WEEK3

Heleine Fouda

2023-07-28

The research question:

Is the difference observed in life expectancy between Africa and Europe between 1952 and 1977 statistically significant?

1. Data Exploration

First let's take a broad look at the data frame

```
glimpse(gapminder)
```

```
## Rows: 1,704
## Columns: 6
## $ country   <fct> "Afghanistan", "Afghanistan", "Afghanistan", "Afghanistan", ~
## $ continent <fct> Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, ~
## $ year      <int> 1952, 1957, 1962, 1967, 1972, 1977, 1982, 1987, 1992, 1997, ~
## $ lifeExp   <dbl> 28.801, 30.332, 31.997, 34.020, 36.088, 38.438, 39.854, 40.8~
## $ pop       <int> 8425333, 9240934, 10267083, 11537966, 13079460, 14880372, 12~
## $ gdpPercap <dbl> 779.4453, 820.8530, 853.1007, 836.1971, 739.9811, 786.1134, ~
```

Let's also check the names, types,nature and structure of the variables contained in the data frame.

```
head(gapminder)
```

```
## # A tibble: 6 x 6
##   country     continent  year lifeExp      pop gdpPercap
##   <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
## 1 Afghanistan Asia       1952    28.8  8425333      779.
## 2 Afghanistan Asia       1957    30.3  9240934      821.
## 3 Afghanistan Asia       1962    32.0 10267083      853.
## 4 Afghanistan Asia       1967    34.0 11537966      836.
## 5 Afghanistan Asia       1972    36.1 13079460      740.
## 6 Afghanistan Asia       1977    38.4 14880372      786.
```

```
names(gapminder)
```

```
## [1] "country"   "continent" "year"      "lifeExp"   "pop"       "gdpPercap"
```

```
str(gapminder)
```

```
## tibble [1,704 x 6] (S3: tbl_df/tbl/data.frame)
##  $ country  : Factor w/ 142 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ continent: Factor w/ 5 levels "Africa","Americas",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ year     : int [1:1704] 1952 1957 1962 1967 1972 1977 1982 1987 1992 1997 ...
##  $ lifeExp  : num [1:1704] 28.8 30.3 32 34 36.1 ...
##  $ pop      : int [1:1704] 8425333 9240934 10267083 11537966 13079460 14880372 12881816 13867957 163
```

```
##  $ gdpPercap: num [1:1704] 779 821 853 836 740 ...
```

The structure function reveals the presence of two numerical variables: lifeExp and gdpPercap; two integers:pop and year variables; two factors: country and continent variables. Let's visualize the numerical variables

```
class(gapminder$lifeExp)
```

```
## [1] "numeric"
```

```
class(gapminder$gdpPercap)
```

```
## [1] "numeric"
```

```
class(gapminder$pop)
```

```
## [1] "integer"
```

```
class(gapminder$pop)
```

```
## [1] "integer"
```

2. Descriptive statistics

```
summary(gapminder)
```

```
##         country        continent        year         lifeExp
##  Afghanistan:  12   Africa  :624   Min.   :1952   Min.   :23.60
##  Albania    :  12   Americas:300   1st Qu.:1966   1st Qu.:48.20
##  Algeria    :  12   Asia    :396   Median :1980   Median :60.71
##  Angola     :  12   Europe  :360   Mean   :1980   Mean   :59.47
##  Argentina  :  12   Oceania : 24   3rd Qu.:1993   3rd Qu.:70.85
##  Australia  :  12                  Max.   :2007   Max.   :82.60
##  (Other)    :1632
##       pop              gdpPercap
##  Min.   :6.001e+04   Min.   :   241.2
##  1st Qu.:2.794e+06   1st Qu.:  1202.1
##  Median :7.024e+06   Median :  3531.8
##  Mean   :2.960e+07   Mean   :  7215.3
##  3rd Qu.:1.959e+07   3rd Qu.:  9325.5
##  Max.   :1.319e+09   Max.   :113523.1
##
```

```
var(gapminder$lifeExp,gapminder$gdpPercap, na.rm= TRUE)
```

```
## [1] 74323.2
```

