Exploratory Data Analysis of the Relationship between Life Expectancy and Income per Person

Group Team Hungary: Jane Choi, Luis Miguel Mestre, Katharine Watson

Introduction

Life expectancy is the average period a person is expected to live if the death rate the year they were born is stable. Normally, life expectancy is used to measure how long a person can live in a particular country and it is most associated with the health of individuals. Prior studies have found that there is a relationship between life expectancy and income per person. The income per person is the average income a person earns in a particular period. Normally, it is measured with the gross domestic product (GDP) and the amount of national production in a period. In this project, we want to learn about the relationship between life expectancy and income per person. In particular, we are interested in this relationship since World War 2 (WW2). We used the Gapminder website (www.gapminder.org/data) to collect information about life expectancy per country, inflation-adjusted income per person through the GDP per capita, and the total population of countries from 1800 until 2018.

Objectives

To understand this relationship **three main objectives (MO)** will be achieved: (1) to analyze the relationship between GDP and life expectancy in 2018, (2) to determine life expectancy over time by continent, and (3) to determine changes in the relationship between GDP and life expectancy over time. These three main objectives will be accomplished answering three sets of questions/aims summarize in Table 1.

Table 1 - Set of questions/aims per main objectives

MO	Set of questions/aims
1	a) How does life expectancy vary with GDP per capita in 2018? b) Can the trends be well-described by a simple model such as a linear model, or is a more complicated model required? c) Is there pattern the same of different for every continent? If some continents are different, which ones, and how is the relationship different in those continents?
2	a) How has average life expectancy changed since WW2 in each continent? b) Have some continents caught up (or at least partially) to others? If so, is this just because of some countries in the continent, or is it more general? c) Have the changes been linear or has it been faster/slower in some periods for some continents? d) What might explain periods of faster/slower change?
3	a) How has the relationship between GDP and life expectancy changed in each continent since WW2? b) Can changes in life expectancy be entirely explained by changes in GDP per capita? c) Does it look like there's a time effect on life expectancy in addition to a GDP effect? d) Has there been "convergence" in the sense that perhaps GDP and/or continent don't matter as much as the used to? e) Are the exceptions to the general patterns?

Methods

The methods are described by the main objectives. All of our analyses were done using R software. For all objectives, we used a base 10 logarithmic transformation for the GDP per capita to understand some patterns clearly. The transformation of predictor variables (in our case GDP per capita) is a common practice among statisticians when there is a clear nonlinear pattern (e.g., logarithmic, quadratic, etc.) to "linearize" the trend. For the main objective (1) we used the *linear method* to measure the regression line of our data. The linear method (*lm*) of the *ggplot2 package*, fits a linear model that is closest to all the observations compared with the other possible lines. For the main objective (2) we calculated: (a) the population-*weighted means* of the life expectancy. The weights used accounted for the different populations of the countries within a continent. The weighted means are important when we need to calculate a mean as some data values seem to be more important than others' values (i.e. its contribution needs to have more "emphasis"). The standard deviations are used to see the dispersion among observations respect to the mean. In our case, we used the standard deviation

to verify the changes in the trends of continent wtih respect to the years. Changes in the dispersion might represent changes in the trend of the variable of interest. The regression lines for our plots were produced using the *loess* method of the ggplot2 package. The loess method (Locally Weighted Least Square) uses more local data to estimate the response variable.

For the main objective (3) we *categorized years since WW2 into decades* (1940s, 1950s, etc.) to measure the time effect in the relationship of GDP per capita and life expectancy. The regression line was calculated using the loess method. Because of the complexity of the graphs for the main objective (3), *the residuals plots* were verified for our models. The residuals plots show the residuals (error between the observed response variable and the predicted response variable) with respect to the predictor variable. If they were randomly dispersed around the x-axis then three assumptions might hold, (a) independence in the errors, (b) homoscedasticity (variances are constant across the error observations), and (c) normality (residuals follows a standard normal distribution).

Results: Results will be given per objective per aim.

Analyze the relationship between GDP and life expectancy in 2018

How does life expectancy vary with GDP per capita in 2018?

Life expectancy increases rapidly when the GDP per capita is lower than or equal to $25,000. However, above GDP per capita of $25,000, the rate of life expectancy increase starts to slow. Figure 1 suggests that the association between life expectancy and GDP per capita is logarithmic. Therefore, GDP per capita was logarithmically transformed. If they have a logarithmic relationship, the log transform of the GDP per capita should have a linear relationship with life expectancy and this is evident in Figure 2.

