Emily Rayens Coding Exercise 1

I first pulled up the gapminder dataset from dslabs using the library() and help() functions.

library("dslabs")  
data(gapminder)  
help(gapminder)

To get a good look at my data, I then used the str() and summary() functions for structure and summary.

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

I then used the class() function to check what type of object gapminder is.

class(gapminder)

## [1] "data.frame"

From this output, we can see it is a data.frame object.

At this point, we want to work with our data, starting by assigning only the African countries to a new object called africadata. We can do this by selecting the rows with the the continent set as “Africa”. In order to assess the success of this selection, I ran a new structure and summary assessment of the object.

africadata <- subset(gapminder, continent == "Africa")  
str(africadata)

## 'data.frame': 2907 obs. of 9 variables:  
## $ country : Factor w/ 185 levels "Albania","Algeria",..: 2 3 18 22 26 27 29 31 32 33 ...  
## $ year : int 1960 1960 1960 1960 1960 1960 1960 1960 1960 1960 ...  
## $ infant\_mortality: num 148 208 187 116 161 ...  
## $ life\_expectancy : num 47.5 36 38.3 50.3 35.2 ...  
## $ fertility : num 7.65 7.32 6.28 6.62 6.29 6.95 5.65 6.89 5.84 6.25 ...  
## $ population : num 11124892 5270844 2431620 524029 4829291 ...  
## $ gdp : num 1.38e+10 NA 6.22e+08 1.24e+08 5.97e+08 ...  
## $ 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(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

As this output has the anticpated 2907 observations (down from 10545), we can assume that the subset has been correctly generated.

Now that we’ve had a practice round, we want to create more complex variables, starting with “death” which will contain the infant\_mortality and life\_expectancy subsets from africa data. In order to achieve this, we will use the c() function within that subset selection we used in filtering out the African countries. And as before, we will check the accuracy of the variable generation by its structure and summary.

death <- subset(africadata, select = c(infant\_mortality, life\_expectancy))  
str(death)

## '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(death)

## 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

We will repeat this by creating a “life” object that contains population and life\_expectancy.

life <- subset(africadata, select = c(population, life\_expectancy))  
str(life)

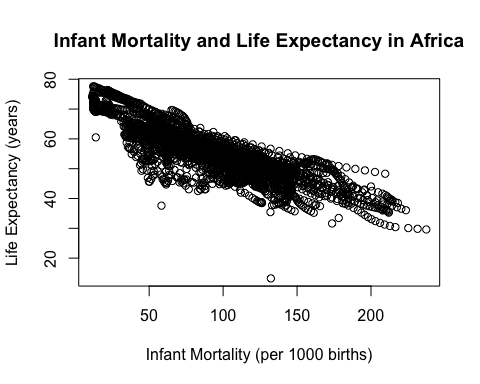
## 'data.frame': 2907 obs. of 2 variables:  
## $ population : num 11124892 5270844 2431620 524029 4829291 ...  
## $ life\_expectancy: num 47.5 36 38.3 50.3 35.2 ...

summary(life)

## population life\_expectancy  
## Min. : 41538 Min. :13.20   
## 1st Qu.: 1605232 1st Qu.:48.23   
## Median : 5570982 Median :53.98   
## Mean : 12235961 Mean :54.38   
## 3rd Qu.: 13888152 3rd Qu.:60.10   
## Max. :182201962 Max. :77.60   
## NA's :51

At this point, we want to graph our new variables and see if we can visualize any relationships between them. We will have two plots in total (as we created new “death” and “life” object), starting with life expectancy as a function of infant mortality, which will place infant mortality on the x axis and life expectancy on the y axis.

plot(death$infant\_mortality, death$life\_expectancy,  
 xlab= "Infant Mortality (per 1000 births)", ylab="Life Expectancy (years)",  
 main="Infant Mortality and Life Expectancy in Africa")



From here, we can see that as infant mortality increases, the overall life expectancy decreases, meaning there is a negative correlation between these two variables.