2a. First, Examining the general correlation between lifeExp and gdpPercap across time and space:

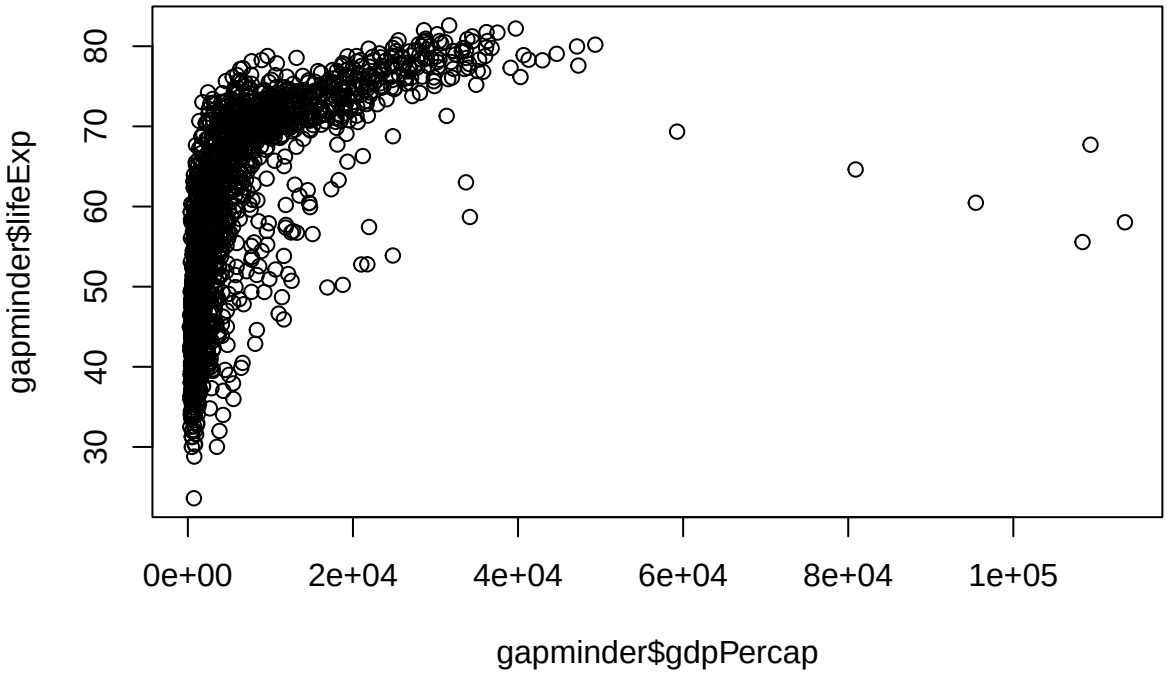
Below, a correlation test, a plot and a boxplot all reveal a statistically significant correlation between lifeExp and gdpPercap across continents at a 95% level of confidence. The higher the gdpPercap, the longer the lifeExp and vice versa.

```
cor.test(gapminder$lifeExp,gapminder$gdpPercap)
```