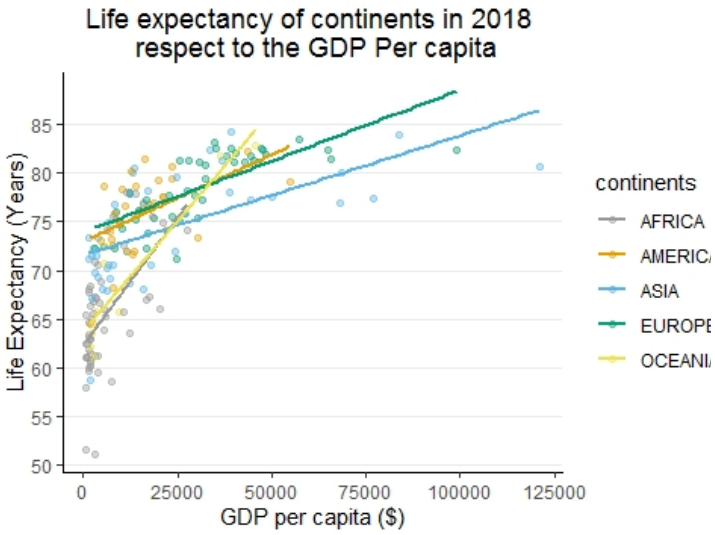


Figure 1 - Life expectancy of continents in 2018 respect to the GDP per capita

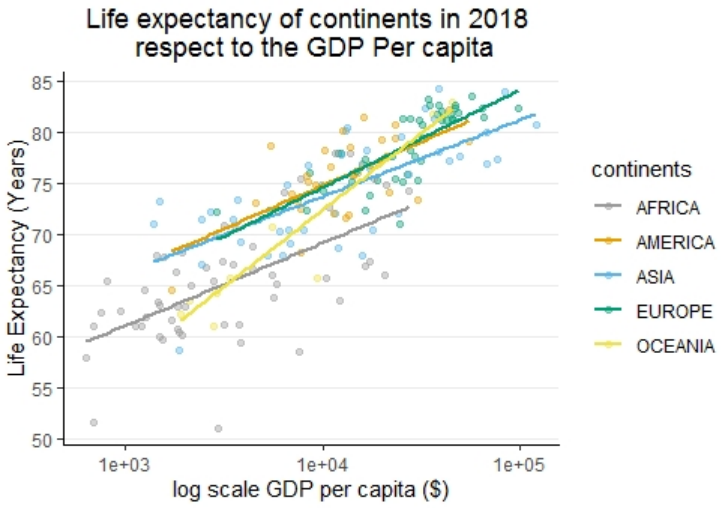


Figure 2 - Life expectancy of continents in 2018 respect to the GDP per capita in log scale

Can the trends be well-described by a simple model such as a linear model, or is a more complicated model required?

After the log transformation of the GDP per capita, the trends can be well described by a simple linear model (Figure 2).

Is there pattern the same of different for every continent? If some continents are different, which ones, and how is the relationship different in those continents?

The pattern is similar between Europe, Asia, and America. Their log scale GDP per capita increases linearly as life expectancies increase. Africa also increases in the same linear manner but the life expectancy curve is

shifted lower. Oceania, has a greater increase in life expectancy than any other continent for each incremental change in GDP per capita.

Determine life expectancy over time by continent

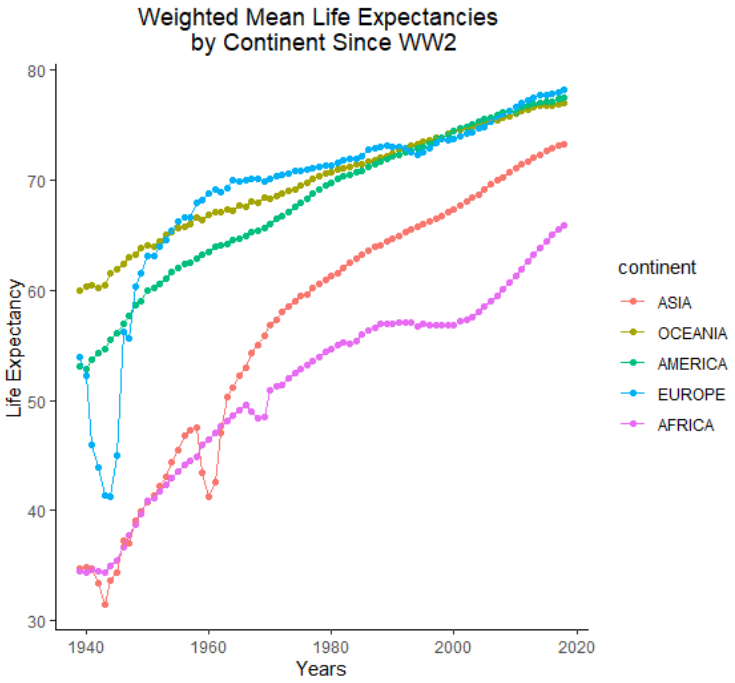
How has average life expectancy changed since WW2 in each continent?

The weighted mean life expectancy (MLE) in every continent has increased since 1939. Between 1939 and 2018, the unweighted MLE in Africa increased from 34.5 to 65.9 years (Table 3). From 1985 to 2000, the increase in weighted MLE was minimal in all continents except Africa, which has some small decreasing period of life expectancy. Between 1939 and 2018, the weighted MLE in Oceania increased from 60.0 years to 77.0 years. This is the smallest change of all continents over time (Table 3 and Figure 2). Oceania had the highest weighted MLE in 1939 while the 2018 MLE was similar to that of Europe and America. Between 1939 and 2018, the weighted MLE in Asia increased from 34.7 to 73.3, the largest gain of any continent. The weighted MLE in Europe increased from 54.0 to 78.2 years. In 2018, Europe had the second highest weighted MLE in 1939 and is now the continent with the highest weighted MLE. In the Americas, between 1939 and 2018, the weighted MLE increased from 53.1 to 77.5 years.

Table 2 – *Weighted MLEs* since WW2 per continent

	Weighted MLE (years)		
Continent	**1939**	**2018**	**Change**
Africa	34.5	65.9	31.4
Oceania	60.0	77.0	17.0
Asia	34.7	73.3	38.6
Europe	54.0	78.2	24.2
Americas	53.1	77.5	24.4

Figure 3 - Mean life expectancy since WW2 per continent



Have some continents caught up (or at least partially) to others? If so, is this just because of some countries in the continent, or is it more general?

Some continents like Oceania, America and Oceania are very similar in 2018. Asia appears to be catching up. However, Africa is the lowest and it appears that it make take longer for the life expectancy to catch up with Europe, Americas and Oceania. Oceania has caught up and this is, largely, because about 70% of the population of that continent is Australia and New Zealand, which had the highest life expectancy worldwide. However, most of the countries in Oceania are consider developing countries with respect to their GDP. As previously discussed, low GDP is associated with lower life expectancies. Most of the countries in Asia have relatively similar life expectancies. Those countries that have higher life expectancies like Japan, South Korea and Taiwan do not represent a significant proportion of the Asian population. As the larger countries in Asia (e.g. China) are increasing their life expectancies, the population-weighted life expectancy is approaching that of Oceania, Europe and the Americas.

