Mini Project 1

Yexin Lu, Zining Qi, Mengchu Li
luyex@iu.edu, qizin@iu.edu, mli3@iu.edu

Introduction

A researcher for a think tank wants to learn about the relationship between life expectancy and GDP per capita. In order to use exploratory data analysis to answer this problem, we need some data set and statistical tools to help us. The data set comes from an R package called gapminder. In that package, a data set called gapminder giving us the information about the GDP per capita and life expectancy in 142 countries for a specific period from 1952 to 2007. In this project, our main research question is: can the increase in life expectancy since World War 2 be largely explained by increases in GDP per capita? However, this question is difficult to answer in a straight way. Thus, the research question is separated into three parts:

- GDP and life expectancy in 2007
- Life expectancy over time by continent
- Changes in the relationship between GDP and life expectancy over time

Each part contains several small questions that can lead us to find the final answer. And in order to answer these questions, we need to visualize the data and fit some models. The packages that we use in this project are:

- ggplot2
- tidyverse
- broom
- GGally

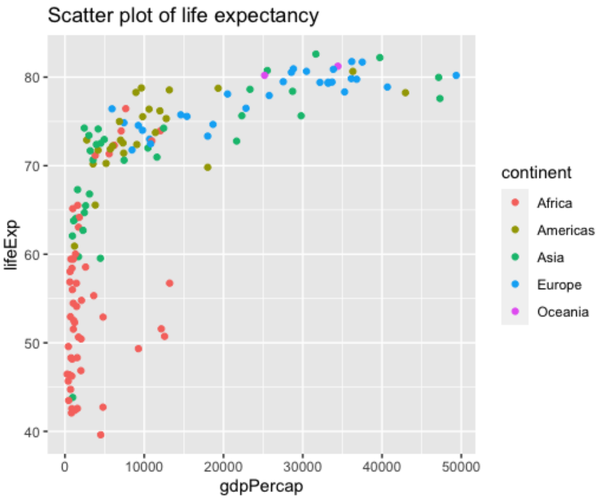
In order to visualize the data that we get from gapminder, plots that we used are scatter plot, facet plot, coplot. In order to explore the relationship between each variable, the model that we used in this project includes log transformed linear model and a more complicated loess model.

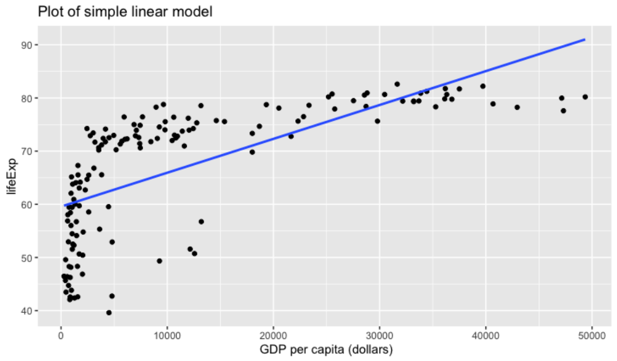
For the quantity appear in the report, here we assumed the unit of the quantities are:

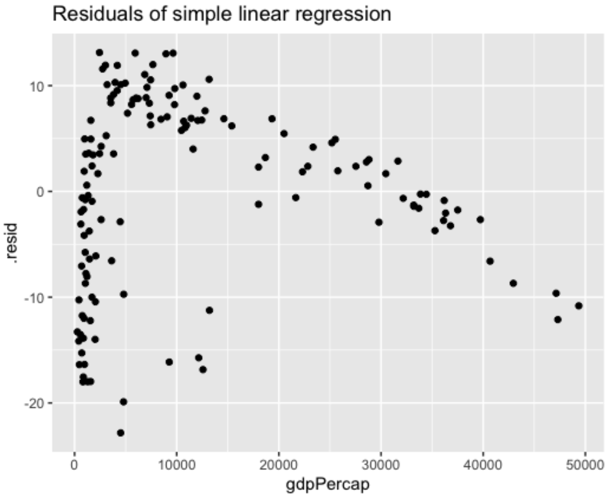
- GDP per capita(gdpPercap): dollar ($)
- life expectancy: year

