MiniProject_Sudan

Jeevan Reddy, Pavan Madineni, Pramod Duvvuri

February 23, 2018

Libraries

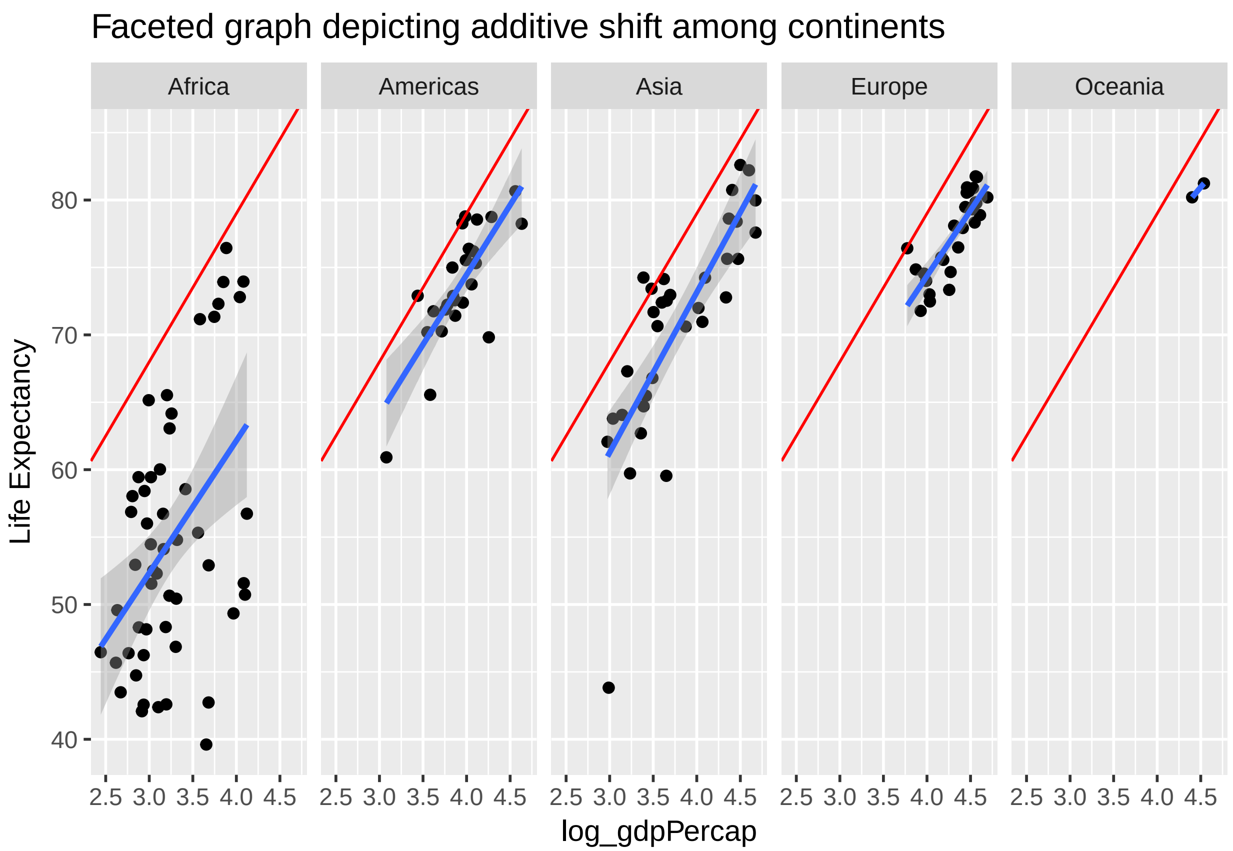
```r
library(gapminder)
library(MASS)
library(ggplot2)
library(broom)
library(tidyr)
```

Question-1

```r
data = gapminder
data_2007 = data[data$year == 2007,]
data_2007$log_gdpPercap = log10(data_2007$gdpPercap)
```

