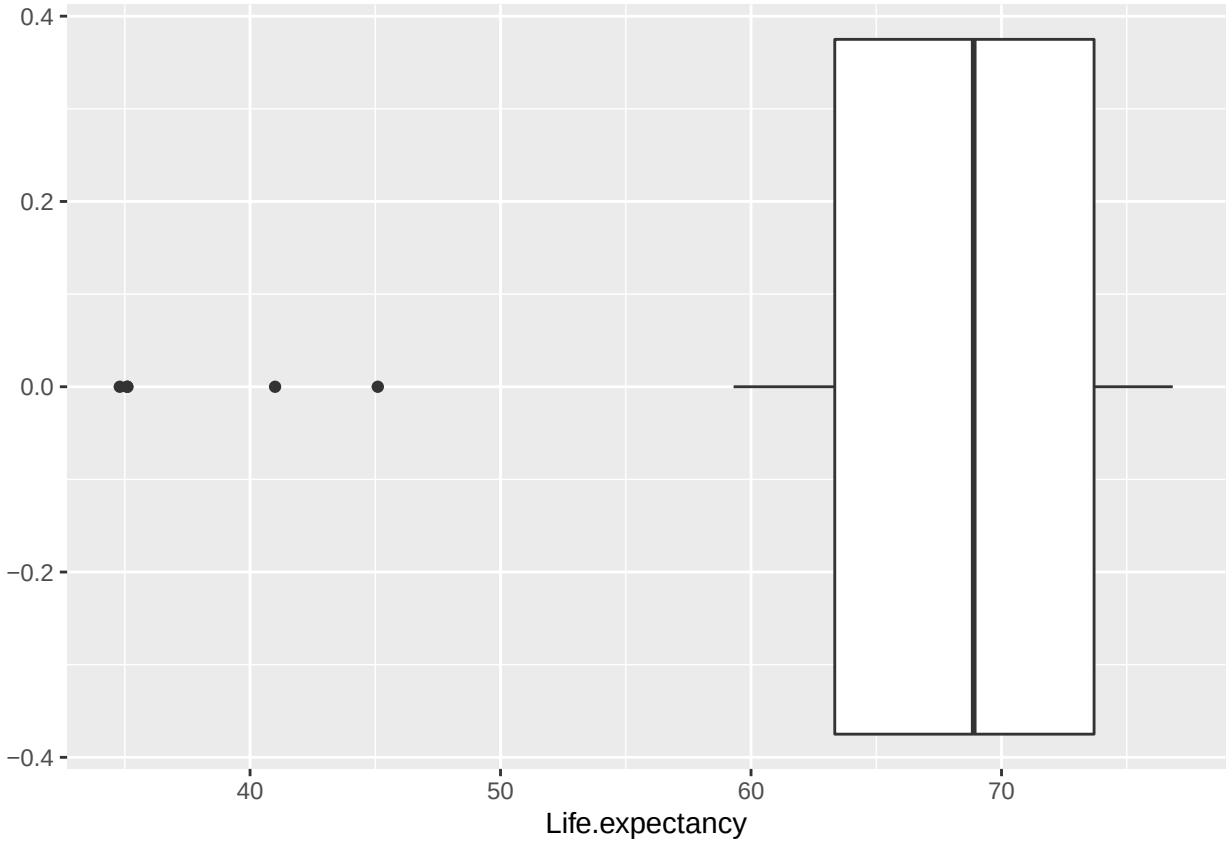
```
outlier = function(x,y){
  vec = c()
  out = c()
    for(i in unique(y)){
   get =   x%>%filter(Entity == i)
```

```
        for(j in get$Life.expectancy){

          q1 = quantile(get$Life.expectancy, 0.25)
          q3 = quantile(get$Life.expectancy, 0.75)
          iqr = q3 - q1
          if(j < q1 - 1.5*iqr | j > q3 + iqr*1.5){
            out = append(out,j)
            if(! i %in% vec){
              vec = append(vec,i)

            }

            }

        }
    }
  return(list("coun" = vec, "life"  = out))
}

out = outlier(clean_life, clean_life$Entity)
out$coun
```

```
##  [1] "Africa"              "Americas"            "Angola"
##  [4] "Argentina"           "Aruba"               "Australia"
##  [7] "Austria"             "Barbados"            "Belarus"
## [10] "Brazil"              "Bulgaria"            "Burundi"
## [13] "Cameroon"            "Canada"              "Colombia"
## [16] "Congo"               "Costa Rica"          "Cuba"
## [19] "Cyprus"              "Czech Republic"      "Dominican Republic"
## [22] "El Salvador"         "Estonia"             "Europe"
## [25] "Fiji"                "Georgia"             "Germany"
## [28] "Ghana"               "Greece"              "Greenland"
## [31] "Guyana"              "Hungary"             "Iraq"
## [34] "Ireland"             "Japan"               "Kazakhstan"
## [37] "Kenya"               "Kuwait"              "Latvia"
## [40] "Lithuania"           "Malawi"              "Mauritius"
## [43] "Moldova"             "Montenegro"          "North Korea"
## [46] "Oceania"             "Pakistan"            "Panama"
## [49] "Paraguay"            "Philippines"         "Poland"
## [52] "Puerto Rico"         "Romania"             "Russia"
## [55] "Rwanda"              "Serbia"              "Sierra Leone"
## [58] "Slovakia"            "South Korea"         "Sri Lanka"
## [61] "Tanzania"            "Trinidad and Tobago" "Uganda"
## [64] "Ukraine"             "Uruguay"             "Venezuela"
## [67] "World"               "Zambia"
```

So there are 68 countries with outliers. Let's take a country and do it's boxplot to check if our function is working

