Life Expectancy in Developing Countries has Increased between 2000 and 2015

Thomas Fiore

May 5, 2019

Overview

This abbreviated file illustrates what a Reproducibility Appendix looks like, and illustrates some things students will be able to do after completing this course. After completing this course, students will be able to: produce a complete record of a data analysis like this file, make good data visualizations, and clean up data.

External Requirements

```
# install.packages("ggplot2") # This is a reminder on how to install
                              # the package, but we do not install in
                              # an Rmd file
library(ggplot2)  # needed library call for the ggplot2 plotting package
library(gridExtra) # needed library call to place figures next to each other

# The following data set is a prepocessed version of
# https://www.kaggle.com/kumarajarshi/life-expectancy-who
LED_complete = read.csv("LED_complete.csv")
# Ordinarily we prefer to read directly from the web, and make a complete
# record of the processing in this file, but I don't do that
# this time because others did it for me
```

Data Description

Source, number of rows and columns, description of variables....

Each row is a country for a specific year and includes a single life expectancy number, per capita GDP for that year, and the country's classification as a developing country or developed country.

Further explanation should be done here....

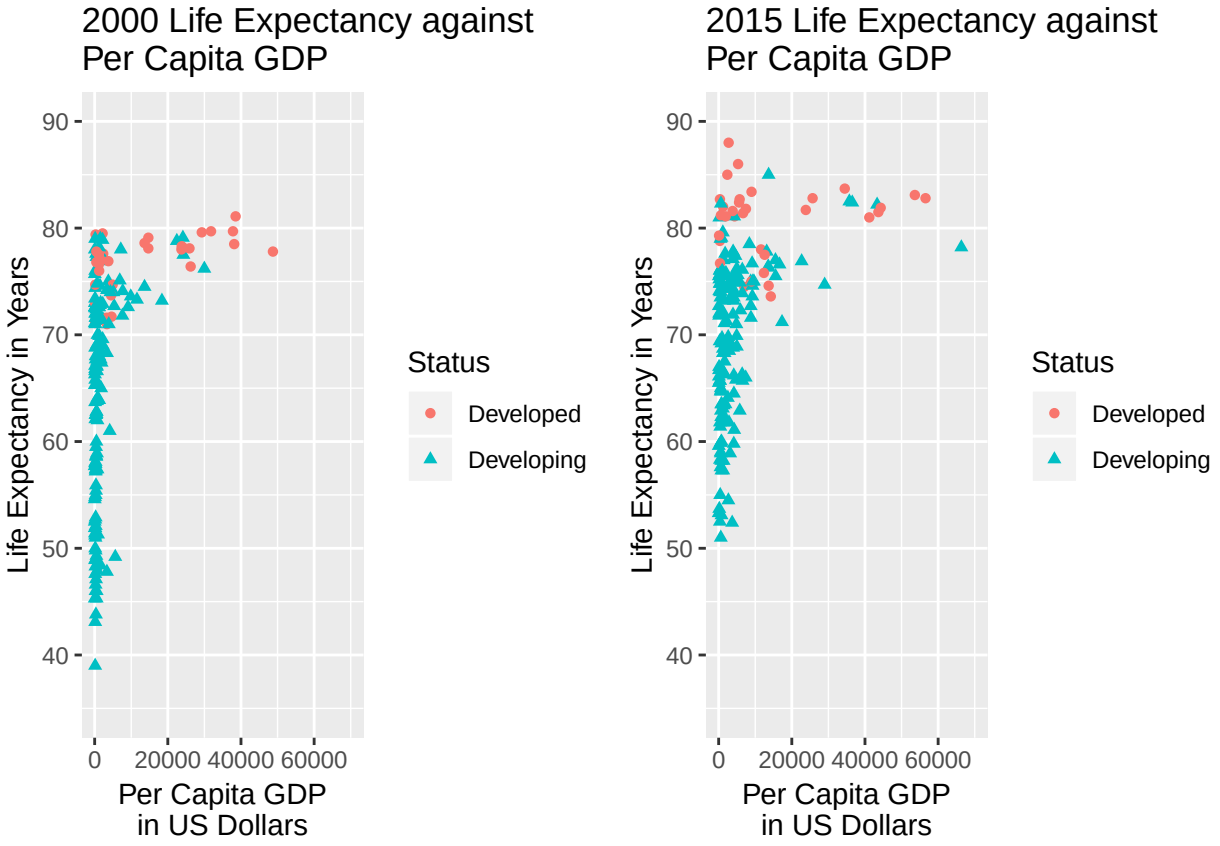
Data Pre-Processing Record

Not done here....

Abbreviated Exploratory Data Analysis

```
LED_forplot_2000 = subset( LED_complete, subset=(Year == 2000),
                           select = c(Country,GDP,Life_expectancy,Status))

plot2000 = ggplot(LED_forplot_2000,
                  aes(x=GDP, y=Life_expectancy, shape=Status, color=Status)) +
     geom_point() +
     xlim(0,70000) +
     ylim(35,90) +
     ggtitle('2000 Life Expectancy against \nPer Capita GDP') +
```

```
     labs(x="Per Capita GDP\nin US Dollars", y="Life Expectancy in Years")


LED_forplot_2015 = subset( LED_complete, subset=(Year == 2015),
                           select = c(Country,GDP,Life_expectancy,Status))

plot2015 = ggplot(LED_forplot_2015,
                  aes(x=GDP, y=Life_expectancy, shape=Status, color=Status)) +
     geom_point() +
     xlim(0,70000) +
     ylim(35,90) +
     ggtitle('2015 Life Expectancy against \nPer Capita GDP') +
     labs(x="Per Capita GDP\nin US Dollars", y="Life Expectancy in Years")

grid.arrange(plot2000, plot2015, nrow=1)
