Megan Lott Coding Exercise and William Norfolk Tidyverse Exercise

We’re going to learn to code by playing around with some of the data in the dslabs package.

library(dslabs)  
  
#use the help function tosee what the dataset gapminder contains  
  
help(gapminder)  
#?gapminder would also work  
  
#inspect the data   
  
str(gapminder)

## 'data.frame': 10545 obs. of 9 variables:  
## $ country : Factor w/ 185 levels "Albania","Algeria",..: 1 2 3 4 5 6 7 8 9 10 ...  
## $ year : int 1960 1960 1960 1960 1960 1960 1960 1960 1960 1960 ...  
## $ infant\_mortality: num 115.4 148.2 208 NA 59.9 ...  
## $ life\_expectancy : num 62.9 47.5 36 63 65.4 ...  
## $ fertility : num 6.19 7.65 7.32 4.43 3.11 4.55 4.82 3.45 2.7 5.57 ...  
## $ population : num 1636054 11124892 5270844 54681 20619075 ...  
## $ gdp : num NA 1.38e+10 NA NA 1.08e+11 ...  
## $ continent : Factor w/ 5 levels "Africa","Americas",..: 4 1 1 2 2 3 2 5 4 3 ...  
## $ region : Factor w/ 22 levels "Australia and New Zealand",..: 19 11 10 2 15 21 2 1 22 21 ...

summary(gapminder)

## country year infant\_mortality  
## Albania : 57 Min. :1960 Min. : 1.50   
## Algeria : 57 1st Qu.:1974 1st Qu.: 16.00   
## Angola : 57 Median :1988 Median : 41.50   
## Antigua and Barbuda: 57 Mean :1988 Mean : 55.31   
## Argentina : 57 3rd Qu.:2002 3rd Qu.: 85.10   
## Armenia : 57 Max. :2016 Max. :276.90   
## (Other) :10203 NA's :1453   
## life\_expectancy fertility population gdp   
## Min. :13.20 Min. :0.840 Min. :3.124e+04 Min. :4.040e+07   
## 1st Qu.:57.50 1st Qu.:2.200 1st Qu.:1.333e+06 1st Qu.:1.846e+09   
## Median :67.54 Median :3.750 Median :5.009e+06 Median :7.794e+09   
## Mean :64.81 Mean :4.084 Mean :2.701e+07 Mean :1.480e+11   
## 3rd Qu.:73.00 3rd Qu.:6.000 3rd Qu.:1.523e+07 3rd Qu.:5.540e+10   
## Max. :83.90 Max. :9.220 Max. :1.376e+09 Max. :1.174e+13   
## NA's :187 NA's :185 NA's :2972   
## continent region   
## Africa :2907 Western Asia :1026   
## Americas:2052 Eastern Africa : 912   
## Asia :2679 Western Africa : 912   
## Europe :2223 Caribbean : 741   
## Oceania : 684 South America : 684   
## Southern Europe: 684   
## (Other) :5586

class(gapminder)

## [1] "data.frame"

names(gapminder)

## [1] "country" "year" "infant\_mortality"  
## [4] "life\_expectancy" "fertility" "population"   
## [7] "gdp" "continent" "region"

By inspecting the data, we can see that gapminder is a data frame that consists of demographic information of the world’s counties. The information includes: infant mortality, life expectancy, fertility, populayion, gdp, continent, region.

We want to look more closely at countries from Africa. We are going to extract that data using the subset() function in base R.