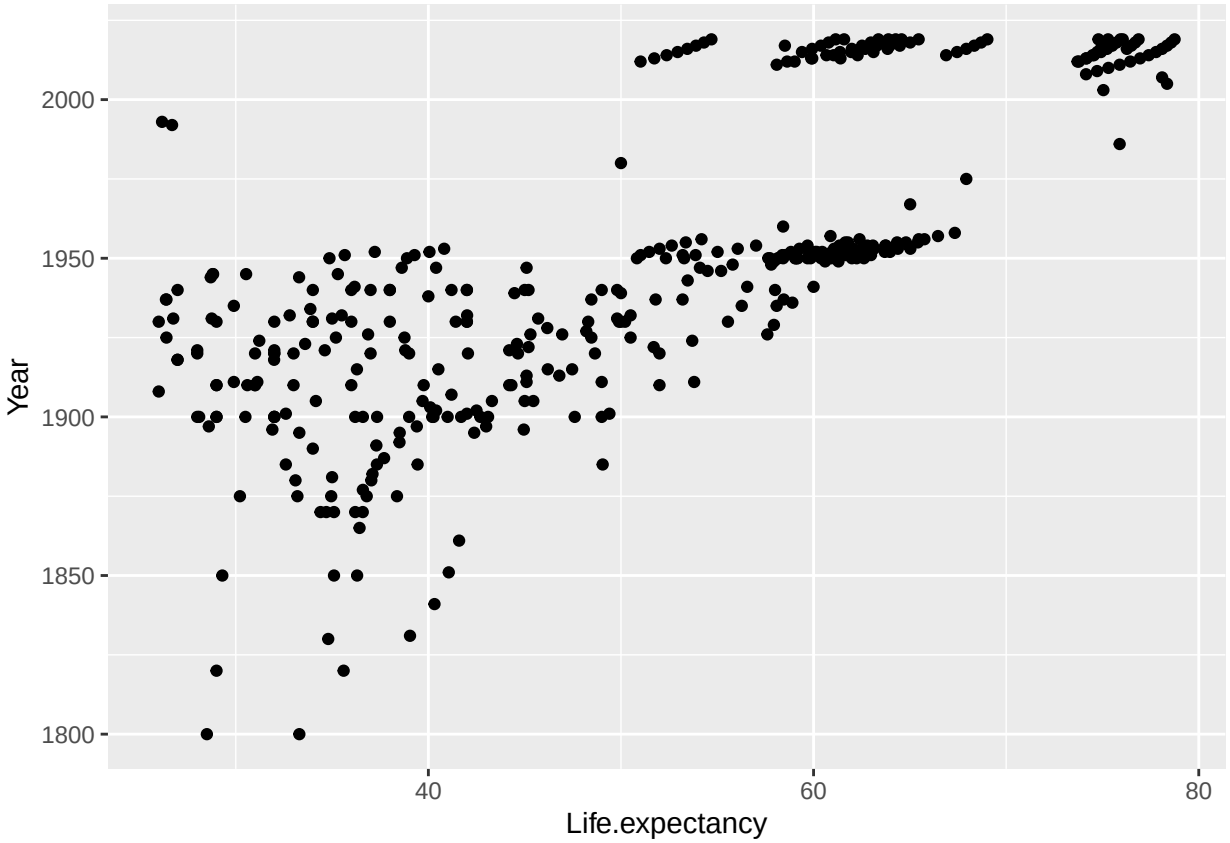
```
clean_life%>%
  filter(Entity %in% c("Americas"))%>%
  ggplot(data = ., aes(x = Life.expectancy))+
  geom_boxplot()
```



And yes there are 4 outliers. So, our function is working.

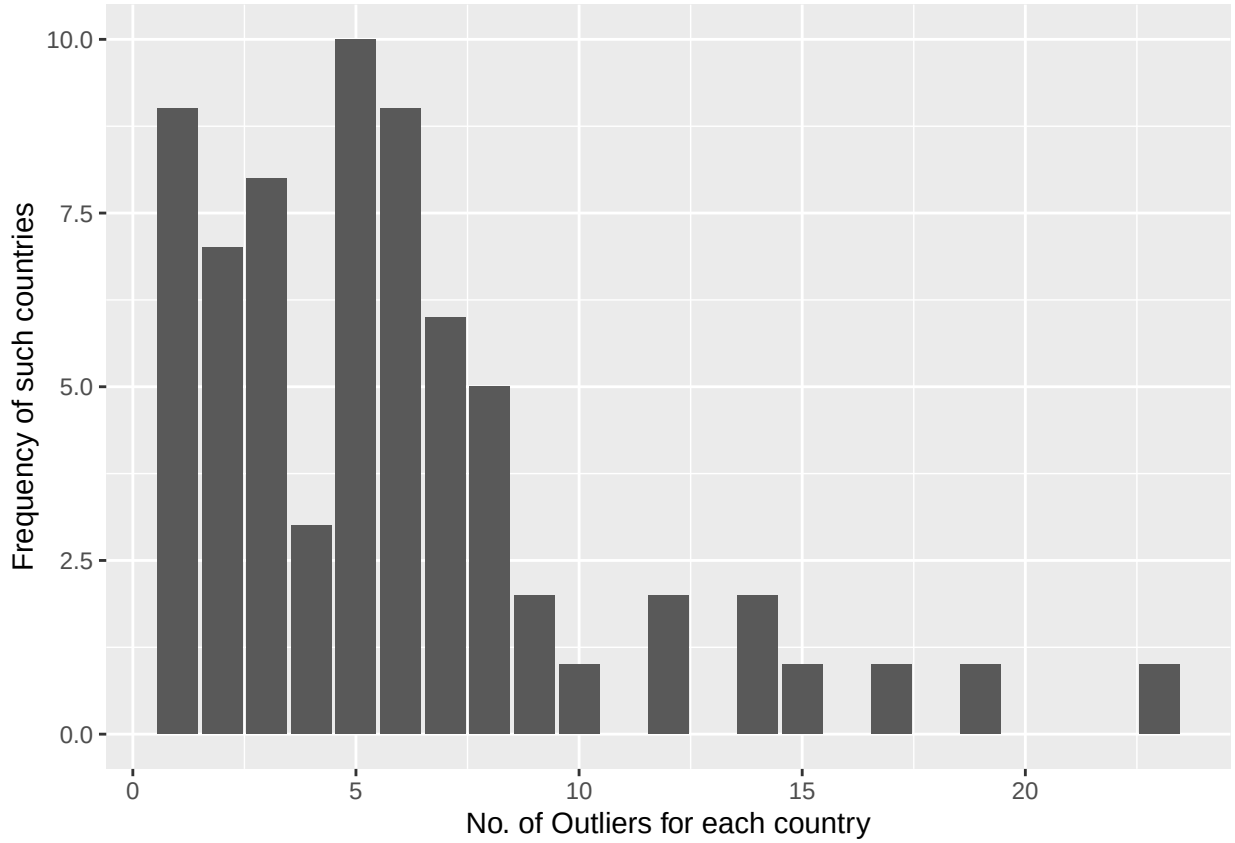
Let us do a plot of the outliers on the basis of year.

```
clean_life%>%
  filter(Entity %in% out$coun & Life.expectancy %in% out$life)%>%
  ggplot(data = ., mapping = aes(x = Life.expectancy, y = Year)) +
  geom_point()
```