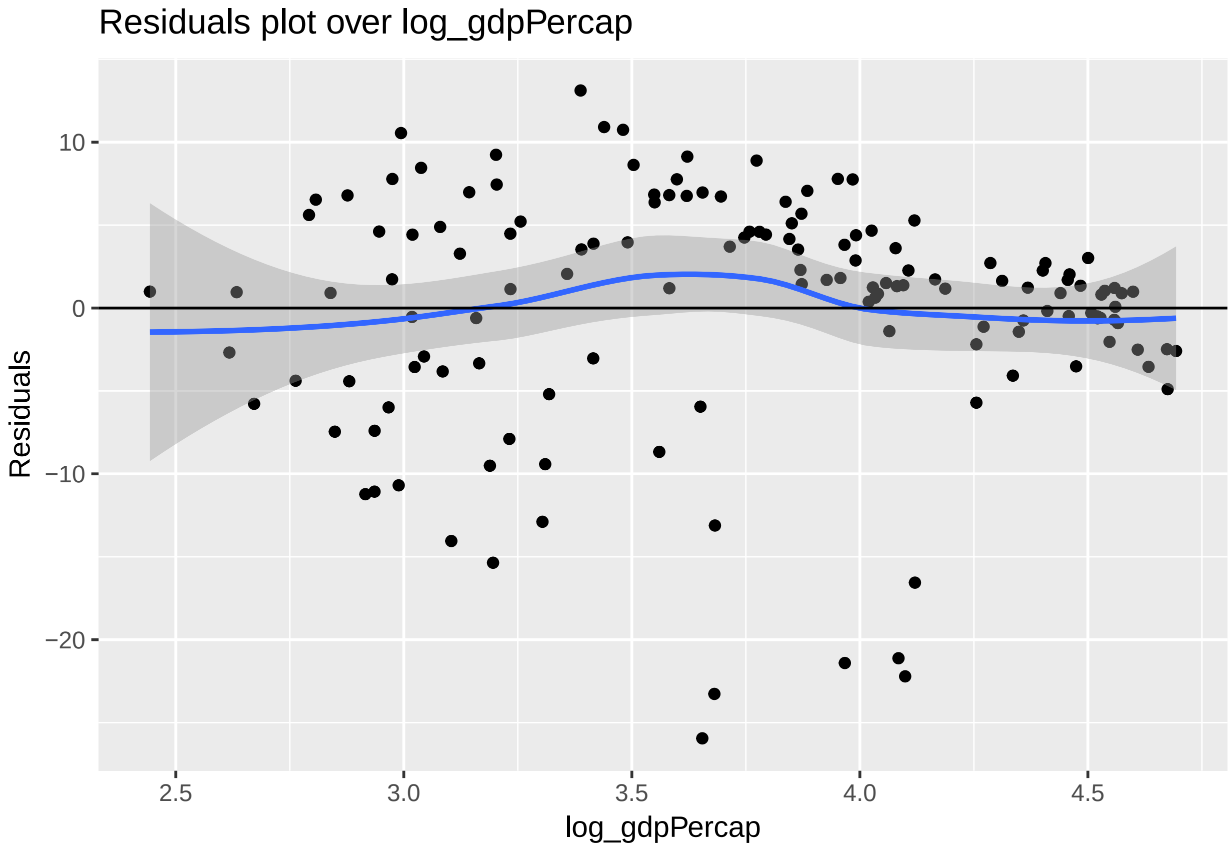
Introduction: The Gapminder is a dataset with 1704 observations predominantly focused on estimating life expectancy in relation with GDP per Capita. This set is mainly concerned about the data for the year 2007 which has 142 observations. Though this dataset comprises data pertaining to countries of different continents with different populations, it does not seem appropriate to consider weighted GDP since gdpPercapita is already computed for every 1000 people of the country.The expected results are :

```r
ggplot(data_2007, aes(x = log_gdpPercap, y = lifeExp)) +
  geom_point() + geom_smooth(method = "lm") +
  facet_grid(~continent) +
  geom_abline(data = data_2007, aes(intercept = 35, slope = 11), color = "red") +
  ylab("Life Expectancy") + ggtitle("Faceted graph depicting additive shift among continents")
```