GDP and life expectancy in 2007

- **Simple Linear Regression**



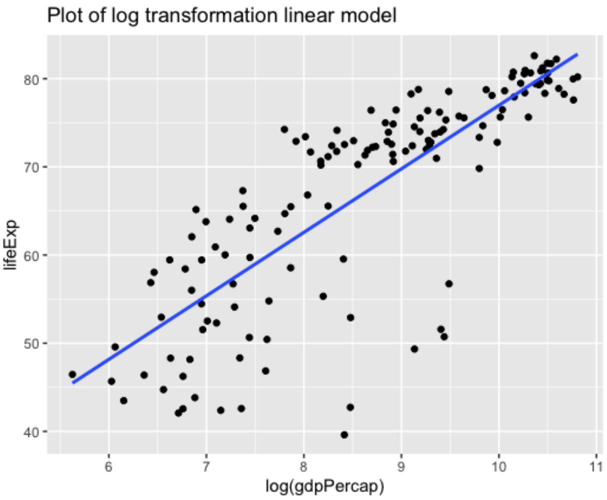


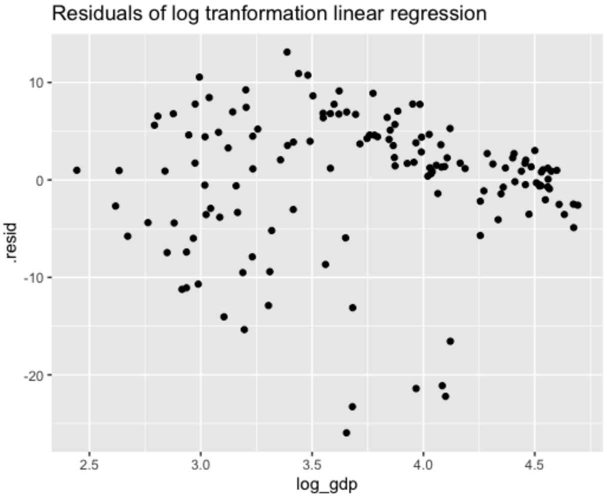


From the plot of points, the overall trend is positive as GDP per capita increase. And the linear regression line also proves that. But the data can't be well described by the simple linear regression model. Because the basic trend of data is not linear from the plot.

And from the residual plot, there is a obvious trend of residuals with gdp per capita. This also indicates that the simple linear regression doesn't work well for the data.
So more complicated model is required.

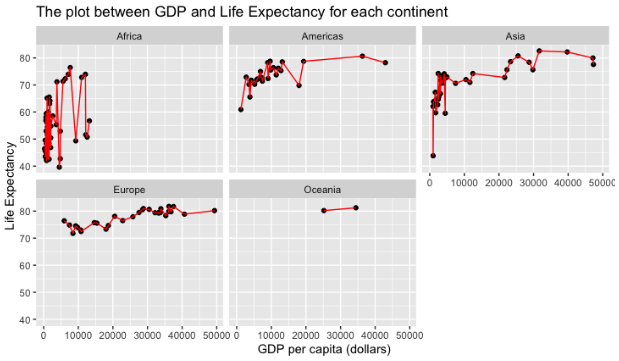
- **Linear Regression with Log Transformation**



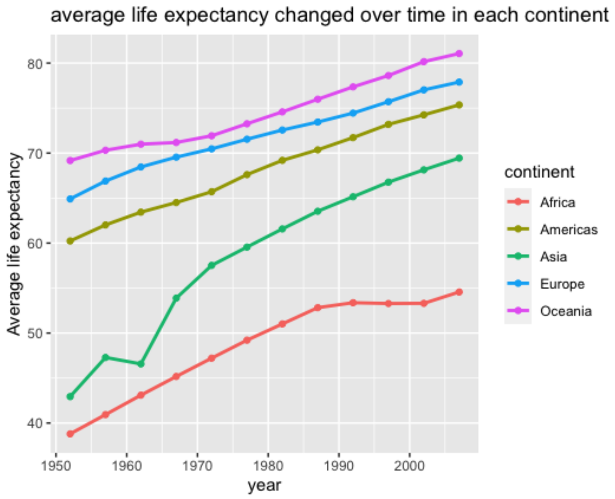


From the plot of log transformed linear model, it shows that the log transformation model does a much better work than simple linear model. The overall trend just looks like there is a positive linear relationship between two variables. Although the data which gdp less than 8.5 is looser than the other part, it is still much better than the simple linear model.
And from the residual plot of log transformed linear model, overall, the residuals plot around 0. The specific trend for residuals is much less than simple regression model.
So, this complicated model is better.

