R coding Exercise

#First, load the “dslabs” package (install if needed). Also, load any other packages needed.

library("dslabs")  
library("tidyverse")  
library(dplyr)

#Look at help file for gapminder data.

help("gapminder")

#Overview of data structure.

str("gapminder")

## chr "gapminder"

#Summary of gapminder data.

summary("gapminder")

## Length Class Mode   
## 1 character character

#Determine object type of gapminder data.

class("gapminder")

## [1] "character"

#Assigns only the African countries to a new object/variable called “africadata”.

africadata <- gapminder %>%  
dplyr::filter(continent == "Africa")

#Check to see if you did it correclty (should have 2907 observations)

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

#Create new object with the Life expectancy and infant mortality of African Countries.

africa\_life\_infant <- africadata %>%  
 select("country", "life\_expectancy", "infant\_mortality")

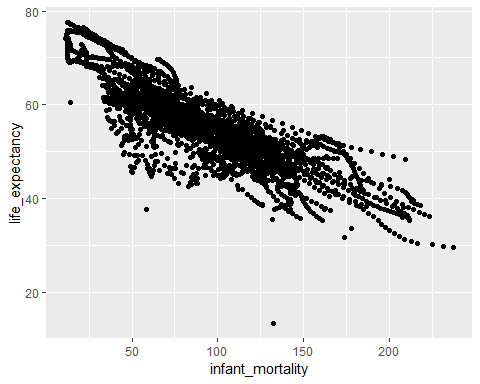
#Create new object with the Life expectancy and population of African Countries.

africa\_life\_pop <- africadata %>%  
 select("country", "life\_expectancy", "population")

#Plot life expectancy as a function of infant mortality.

ggplot(data = africa\_life\_infant) +   
 geom\_point(mapping = aes(infant\_mortality, life\_expectancy))

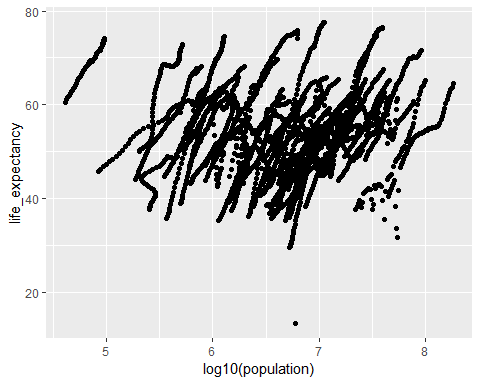
## Warning: Removed 226 rows containing missing values (geom\_point).



#Plot life expectancy as a function of infant mortality.

ggplot(data = africa\_life\_pop) +   
 geom\_point(mapping = aes(log10(population), life\_expectancy))

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



#Identify missing Data for infant mortality.

Missing1 <- africadata %>%   
 select(year,infant\_mortality) %>%   
 filter(is.na(infant\_mortality)) %>%   
 count(year) %>%   
 arrange(-n)

#Select data from Y2K

Y2K\_africadata <- africadata %>%  
 filter(year == "2000")

#Create new object with the Life expectancy and infant mortality of African Countries for Y2K.

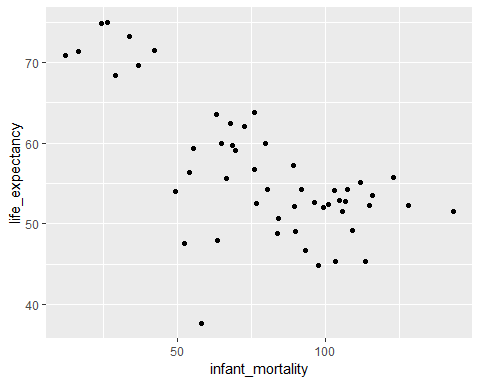
Y2K\_africa\_life\_infant <- Y2K\_africadata %>%  
 select("country", "life\_expectancy", "infant\_mortality")

#Create new object with the Life expectancy and population of African Countries.