#Tidyverse Exercise   
  
#Using further packages we can make our data wrangling a bit easier. Load the tidyverse and skimr packages.  
  
library(tidyverse)  
library(skimr)

#Tidyverse Exercise  
  
#Use the glimpse function from dplyr to look at the gapminder data.  
  
glimpse(gapminder)

## Observations: 10,545  
## Variables: 9  
## $ country <fct> Albania, Algeria, Angola, Antigua and Barbuda, …  
## $ year <int> 1960, 1960, 1960, 1960, 1960, 1960, 1960, 1960,…  
## $ infant\_mortality <dbl> 115.40, 148.20, 208.00, NA, 59.87, NA, NA, 20.3…  
## $ life\_expectancy <dbl> 62.87, 47.50, 35.98, 62.97, 65.39, 66.86, 65.66…  
## $ fertility <dbl> 6.19, 7.65, 7.32, 4.43, 3.11, 4.55, 4.82, 3.45,…  
## $ population <dbl> 1636054, 11124892, 5270844, 54681, 20619075, 18…  
## $ gdp <dbl> NA, 13828152297, NA, NA, 108322326649, NA, NA, …  
## $ continent <fct> Europe, Africa, Africa, Americas, Americas, Asi…  
## $ region <fct> Southern Europe, Northern Africa, Middle Africa…

#Glimpse shows a data set with 10,545 observations and 9 variables. Variables are listed in rows containing the variable name, class, and a few early observations from the set. Glimpse appears to be similar to a "cleaner" version of the str function.

#Tidyverse Exercise  
  
#Use the skim function from skimr to look at the gapminder data.  
  
skim(gapminder)

## Skim summary statistics  
## n obs: 10545   
## n variables: 9   
##   
## ── Variable type:factor ────────────────────────  
## variable missing complete n n\_unique  
## continent 0 10545 10545 5  
## country 0 10545 10545 185  
## region 0 10545 10545 22  
## top\_counts ordered  
## Afr: 2907, Asi: 2679, Eur: 2223, Ame: 2052 FALSE  
## Alb: 57, Alg: 57, Ang: 57, Ant: 57 FALSE  
## Wes: 1026, Eas: 912, Wes: 912, Car: 741 FALSE  
##   
## ── Variable type:integer ───────────────────────  
## variable missing complete n mean sd p0 p25 p50 p75 p100  
## year 0 10545 10545 1988 16.45 1960 1974 1988 2002 2016  
## hist  
## ▇▇▇▇▇▇▇▇  
##   
## ── Variable type:numeric ───────────────────────  
## variable missing complete n mean sd p0  
## fertility 187 10358 10545 4.08 2.03 0.84  
## gdp 2972 7573 10545 1.5e+11 7e+11 4e+07   
## infant\_mortality 1453 9092 10545 55.31 47.73 1.5   
## life\_expectancy 0 10545 10545 64.81 10.67 13.2   
## population 185 10360 10545 2.7e+07 1.1e+08 31238   
## p25 p50 p75 p100 hist  
## 2.2 3.75 6 9.22 ▅▇▃▃▅▆▂▁  
## 1.8e+09 7.8e+09 5.5e+10 1.2e+13 ▇▁▁▁▁▁▁▁  
## 16 41.5 85.1 276.9 ▇▃▂▂▁▁▁▁  
## 57.5 67.54 73 83.9 ▁▁▁▂▃▅▇▃  
## 1333486 5e+06 1.5e+07 1.4e+09 ▇▁▁▁▁▁▁▁

#Skim generates a summary of the gapminder data with particular emphasis on the variables in the set. Skim breaks down each of the variables and provides a short summary that is relevant to the data class. Skim also provides the total obsevations and missing values for each variable in the data set.

#assign only the African countries to new objects/variables   
  
africadata = subset(gapminder, continent == "Africa")  
summary(africadata)