- **Pattern for Different Continents**
From the plots of each continents, the pattern for each continent is different than others. And the difference is not small. On the same axis scale, there are some obvious differences among plots. And there is no common trends among these plots. So, there is a good reason to believe that the differences are more complicated than additive or multiplicative shift.

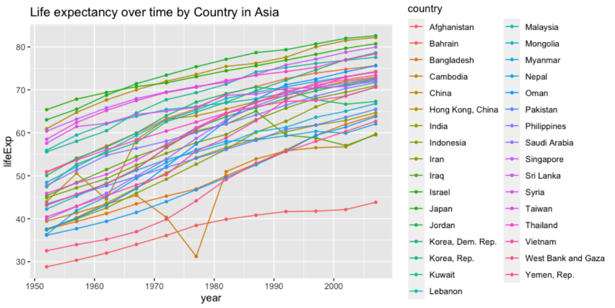


Life expectancy over time by continent



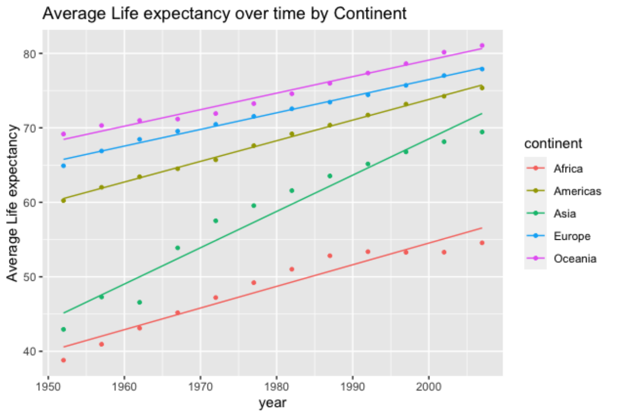
- The average life expectancy increased during the period from 1950s to 2000s overall. However, there exists a small fluctuation in Asia between 1960s to 1970s: a tiny decrease and then increase again.

- There is no continents catching up to other continents because of no intersection in the plot before. But we can see that all continents caught up others partially, especially Asia. Asia has reduced largely the distance of life expectancy of other continents.

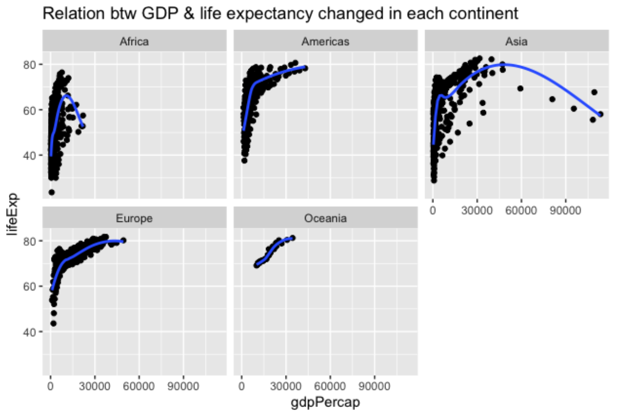


From the plot Life expectancy over time by country in Asia, we can see that almost every Asian country's life expectancy increases in the same level rate except one country Afghanistan with a very low increase rate. Therefore, Asia caught up other continents is because a more general situation not just some countries in Asia.

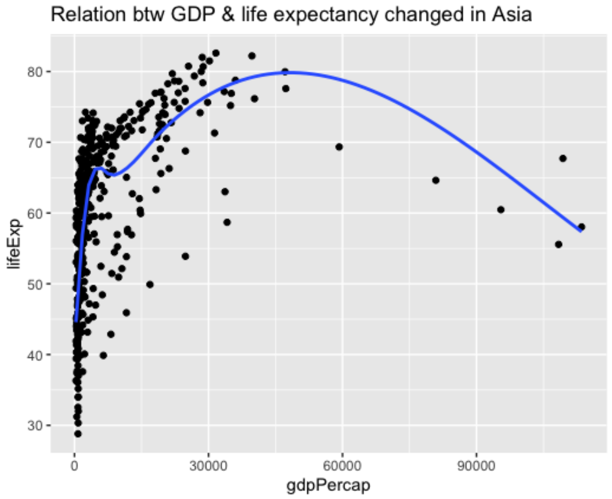
- Since there are only two countries in Oceania, the sample size is too small to observe the changes of life expectancy. Except Oceania, the changes of Europe and America have been linear. The changes of Asia is not linear, and it has been slower between 1960s to 1970s; the reason may be the poverty and wars occurred during that period in Asia. In Asia, it has been faster between 1970s to 1990s, and the reason may be the recovery of world economy and the improvement of medical level. In Africa, it has been faster between 1970s to 1990s since the recovery of world economy and the improve of medical level.