Y2K\_africa\_life\_pop <- Y2K\_africadata %>%  
 select("country", "life\_expectancy", "population")

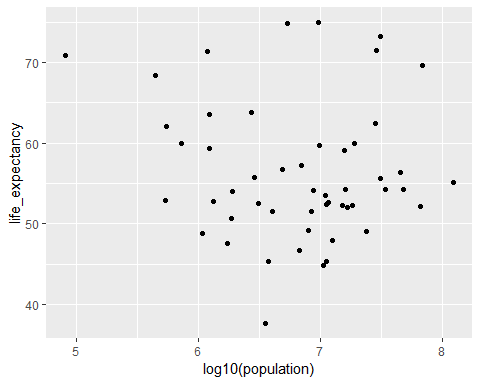
#Plot life expectancy as a function of infant mortality.

ggplot(data = Y2K\_africa\_life\_infant) +   
 geom\_point(mapping = aes(infant\_mortality, life\_expectancy))



#Plot life expectancy as a function of infant mortality.

ggplot(data = Y2K\_africa\_life\_pop) +   
 geom\_point(mapping = aes(log10(population), life\_expectancy))



#Linear Model: Infant Mortality, Life Expectancy for Y2K.

fit\_infant\_life <- lm(life\_expectancy~infant\_mortality, data = Y2K\_africa\_life\_infant)  
summary(fit\_infant\_life)

##   
## Call:  
## lm(formula = life\_expectancy ~ infant\_mortality, data = Y2K\_africa\_life\_infant)  
##   
## 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 \*\*\*  
## 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

#p-value = 2.826e-08; Based on the p-value, there is a significant correlation between infant mortality and life expectancy. As such, the slope of the relationship is negative, indicating a significant negative relationship between infant mortlaity and life expectancy. As infant mortality increases, life expectancy decreases.

#Linear Model: population, Life Expectancy for Y2K.

fit\_pop\_life <- lm(life\_expectancy~population, data = Y2K\_africa\_life\_pop)  
summary(fit\_pop\_life)

##   
## Call:  
## lm(formula = life\_expectancy ~ population, data = Y2K\_africa\_life\_pop)  
##   
## 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 \*\*\*  
## 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

#P-value = 0.6159; Based on the p-value, no significant relationship between population size and life expectancy exists.

2021.09.09: Additions by MYC

install broom

library(broom)

#Attempt to print lm outputs into table (populations and life expectancy)

print(tidy(fit\_pop\_life))

## # A tibble: 2 x 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) 55.9 1.47 38.1 4.51e-38  
## 2 population 0.0000000276 0.0000000546 0.505 6.16e- 1

#Attempt to print lm outputs into table (infant mortality and life expectancy)

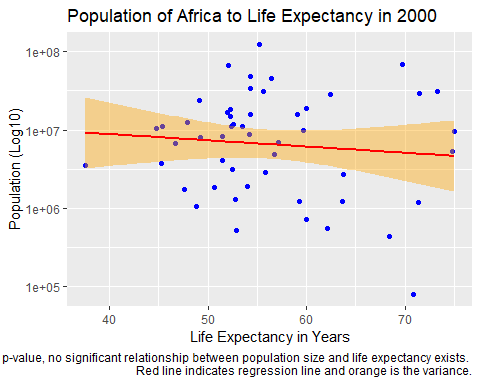
print(tidy(fit\_infant\_life))

## # A tibble: 2 x 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) 71.3 2.43 29.4 8.91e-33  
## 2 infant\_mortality -0.189 0.0287 -6.59 2.83e- 8

