**EXPLORATORY DATA ANALYSIS MINI PROJECT 1**

*Authors: Vikrant Deshpande, Tanvi Kolhatkar, Saishree Godbole*

# INTRODUCTION

In this project, we want to understand life expectancy and assess its relationship to GDP per capita. For this, we have used the gapminder R package which has information about the life expectancy in 142 countries for a selection of years from 1952 to 2007. To understand the type of relationship between life expectancy and GDP, we plan to follow the below steps :

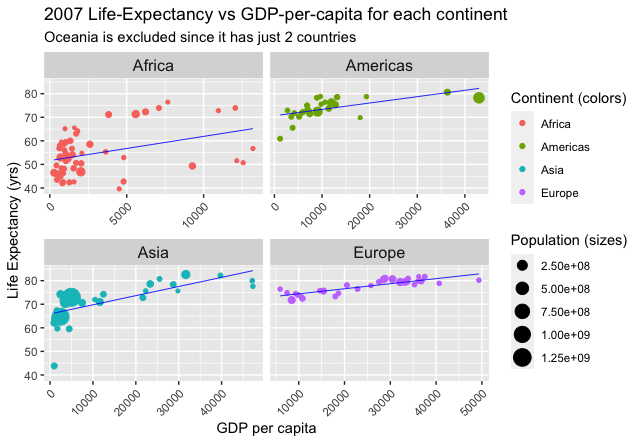
1. Check if the relationship can be fitted by a linear model and explaining any differences
2. Analyzing the trend of life expectancy over time for individual continents and investigating changes caused by respective countries
3. Checking if there are any other factors affecting the life expectancy apart from the GDP

We will now go over these steps in detail.

# LIFE EXPECTANCY AND GDP TREND IN 2007

*\*\*Can the increase in life expectancy since World War 2 be largely explained by increases in GDP per capita?\*\**

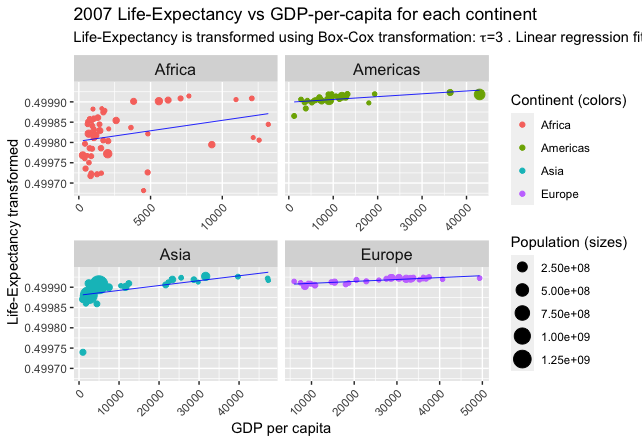
We have skipped Oceania from this analysis: there are just 2 countries with great GDP-per-capita values and correspondingly good Life-expectancies.



Africa : We see an unfortunate scatterplot for the countries here. Most have very low GDP values, and life-expectancies ~40-50 years. There is no clear observable linear relationship between GDP and Life-Expectancy. Hence, it doesn't make sense to use a simple linear-model here.

Americas, Asia: We see a fairly linear relationship between GDP per capita, and the Life-Expectancy. There are a few observable outliers for Americas and Asia. If we fit a simple linear-regression line, the predicted Life-Expectancy increases as GDP-per-capita increases.

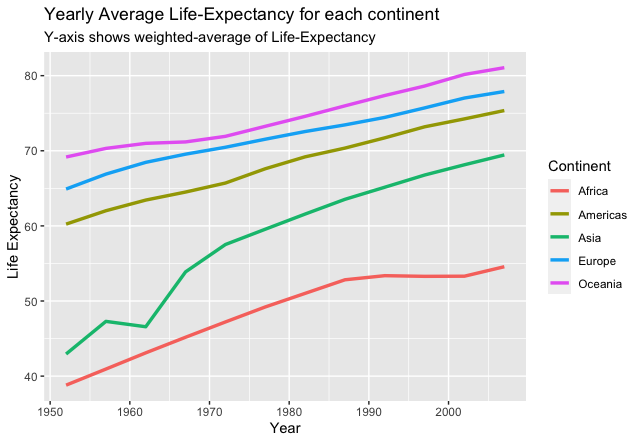
Europe: These plots look slightly hollow-down, and thus applying a Box-Cox transformation with a large value of T (like T=3), should give us better linear relationships. A weighted-linear regression model which uses population as weights, for the Life-expectancy of a country, makes sense.