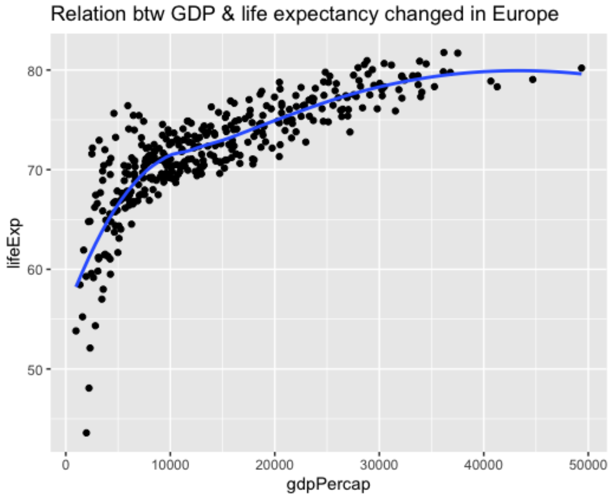
Changes in the relationship between GDP and life expectancy over time



- From the above plot, we observed three variables at the same time: GDP, life expectancy, and time (year). However, since the level of life expectancy of every continent is different and various, to observe the relationship more carefully, we tried to separate continents to see the relationship.

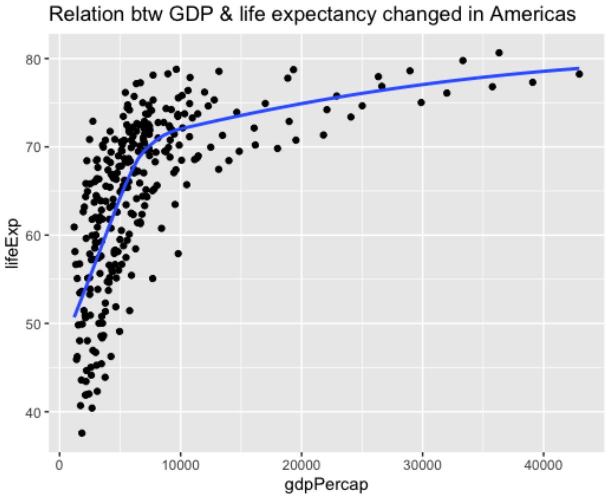


- In Asia, the relationship between GDP and life expectancy can be divided into two parts: first, the interval(0, 45000), the life expectancy increased overall with the GDP increasing with a positive slope; second, the interval with GDP greater than 45000, the change of life expectancy became slow with a negative slope. However, the sample size of GDP greater than 60000 is too small to conclude something with considering the influence of overfitting.

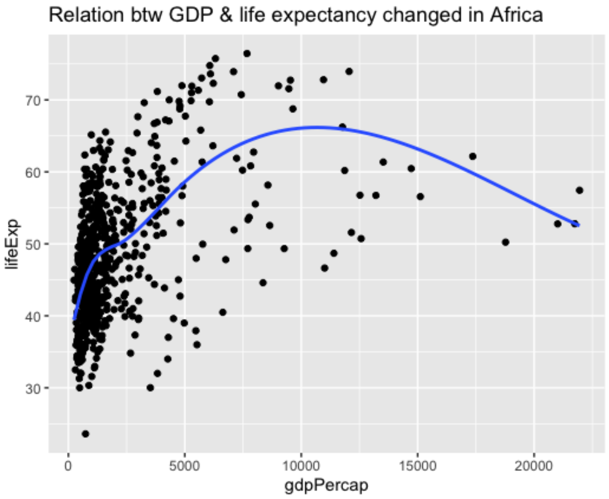


- In Europe, the relationship between GDP and life expectancy was fast with a positive and larger slope when GDP is smaller than 10000; when GDP is greater than 10000, the relationship between GDP and life expectancy was much slower with positive and smaller slope.