#Population of Africa to Life Expectancy in 2000

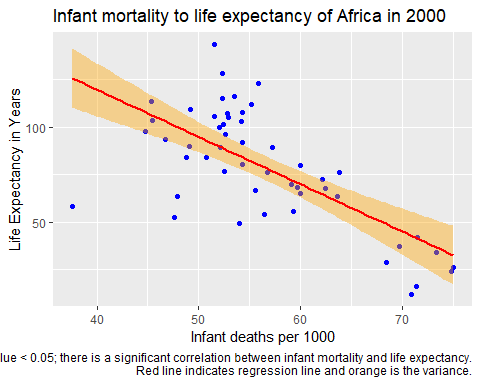
Y2K\_africa\_life\_pop%>%  
 ggplot(aes(life\_expectancy, population))+  
 scale\_y\_log10()+  
 geom\_point(color="blue")+  
 geom\_smooth(method = "lm", se=TRUE, fill="orange", color="red")+  
# stat\_smooth(fill="orange", color="orange")+  
 labs(x="Life Expectancy in Years", y="Population (Log10)", title = "Population of Africa to Life Expectancy in 2000", caption = "P-value > 0.05; Based on the p-value, no significant relationship between population size and life expectancy exists. \n Red line indicates regression line and orange is the variance.")

## `geom\_smooth()` using formula 'y ~ x'

 #Infant mortality to life expectancy of Africa in 2000

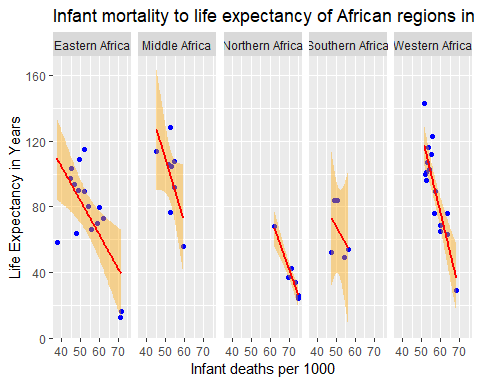
Y2K\_africa\_life\_infant%>%  
 ggplot(aes(life\_expectancy, infant\_mortality))+  
 geom\_point(color="blue")+  
 geom\_smooth(method = "lm", se=TRUE, fill="orange", color="red")+  
 labs(x="Infant deaths per 1000", y="Life Expectancy in Years", title = "Infant mortality to life expectancy of Africa in 2000", caption = "P-value < 0.05; there is a significant correlation between infant mortality and life expectancy.\n Red line indicates regression line and orange is the variance.")

## `geom\_smooth()` using formula 'y ~ x'

 #Infant Mortality by region

Y2K\_africadata%>%  
 ggplot(aes(life\_expectancy, infant\_mortality))+  
 geom\_point(color="blue")+  
 geom\_smooth(method = "lm", se=TRUE, fill="orange", color="red")+  
 facet\_grid(.~region)+  
 labs(x="Infant deaths per 1000", y="Life Expectancy in Years", title = "Infant mortality to life expectancy of African regions in 2000")

## `geom\_smooth()` using formula 'y ~ x'

 Batch Tidy linear regression by African regions in 2000

Y2K\_africadata%>%  
 group\_by(region)%>%  
 do(tidy(lm(life\_expectancy~infant\_mortality,.)))

## # A tibble: 10 x 6  
## # Groups: region [5]  
## region term estimate std.error statistic p.value  
## <fct> <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 Eastern Africa (Intercept) 69.6 5.17 13.5 2.12e- 9  
## 2 Eastern Africa infant\_mortality -0.210 0.0637 -3.29 5.37e- 3  
## 3 Middle Africa (Intercept) 63.7 5.25 12.1 1.90e- 5  
## 4 Middle Africa infant\_mortality -0.111 0.0523 -2.13 7.77e- 2  
## 5 Northern Africa (Intercept) 82.3 1.50 54.8 6.65e- 7  
## 6 Northern Africa infant\_mortality -0.290 0.0365 -7.94 1.36e- 3  
## 7 Southern Africa (Intercept) 57.2 7.20 7.95 4.15e- 3  
## 8 Southern Africa infant\_mortality -0.0884 0.108 -0.819 4.73e- 1  
## 9 Western Africa (Intercept) 70.9 2.37 29.9 4.32e-14  
## 10 Western Africa infant\_mortality -0.153 0.0248 -6.19 2.34e- 5