So, max outliers are taken around or less than 1950. So, we can consider eliminating them. But before that let us do a quick analysis of number of outliers on basis of a country.

```
clean_life%>%
  filter(Entity %in% out$coun & Life.expectancy %in% out$life)%>%
  group_by(Entity)%>%
  summarize(count = n()) %>%
  ggplot(data = ., aes(x = count))+ labs(x = "No. of Outliers for each country", y = "Frequency of such
  geom_bar()
```



This is surprising because there are some countries which has more than 10 outliers. Let us filter and remove those entries which has less than 10 outliers and focus more on the bigger counterparts.

```
coun_out = clean_life%>%
  filter(Entity %in% out$coun & Life.expectancy %in% out$life)%>%
  group_by(Entity)%>%
  summarize(count = n())%>%
  filter(count < 10)

clean_life = clean_life%>%
  filter(!(Entity %in% coun_out$Entity & Life.expectancy %in% out$life))

outlier = clean_life%>%
  filter(Entity %in% out$coun & Life.expectancy %in% out$life)%>%
  group_by(Entity)%>%
  summarize(count = n())%>%
  filter(count > 10)

outlier
```

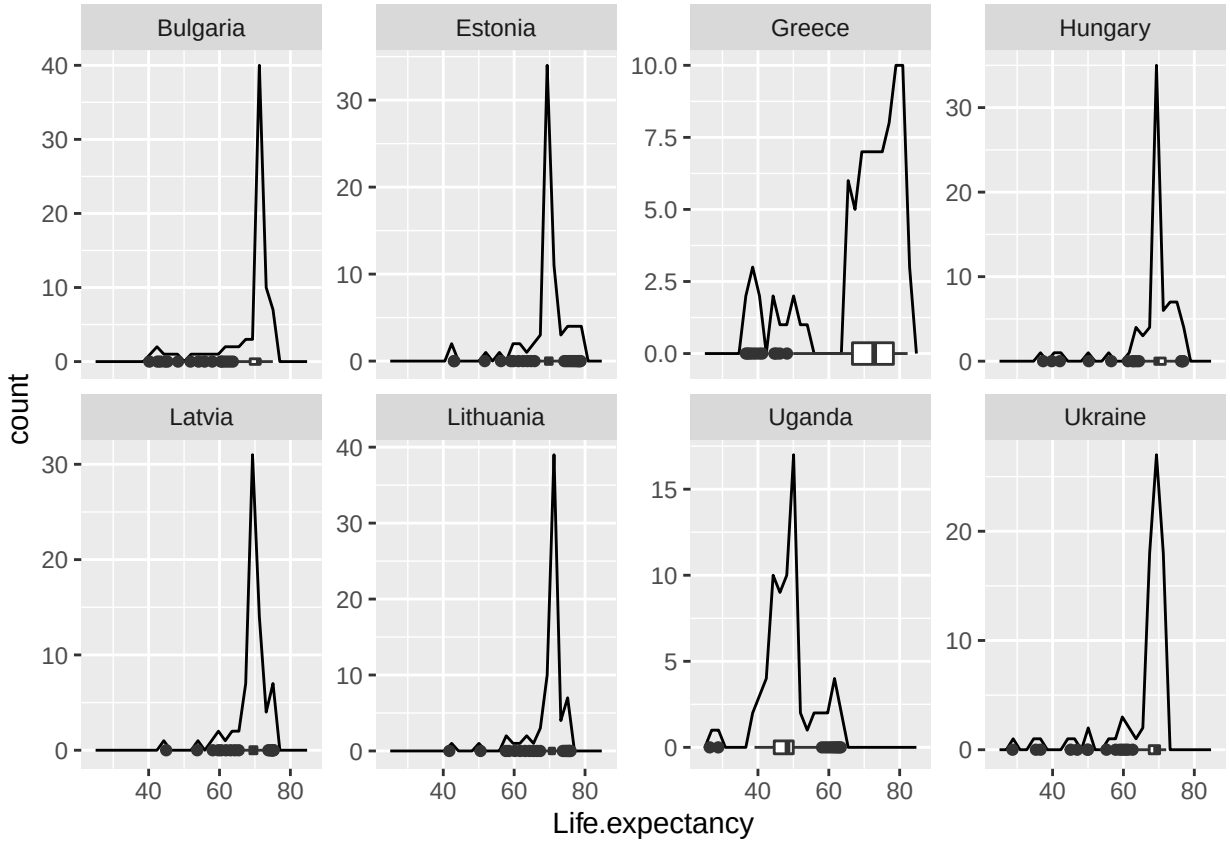
```
## # A tibble: 8 x 2
##   Entity    count
##   <chr>     <int>
## 1 Bulgaria     15
## 2 Estonia      23
## 3 Greece       12
## 4 Hungary      14
```

```
## 5 Latvia       17
## 6 Lithuania    19
## 7 Uganda       12
## 8 Ukraine      14
```