- In Americas, the relationship between GDP and life expectancy was fast with a positive and larger slope when GDP is smaller than 7000(estimated value); when GDP is greater than 7000, the relationship between GDP and life expectancy was much slower with a positive and smaller slope.

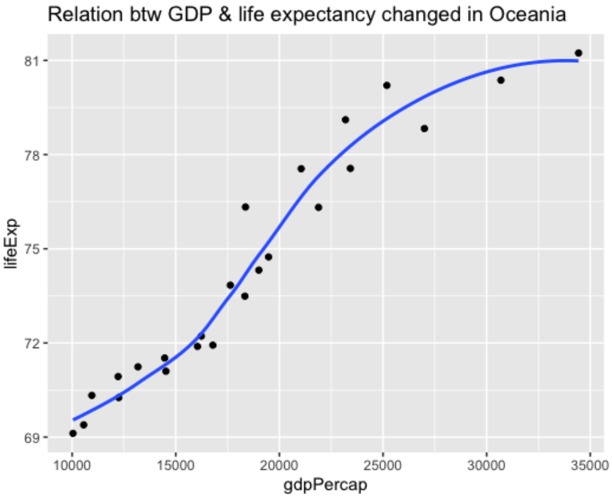


- In Africa, the relationship between GDP and life expectancy can be divided into two parts: first, the interval(0, 10000), the life expectancy increased overall with the GDP increasing with a positive slope; second, the interval with GDP greater than 10000, the change of life expectancy became slow with a negative slope. However, the sample size of GDP greater than 15000 is too small to conclude something with considering the influence of overfitting.

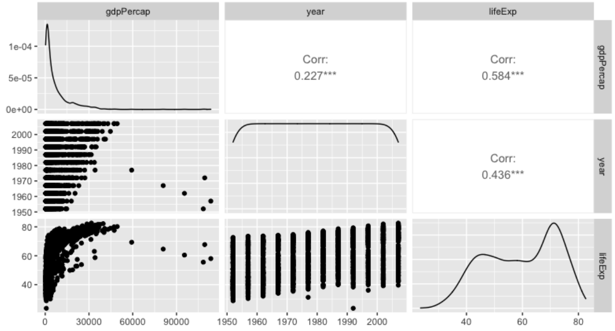


- In Oceania, the relationship between GDP and life expectancy can also be divided into two parts. When GDP is smaller than 20000(estimated value), the slope increased from 0 to a value (curvature > 0); when GDP is greater than 20000(estimated value), the slope decreased from this value to 0 (curvature < 0).

- The indication of span and degree of LOESS fit
  Since we use LOESS fit in the above six plots, here the span value we use is 0.75 and the degree we use is 2.



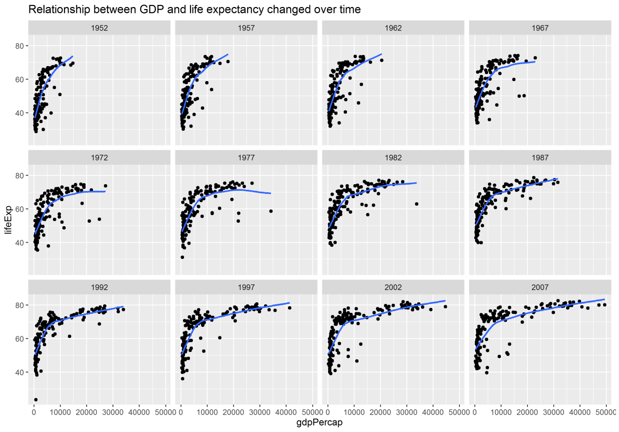
- The changes in life expectancy cannot be entirely explained by changes in GDP per capita, because from the following coplot, there are lots of variations and we cannot find a common trend to describe the relationship between life expectancy and GDP per capita for each continent. Thus, we cannot just use changes in GDP per capita to explain the changes in life expectancy, and there might be other factors also have effects on life expectancy.