## country year infant\_mortality life\_expectancy  
## Algeria : 57 Min. :1960 Min. : 11.40 Min. :13.20   
## Angola : 57 1st Qu.:1974 1st Qu.: 62.20 1st Qu.:48.23   
## Benin : 57 Median :1988 Median : 93.40 Median :53.98   
## Botswana : 57 Mean :1988 Mean : 95.12 Mean :54.38   
## Burkina Faso: 57 3rd Qu.:2002 3rd Qu.:124.70 3rd Qu.:60.10   
## Burundi : 57 Max. :2016 Max. :237.40 Max. :77.60   
## (Other) :2565 NA's :226   
## fertility population gdp continent   
## Min. :1.500 Min. : 41538 Min. :4.659e+07 Africa :2907   
## 1st Qu.:5.160 1st Qu.: 1605232 1st Qu.:8.373e+08 Americas: 0   
## Median :6.160 Median : 5570982 Median :2.448e+09 Asia : 0   
## Mean :5.851 Mean : 12235961 Mean :9.346e+09 Europe : 0   
## 3rd Qu.:6.860 3rd Qu.: 13888152 3rd Qu.:6.552e+09 Oceania : 0   
## Max. :8.450 Max. :182201962 Max. :1.935e+11   
## NA's :51 NA's :51 NA's :637   
## region   
## Eastern Africa :912   
## Western Africa :912   
## Middle Africa :456   
## Northern Africa :342   
## Southern Africa :285   
## Australia and New Zealand: 0   
## (Other) : 0

Now, we have only 2907 observations. We are interested in examining the infant mortality, life expectancy, and population of the countries in Africa.

#Tidyverse Exercise  
  
#Extract only the African countries from the gapminder data set.   
  
africancountries <- filter(gapminder, continent == "Africa")  
  
#The object africancountries is used to store data for this exercise to distinguish itself from the previous object africadata. It should be noted that both objects contain the same data with 2907 observations and 9 variables.   
  
#To convert the data to a "friendly" viewing format convert the object africancountries into a tibble. This step is not mandatory, however it formats the data into a clean view and prevents R from printing all of the data into the console if you view the object. Note the number of observations and variables remains the same.  
  
africatibble <- tbl\_df(africancountries)  
africatibble

## # A tibble: 2,907 x 9  
## country year infant\_mortality life\_expectancy fertility population  
## <fct> <int> <dbl> <dbl> <dbl> <dbl>  
## 1 Algeria 1960 148. 47.5 7.65 11124892  
## 2 Angola 1960 208 36.0 7.32 5270844  
## 3 Benin 1960 187. 38.3 6.28 2431620  
## 4 Botswa… 1960 116. 50.3 6.62 524029  
## 5 Burkin… 1960 161. 35.2 6.29 4829291  
## 6 Burundi 1960 145. 40.6 6.95 2786740  
## 7 Camero… 1960 167. 43.5 5.65 5361367  
## 8 Cape V… 1960 NA 50.1 6.89 202316  
## 9 Centra… 1960 166. 37.4 5.84 1503501  
## 10 Chad 1960 NA 41.0 6.25 3002596  
## # … with 2,897 more rows, and 3 more variables: gdp <dbl>,  
## # continent <fct>, region <fct>

#make two new variables: one that contains only infant\_mortality and life\_expectancy and one that contains only population and life\_expectancy. The c() function might be useful to efficiently pull out the variables you want.   
  
africa\_data\_set1 = subset(africadata, select=c(infant\_mortality, life\_expectancy))  
africa\_data\_set2 = subset(africadata, select=c(population, life\_expectancy))  
  
#You should have two new objects/variables with 2907 rows and two columns.  
#NOTE: We no longer have the country information.   
#what are the units on infant mortality?  
  
#Use the str, and summary commands to take a look at both variables.  
  
str(africa\_data\_set1)

## 'data.frame': 2907 obs. of 2 variables:  
## $ infant\_mortality: num 148 208 187 116 161 ...  
## $ life\_expectancy : num 47.5 36 38.3 50.3 35.2 ...

summary(africa\_data\_set1)

## infant\_mortality life\_expectancy  
## Min. : 11.40 Min. :13.20   
## 1st Qu.: 62.20 1st Qu.:48.23   
## Median : 93.40 Median :53.98   
## Mean : 95.12 Mean :54.38   
## 3rd Qu.:124.70 3rd Qu.:60.10   
## Max. :237.40 Max. :77.60   
## NA's :226