Ad so we got the countries with max outliers. Let us do a box plot of these 8 countries to understand more about them.

```
clean_life %>%
  filter(Entity %in% outlier$Entity)%>%
  ggplot(data = ., aes(x = Life.expectancy))+
  geom_freqpoly()+
  geom_boxplot()+
  facet_wrap(~Entity, ncol = 4, scales = "free_y")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



These boxplots shows something very fishy kind of situation which demands some practical checking for those fluctuations among those countries.

3.0.1 EDA

Since, this is a life expectancy dataset, let us dive some deeper into the life expectancy values and try to know what is the scenario of life all over the world.

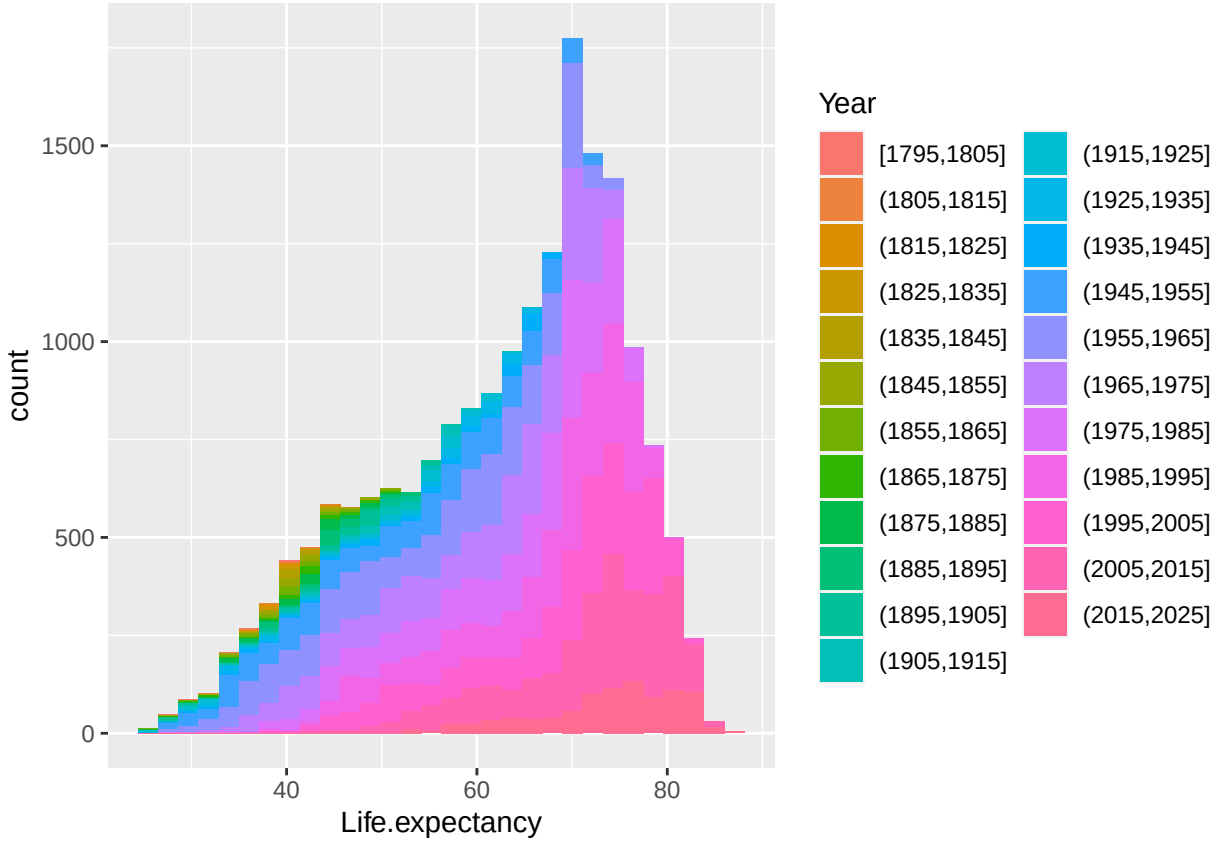
```
clean_life%>%
  count(cut_interval(as.integer(Year),10))
```

```
##    cut_interval(as.integer(Year), 10)    n
```

```
## 1                  [1.8e+03,1.82e+03]   34
## 2                 (1.82e+03,1.84e+03]   70
## 3                 (1.84e+03,1.87e+03]  169
## 4                 (1.87e+03,1.89e+03]  235
## 5                 (1.89e+03,1.91e+03]  331
## 6                 (1.91e+03,1.93e+03]  426
## 7                 (1.93e+03,1.95e+03] 1373
## 8                 (1.95e+03,1.98e+03] 5331
## 9                    (1.98e+03,2e+03] 5339
## 10                   (2e+03,2.02e+03] 5298
```

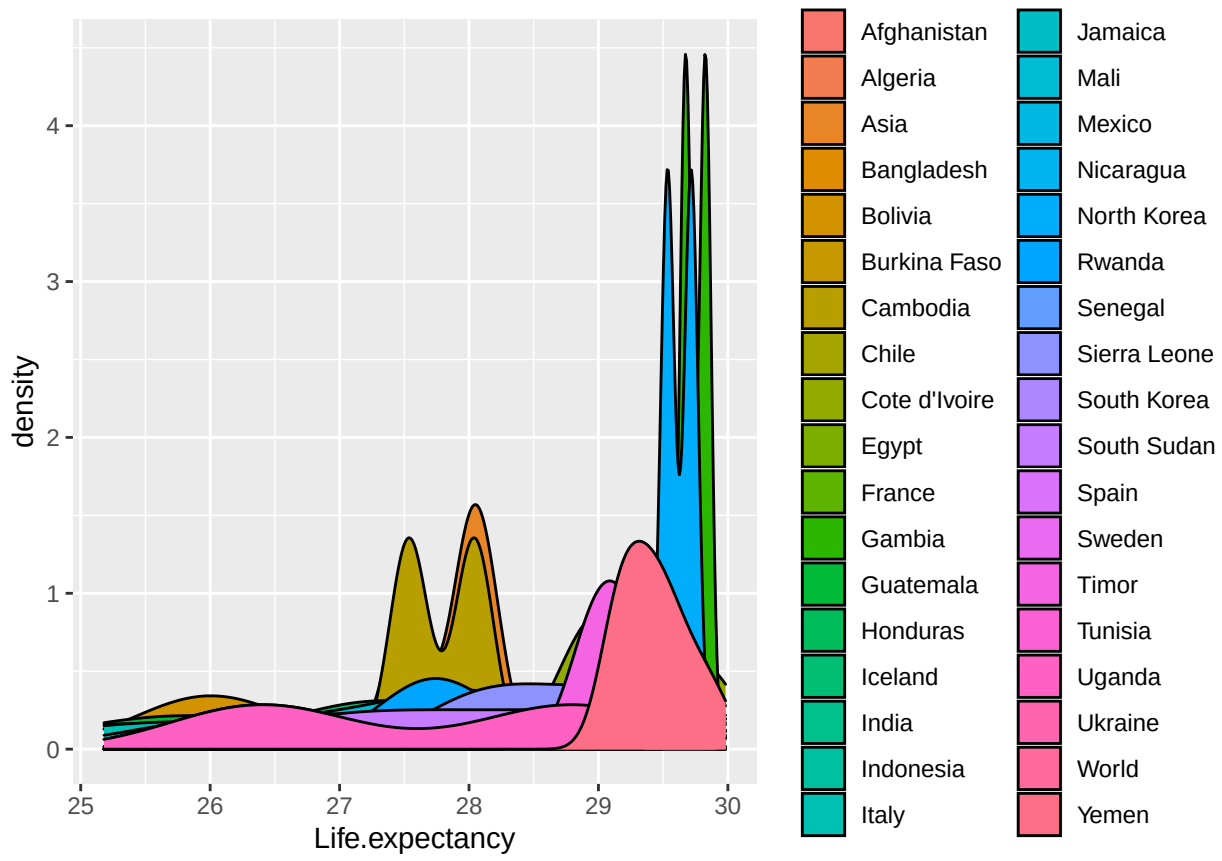
```
ggplot(data = clean_life, aes(x= Life.expectancy, fill = cut_width(Year,10)))+
  geom_histogram()+
  labs(fill = "Year")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



As the plot suggests, as time progressed, people's life expectancy increased and tend to concentrate more around 70s. But I am more interested on least and most life expectant countries. So let's do a closer look.