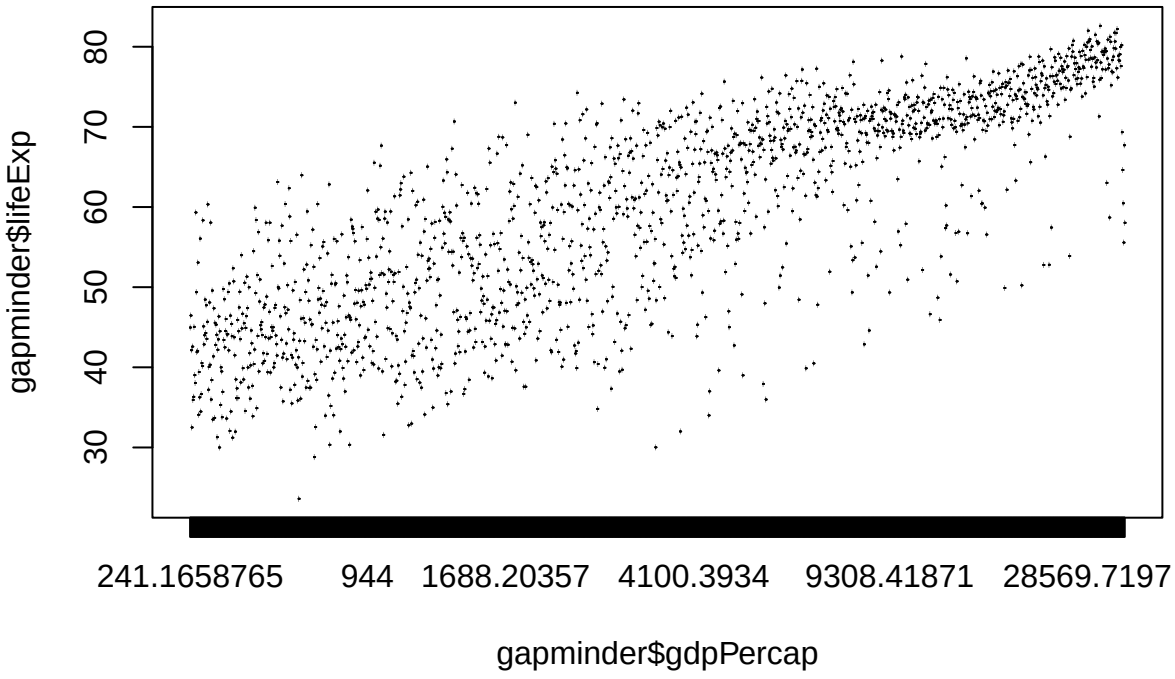
```
##
##  Pearson's product-moment correlation
##
## data:  gapminder$lifeExp and gapminder$gdpPercap
```

```
## t = 29.658, df = 1702, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5515065 0.6141690
## sample estimates:
##       cor
## 0.5837062
```

```
plot(gapminder$lifeExp~gapminder$gdpPercap)
```

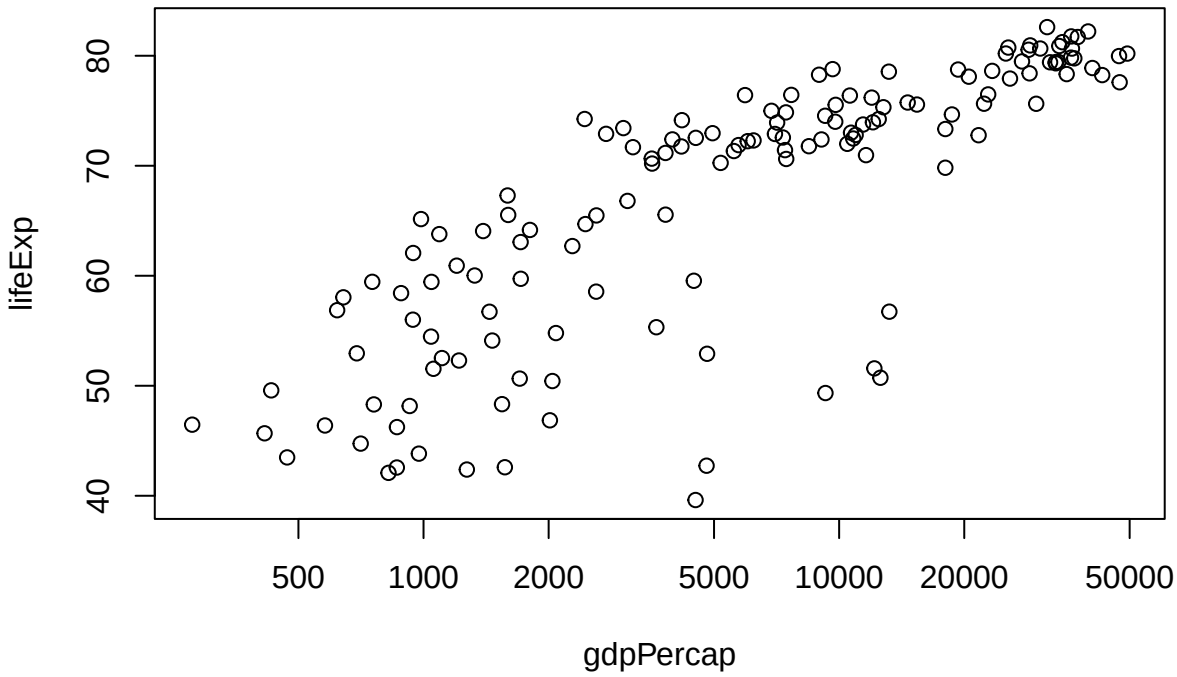



```
boxplot(gapminder$lifeExp~gapminder$gdpPercap)
```