The trends for Americas, Asia and Europe appear more or less similar with a linear trend lying underneath where the Life Expectancy increases with GDP per Capita barrring a few exceptions. The relation between these three continents can also be explained by an additive shift.Africa also has an underlying linear trend but beyond a certain point, the lifeExp randomly varies with incresing gdpPercapita.

```
data.lm = lm(lifeExp ~ log_gdpPercap, data = data_2007)
data_2007.df = augment(data.lm)
ggplot(data_2007.df, aes(x = log_gdpPercap, y = .resid)) + geom_point() +geom_smooth() +
  geom_abline(slope = 0, intercept = 0) + ggtitle("Residuals plot over log_gdpPercap")+
  ylab("Residuals")
```



The residuals move just around the Zero line implying that the linear model fits the data well

```
df = data.frame(gapminder)
names(df) = c("Country", "Continent", "Year", "Life_Expectancy",
              "Population", "GDP_per_Capita")
```

Loess model looks for a small amount of data around its vicinity to fit a linear model or polynomial model where we can change this neighbourhood by changing the parameter "Span" Span means how much percentage of the data is to be considered while building a model Robust Linear model is robust to extreme outliers as it won't take the OLS method to find the best fitting curve. This is useful when our data has many outliers So, this can be converted to a timeseries data as each of the row(year) is dependent and are evenly spaced.

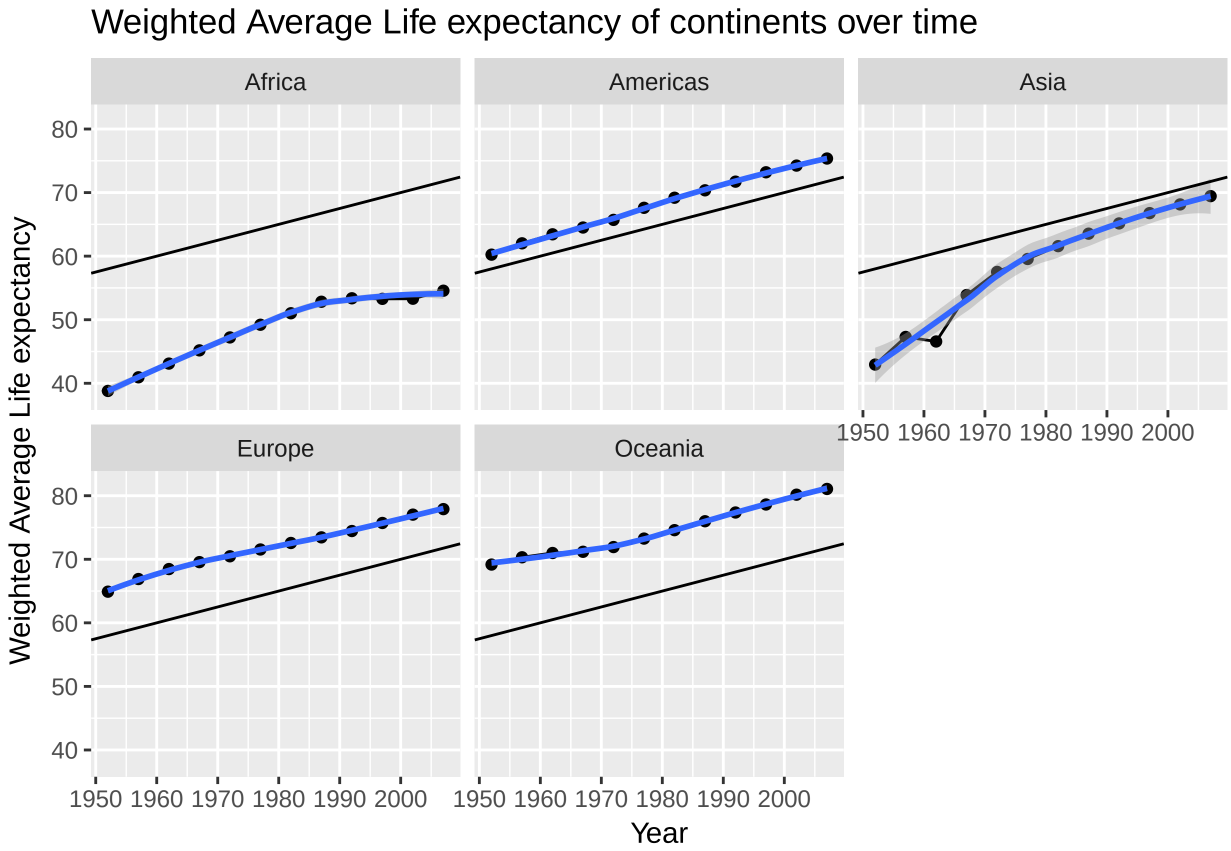
```
# Calculating the weighted GDP percapita for each year
#so that it will be easier to plot using ggplot
df$tw = df$GDP_per_Capita * df$Population #Total wealth per country
ts1 = aggregate(df$tw, by = list(df$Year,df$Continent),"sum")
#Divide it by 1000 to take care of integer overflow error
ts2 = aggregate(df$Population/1000, by = list(df$Year,df$Continen),"sum")
ts3 = ts1$x/(ts2$x*1000)
ts3= data.frame(ts3)
ts3$Year = ts1$Group.1
ts3$Continent = ts1$Group.2
ts3$Avg_GDP_PCI = ts3$ts3
ts3 = ts3[,c(2:4)]
#Calculating the weighted life expectancy for each year
df$tle = df$Population * df$Life_Expectancy #Total wealth per country
te1 = aggregate(df$tle, by = list(df$Year,df$Continent),"sum")
#Divide it by 1000 to take care of integer overflow error
```

```
te2 = aggregate(df$Population/1000, by = list(df$Year,df$Continent),"sum")
ts3$Avg_Le = te1$x/(te2$x*1000)
```