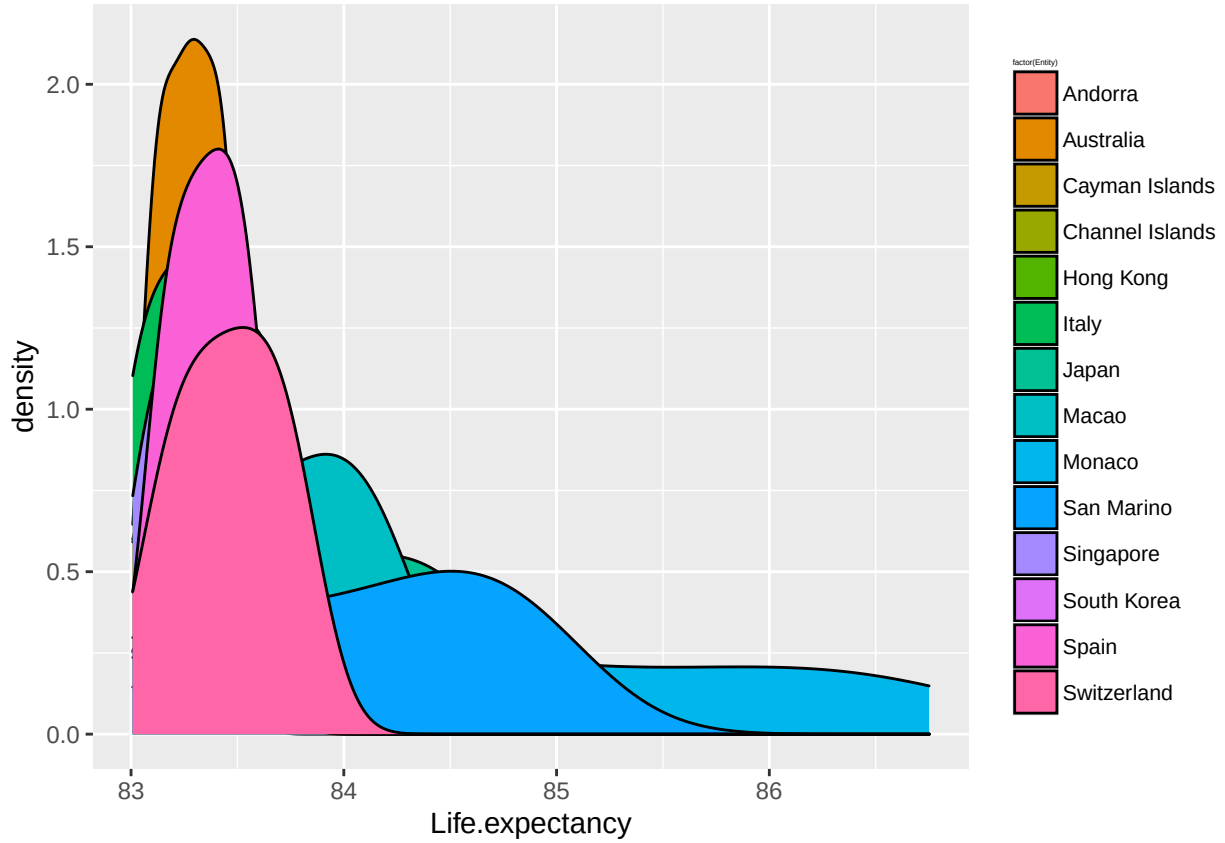
```
clean_life%>%
  filter(Life.expectancy < 30)%>%
  ggplot(data = ., aes(x= Life.expectancy, fill = factor(Entity)))+
  geom_density()
```



So, some interesting results like Bangladesh, India on the list with a low desity sweeping around the low life mark whereas that of Asia and World have moderately good density around 30 years.

So we get to know that at certain periods of time, life was our continent/world was pretty tough to a high extent.

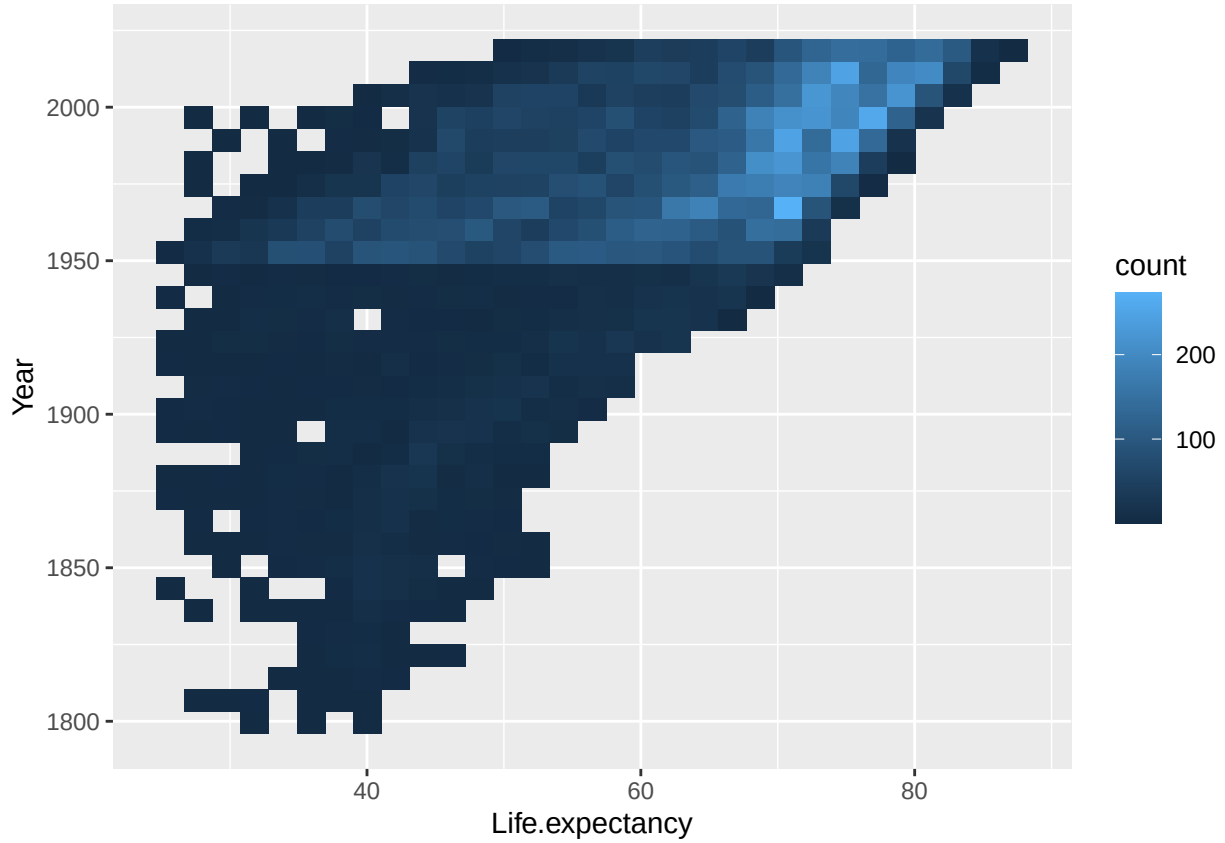
```
clean_life%>%
  filter(Life.expectancy > 83)%>%
  ggplot(data = ., aes(x= Life.expectancy, fill = factor(Entity)))+
  geom_density()+
  guides(shape = guide_legend(override.aes = list(size = 0.5)))+
  theme(legend.title = element_text(size = 3), legend.text = element_text(size = 8))
```