Have the changes been linear or has it been faster/slower in some periods for some continents?

The MLE change in Europe and the Americas has been somewhat logistic, while in Africa, Oceania and Asia the change has been somewhat linear (Figure 3). However, there are some periods where life expectancy does not follow the trend of the changes. This includes Europe between 1939 and 1945, Asia during the 1960s, and Africa between the mid-1980s until the mid-2000s.

What might explain periods of faster/slower change?

Asia and Africa did not have much MLE change from 1939 to 1945 when there was an abrupt 8-10 year increase in these continents. The rate of increase in Asia has been relatively logistic since then. In Africa, the rate

plateaued from 1985 to 2005, then began to increase again, likely corresponding to the AIDS epidemic and the Rwanda genocide. The MLE change in Europe was irregular between 1939 and 1945, likely due to the Holocaust. After the war ended, the MLE followed a logistic form.

The weighted MLE change in Europe was irregular between 1939 and 1945, likely due to the Holocaust. In Asia there were an irregular behavior in the pattern during the 1960s, likely due to the internal conflicts in countries in the continent such as India, or Iran or wars such as the Vietnam War. The change of pattern during early 2000s in Africa might be due to the Rwanda genocide. After the war ended, the MLE followed a somewhat logistic form. The high weighted life expectancies in Europe, Oceania, and the Americas may be related to healthier lifestyles than other continents, higher GDP per capita, and better access to healthcare.

Determine changes in the relationship between GDP and life expectancy over time

How has the relationship between GDP and life expectancy changed in each continent since WW2?

Since WW2, for all continents, the log scale GDP has increased over time as life expectancy has increased (Figure 4). In other words, the more time has passed since the end of WW2, the more the life expectancy and the log scale GDP per capita has increased.

Can changes in life expectancy be entirely explained by changes in GDP per capita?

Changes in life expectancy cannot be entirely explained by changes in the log scale GDP per capita. It seems time might play a role in the relationship between log scale GDP per capita and life expectancy. Figure 4 shows that some decades have different trends from others, even though they are all linear for all the continents. The discrepancies in trend or change between the decades might be due to events during those periods within a continent. Therefore, we need time, GDP per capita and also the continent to explain life expectancy.

Does it look like there's a time effect on life expectancy in addition to a GDP effect?

It appears that there is a time effect that might influence the life expectancy in addition to a GDP effect or continent effect (Figure 4). There is a pattern where the year increases and the life expectancy also increases.

Has there been "convergence" in the sense that perhaps GDP and/or continent don't matter as much as the used to?

GDP continues to be an important contributor to life expectancy. Continents appear to be becoming more similar, although Africa continues to have a lower life expectancy than other continents.