We will now graph the “death” object with life expectancy as a function of population size, which will place population size on the x axis and life expectancy on the y axis. Due to the variability and size of values for population size, it will help us to adjust the x axis to a log scale.

plot(log(life$population, 10), y = life$life\_expectancy,   
 xlab="Population Size (Log10)", ylab="Life Expectancy (years)",  
 main = "Population Size and Life Expectancy in Africa")



This plot ends up looking quite a bit different as there are many individual lines of association. However, all these show a positive correlation between the variables, meaning as population size increases, life expectancy increases. We can look back at the death plot and see these individual lines of association (however, they overlap to a greater degree). These individual associations most likely represent sampling within single regions or countries, distinct by a number of different factors but with the positive correlation in population size and life expectancy remaining the same.

As we can see the change in individual countries and these countries increase in population size (and life expectancy), we want to look within a single year to look for pattterns. To help us make the most applicable conclusions, we want to select the largest possible data set within a single year and the best way to start narrowing this down is by removing incomplete data sets for infant mortality.

To do this, we need to organize which years are missing data and we can do this by generating a table that displays the years and their corresponding midding infant mortality data.

table(as.factor(africadata$year), is.na(africadata$infant\_mortality),  
 dnn = c("year", "Missing infant mortality"))

## Missing infant mortality  
## year FALSE TRUE  
## 1960 41 10  
## 1961 34 17  
## 1962 35 16  
## 1963 35 16  
## 1964 36 15  
## 1965 37 14  
## 1966 38 13  
## 1967 40 11  
## 1968 40 11  
## 1969 44 7  
## 1970 46 5  
## 1971 45 6  
## 1972 45 6  
## 1973 45 6  
## 1974 46 5  
## 1975 46 5  
## 1976 48 3  
## 1977 48 3  
## 1978 49 2  
## 1979 49 2  
## 1980 50 1  
## 1981 50 1  
## 1982 51 0  
## 1983 51 0  
## 1984 51 0  
## 1985 51 0  
## 1986 51 0  
## 1987 51 0  
## 1988 51 0  
## 1989 51 0  
## 1990 51 0  
## 1991 51 0  
## 1992 51 0  
## 1993 51 0  
## 1994 51 0  
## 1995 51 0  
## 1996 51 0  
## 1997 51 0  
## 1998 51 0  
## 1999 51 0  
## 2000 51 0  
## 2001 51 0  
## 2002 51 0  
## 2003 51 0  
## 2004 51 0  
## 2005 51 0  
## 2006 51 0  
## 2007 51 0  
## 2008 51 0  
## 2009 51 0  
## 2010 51 0  
## 2011 51 0  
## 2012 51 0  
## 2013 51 0  
## 2014 51 0  
## 2015 51 0  
## 2016 0 51

From here, we can see that there are a number of years without missing values, including the year 2000. To look for patterns, we will extracting on the data from the year 2000 by creating a new Year2000 subset.

Year2000 <- subset(africadata, year =="2000")  
str(Year2000)

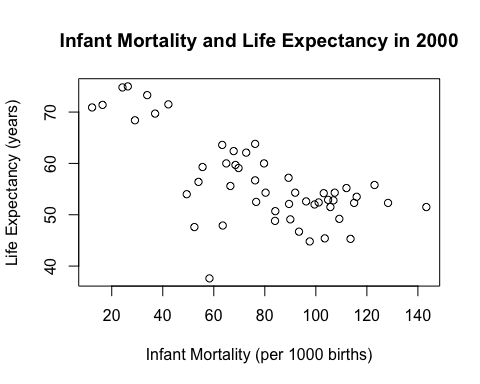
## 'data.frame': 51 obs. of 9 variables:  
## $ country : Factor w/ 185 levels "Albania","Algeria",..: 2 3 18 22 26 27 29 31 32 33 ...  
## $ year : int 2000 2000 2000 2000 2000 2000 2000 2000 2000 2000 ...  
## $ infant\_mortality: num 33.9 128.3 89.3 52.4 96.2 ...  
## $ life\_expectancy : num 73.3 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(Year2000)