Question-2

```
ggplot(ts3, aes(x=Year, y= Avg_Le)) + geom_point() + geom_line() + geom_smooth() +
  facet_wrap(~Continent) +geom_abline(slope= 0.25,intercept = -430) +
  ggtitle("Weighted Average Life expectancy of continents over time") +
  ylab("Weighted Average Life expectancy")
```

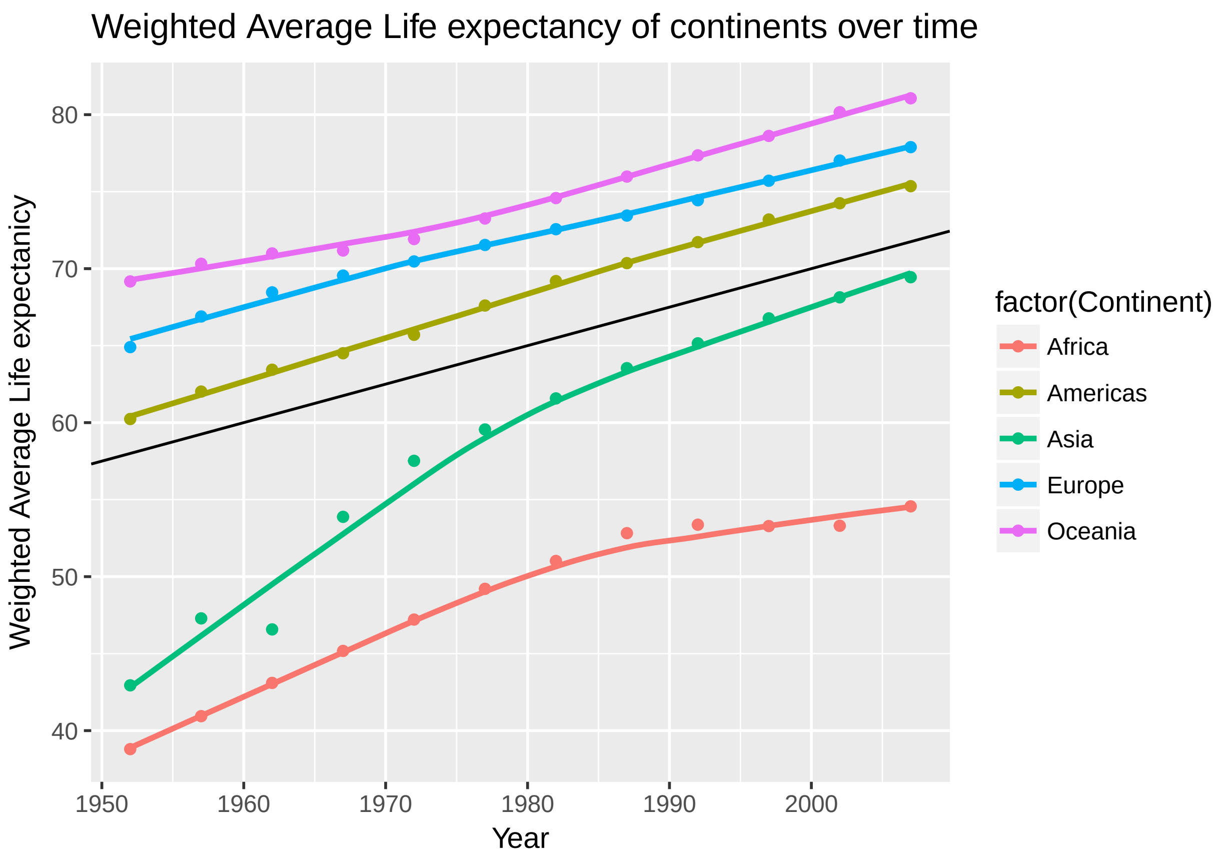
```
## `geom_smooth()` using method = 'loess'
```