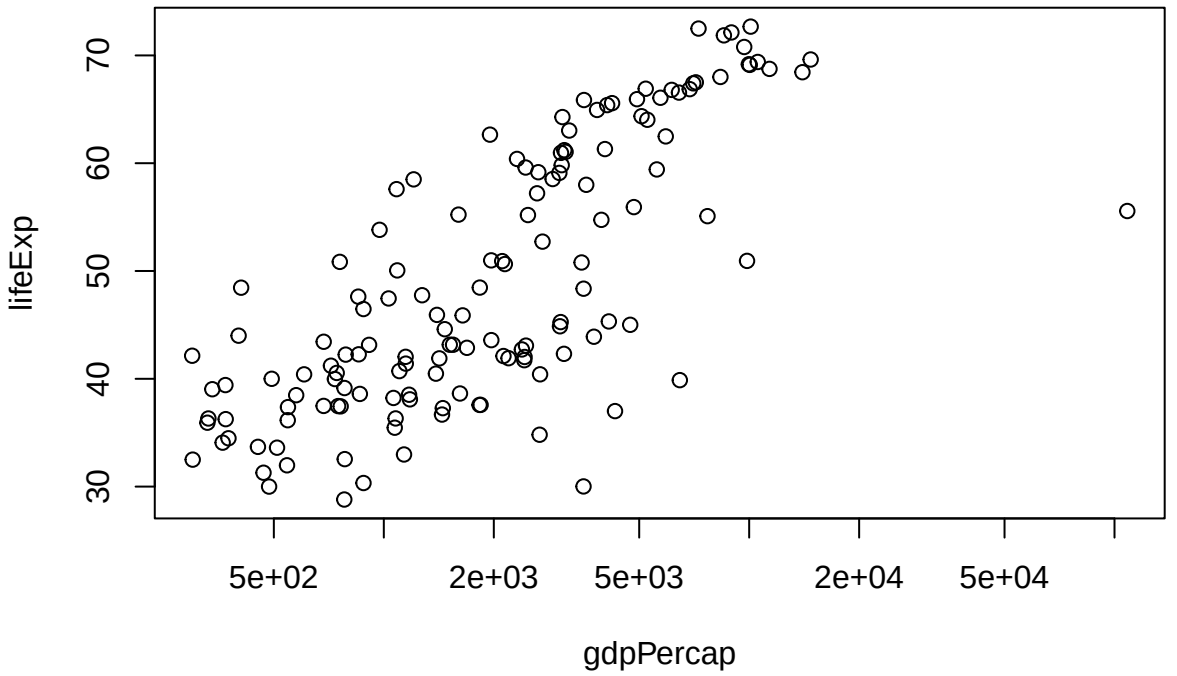


There is a correlation across time between GDP per capita and life expectancy as shown in the boplot for 2007 and 1952