## 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

With the structure and summary, we can see that there are only 51 observations and 9 variables in this Year2000 subset. And now that we’ve specified this dataset, we can plot this smaller population with the same comparisons as we set previously.

plot(Year2000$infant\_mortality, Year2000$life\_expectancy,  
 xlab = "Infant Mortality (per 1000 births)", ylab = "Life Expectancy (years)",   
 main = "Infant Mortality and Life Expectancy in 2000")



We can see that overall the same negative correlation is maintained between life expectancy and infant mortality.

plot(log(Year2000$population), Year2000$life\_expectancy,  
 xlab = "Population Size (log10)", ylab = "Life Expectancy (years)",   
 main = "Population Size and Life Expectancy in 2000")



Unlike we saw in the plot inclusing all African countries, it is not clear there is a correlation between these variables. This is where statistics to determine a non-random relationship comes into play. We will first look at the relationship between life expectancy and infant mortality.

fit1 <- lm(Year2000$life\_expectancy ~ Year2000$infant\_mortality)  
summary(fit1)

##   
## Call:  
## lm(formula = Year2000$life\_expectancy ~ Year2000$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 \*\*\*  
## Year2000$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

With a p-value <0.01, we can reject the null hypothesis that there is no relationship between life expectancy and infant mortality (F=43.48, d.f.=1,49, p=2.826e-08). The negative correlation is confrimed with a slope of -0.18916 which means for every year increase in life expectancy, we expect a decrease of 0.18916 in infant mortality per 1000 births.

To examine the realtionship between life expectancy and population in the year 2000, I am going to examine with population both in and out of the log format to look for differences.

fit2 <- lm(Year2000$life\_expectancy ~ Year2000$population)  
fit3 <- lm(Year2000$life\_expectancy ~ log(Year2000$population, 10))  
summary(fit2)

##   
## Call:  
## lm(formula = Year2000$life\_expectancy ~ Year2000$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 \*\*\*  
## Year2000$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

summary(fit3)

##   
## Call:  
## lm(formula = Year2000$life\_expectancy ~ log(Year2000$population,   
## 10))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19.113 -4.809 -1.554 3.907 18.863   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 65.324 12.520 5.217 3.65e-06 \*\*\*  
## log(Year2000$population, 10) -1.315 1.829 -0.719 0.476   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.502 on 49 degrees of freedom  
## Multiple R-squared: 0.01044, Adjusted R-squared: -0.009755   
## F-statistic: 0.517 on 1 and 49 DF, p-value: 0.4755

While we can see that analyzing the correlation in log form improves the p-value (from p=0.616 to 0.4755), we cannot reject the null hypothesis that there is no significant relationship between life expectancy and population size (F=0.517, d.f.=1,49, p=0.4755).

##Tidyverse Exercise by Megan Robertson

Now we can look at the data using tidyverse. Let’s load the tidyverse and skimr packages.

library(tidyverse)

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

## ✔ ggplot2 3.2.1 ✔ purrr 0.3.2  
## ✔ tibble 2.1.3 ✔ dplyr 0.8.3  
## ✔ tidyr 0.8.3 ✔ stringr 1.4.0  
## ✔ readr 1.3.1 ✔ forcats 0.4.0

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

library(skimr)

##   
## Attaching package: 'skimr'

## The following object is masked from 'package:stats':  
##   
## filter

Let’s look at gapminder data again using these packages.

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…

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 ▇▁▁▁▁▁▁▁

The summary from the glimpse function shows us 10,545 observations and 9 variables (similar to what is above wih the base R exercise). The skim function also breaks down by summary statistics by variable type and tells us the total number of missing values.