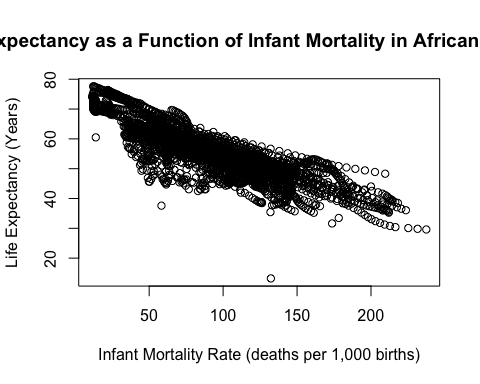
#Tidyverse Exercise  
  
#Using only African countries select the following variables to keep: infant\_mortality, life\_expectancy, population, and country. Create a new object using the previously made africatibble and use the select function to choose the variables of interest.   
  
africa\_plot\_data <- select(africatibble, life\_expectancy, infant\_mortality, population, country)  
  
africa\_plot\_data

## # A tibble: 2,907 x 4  
## life\_expectancy infant\_mortality population country   
## <dbl> <dbl> <dbl> <fct>   
## 1 47.5 148. 11124892 Algeria   
## 2 36.0 208 5270844 Angola   
## 3 38.3 187. 2431620 Benin   
## 4 50.3 116. 524029 Botswana   
## 5 35.2 161. 4829291 Burkina Faso   
## 6 40.6 145. 2786740 Burundi   
## 7 43.5 167. 5361367 Cameroon   
## 8 50.1 NA 202316 Cape Verde   
## 9 37.4 166. 1503501 Central African Republic  
## 10 41.0 NA 3002596 Chad   
## # … with 2,897 more rows

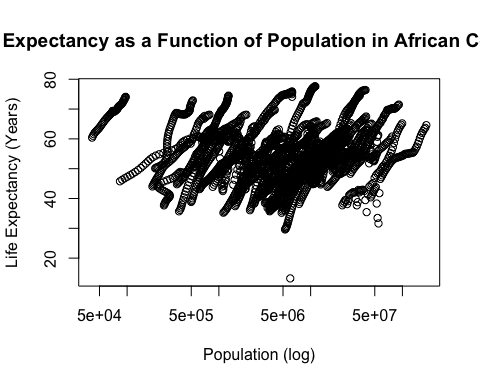
#The result is a tibble with 2907 observations and 4 variables. Note this outcome could also be acheived by selecting all variables that are not of interest and placing a - symbol in front of each of their names.

We are going to examine the data on infant mortality, life expectancy, and population by plotting this data.

#plot life expectancy as a function of infant mortality   
  
plot(africa\_data\_set1$infant\_mortality, africa\_data\_set1$life\_expectancy, xlab = "Infant Mortality Rate (deaths per 1,000 births)", ylab = "Life Expectancy (Years)", main = "Life Expectancy as a Function of Infant Mortality in African Countries")



#plot life expectancy as a function of population size   
  
plot(africa\_data\_set2$population, africa\_data\_set2$life\_expectancy, log = "x", xlab = "Population (log)", ylab = "Life Expectancy (Years)", main = "Life Expectancy as a Function of Population in African Countries")

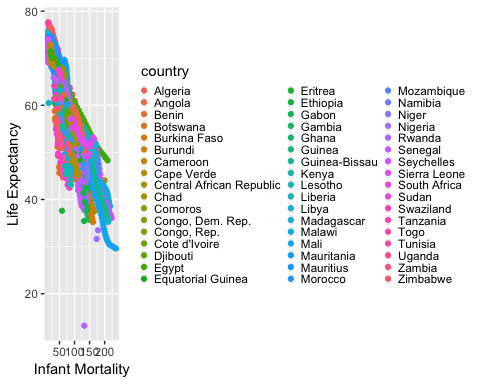


We see a negative correlation between infant mortality and life expectancy. We see a positive correlation between population size and life expectancy, but this data has streaks. **Why is this?**

We have different years for individual countries. Overtime, these countries increase in population size and in life expectancy. To see the relationship between the two variables in focus, we will tease out the data from a single year of interest. We will look at the year for which we have the most amount of data.