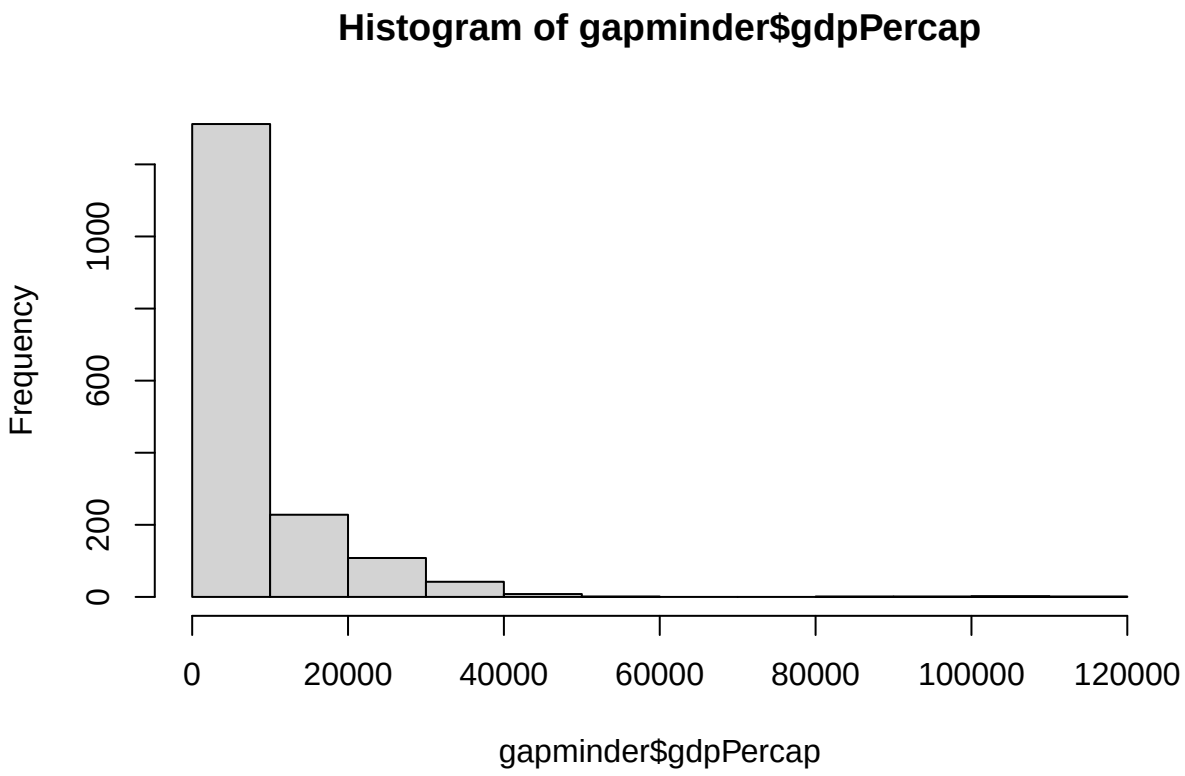
```
plot(lifeExp ~ gdpPercap, gapminder, subset = year == 2007, log = "x")
```