Let’s only look at the African countries. Let’s a new data frame called TidyAfrica and filter for the countries on Africa.

TidyAfrica <- filter(gapminder, continent == "Africa")

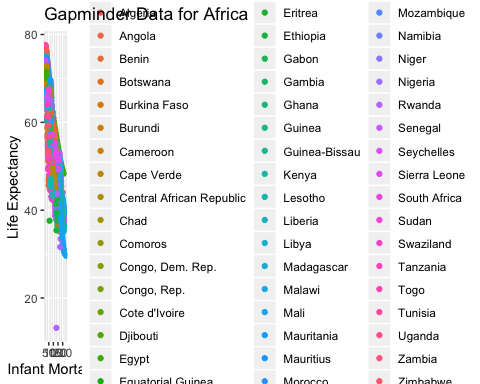
This TidyAfrica data frame shows up in the global environment with the same observations and variables as base R Africa data frame. Next, let’s only keep the variables we are interested in. We can use the select function to pull out only the columns we want.

TidyAfrica2 <- select(TidyAfrica, infant\_mortality, life\_expectancy, population, country )

Now let’s use ggplot to a plot of life expectancy as a function of infant mortality.

ggplot(TidyAfrica2, aes(x = infant\_mortality, y = life\_expectancy)) +  
 geom\_point(aes(color = country)) +   
 labs(title = "Gapminder Data for Africa", x = "Infant Mortality", y = "Life Expectancy")

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

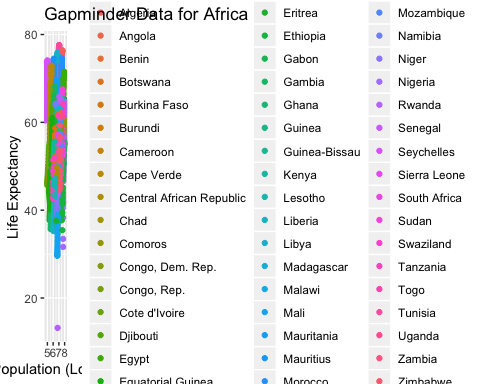


As above, you can see that as infant mortality increases, life expectancy decreases.

Let’s look at population vs life expectancy.

ggplot(TidyAfrica2, aes(x = log10(population), y = life\_expectancy)) +  
 geom\_point(aes(color = country)) +  
 labs(title = "Gapminder Data for Africa", x = "Population (Log 10)", y = "Life Expectancy")

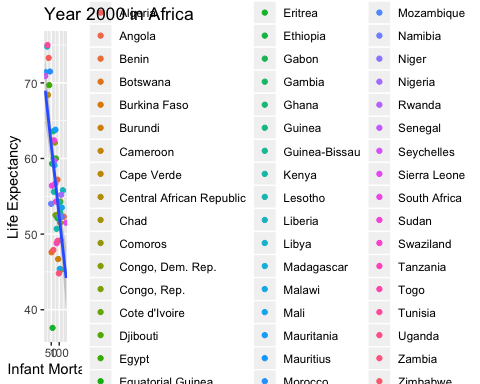
## Warning: Removed 51 rows containing missing values (geom\_point).



As you can see from above, as population increases, life expectancy also increases.

Let’s look specifically at Year 2000 in African countries.

gapminder %>%  
 filter(continent == "Africa" & year == 2000) %>%  
 ggplot(aes(x = infant\_mortality, y = life\_expectancy)) +  
 geom\_point(aes(color = country)) +  
 geom\_smooth(method = "lm") +  
 labs(title = "Year 2000 in Africa", x = "Infant Mortality", y = "Life Expectancy")



There is a negative association between infant mortality and life expectancy.

Overall, tidyverse does what base R does, but much faster! This coding exercise was definitely easier to do, however gapminder is a tidy dataset and didn’t require any cleaning.

##End of Tidyverse Exercise