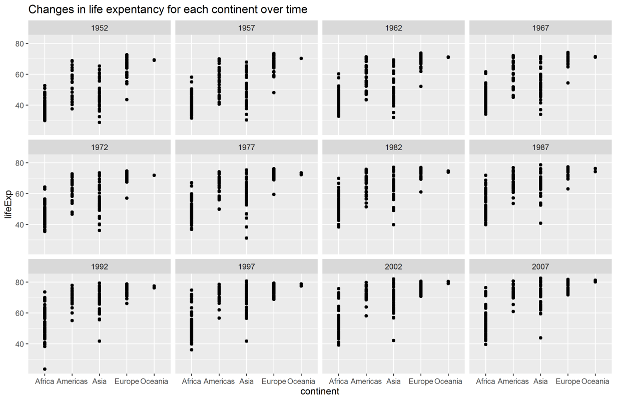
There is a time effect on life expectancy in addition to a GDP effect, because when year increase the life expectancy increase overall. Also, from the pair plots for variable "gdpPercap", "year"and "lifeExp", we can see that the correlation between year and life expectancy is about 0.584, which means that there is a positive relationship between year and life expectancy.

- **Facet plot of Relationship between GDP and life expectancy changed over time without outlier Kuwait:**
  From the facet plot of relationship between GDP and life expectancy changed over time, we can see that there are some effects on the relationship between GDP per capita and life expectancy. In the early years, like 1952 and 1957, changes in GDP per capita can lead to a huge change in life expectancy. In the later period, changes in GDP per capita only lead to a huge change in life expectancy in certain range of GDP. For example, for the GDP per capita within the range 0 to 10000, the life expectancy increases very quickly as long as the increase of GDP. However,

when the GDP per capita goes larger than 10000, the life expectancy only has very small change or even doesn't have changes. Thus, there is a "convergence"in the sense that the variable GDP per capita doesn't matter as much as it used to.

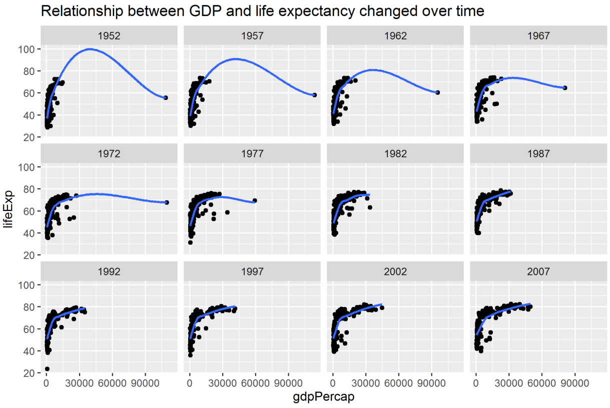


That result might because in the early years GDP per capita for each country don't have very large difference, because the data points in the plot are more clustered. However, in the later years, the data points becomes more spread out, which means the difference of the GDP per capita for each country becomes larger.



- From the above plot, we can see that in the early years, the differences of life expectancy between each continent are very large. For example, in 1952, Europe has the highest life expectancy level. And the highest life expectancy in Europe is about 75 years. Africa has the lowest life expectancy level. And the highest life expectancy in Africa is about 55 years. The difference of life expectancy between these two countries which have the highest life expectancy level among these two continents is about 20 years. Thus, the continent does matter to explain the changes in life expectancy in the period around 1952 to 1987. However, in the later years, the difference of life expectancy between each continent gradually reduced. For example, in 2007, the life expectancy level for Europe, Asia, America, and even Oceania have almost the same level, although Oceania needs to be considered as an outlier. The only one continent left is Africa. However, although Africa still has the lowest life expectancy level, the difference of life expectancy of Africa between other continents is very small. The difference between the highest level of life expectancy for Africa and Europe is reduced to 2-3 years. Thus, the continent doesn't matter in the period around 1987 to 2007.

- **Facet plot of Relationship between GDP and life expectancy changed over time include outlier Kuwait:**



There are some exceptions to general pattern:
The facet plot of relationship between GDP and life expectancy changed over time that we used in our analysis displaying all countries' data points except Kuwait, because that country has very high GDP per capita in the period between 1952 to 1977, which can be considered as an outlier.
Also, the facet plot of changes in life expectancy for each continent over time displays that Oceania has very few data points comparing with other continents, because Oceania only has two countries. Thus, it is very difficult for us to compare the life expectancy with other continents fairly. So, we can exclude that continent in our analysis.