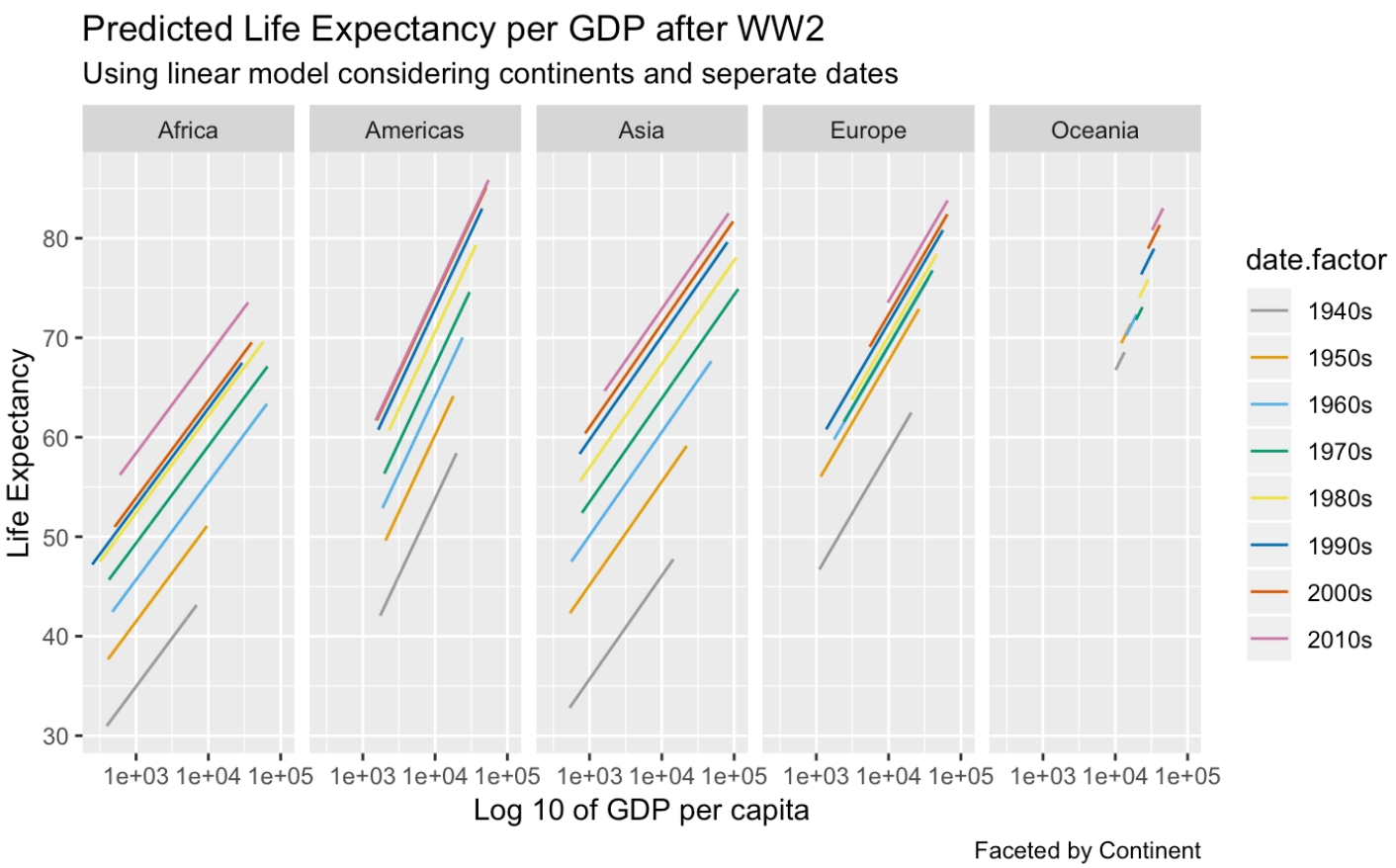


Figure 4 - Life expectancy respect to log scale GDP per capita per continent since WW2

Are there exceptions to the general pattern?

Africa continues to have the lowest life expectancy with respect to log GDP per capita per continent across all decades. Oceania has the smallest range of GDP. They also have the smallest range of life expectancy.

Discussion

The main objectives of this project were different approaches to understanding and exploring the relationship between GDP per capita and life expectancy. The relationship between GDP per capita and life expectancy from 1939 to 2018 was logarithmic. The log transformation of GDP per capita showed that life expectancy has a linear relationship with log scale GDP per capita. This linear relationship also occurs among the continents. America, Africa, and Asia have a very similar linear trend. Oceania had a similar linear trend, but the life expectancy increase was greater with each increase in GDP. Africa had the lowest life expectancies with respect to log GDP per capita.

Considering the relationship between GDP per capita and life expectancies among continents since WW2, America, Oceania, and Europe started in different weighted MLE but have converged. However, Asia and Africa are somewhat behind, despite that Asia was the continent with the highest change in life expectancy since WW2. Another interesting finding was that Europe, Asia, and Africa had periods that did not follow the general pattern they had through time. Those periods correspond to the end of WW2, the AIDS pandemic, the Rwanda Genocide, the Vietnam Wars, and revolutions among the western world and Middle East in Asia.

The time effect is one of the main variables that influenced life expectancy. As more time has passed since the end of WW2, life expectancy has increased in all the continents. However, the time effect alone does not explain the life expectancy phenomena. Time with GDP per capita and continents might better explain the life expectancy behavior. In conclusion, the relationship between GDP per capita and life expectancy is affected by time effect

and the continents. Social and global events through time and place might affect a country and the GDP per capita of that country may play a role in explaining the behavior of life expectancy since WW2.

Limitations

We did not consider the missing data for small countries as Holy See, St. Lucia and others, which may result in information “bias” in continent. It was detected that time affects the relationship between GDP per capita and life expectancy. However, which component of time that directly influences the relationship is still unknown. Finally, the residual plots for Figure 4 show that residuals do not behave ideally. Rather than using a linear line through the linear method, the loess method was used to develop a “line” with some curves to adjust all the residuals (see appendix). This means that our model slightly does not meet all assumptions. Therefore, interpretations of this work should take into consideration the lack of information for some small or some developing countries and the limitations of our models.

Appendix

```
setwd("C:\\Users\\lmestre\\Documents")


#Libraries
library(pastecs)
library(egg)
library(cowplot)
library(car)
library(tidyr)
library(reshape2)
library(tidyverse)
library(ggplot2)
library(gapminder)
library(reshape2)
library(data.table)


dt.lifexpect <- fread("life_expectancy_years.csv", header = TRUE) #life expectancy of countries through years
dt.incomepercapita <- fread("income_per_person_gdppercapita_ppp_inflation_adjusted.csv", header = TRUE)
dt.population <- fread("population_total.csv", header = TRUE)
dt.continent <- fread("Country  with its continent MP1 Team Hungary.csv",header = TRUE)



#Adding a continent column to all datasets
dt.lifexpect$continents <- rep(NA, dim(dt.lifexpect)[1])
```

```
dt.incomepercapita$continents <- rep(NA, dim(dt.incomepercapita)[1])
dt.population$continents <- rep(NA, dim(dt.population)[1])

#Organizing the columns so continent is the first one
dt.lifexpect <- dt.lifexpect[,c(221, 1:220)]
dt.incomepercapita <- dt.incomepercapita[,c(243, 1:242)]
dt.population <- dt.population[,c(303, 1:302)]

#Establishing country per continent
k.2018 <- dim(dt.lifexpect)[2]

#Redoing new data based on files. We only want data until 2018
dt.lifexpect <- dt.lifexpect[,1:k.2018]
dt.incomepercapita <- dt.incomepercapita[,1:k.2018]
dt.population <- dt.population[,1:k.2018]
# dt.continentmat <- dt.continentmat[,1:k.2018]
dt.continent <- dt.continent[,1:k.2018]

#To rbind they need to have the same column names.
#names(dt.continent) == names(dt.population)

l <- dim(dt.lifexpect)[1]
i <- dim(dt.incomepercapita)[1]
p <- dim(dt.population)[1]
# c <- dim(dt.continentmat)[1]
c <- dim(dt.continent)[1]
#Merging all datasets
# dt.analysis <- data.frame(rbind(dt.lifexpect, dt.incomepercapita, dt.population, dt.continentmat))
dt.analysis <- data.frame(rbind(dt.lifexpect, dt.incomepercapita, dt.population, dt.continent))
dt.analysis$variable <- rep(0, length(dt.analysis$country))
dt.analysis$variable[1:l] <- "LE" #Life expectancy dataset
dt.analysis$variable[sum(l,1):sum(i,l)] <- "GDPpercap" #Incomepercapita dataset
dt.analysis$variable[sum(l,i,1): sum(l,i,p)] <- "POP" #Population dataset
```

```
dt.analysis$variable[sum(l,i,p,1): sum(l,i,p,c)] <- "continents"
dt.analysis <- dt.analysis[order(dt.analysis$country),]
dt.analysis <- dt.analysis[,-c(3:141)] #Only data since WW2 (1939 - 2018). In other words, years from 1800 until
#1938 were deleted because they will not be used in the analsysis

dt.analysis$year <- rep(0, length(dt.analysis$country))

dt.analysis <- dt.analysis[,c(1, 2, 83,84, 3:82)] #re arranging the columns
# write.csv(dt.analysis, "dt.analysiswrite.csv")
# dt.analysis <- fread("dt.analysiswrite.csv", header = TRUE)

# dt.analysis$continent <- dt.continentlong$continent

col.val <- c("X1939",
             "X1940",
             "X1941",
             "X1942",
             "X1943",
             "X1944",
             "X1945",
             "X1946",
             "X1947",
             "X1948",
             "X1949",
             "X1950",
             "X1951",
             "X1952",
             "X1953",
             "X1954",
             "X1955",
             "X1956",
             "X1957",
             "X1958",
```