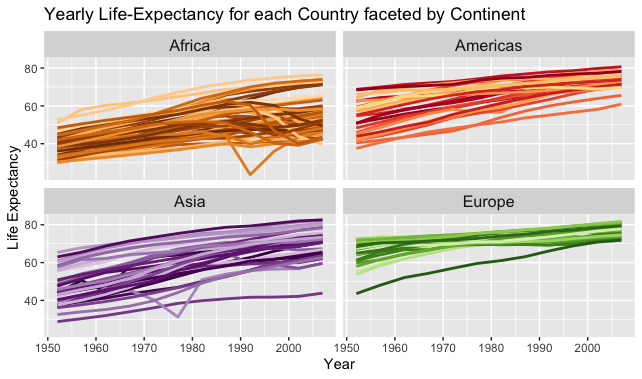


Africa: There is still no observable linear relationship between the features. We shouldn't model the distribution for Africa using a Linear-model.

Americas, Asia, and Europe: We see a better linear relationship between the transformed Life-Expectancy, and the GDP-per-capita. Now we see fewer observable outliers for Americas and Asia, if we fit a linear-regression model.

# LIFE EXPECTANCY OVER TIME

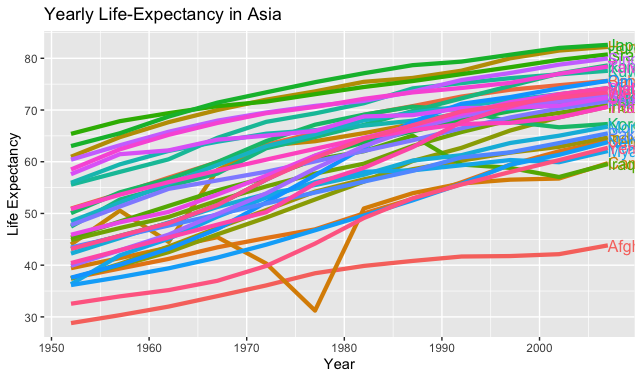




To answer this question we have plotted the average life expectancy for all continents over the years 1950-2010. On looking at the plot, we can see the life expectancies in most continents have been linearly increasing with time.

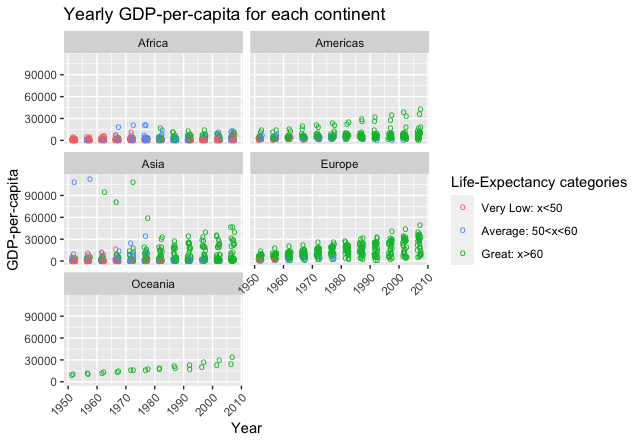
Continents of Europe and Americas started at high average life expectancies and have had a steady increase over the years. Asia and Africa were the two continents with very low life expectancy in 1950. Over time, notwithstanding a dip around 1962, Asia has caught up with the other continents at a fast rate.

Average life expectancy in Africa has had a steady but relatively small increase from 30-50 to 40-60.



In Asia, the dip in average life expectancy (around 1962) can be attributed to a dip in China. China has the largest population in Asia and contributes heavily to the weighted average life expectancy of the entire continent. The causes of these changes could be attributed socio-economic changes and natural disasters such as famines occurring around 1962. Apart from this, the majority of the countries have had a steady growth in their life expectancy which has contributed to the overall growth of Asia.

# CHANGES IN THE RELATIONSHIP BETWEEN GDP AND LIFE EXPECTANCY OVER TIME



Africa: From Ans.1, we already know that there is not much of a linear-relationship between GDP and Life-Expectancy, and in Ans.2 we saw that there

Americas, Asia, Europe, Oceania: GDP and life expectancy have a more or less linear relationship. Looking at the example of Africa, we cannot entirely attribute the changes to be due to the GDP. We can see that time also contributes to the increase in life expectancy.