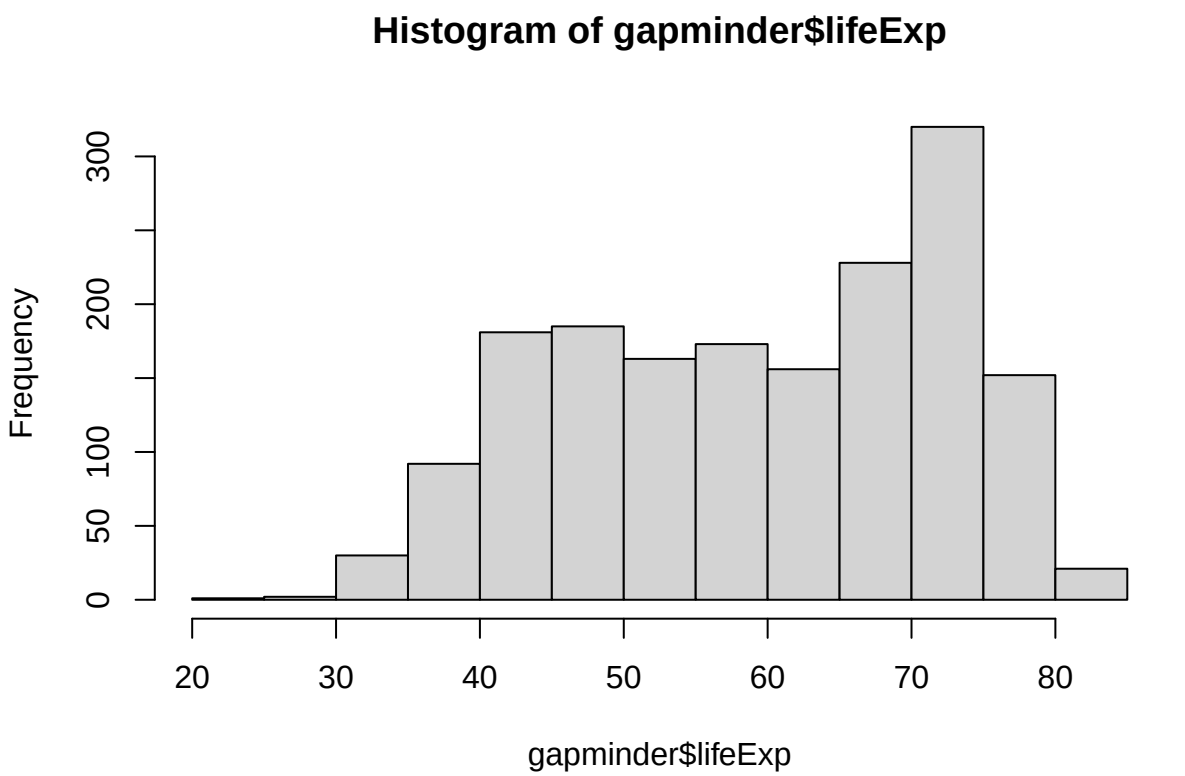

```
plot(lifeExp ~ gdpPercap, gapminder, subset = year == 1952, log = "x")
```




```
hist(gapminder$gdpPercap)
```




```
hist(gapminder$lifeExp)
```




```
library (gapminder)
lifeExp<- c(gapminder$lifeExp)
aggregate(lifeExp~continent,gapminder, median)
```

```
##   continent lifeExp
## 1    Africa 47.7920
## 2  Americas 67.0480
## 3      Asia 61.7915
## 4    Europe 72.2410
## 5   Oceania 73.6650
```

2b. Second, Examining the correlation between lifeExp and gdpPercap between Africa and Europe

METHOD:

Creating a subset of the main data frame with new columns names.

The subset will show data for only Europe and Africa. LifeExp and gdpPercap will also be renamed as a second step

```
library(dplyr)
library(ggplot2)
library(tidyverse)
library(gapminder)

data_subset <- gapminder[13:48,2:6]
summary(data_subset)
```

```
##     continent       year         lifeExp          pop             gdpPercap
##  Africa  :24   Min.   :1952   Min.   :30.02   Min.   : 1282697   Min.   :1601
##  Americas: 0   1st Qu.:1966   1st Qu.:40.88   1st Qu.: 3402653   1st Qu.:2725
##  Asia    : 0   Median :1980   Median :56.62   Median : 6589530   Median :3527
##  Europe  :12   Mean   :1980   Mean   :55.12   Mean   : 9921682   Mean   :3763
##  Oceania : 0   3rd Qu.:1993   3rd Qu.:68.99   3rd Qu.:12505482   3rd Qu.:4826
##                Max.   :2007   Max.   :76.42   Max.   :33333216   Max.   :6223
```