"X1959",
"X1960",
"X1961",
"X1962",
"X1963",
"X1964",
"X1965",
"X1966",
"X1967",
"X1968",
"X1969",
"X1970",
"X1971",
"X1972",
"X1973",
"X1974",
"X1975",
"X1976",
"X1977",
"X1978",
"X1979",
"X1980",
"X1981",
"X1982",
"X1983",
"X1984",
"X1985",
"X1986",
"X1987",
"X1988",
"X1989",
"X1990",
"X1991",

```
        "X1992",
        "X1993",
        "X1994",
        "X1995",
        "X1996",
        "X1997",
        "X1998",
        "X1999",
        "X2000",
        "X2001",
        "X2002",
        "X2003",
        "X2004",
        "X2005",
        "X2006",
        "X2007",
        "X2008",
        "X2009",
        "X2010",
        "X2011",
        "X2012",
        "X2013",
        "X2014",
        "X2015",
        "X2016",
        "X2017",
        "X2018"
)

dt.analysislong <- melt(dt.analysis, id.vars = c("variable", "country"), measure.vars = col.val, variable.name = "year")

# dt.continentlong <- melt(dt.continent, id.vars = c("country"), measure.vars = c("continent1", "continent2", "continent3"), variable.name = "continent")

# dt.continentlong <- dt.continentlong[order(dt.continentlong$country),]
```

```
dt.analysiswide <- spread(dt.analysislong, key = variable, value = value)
dt.analysiswide$LE <- as.numeric(dt.analysiswide$LE)
dt.analysiswide$GDPpercap <- as.numeric(dt.analysiswide$GDPpercap)
dt.analysiswide$POP <- as.numeric(dt.analysiswide$POP)



#------------------------------Trying to create the weight means

# gapminder <- gapminder
dfdummy <- dt.analysiswide
dfdummy$LEPop <- dfdummy$POP*dfdummy$LE
dfdummy$continents <- as.factor(dfdummy$continents)
dfdummy$year <- as.numeric(gsub("X", " ", dfdummy$year))



Asia <- subset(dfdummy, continents == "ASIA")
Americas <- subset(dfdummy, continents == "AMERICA")
Oceania <- subset(dfdummy, continents == "OCEANIA")
Africa <- subset(dfdummy, continents == "AFRICA")
Europe <- subset(dfdummy, continents == "EUROPE")

dfAsia <- NULL
dfOceania <- NULL
dfAmericas <- NULL
dfEurope <- NULL
dfAfrica <- NULL

k <- 1938 #initial year (1939-1)
l <- 2018-1939+1

for(i in 1:l)
{
```

```
#-----------Asia
m <- Asia$year == k+i
dfAsia$year[i] <- k+i
dfAsia$POP[i] <- sum(Asia$POP[m], na.rm = TRUE)
if(dfAsia$POP[i] == 0)
{
  dfAsia$weightMLE[i] <- 0
}
else
{
  dfAsia$weightMLE[i] <- sum(Asia$LEPop[m], na.rm = TRUE)/dfAsia$POP[i]
}



#------------Oceania
m <- Oceania$year == k+i
dfOceania$year[i] <- k+i
dfOceania$POP[i] <- sum(Oceania$POP[m], na.rm = TRUE)
if(dfOceania$POP[i] == 0)
{
  dfOceania$weightMLE[i] <- 0
}
else
{
  dfOceania$weightMLE[i] <- sum(Oceania$LEPop[m], na.rm = TRUE)/dfOceania$POP[i]
}

#---------- Americas
m <- Americas$year == k+i
dfAmericas$year[i] <- k+i
dfAmericas$POP[i] <- sum(Americas$POP[m], na.rm = TRUE)
if(dfAmericas$POP[i] == 0)
{
```

```
  dfAmericas$weightMLE[i] <- 0
 }
 else
 {
  dfAmericas$weightMLE[i] <- sum(Americas$LEPop[m], na.rm = TRUE)/dfAmericas$POP[i]
 }

 # ----------- Europe
 m <- Europe$year == k+i
 dfEurope$year[i] <- k+i
 dfEurope$POP[i] <- sum(Europe$POP[m], na.rm = TRUE)
 if(dfEurope$POP[i] == 0)
 {
  dfEurope$weightMLE[i] <- 0
 }
 else
 {
  dfEurope$weightMLE[i] <- sum(Europe$LEPop[m], na.rm = TRUE)/dfEurope$POP[i]
 }

 #---------- Africa
 m <- Africa$year == k+i
 dfAfrica$year[i] <- k+i
 dfAfrica$POP[i] <- sum(Africa$POP[m],na.rm = TRUE)
 if(dfAfrica$POP[i] == 0)
 {
  dfAfrica$weightMLE[i] <- 0
 }
 else
 {
  dfAfrica$weightMLE[i] <- sum(Africa$LEPop[m],na.rm = TRUE)/dfAfrica$POP[i]
 }
}
```

```
dfAsia$continent <- "ASIA"
dfAsia$continent <- as.factor(dfAsia$continent)
dfAsia <- as.data.frame(dfAsia)

dfOceania$continent <- "OCEANIA"
dfOceania$continent <- as.factor(dfOceania$continent)
dfOceania <- as.data.frame(dfOceania)

dfAmericas$continent <- "AMERICA"
dfAmericas$continent <- as.factor(dfAmericas$continent)
dfAmericas <- as.data.frame(dfAmericas)

dfEurope$continent <- "EUROPE"
dfEurope$continent <- as.factor(dfEurope$continent)
dfEurope <- as.data.frame(dfEurope)

dfAfrica$continent <- "AFRICA"
dfAfrica$continent <- as.factor(dfAfrica$continent)
dfAfrica <- as.data.frame(dfAfrica)

dtgg <- rbind(dfAsia, dfOceania, dfAmericas, dfEurope, dfAfrica)



#----Plot change by Katie