Post the world war 2 even though the life expectancies of different continents are different but the rate of change of life expectancies over time (the slope in the above plot explains that) of Asia is highest and for Americas, Europe and Africa is almost the same (there is an additive shift between these continents). We cannot generalize this to Oceania as there are only 2 countries in it.

```
ggplot(ts3,aes(y=Avg_Le,x=Year, color = factor(Continent))) +
  geom_point() +
  geom_smooth(method.args = list(degree = 1), se = FALSE) +
  geom_abline(slope= 0.25,intercept = -430) +
  ggtitle("Weighted Average Life expectancy of continents over time") +
  ylab("Weighted Average Life expectanicy")
```

```
## `geom_smooth()` using method = 'loess'
```

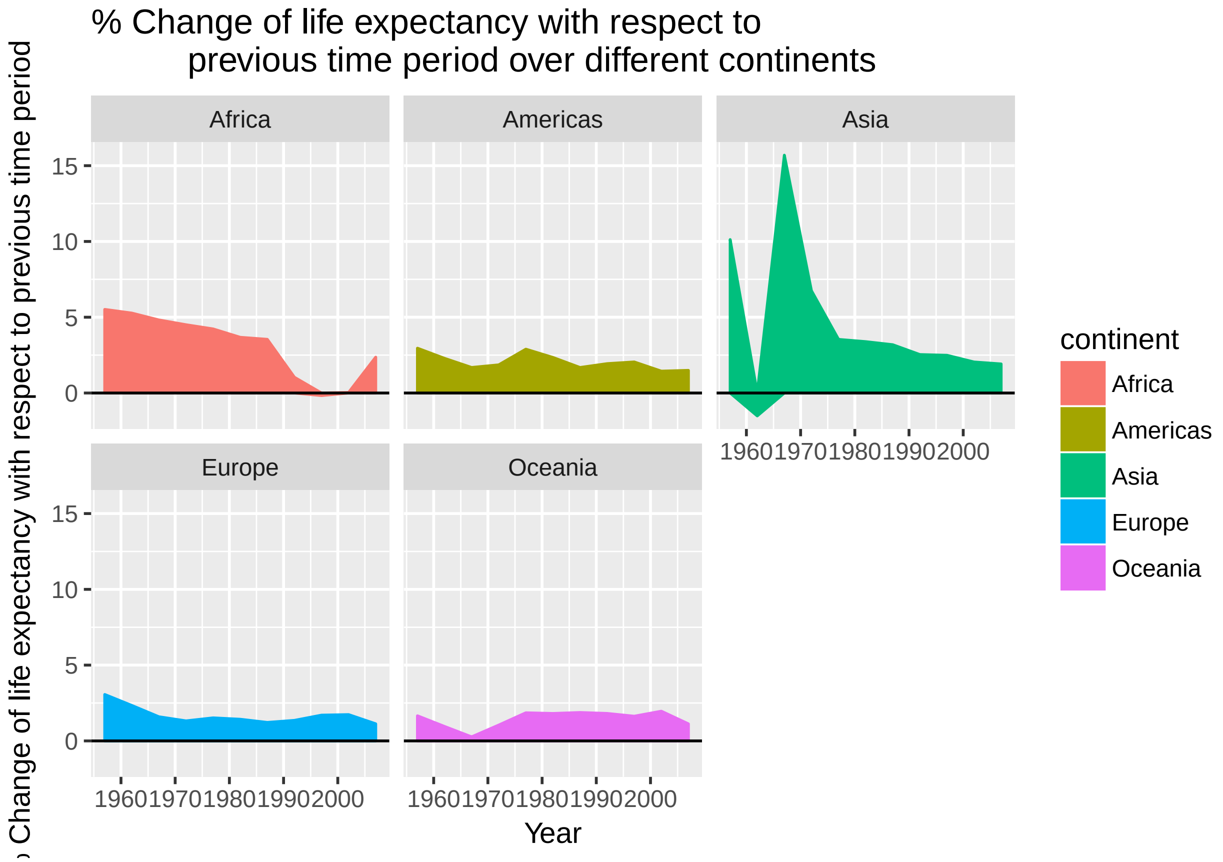


```
# We shall convert the data frame into short form and then convert into
#time series to find out the percentage changes of Average life expectancy
#for the next five years
tsa = ts3[,c("Year", "Continent", "Avg_Le")]
tsa$Year = factor(tsa$Year)
library(tidyr)

#wide form
tsaw = spread(tsa,Continent,Avg_Le)
tsawt = ts(tsaw, start = 1952, end = 2007,frequency = 1/5)
tsawtd = 100*(tsawt-lag(tsawt,-1))/lag(tsawt,-1)

perc = data.frame(x = as.matrix(tsawtd), year = time(tsawtd))
names(perc) = c("Yearc","Africa", "Americas", "Asia", "Europe", "Oceania", "Year")
perc = perc[,c("Year","Africa","Americas","Asia","Europe","Oceania")]
#the % change of Average life expectancy for each time period
#Converting to Long form for ggplot2
percc = gather(perc,key=continent, measurement,Africa:Oceania)
```

```
ggplot(percc,aes(x=Year, y= measurement, color = continent,fill = continent)) +
  geom_area() +  facet_wrap(~continent) + geom_hline(yintercept = 0.0) +
  ylab("% Change of life expectancy with respect to previous time period") +
  ggtitle("% Change of life expectancy with respect to
          previous time period over different continents")
```