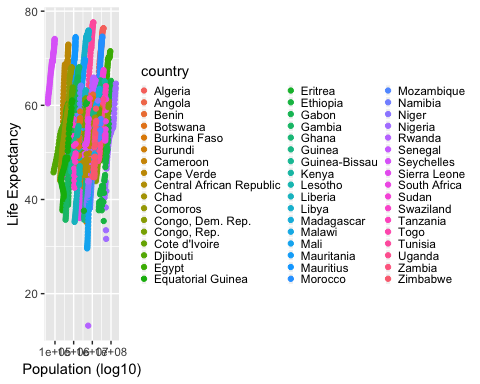
#Tidyverse Exercise   
  
#Make two plots using ggplot for life expectancy as a function of infant mortality and population. Assign different colors for each country in the data set.   
  
#Note there are two different plotting functions within ggplot2: qplot (quick plot) and ggplot. qplot is streamlined and useful for simple figures and ggplot is ideal for more complex figures. I will use qplot for my first two plots and ggplot for the third.  
  
#Make a plot of life expectancy vs. infant mortality.  
  
#Using the qplot function input the desired variables starting with x then y, color defines the data point color, data assigns the africa\_plot\_data object, the labs function creates professional labels for the x and y axes, and the theme function sets formatting to the figure legend.   
  
qplot(infant\_mortality, life\_expectancy, color = country, data = africa\_plot\_data) + labs(y = "Life Expectancy", x = "Infant Mortality") + theme(legend.key.size = unit(0.2, "cm"), legend.key.width = unit(0.1, "cm"))

## Warning: Removed 226 rows containing missing values (geom\_point).



#The resulting scatterplot shows the same negative correlation seen in the previous exercise with the addition of a color coded legend to illistrate different countries. The warning of the removal of 226 rows is consistent with the measure of NA values for infant mortality and was expected with the creation of this plot.   
  
#Make a plot of life expectancy vs. population. Remember to set the population size to a log scale.  
  
qplot(population, life\_expectancy, color = country, data = africa\_plot\_data) + labs(y = "Life Expectancy", x = "Population (log10)") + scale\_x\_log10() + theme(legend.key.size = unit(0.2, "cm"), legend.key.width = unit(0.1, "cm"))

## Warning: Removed 51 rows containing missing values (geom\_point).



#The resulting scatter plot shows the same "streaks" seen in the first coding exercise with the addition of color coded countries. With the addition of color it is easier to see that as population within a country increases life expectancy increases also. The warning of the removal of 51 rows is consistent with the measure of NA values for population and was expected with the creation of this plot.

#Write some base R code that figures out which years have missing data for infant mortality. The is.na() function might be helpful. You can use the print() function to print the missing years to the console.  
  
  
#check\_na = is.na(africadata$infant\_mortality[8])  
  
#print(africadata$year[8])  
  
  
years\_missing\_data = data.frame()  
  
x = 2907  
  
for(x in 1:2907){  
if(is.na(africadata$infant\_mortality[x] == "TRUE")){years\_missing\_data = rbind(years\_missing\_data, africadata$year[x])}}  
  
  
#You should find that there is missing up to 1981 and then again for 2016. So we’ll avoid those years and go with 2000 instead.   
  
