R Notebook

# Download tidyverse   
# tidyverse is already installed and dplyr and ggplot2 are included in the tidyverse package.  
library(tidyverse)

## ── Attaching packages ────────────────────────────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ ggplot2 3.1.0 ✔ purrr 0.2.5  
## ✔ tibble 1.4.2 ✔ dplyr 0.7.7  
## ✔ tidyr 0.8.2 ✔ stringr 1.3.1  
## ✔ readr 1.1.1 ✔ forcats 0.3.0

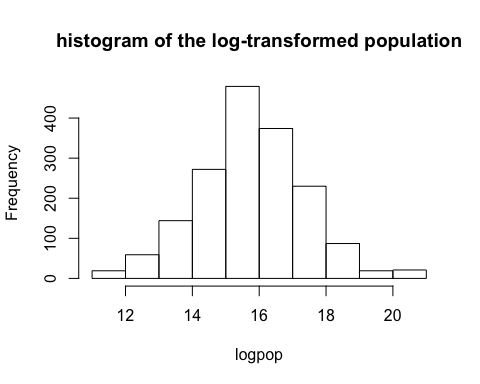
## ── Conflicts ───────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

# Load the full gapminder dataset into a data frame  
# install.packages("gapminder")  
library(gapminder)  
data(gapminder)  
gm <- gapminder  
gm1<- as.data.frame(gm) ### gm1=gm

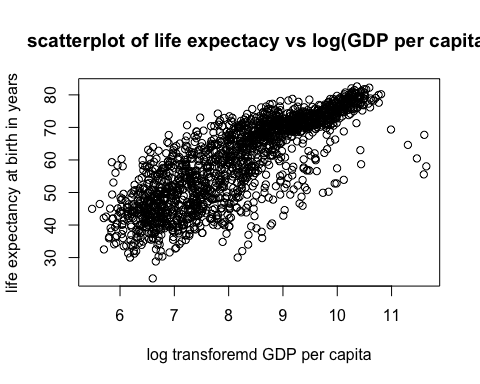
# summary of the gapminder   
summary(gm1)

## 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   
##

# histogram of log(pop)  
logpop <- log(gm1$pop)  
hist(logpop, main="histogram of the log-transformed population")



# scatterplot of life expectancy vs log(gdp per capita)  
loggdp<- log(gm1$gdpPercap)  
plot(loggdp,gm1$lifeExp, main="scatterplot of life expectacy vs log(GDP per capita)", xlab="log transforemd GDP per capita",ylab="life expectancy at birth in years")

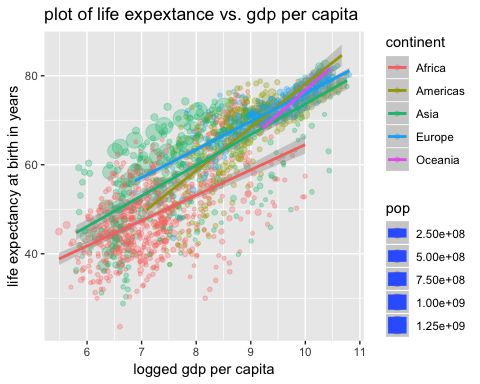


# The null hypothesis is that the life-expectancy mean difference between South Africa and Ireland is 0.  
# The alternative hypothesis is that the life-expectancy means difference is not 0.  
# Set the level of significance equal to 0.05. If the p-value is less than 0.05, reject the null hypothesis.  
# First, do the pipe operation to select country and life expectancy variables and then, filter the countries if countries are South Africa or Ireland.   
df <- gm1 %>%  
 select(country, lifeExp) %>%  
 filter(country == "South Africa" |   
 country == "Ireland") %>%  
 group\_by(country)   
  
# Then, do the t-test.   
t.test(data=df, lifeExp ~ country, alternative="two.sided",mu=0,conf.level =0.95)

##   
## Welch Two Sample t-test  
##   
## data: lifeExp by country  
## t = 10.067, df = 19.109, p-value = 4.466e-09  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 15.07022 22.97794  
## sample estimates:  
## mean in group Ireland mean in group South Africa   
## 73.01725 53.99317

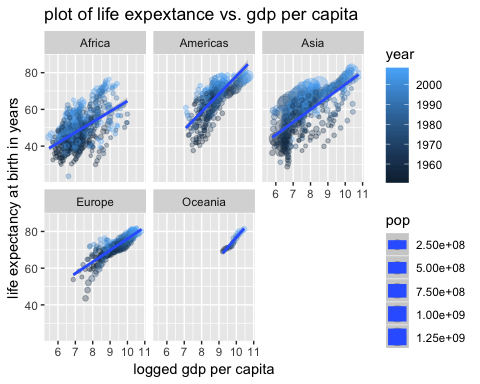
# Since the p-value =4.466e-09 is less than 0.05, we can conclude that the true difference in means is not 0.

# ggplot of life exp vs. gdpPercap   
plot\_lifeExpandgdp <- gm1 %>%  
 # grap gdp-per-capita less than 50000  
 filter(gdpPercap < 50000) %>%  
 mutate(log\_gdp=log(gdpPercap))  
ggplot(data=plot\_lifeExpandgdp, aes(x=log\_gdp,y=lifeExp,color=continent,size=pop)) +  
 geom\_point(alpha=0.3) +  
 geom\_smooth(method=lm) +  
 ylab("life expectancy at birth in years") +  
 xlab("logged gdp per capita") +  
 ggtitle("plot of life expextance vs. gdp per capita")



I dont know why the plots do not show here above and below…

# faceting ggplot  
ggplot(data=plot\_lifeExpandgdp, aes(x=log\_gdp,y=lifeExp,color=year,size=pop)) +  
 geom\_point(alpha=0.3) +  
 geom\_smooth(method=lm) +  
 ylab("life expectancy at birth in years") +  
 xlab("logged gdp per capita") +  
 ggtitle("plot of life expextance vs. gdp per capita") +  
 facet\_wrap(~ continent)



lm\_lifeExpbygdp <- lm(lifeExp ~ gdpPercap + pop,data=gm1)  
summary(lm\_lifeExpbygdp)

##   
## Call:  
## lm(formula = lifeExp ~ gdpPercap + pop, data = gm1)  
##   
## 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 \*\*\*  
## gdpPercap 7.676e-04 2.568e-05 29.89 < 2e-16 \*\*\*  
## 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