ggplot(dtgg, aes(x = year, y = weightMLE, color = continent)) + geom_point() + geom_line() +
  theme_classic() + ggtitle("Weighted Mean Life Expectancies \n by Continent Since WW2") +
  xlab("Years") + ylab("Life Expectancy") +
  theme(panel.grid.minor.y = element_blank(), panel.grid.minor.x =
          element_blank(),panel.grid.major.x =
          element_blank(), panel.border = element_blank(), axis.line.x = element_line(size = 0.5,
          linetype = "solid", colour = "black"), axis.line.y = element_line(size = 0.5,
```

```
        linetype = "solid", colour = "black"), plot.title = element_text(hjust = 0.5))

# ----- Descriptive statistics for the MLE of countries in 1939 and 2018

#Descriptive Statistics for 1939
dtgg[dtgg$year == 1939,]
summary(dtgg[dtgg$year == 1939,])
stat.desc(dtgg[dtgg$year == 1939,])

#Descriptive Statistics for 2018
dtgg[dtgg$year == 2018,]
summary(dtgg[dtgg$year == 2018,])
stat.desc(dtgg[dtgg$year == 2018,])

##########
setwd("/Users/janechoi/Desktop/IU/DATA/miniproject")
library(gapminder)
library(tidyverse)
library(rio)
library(data.table)
library(GGally)
library(mgcv)
library(modelr)
library(broom)
library(arm)
library(plotly)



cb_palette <- c("#999999", "#E69F00", "#56B4E9", "#009E73", "#F0E442", "#0072B2", "#D55E00", "#CC79A7")

life.expectancy =fread('life_expectancy_years.csv',header = T)
population<-fread('population_total.csv',header = T)
gdp<-fread('income_per_person_gdppercapita_ppp_inflation_adjusted.csv',header = T)
```

```
gapminder <- gapminder



#unique(population$country) #193

#unique(gapminder$country) #142

#unique(gdp$country) #193

#unique(life.expectancy$country) #187


gapminder.new<- gapminder[,c(1,2)]


gapminder.new<- unique(gapminder.new)


new.gdp <- merge(gdp, gapminder.new, by='country')

new.life.expectancy <- merge(life.expectancy, gapminder.new, by='country')

new.population <- merge(population, gapminder.new, by='country')




life.expectancy.long = gather(new.life.expectancy, key = life.expectancy.date , value = life.expectancy.number, 2:220)

gdp.long <- gather(new.gdp, key= gdp.date, value= gdp.number , 2:242)


population.long <- gather(new.population, key=population.date, value=population.number, 2:302)



gdp.long$key <- paste(gdp.long$country , by='-',gdp.long$gdp.date)

population.long$key <- paste(population.long$country , by='-',population.long$population.date)

life.expectancy.long$key <- paste(life.expectancy.long$country , by='-',life.expectancy.long$life.expectancy.date)



data<- full_join(gdp.long, life.expectancy.long, by='key')

data<- full_join(data, population.long, by='key')
```

```
data<-data[,-c(6:8,10:12)]

colnames(data) <- c("country" , "continent","date","gdp.number", "key" ,"life.expectancy.number", "population.number")



data<- data[,c(5,3,1,2,4,6,7)]

#head(data)

colSums(is.na(data))

data <- data[is.na(data$life.expectancy.number) == F,]
colSums(is.na(data))




#write.csv(data, file='data.csv',row.names=F)
#data.new<-read.csv('data.csv')

data.ww2 <- data %>%
  filter(data$date > 1939)
#####making date.factor
data.ww2$date.factor <- 0

i=0

for ( i in 1:10586){
  if (data.ww2[i,2] <= 1949 ) {
    data.ww2[i,8] <- '1940s'}

  else if (data.ww2[i,2] <= 1959){
```

```
    data.ww2[i,8] <- '1950s'
  }
   else if (data.ww2[i,2] <= 1969){
    data.ww2[i,8] <- '1960s'
   }
   else if (data.ww2[i,2] <= 1979){
    data.ww2[i,8] <- '1970s'
   }
   else if (data.ww2[i,2] <= 1989){
    data.ww2[i,8] <- '1980s'
   }
   else if (data.ww2[i,2] <= 1999){
    data.ww2[i,8] <- '1990s'
   }
   else if (data.ww2[i,2] <= 2009){
    data.ww2[i,8] <- '2000s'
   }
   else{ data.ww2[i,8] <- '2010s'
  }
}



data.ww2$date<-as.factor(data.ww2$date)


data.ww2$date.factor<-as.factor(data.ww2$date.factor)


str(data.ww2)
####seeing plots
#plot_ly(data.ww2, x= ~log(gdp.number) ,y=~life.expectancy.number , z=~date, type= 'scatter3d' , marker= list(size=1.5))
ggpairs(data.ww2[,c(5:6,8)])
```

```
####1. Method: Modeling without continent information:

###(1-1) model with linear:

data.ww2.lm = lm(life.expectancy.number ~ log10(gdp.number) + date, data = data.ww2 )

data.ww2.lm.df = augment(data.ww2.lm)


colnames(data.ww2.lm.df) = c("life.expectancy.number" , "gdp.number" ,  "date"  ,   ".fitted"  , ".se.fit" ,  ".resid" ,