#create a new object by extracting only the data for the year 2000 from the africadata object. You should end up with 51 observations and 9 variables. Check it with str and summary  
  
year\_2000 = subset(africadata, year == 2000)  
  
str(year\_2000)

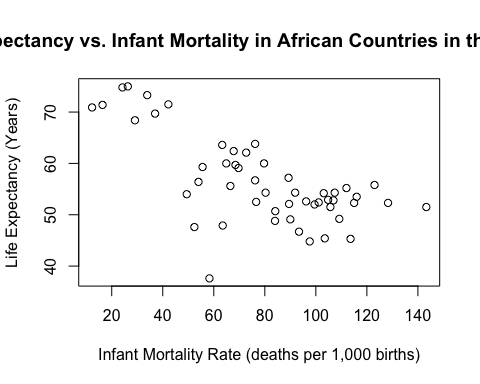
## 'data.frame': 51 obs. of 9 variables:  
## $ country : Factor w/ 185 levels "Albania","Algeria",..: 2 3 18 22 26 27 29 31 32 33 ...  
## $ year : int 2000 2000 2000 2000 2000 2000 2000 2000 2000 2000 ...  
## $ infant\_mortality: num 33.9 128.3 89.3 52.4 96.2 ...  
## $ life\_expectancy : num 73.3 52.3 57.2 47.6 52.6 46.7 54.3 68.4 45.3 51.5 ...  
## $ fertility : num 2.51 6.84 5.98 3.41 6.59 7.06 5.62 3.7 5.45 7.35 ...  
## $ population : num 31183658 15058638 6949366 1736579 11607944 ...  
## $ gdp : num 5.48e+10 9.13e+09 2.25e+09 5.63e+09 2.61e+09 ...  
## $ continent : Factor w/ 5 levels "Africa","Americas",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ region : Factor w/ 22 levels "Australia and New Zealand",..: 11 10 20 17 20 5 10 20 10 10 ...

summary(year\_2000)

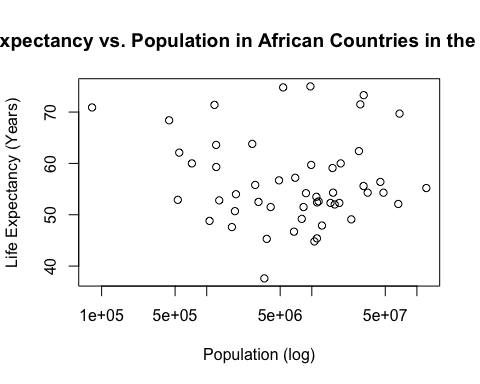
## country year infant\_mortality life\_expectancy  
## Algeria : 1 Min. :2000 Min. : 12.30 Min. :37.60   
## Angola : 1 1st Qu.:2000 1st Qu.: 60.80 1st Qu.:51.75   
## Benin : 1 Median :2000 Median : 80.30 Median :54.30   
## Botswana : 1 Mean :2000 Mean : 78.93 Mean :56.36   
## Burkina Faso: 1 3rd Qu.:2000 3rd Qu.:103.30 3rd Qu.:60.00   
## Burundi : 1 Max. :2000 Max. :143.30 Max. :75.00   
## (Other) :45   
## fertility population gdp continent   
## Min. :1.990 Min. : 81154 Min. :2.019e+08 Africa :51   
## 1st Qu.:4.150 1st Qu.: 2304687 1st Qu.:1.274e+09 Americas: 0   
## Median :5.550 Median : 8799165 Median :3.238e+09 Asia : 0   
## Mean :5.156 Mean : 15659800 Mean :1.155e+10 Europe : 0   
## 3rd Qu.:5.960 3rd Qu.: 17391242 3rd Qu.:8.654e+09 Oceania : 0   
## Max. :7.730 Max. :122876723 Max. :1.329e+11   
##   
## region   
## Eastern Africa :16   
## Western Africa :16   
## Middle Africa : 8   
## Northern Africa : 6   
## Southern Africa : 5   
## Australia and New Zealand: 0   
## (Other) : 0

Now, we can examine the relationship between infant mortality, life expectancy, and population in the Year 2000.