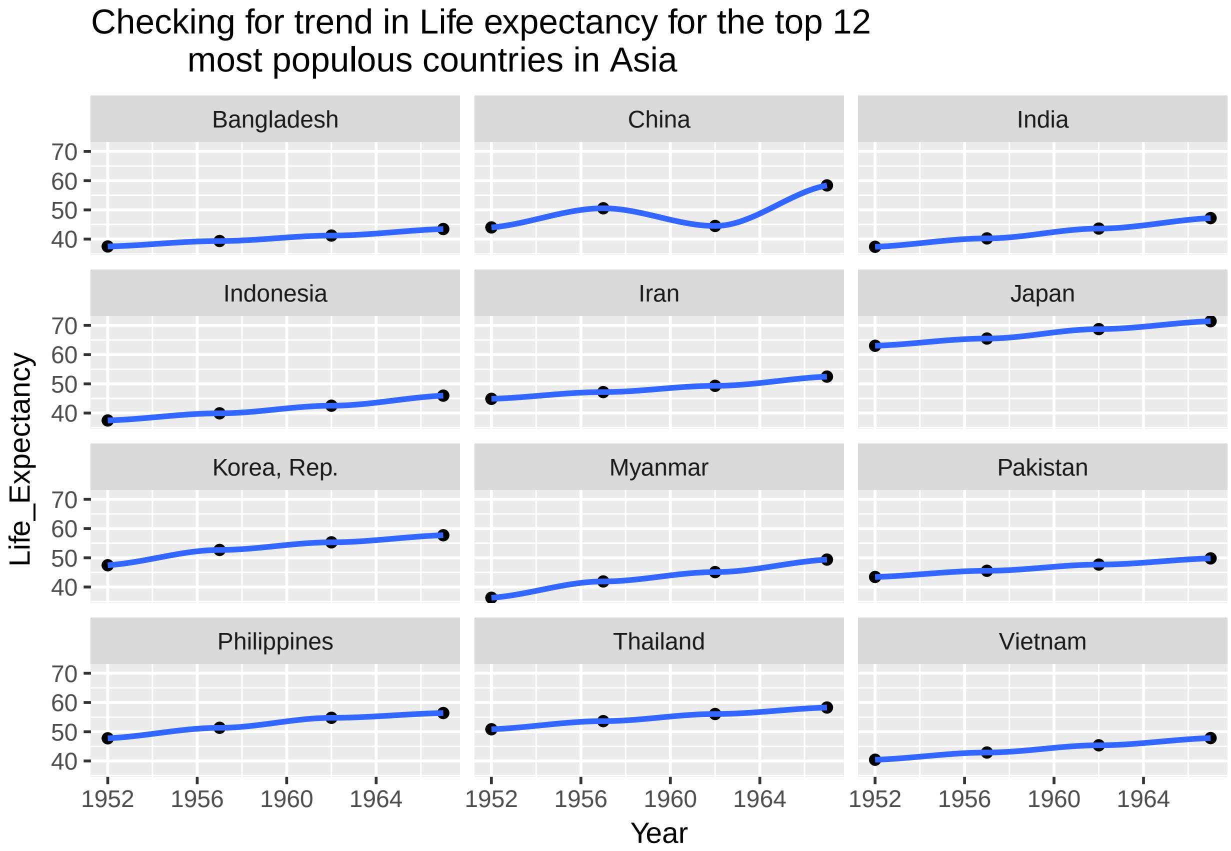


From this graph, we could say that the percentage increase in life expectancy for all the time periods of all the continents is positive except for Asia in 1962 and for Africa in 1992, but this is negligible when compared to that of Asia's.

```
d62 = subset(df,Year < 1972 ,
             select = c(Year,Country,Continent,Life_Expectancy,Population))
d62 = d62[d62$Continent=="Asia",]

ggplot(d62[d62$Population > 17000000,],aes(y=Life_Expectancy, x=Year)) +
  geom_point() +  geom_smooth() +
  facet_wrap(~Country,ncol=3) +
  ggtitle("Checking for trend in Life expectancy for the top 12
          most populous countries in Asia")
```

```
## `geom_smooth()` using method = 'loess'
```