".hat" ,   ".sigma" ,".cooksd"  ,   ".std.resid"  )


ggplot(data.ww2.lm.df, aes(x = gdp.number, y = .resid)) +
  geom_smooth(se=F) +geom_point(alpha=0.1) +
  ggtitle('Residual Plot for  Linear Model ') +
  labs(x='log10 of gdp')




ggplot(data.ww2.lm.df, aes(x = .fitted, y = abs(.resid))) +
  geom_smooth(se=F) +geom_point(alpha=0.1)+
  ggtitle('Residual & Fitted Plot for Linear Model ') +
  labs(x='log10 of gdp')

  ##making prediction

  ww2.lm.grid <-expand.grid(gdp.number = seq(5.5, 11.6, 0.1), date = seq(1940,2018,1))

  ww2.lm.grid$date<-as.factor(ww2.lm.grid$date)



  ww2.lm.grid.predict = predict(data.ww2.lm, newdata = ww2.lm.grid)
  ww2.lm.grid.predict = data.frame(ww2.lm.grid, life.expectancy = as.vector(ww2.lm.grid.predict))



  #ggplot(ww2.lm.grid.predict, aes(x = gdp.number, y = life.expectancy)) +  geom_line() +facet_wrap(~date)
```

```
  ggplot(ww2.lm.grid.predict, aes(x = gdp.number, y = life.expectancy, group = date, color = date)) + geom_line()
+

    ggtitle('Predicted Life Expectancy per GDP after WW2' ) +

  labs(x= 'Log 10 of GDP per capita' , y= 'Life Expectancy', subtitle = 'Using linear model with seperate dates')




  ###(1-2) model with liner : using date.factors

  data.ww2.lm.f = lm(life.expectancy.number ~ log10(gdp.number) + date.factor, data = data.ww2)

  data.ww2.lm.f.df = augment(data.ww2.lm.f)


  colnames(data.ww2.lm.f.df) = c("life.expectancy.number" , "gdp.number" ,  "date"  ,   ".fitted"  , ".se.fit" ,  ".resid"
, ".hat" ,   ".sigma" ,".cooksd"  ,   ".std.resid"  )


  ggplot(data.ww2.lm.f.df, aes(x = gdp.number, y = .resid)) +

    geom_smooth(se=F) +geom_point(alpha=0.1) +

    ggtitle('Residual Plot for  Linear Model ') +

    labs(x='log10 of gdp')




  ggplot(data.ww2.lm.df, aes(x = .fitted, y = abs(.resid))) +

    geom_smooth(se=F) +geom_point(alpha=0.1)+

    ggtitle('Residual & Fitted Plot for Linear Model ') +

    labs(x='log10 of gdp')


  ###predictions

  ww2.lm.f.grid <-expand.grid(gdp.number = seq(5.5, 11.6, 0.1), date.factor =
c('1940s','1950s','1960s','1970s','1980s','1990s','2010s'))


  ww2.lm.f.grid$date.factor<-as.factor(ww2.lm.f.grid$date.factor)
```

```
ww2.lm.f.grid.predict = predict(data.ww2.lm.f, newdata = ww2.lm.f.grid)
ww2.lm.f.grid.predict = data.frame(ww2.lm.f.grid, life.expectancy = as.vector(ww2.lm.f.grid.predict))



#ggplot(ww2.lm.f.grid.predict, aes(x = gdp.number, y = life.expectancy)) + geom_line()
+facet_wrap(~date.factor)


ggplot(ww2.lm.f.grid.predict, aes(x = gdp.number, y = life.expectancy, group = date.factor, color = date.factor))
+ geom_line()+
  ggtitle('Predicted Life Expectancy per GDP after WW2' ) +
  labs(x= 'Log 10 of GDP per capita' , y= 'Life Expectancy', subtitle = 'Using linear model with dates together')+
  scale_color_manual(values=cb_palette)



## Check the R-squared value for each model:


summary(data.ww2.lm) #Adjusted R-squared:  0.7931
summary(data.ww2.lm.f) #Adjusted R-squared:  0.7877



##2. Method: Modeling with different continent information:


###(2-0) not using date when modeling with linear:
continent.lm.no.date.f = function(data){
  lm(life.expectancy.number ~ log10(gdp.number), data = data)
}
continent.lm.m.no.date.f = data.ww2 %>%
  group_by(continent) %>%
  nest()



continent.lm.m.no.date.f.m = map(continent.lm.m.no.date.f$data, continent.lm.no.date.f)
```

```
continent.lm.m.no.date.f = mutate(continent.lm.m.no.date.f, model = continent.lm.m.no.date.f.m)

continent.lm.m.no.date.f = mutate(continent.lm.m.no.date.f, .fitted = map2(data, model, add_predictions))

continent.lm.m.no.date.f = mutate(continent.lm.m.no.date.f, .resid = map2(data, model, add_residuals))

continent.lm.m.no.date.f.fitted = unnest(continent.lm.m.no.date.f, .fitted,.resid)



continent.lm.m.no.date.f.fitted<-continent.lm.m.no.date.f.fitted[,-c(10:16)]


ggplot(continent.lm.m.no.date.f.fitted , aes(x=(gdp.number),y= resid))+
  geom_point()+ geom_smooth(method='loess',se=F)+
  scale_x_log10() +
  ggtitle("Residual Plot for Gdp per Capita and Life Expectancy")+
  labs(x='log 10 of gdp per capita')


ggplot(continent.lm.m.no.date.f.fitted , aes(x=(gdp.number),y= abs(resid))) +