3. Renaming LifeExp and gdpPercap to Life_Expectancy and GDP

```
data_subset = data("gapminder")
gapminder %>%
  select(continent,lifeExp,gdpPercap) %>%
  filter(continent%in%
  c("Africa","Europe"))%>%
  rename(Life_Expectancy=lifeExp,
         GDP=gdpPercap,) %>%
  arrange()
```

```
## # A tibble: 984 x 3
##    continent Life_Expectancy   GDP
##    <fct>               <dbl> <dbl>
##  1 Europe               55.2 1601.
##  2 Europe               59.3 1942.
##  3 Europe               64.8 2313.
##  4 Europe               66.2 2760.
##  5 Europe               67.7 3313.
##  6 Europe               68.9 3533.
##  7 Europe               70.4 3631.
```

```
##  8 Europe                72   3739.
##  9 Europe                71.6 2497.
## 10 Europe                73.0 3193.
## # i 974 more rows
```

```
data_subset = data("gapminder")
gapminder %>%
  select(continent,lifeExp,gdpPercap) %>%
  filter(continent%in%
  c("Africa","Europe"))%>%
  rename(Life_Expectancy=lifeExp,
         GDP=gdpPercap,) %>%
  names()
```

```
## [1] "continent"       "Life_Expectancy" "GDP"
```

4. Samples Findings:

4a. Statical significance

How statistically significant are the sample findings? Assuming (H0), against the samples findings, that there is absolutely no difference between the life expectancy in Africa and the life expectancy in Europe, how likely is it that one can randomly come across a sample showing a difference of the magnetude seen in the data sets findings?

4b. Samples summaries from both the main and the subset data frames reveal an average of 25 to 30 years of difference in life expectancy between Europe and Africa.

4c. Conducting a Two sample t-test

```
data("gapminder")
gapminder %>%
filter(continent %in% c("Africa","Europe")) %>%
  t.test(lifeExp~continent,data =.,
         alternative= "two.sided",
         paired=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  lifeExp by continent
## t = -49.551, df = 981.2, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Africa and group Europe is not equal t
## 95 percent confidence interval:
##  -23.95076 -22.12595
## sample estimates:
## mean in group Africa mean in group Europe
##             48.86533             71.90369
```