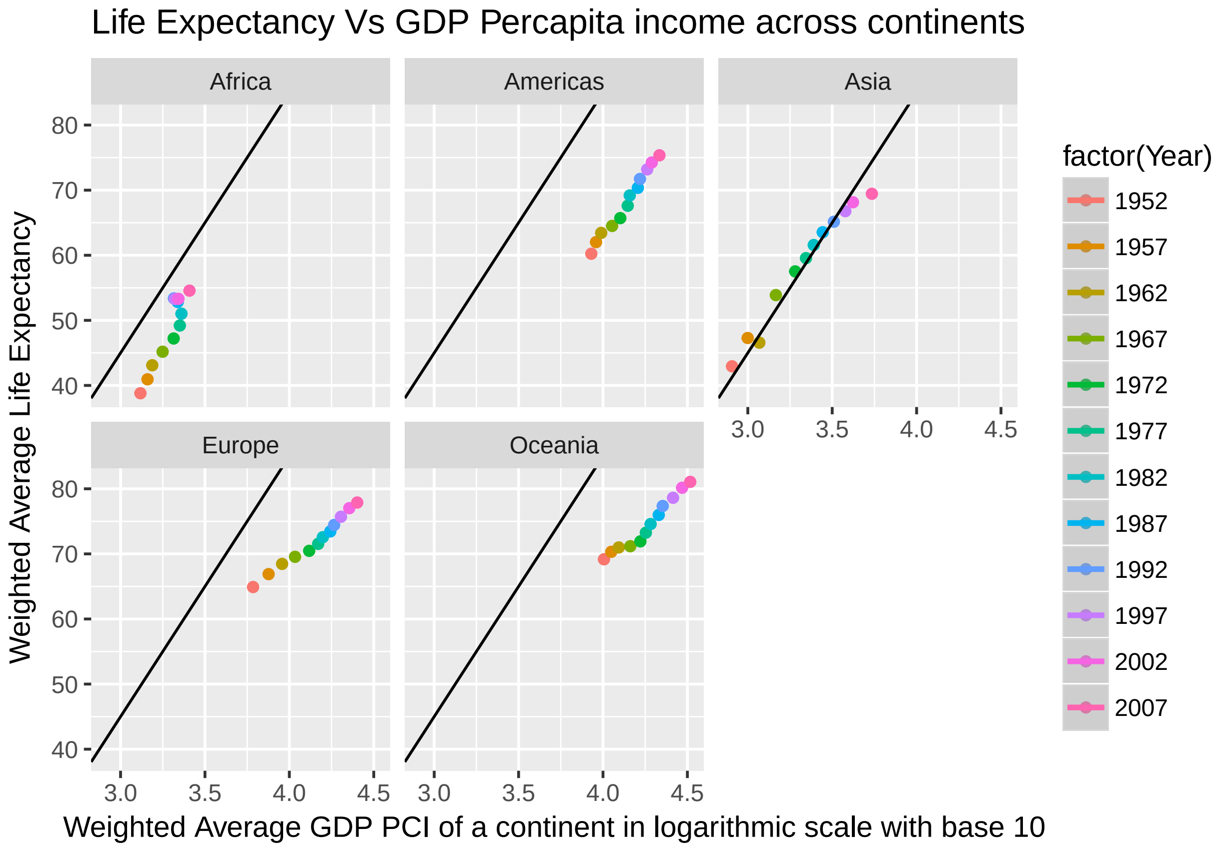
From this plot, we see that the most populous countries in Asia have a decline in Life expectancy in 1962. So, this might be a major reason as to why there is a decrease in the overall life expectancy in Asia as we have taken weighted average life expectancy.

Question-3

```r
ggplot(ts3,aes(y= Avg_Le, x= log10(Avg_GDP_PCI),group = factor(Year),
               color= factor(Year))) +
  geom_point() + geom_line() + geom_smooth() +
  facet_wrap(~Continent) + geom_abline(slope = 40, intercept = -75) +
  xlab("Weighted Average GDP PCI of a continent in logarithmic scale with base 10") +
  ylab("Weighted Average Life Expectancy") +
  ggtitle("Life Expectancy Vs GDP Percapita income across continents")
```

```
## `geom_smooth()` using method = 'loess'
```

```
## geom_path: Each group consists of only one observation. Do you need to
## adjust the group aesthetic?
## geom_path: Each group consists of only one observation. Do you need to
## adjust the group aesthetic?
## geom_path: Each group consists of only one observation. Do you need to
## adjust the group aesthetic?
## geom_path: Each group consists of only one observation. Do you need to
## adjust the group aesthetic?
## geom_path: Each group consists of only one observation. Do you need to
## adjust the group aesthetic?
```



This plot has a lot of information embedded in it. Each color of a point explains the year within which the measurment has been taken. If we look at different points through the X axis (if we project the points onto X-axis) we can say that the AVG GDP PCI for different points has increased overtime. If we look at different points through the Y axis (if we project the points onto Y-axis) we can say that AVG Life Expectancy for different points has been increased overtime. We can observe from the above plot that for every continent when there is an increase in GDPPercapita Income the Life expectancy is increasing (slope of the model defines that). For Africa this change is the highest, For Asia and Americas this rate has been almost equal(there is an additive shift between these two), For Europe it is some what lesser We cannot generalize on Oceania as there are only 2 countries in that Continent. The life expectancy at 1962 for Asia is less than that of 1957 as explained in Question-2.