*#Pending:*

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

*# Are there exceptions to the general patterns?*

# CONCLUSION

# APPENDIX

Code

Q1

# Pending: Make single linear/complex model for all continents

# Make pairwise QQ plots

install.packages(c("gapminder", "directlabels"))

require(gapminder)

require(tidyverse)

require(directlabels)

glimpse(gapminder)

gapminder.2007 <- gapminder %>% filter((year==2007) & (continent!="Oceania"))

gapminder.2007 %>%

ggplot(mapping=aes(x=gdpPercap, y=lifeExp, color=continent)) +

geom\_point(aes(size=pop)) +

geom\_smooth(method="lm", se=F, color="blue", size=0.3) +

facet\_wrap(~continent, scales="free\_x") +

theme(axis.text.x=element\_text(angle=45, hjust=1),strip.text=element\_text(size = rel(1.1))) +

labs(title="2007 Life-Expectancy vs GDP-per-capita for each continent",

subtitle="Oceania is excluded since it has just 2 countries",

x="GDP per capita", y="Life Expectancy (yrs)") +

scale\_color\_discrete("Continent (colors)") +

scale\_size\_continuous("Population (sizes)")

box\_cox\_transformation <- function(data, n){

return (((data^n)-1)/n)

}

gapminder.2007$transformed.lifeExp <- box\_cox\_transformation(gapminder.2007$lifeExp, -2)

gapminder.2007 %>%

ggplot(mapping=aes(x=gdpPercap, y=transformed.lifeExp, color=continent)) +

geom\_point(aes(size=pop)) +

geom\_smooth(method="lm", se=F, color="blue", size=0.3, weight=aes(pop)) +

facet\_wrap(~continent, scales="free\_x") +

theme(axis.text.x=element\_text(angle=45, hjust=1)) +

labs(title="2007 Life-Expectancy vs GDP-per-capita for each continent",

subtitle=expression(paste("Life-Expectancy is transformed using Box-Cox transformation: ",tau,"=3 . Linear regression fits better now.")),

x="GDP per capita", y="Life-Expectancy transformed") +

theme(strip.text=element\_text(size = rel(1.1))) +

scale\_color\_discrete("Continent (colors)") +

scale\_size\_continuous("Population (sizes)")

Q2

#average life exp change over time for each continent

gapminder %>%

group\_by(continent, year) %>%

summarise(lifeExp=weighted.mean(x=lifeExp, w=pop, na.rm=T)) %>%

ggplot(aes(x =year, y=lifeExp, colour=continent)) +

geom\_line(size=1.2) +

labs(title="Yearly Average Life-Expectancy for each continent",

subtitle="Y-axis shows weighted-average of Life-Expectancy",

x="Year", y="Life Expectancy") +

scale\_color\_discrete("Continent")

#faceted plot of each continent showing individual countries.

gapminder %>%

filter(continent!="Oceania") %>%

ggplot(aes(x=year, y=lifeExp, group=country, color=country)) +

geom\_line(lwd=1, show.legend=F)+

facet\_wrap(~continent) +

scale\_color\_manual(values=country\_colors) +

theme(strip.text=element\_text(size = rel(1.1)))

#deep dive into Asia as it has caught up to other continents

library(directlabels)

gapminder %>%

filter(continent=="Asia") %>%

ggplot(aes(x=year, y=lifeExp, group=country, color=country)) +

geom\_line(size=1.5, show.legend=F) +

geom\_dl(aes(label=country), method=list("last.points"))

Q3

gapminder %>%

mutate(

cat.lifeExp=if\_else(lifeExp<50, "Very low", if\_else(lifeExp<60, "Average", "Great")),

color.lifeExp=if\_else(lifeExp<50, "gray", if\_else(lifeExp<60, "red", "green")),

) %>%

ggplot(aes(x=year, y=gdpPercap, color=color.lifeExp)) +

geom\_jitter(shape='o', width=.7, size=2.5) +

theme(axis.text.x=element\_text(angle=45, hjust=1)) +

facet\_wrap(~continent, ncol=2) +

labs(title="Yearly GDP-per-capita for each continent",

x="Year", y="GDP-per-capita", ) +

scale\_colour\_discrete(name="Life-Expectancy categories",

breaks=c("gray", "red", "green"),

labels=c("Very Low: x<50", "Average: 50<x<60", "Great: x>60"))