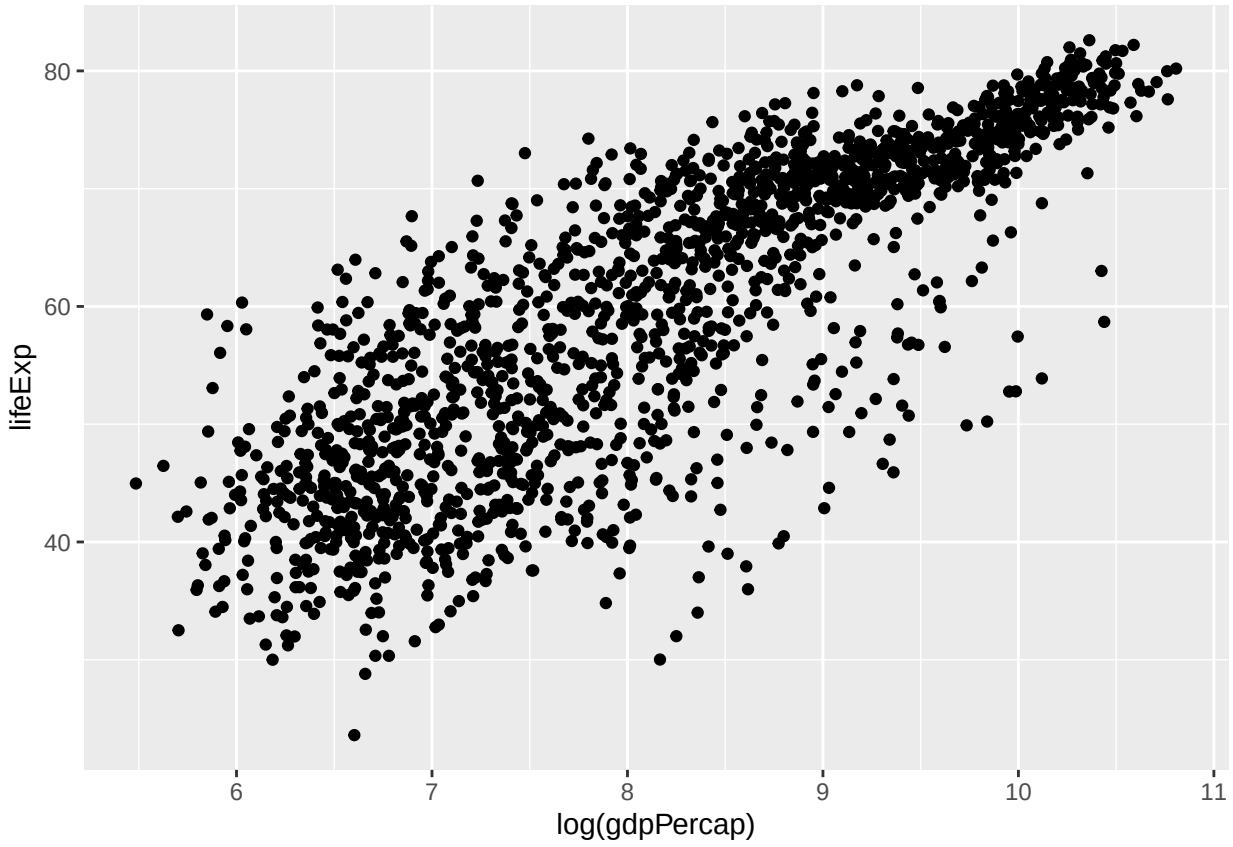
5. Results of the t-test

5a, The t-test results reveal (within a 95 percent confidence interval) that a true, statistically significant difference in life expectancy existed between Africa and Europe from 1952 to 1977.

One must therefore rejects the null hypothesis of zero difference and accept the alternative idea that there is, in truth,a difference and that life expectancy is a function of the gdp per capita.

6. Visualizing the findings

```
library(ggplot2)
gapminder %>%
  filter(gdpPercap < 50000) %>%
  ggplot(aes(x=log(gdpPercap), y=lifeExp))+
  geom_point()
```



7. Regression/Linear modeling

```
library(ggplot2)
library(tidyverse)
library(gapminder)

lm(gapminder$lifeExp~gapminder$gdpPercap)
```

```
##
```

```
## Call:
## lm(formula = gapminder$lifeExp ~ gapminder$gdpPercap)
##
## Coefficients:
##         (Intercept)  gapminder$gdpPercap
##           5.396e+01            7.649e-04
```

```
library(ggplot2)
library(tidyverse)
library(gapminder)

summary(lm(gapminder$lifeExp~gapminder$gdpPercap))
```

```
##
## Call:
## lm(formula = gapminder$lifeExp ~ gapminder$gdpPercap)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -82.754  -7.758   2.176   8.225  18.426
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.396e+01  3.150e-01  171.29   <2e-16 ***
## gapminder$gdpPercap 7.649e-04  2.579e-05   29.66   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.49 on 1702 degrees of freedom
## Multiple R-squared:  0.3407, Adjusted R-squared:  0.3403
## F-statistic: 879.6 on 1 and 1702 DF,  p-value: < 2.2e-16
```

```
library(ggplot2)
library(tidyverse)
library(gapminder)

summary(lm(gapminder$lifeExp~gapminder$gdpPercap +gapminder$pop))
```

```
##
## Call:
## lm(formula = gapminder$lifeExp ~ gapminder$gdpPercap + gapminder$pop)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -82.754  -7.745   2.055   8.212  18.534
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.365e+01  3.225e-01  166.36  < 2e-16 ***
## gapminder$gdpPercap 7.676e-04  2.568e-05   29.89  < 2e-16 ***
## gapminder$pop       9.728e-09  2.385e-09    4.08 4.72e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.44 on 1701 degrees of freedom
```

```
## Multiple R-squared:  0.3471, Adjusted R-squared:  0.3463
## F-statistic: 452.2 on 2 and 1701 DF,  p-value: < 2.2e-16
```

Conclusion:

There is a strong and positive correlation between Life Expectancy and GDP per capita. The strength of that relationship is verified when comparing Africa and Europe, at least for the period 1952 - 1977.