And as we can see from the graph, Monaco consistently keeps high life consistency rate. And as a continent, Australia is doing the best and lot of big developed country names like Japan are also here.

Now, let's measure the growth or the development of countries by analysing Life expectancy with time.

```
ggplot(data = clean_life, aes(x = Life.expectancy, y = Year))+
  geom_bin2d()
```



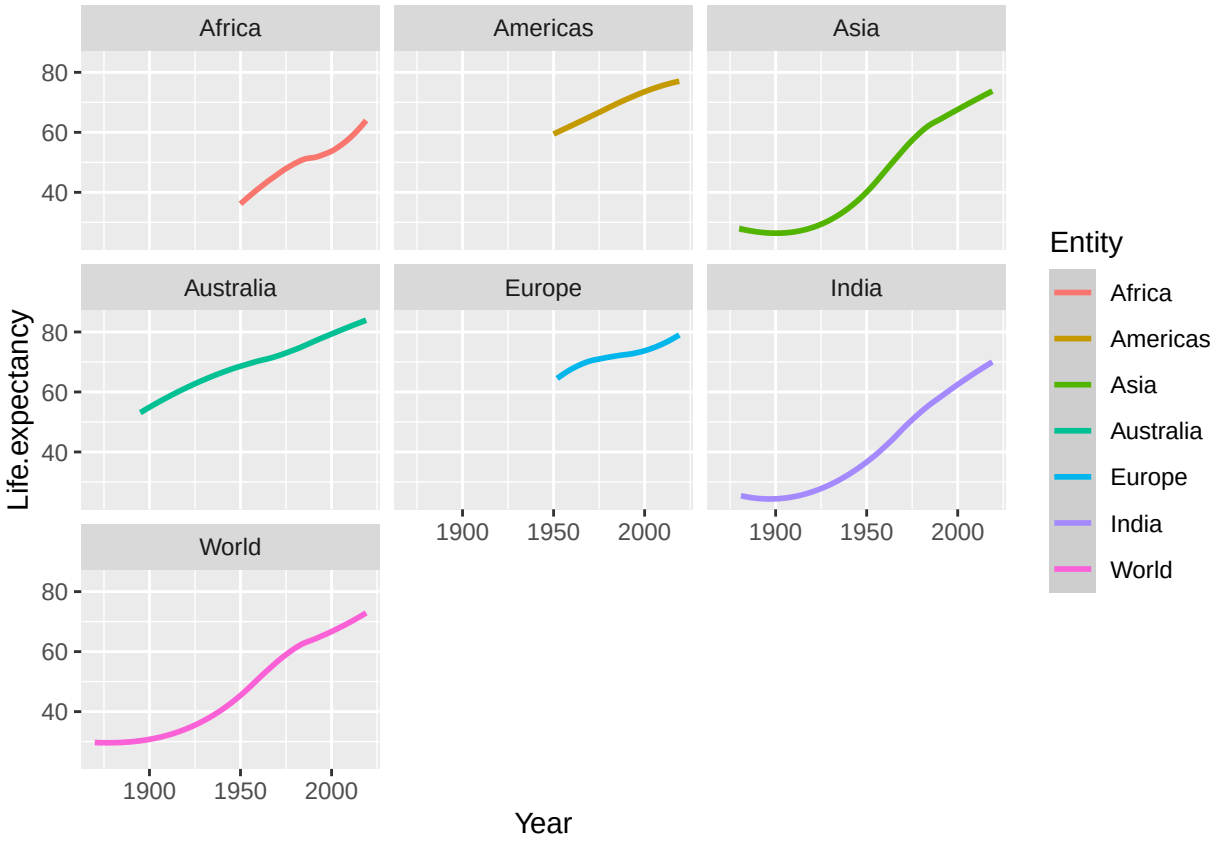
So, this plot clearly shows that at the late 20th Century or the early 21st century led to an huge rise of Life expectancy.

Going out of overall analysis and going into specific analysis of all the Continents, World and India.

Let's start by growth analysis.