#use base R plotting again and do the same two plots again, this time only for the year 2000  
  
plot(year\_2000$infant\_mortality, year\_2000$life\_expectancy, xlab = "Infant Mortality Rate (deaths per 1,000 births)", ylab = "Life Expectancy (Years)", main = "Life Expectancy vs. Infant Mortality in African Countries in the Year 2000")



plot(year\_2000$population, year\_2000$life\_expectancy, log = "x", xlab = "Population (log)", ylab = "Life Expectancy (Years)", main = "Life Expectancy vs. Population in African Countries in the Year 2000")



We see that there is a negative correlation between infant mortality and life expectancy, but no noticeable correlation between population size and life expectancy.

# Use the lm function and fit life expectancy as the outcome, and infant mortality as the predictor. Then use the population size as the predictor.   
  
fit1 = lm(year\_2000$life\_expectancy ~ year\_2000$infant\_mortality)  
  
summary(fit1)

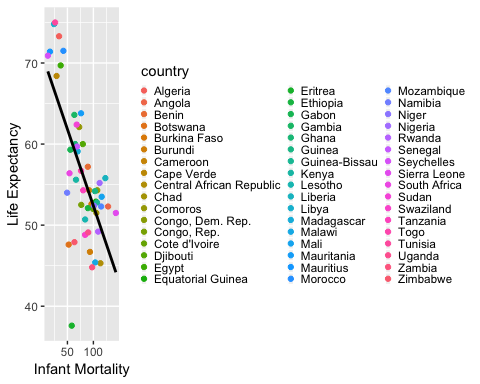
##   
## Call:  
## lm(formula = year\_2000$life\_expectancy ~ year\_2000$infant\_mortality)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -22.6651 -3.7087 0.9914 4.0408 8.6817   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 71.29331 2.42611 29.386 < 2e-16 \*\*\*  
## year\_2000$infant\_mortality -0.18916 0.02869 -6.594 2.83e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.221 on 49 degrees of freedom  
## Multiple R-squared: 0.4701, Adjusted R-squared: 0.4593   
## F-statistic: 43.48 on 1 and 49 DF, p-value: 2.826e-08

fit2 = lm(year\_2000$life\_expectancy ~ year\_2000$population)  
  
summary(fit2)

##   
## Call:  
## lm(formula = year\_2000$life\_expectancy ~ year\_2000$population)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.429 -4.602 -2.568 3.800 18.802   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.593e+01 1.468e+00 38.097 <2e-16 \*\*\*  
## year\_2000$population 2.756e-08 5.459e-08 0.505 0.616   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.524 on 49 degrees of freedom  
## Multiple R-squared: 0.005176, Adjusted R-squared: -0.01513   
## F-statistic: 0.2549 on 1 and 49 DF, p-value: 0.6159

The p-value for fit1 is 2.826e-08. There is a significant correlation between life expectancy and infant mortality. The p-value for fit2 is 0.6159. There is no significant correlation between life expectancy and population.

#Tidyverse Exercise  
  
#Write code that pulls Africa and the year 2000 out of the gapminder data set and then plot life expectancy as a function of infant mortality with a linear fit model added.   
  
#First create an object to select for Africa from the continent variable, and 2000 from the year variable.  
  
africa2000 <- filter(gapminder, continent == "Africa" & year == 2000)  
  
#Plot life expectancy vs. infant mortality with the addition of a linear fit model.   
  
#Using the ggplot function define the africa2000 data and set the axes. geom\_point defines a scatterplot, stat\_smooth applies the linear fit with the method set to "lm", col sets the color of the regression line, se applies or removes the standard error field surrounding the line.   
  
ggplot(africa2000, aes(x = infant\_mortality, y = life\_expectancy, color = country)) + geom\_point() + stat\_smooth(method = "lm", col = "black", se = FALSE) + labs(y = "Life Expectancy", x = "Infant Mortality") + theme(legend.key.size = unit(0.2, "cm"), legend.key.width = unit(0.1, "cm"))



#The resulting plot shows a distinct negative correlation between life expectancy and infant mortality as expected from the previous exercise.