  geom_point()+ geom_smooth(method='loess',se=F)+
  scale_x_log10() +
  ggtitle("Absolute Residual Plot for Gdp per Capita and Life Expectancy")+
  labs(x='log 10 of gdp per capita')

continent.lm.m.no.date.f.fitted.df<-as.data.frame(continent.lm.m.no.date.f.fitted)


ggplot(continent.lm.m.no.date.f.fitted.df, aes(x = gdp.number, y = pred, group=continent , color= continent)) +
  geom_line()+scale_x_log10() +
  ggtitle('Predicted Life Expectancy per GDP after WW2' ) +
```

```
  labs(x= 'Log 10 GDP per capita' , y= 'Life Expectancy', subtitle = 'Using linear model by continents')+
  scale_color_manual(values=cb_palette)


###(2-1) model with linear:
continent.lm = function(data){
  lm(life.expectancy.number ~ log10(gdp.number) + date, data = data)
}
continent.lm.m = data.ww2 %>%
  group_by(continent) %>%
  nest()


continent.lm.m.m = map(continent.lm.m$data, continent.lm)

continent.lm.m = mutate(continent.lm.m, model = continent.lm.m.m)
continent.lm.m = mutate(continent.lm.m, .fitted = map2(data, model, add_predictions))
continent.lm.m = mutate(continent.lm.m, .resid = map2(data, model, add_residuals))
continent.lm.m.fitted = unnest(continent.lm.m, .fitted,.resid)

continent.lm.m.fitted<-continent.lm.m.fitted[,-c(10:16)]


ggplot(continent.lm.m.fitted , aes(x=(gdp.number),y= resid))+
  geom_point()+ geom_smooth(method='loess',se=F)+
  scale_x_log10() +
  ggtitle("Residual Plot for Gdp per Capita and Life Expectancy")+
  labs(x='log 10 of gdp per capita')


ggplot(continent.lm.m.fitted , aes(x=(gdp.number),y= abs(resid))) +
  geom_point()+ geom_smooth(method='loess',se=F)+
  scale_x_log10() +
```

```
  ggtitle("Absolute Residual Plot for Gdp per Capita and Life Expectancy")+
  labs(x='log 10 of gdp per capita')


continent.lm.m.fitted.df<-as.data.frame(continent.lm.m.fitted)


ggplot(continent.lm.m.fitted.df, aes(x = gdp.number, y = pred, group=date , color= date)) +
  geom_line()+scale_x_log10() +
  facet_grid(~continent)+
  ggtitle('Predicted Life Expectancy per GDP after WW2' ) +
  labs(x= 'Log 10 GDP per capita' , y= 'Life Expectancy', subtitle = 'Using linear model faceting by continents')


###(1-2) model with liner : using date.factors
continent.lm.f = function(data){
  lm(life.expectancy.number ~ log10(gdp.number) + date.factor, data = data)
}
continent.lm.m.f = data.ww2 %>%
  group_by(continent) %>%
  nest()

continent.lm.m.f.m = map(continent.lm.m.f$data, continent.lm.f)

continent.lm.m.f = mutate(continent.lm.m.f, model = continent.lm.m.f.m)
continent.lm.m.f = mutate(continent.lm.m.f, .fitted = map2(data, model, add_predictions))
continent.lm.m.f = mutate(continent.lm.m.f, .resid = map2(data, model, add_residuals))
continent.lm.m.f.fitted = unnest(continent.lm.m.f, .fitted,.resid)

continent.lm.m.f.fitted<-continent.lm.m.f.fitted[,-c(10:16)]


ggplot(continent.lm.m.f.fitted , aes(x=(gdp.number),y= resid))+
```

```
  geom_point()+ geom_smooth(method='loess',se=F)+
  scale_x_log10() +
  ggtitle("Residual Plot for Gdp per Capita and Life Expectancy")+
  labs(x='log 10 of gdp per capita')


ggplot(continent.lm.m.f.fitted , aes(x=(gdp.number),y= abs(resid))) +
  geom_point()+ geom_smooth(method='loess',se=F)+
  scale_x_log10() +
  ggtitle("Absolute Residual Plot for Gdp per Capita and Life Expectancy")+
  labs(x='log 10 of gdp per capita')


ggplot(continent.lm.m.f.fitted, aes(x = gdp.number, y = pred, group = date.factor, color = date.factor)) +
  geom_line()+scale_x_log10()+
  facet_grid(~continent)+
  ggtitle('Predicted Life Expectancy per GDP after WW2' ) +
  labs(x= 'Log 10 of GDP per capita' , y= 'Life Expectancy', subtitle = 'Using linear model considering continents and seperate dates', caption = 'Faceted by Continent')+scale_color_manual(values=cb_palette)


ggplot(continent.lm.m.f.fitted, aes(x = gdp.number, y = pred, group = continent, color = continent)) +
  geom_line()+scale_x_log10()+
  facet_grid(~date.factor)+
  ggtitle('Predicted Life Expectancy per GDP after WW2' ) +
  labs(x= 'Log 10 of GDP per capita' , y= 'Life Expectancy', subtitle = 'Using linear model considering continents with dates together', caption = 'Faceted by Date') +scale_color_manual(values=cb_palette)
```