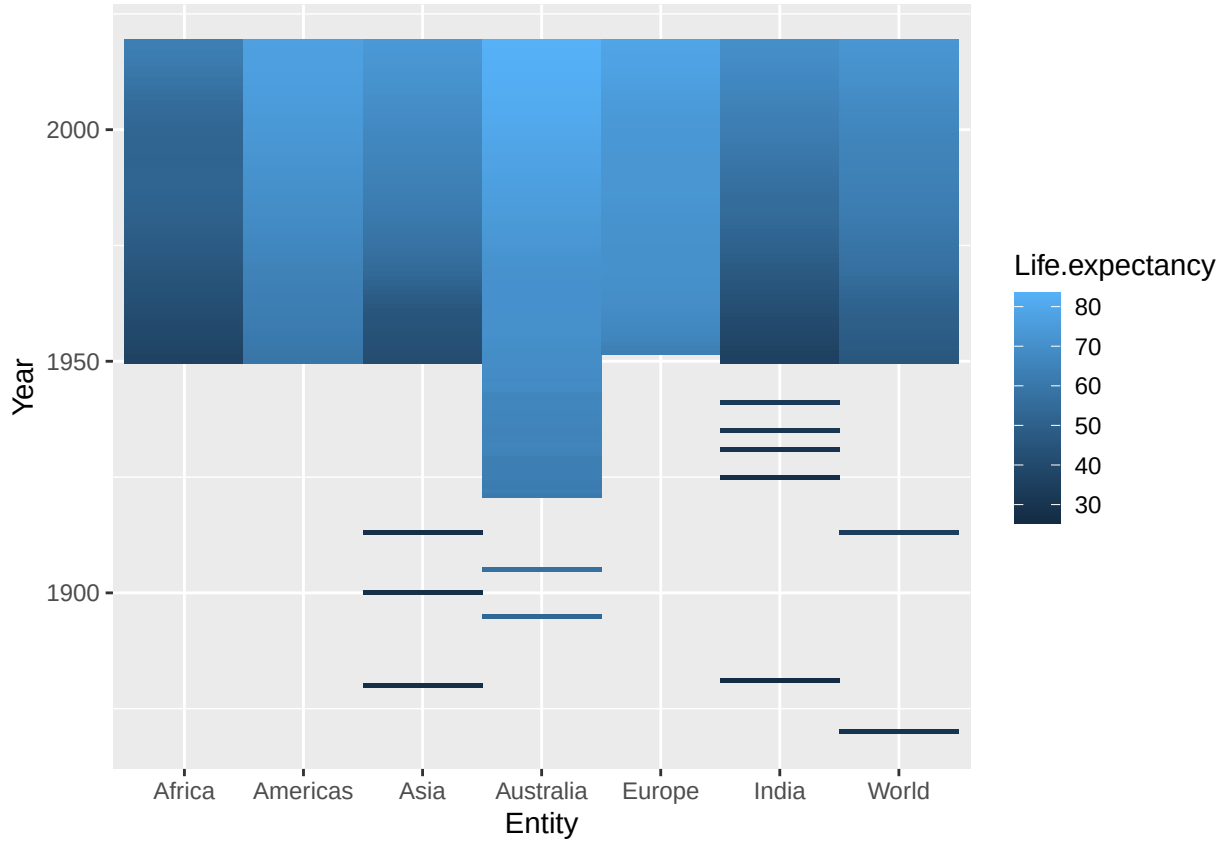
```
country = c("India", "Americas", "Africa", "Asia", "Australia", "Europe", "World")
scale_value = 1
clean_life %>%
  filter(Entity %in% country)%>%
  ggplot(data = ., aes(x = Year, y = Life.expectancy, color = Entity))+
  geom_smooth()+
  facet_wrap(~Entity)
```

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



From a very low expectancy rate to a considerable expecatancy rate, the growth of India shows that it is developing and so is the case all over the world.

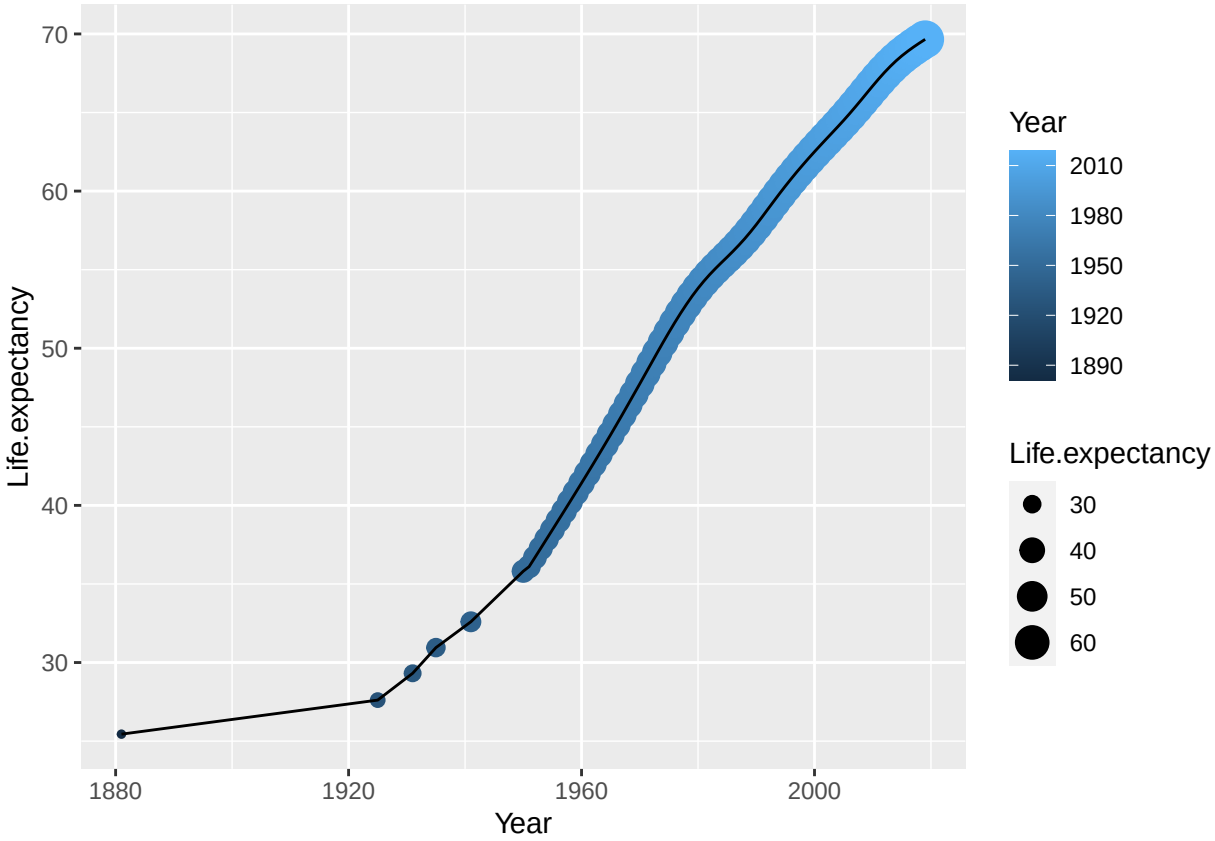
```
clean_life %>%
  filter(Entity %in% country)%>%
  ggplot(data = ., aes(x = Entity, y = Year, fill = Life.expectancy))+
  geom_tile()
```



This plot gives us an idea that most of the records of India and World are mainly influenced by low availability of records in the early 20th century. But as a whole they are below average whereas Americas and Australia has pretty high life expectancy.