```



From the two graphs, it looks like blue triangles below 50 in 2000 have moved above 50 in 2015. This suggests life expectancy has increased, but we do not know if it is due to random variation or not.

Hypothesis Test and Confidence Interval

To check if life expectancy in developing countries really did increase, and the oberserved difference is not due to random variation, we must do a hypothesis test. Hypothesis testing *will not* be a major part of this class, but I include it here to show what a technical appendix can include.

We have here *paired data*: the life expectancy for each developing country in 2000 and 2015. We would like to compare the means of developing countries in 2000 and 2015. The appropriate hypothesis test is a *one-sided*

paired t-test. The code below does the test. The output indicates the p-value is very far below .05, so we reject the null hypothesis of no difference in means, in favor of the alternative hypothesis of higher mean in 2015.

The output also indicates that we are 95% confident that the difference in means is at least 4.46 years.

```
developing_expectancy_2000 = subset(LED_forplot_2000, subset=(Status == 'Developing'),
                                    select=c(Life_expectancy) )

developing_expectancy_2015 = subset(LED_forplot_2015, subset=(Status == 'Developing'),
                                    select=c(Life_expectancy))

length(developing_expectancy_2000) == length(developing_expectancy_2015)
```

```
## [1] TRUE
```

```
t.test(developing_expectancy_2015$Life_expectancy,
       developing_expectancy_2000$Life_expectancy,
       paired = TRUE, alternative = "greater")
```

```
##
##  Paired t-test
##
## data:  developing_expectancy_2015$Life_expectancy and developing_expectancy_2000$Life_expectancy
## t = 13.76, df = 150, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  4.460355      Inf
## sample estimates:
## mean of the differences
##                5.070199
```

Discussion of Assumptions of Hypothesis Test

Can we really trust the hypothesis test? We should check to what degree the data satisfy the 4 assumptions of the paired t-test (for instance normality). Just looking at the plots above, it looks like the 2000 data may be skewed toward lower life expectancies, which could be a problem. A technical appendix would explore this and the other assumptions... But we skip that here because the purpose of this class is statistical programming, not statistical inference. Statistical inference is covered in a different class.