A closer look at India.

```
clean_life %>%
  filter(Entity == "India")%>%
  ggplot(data = ., aes(x = Year, y = Life.expectancy))+
  geom_point(aes(size= Life.expectancy, color = Year))+
  geom_line()
```



This graph proves that our deduction about India is correct.

4 Results and Discussions

The whole descriptive statistical analysis is

```
clean_life %>%
  filter(Entity %in% country)%>%
  group_by(Entity)%>%
  summarise(mean(Life.expectancy), min(Life.expectancy), max(Life.expectancy))
```

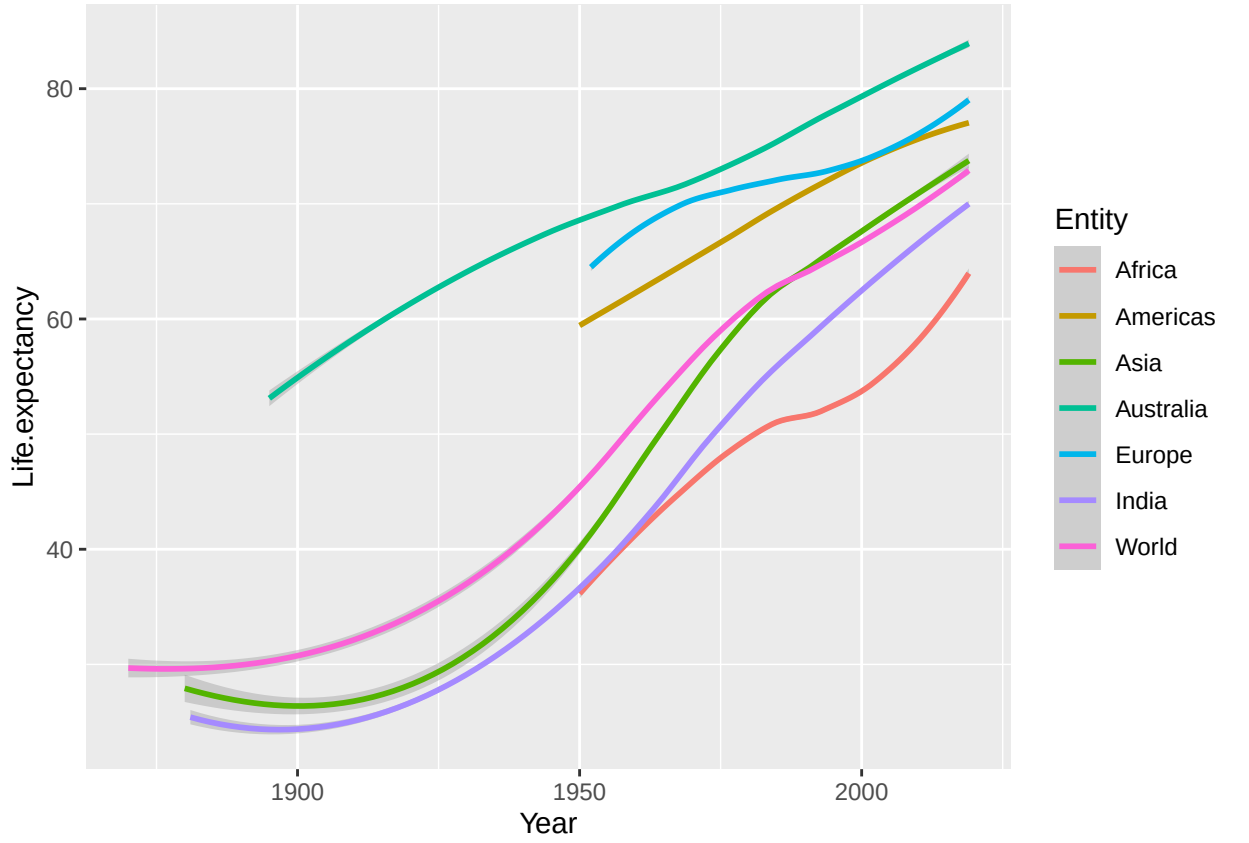
```
## # A tibble: 7 x 4
##   Entity    `mean(Life.expectancy)` `min(Life.expectancy)` `max(Life.expectanc~`
##   <chr>                       <dbl>                  <dbl>                  <dbl>
## 1 Africa                       49.8                   36.4                   63.2
## 2 Americas                     69.1                   59.3                   76.8
## 3 Asia                         58.6                   27.5                   73.6
## 4 Australia                    72.3                   53.0                   83.4
## 5 Europe                       72.1                   63.3                   78.6
## 6 India                        52.8                   25.4                   69.7
## 7 World                        60.2                   29.7                   72.6
```

And the graphical analysis of the findings

```
clean_life %>%
  filter(Entity %in% country)%>%
  ggplot(data = ., aes(x = Year, y = Life.expectancy, color = Entity))+
```

```
  geom_smooth()
```

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



5 Conclusion

The life expectancy of human beings is largely around 70s and varies from country to country. The general trend being, the more developed a country is in terms of medical facilities, the more is its life expectancy.

The highest Life Expectancy is of Australlia, Africa among continents. Asia and India's life expectancy is below the world average and is gradually increasing as we are developing.

The 20th century showed huge development in terms of life expectancy. The developed countries like Japan did extremely good in terms of life expectancy. But our surprise winner here is Monaco. We definitely need to do some background check for why is this true.

Continents like Asia and Africa have good variations of life expectancy, keeping a large part ranging from 30s to 80s. This displays the enormous versatility of these continents to hold some really rich and some really poor countries.

Countries like Bangladesh, Bolivia needs to sort out their low expectancy issues. And some like Hungary, Greece needs to sort their fluctuations in the Life expectancy. And others like Cambodia did something magical to boost their expectancy after 1970s.

Though this is a study on the life expectancy of various countries all over the world but there is a lot of scope in this data to investigate more on certain parameters.

6 References

6.0.0.1 Exploratory Data Analysis: https://r4ds.had.co.nz/exploratory-data-analysis.html