HW2-Data Analysis

Aylin Mumcular

21 09 2019

install.packages("gapminder") install.packages("tidyverse") install.packages("dplyr")

1.1 Use filter() to subset the gapminder data to three countries of your choice in the 1970's.

```
rm(list = ls(all.names = TRUE)) #Clear the environment

knitr::kable((gap70 <-  gapminder %>%
                filter(country == "Afghanistan" | country == "Albania" | country == "Algeria",
                   year >= 1970 & year < 1980)))
```

country	continent	year	lifeExp	pop	gdpPercap
Afghanistan	Asia	1972	36.088	13079460	739.9811
Afghanistan	Asia	1977	38.438	14880372	786.1134
Albania	Europe	1972	67.690	2263554	3313.4222
Albania	Europe	1977	68.930	2509048	3533.0039
Algeria	Africa	1972	54.518	14760787	4182.6638
Algeria	Africa	1977	58.014	17152804	4910.4168

1.2 Use the pipe operator %>% to select "country" and "gdpPercap" from your filtered dataset in 1.1.

```
knitr::kable(gap70 %>%
          select(country, gdpPercap))
```

country	gdpPercap
Afghanistan	739.9811
Afghanistan	786.1134
Albania	3313.4222
Albania	3533.0039
Algeria	4182.6638
Algeria	4910.4168

1.3 Filter gapminder to all entries that have experienced a drop in life expectancy. Be sure to include a new variable that's the increase in life expectancy in your tibble. Hint: you might find the lag() or diff() functions useful.

```
lifeExpDiff <- 0 #Define an empty array
length <- nrow(gapminder)-1 #For loop upper end

#Assign the first element to zero since one cannot make any comparison
#with the first data point
lifeExpDiff[1] <- 0

for (k in 1:length) {lifeExpDiff[k+1] <-
      if (
           select(gapminder[k+1,], country) == select(gapminder[k,], country)) {
        #If countries are the same for subsequent rows
```

```
          select(gapminder[k+1,], lifeExp) - select(gapminder[k,], lifeExp)}
        #Calculate the difference
      else {0}
        #Else, assign the first element of each country to zero
                    }

#Unlist the lifeExpDiff2 array to attach the gapminder dataset with mutate
lifeExpDiff2 <- unlist(lifeExpDiff, use.names=FALSE)

gapminder2 <- mutate(gapminder, lifeExpDiff2) #Combined dataset

DT::datatable(gapminder2 %>% filter(lifeExpDiff2<0)) #Filter the results
```

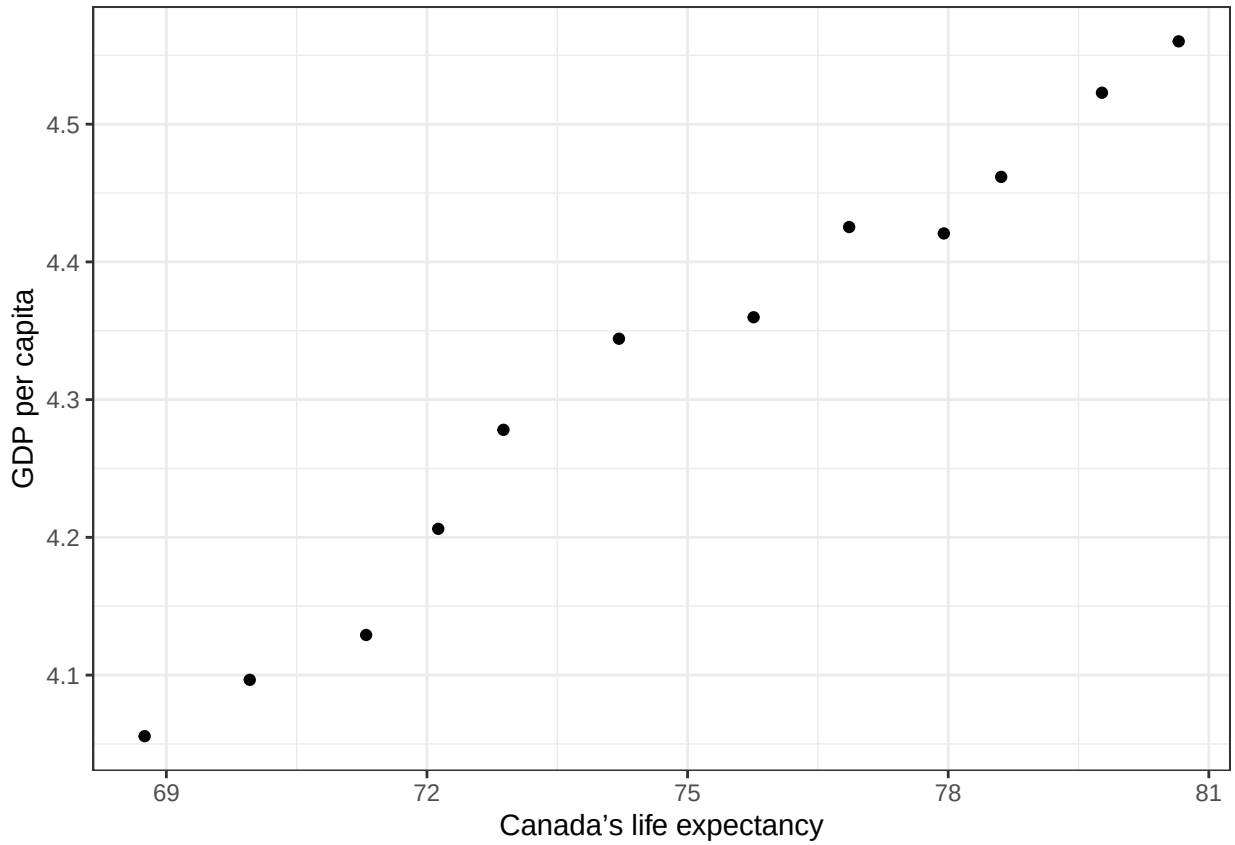
1.4 Filter gapminder to contain six rows: the rows with the three largest GDP per capita, and the rows with the three smallest GDP per capita. Be sure to not create any intermediate objects when doing this (with, for example, the assignment operator). Hint: you might find the sort() function useful, or perhaps even the dplyr::slice() function.

```
#Sort the dataset according to descending gdpPercap
#Concatenate the first and last three data points

knitr::kable(rbind(slice(arrange(gapminder, desc(gdpPercap)), 1:3),

slice(arrange(gapminder, desc(gdpPercap)), (nrow(gapminder)-2):nrow(gapminder))))
```

country	continent	year	lifeExp	pop	gdpPercap
Kuwait	Asia	1957	58.033	212846	113523.1329
Kuwait	Asia	1972	67.712	841934	109347.8670
Kuwait	Asia	1952	55.565	160000	108382.3529
Lesotho	Africa	1952	42.138	748747	298.8462
Congo, Dem. Rep.	Africa	2007	46.462	64606759	277.5519
Congo, Dem. Rep.	Africa	2002	44.966	55379852	241.1659

1.5 Produce a scatterplot of Canada's life expectancy vs. GDP per capita using ggplot2, without defining a new variable. That is, after filtering the gapminder data set, pipe it directly into the ggplot() function. Ensure GDP per capita is on a log scale.

```
gapminder %>%
  filter(country == "Canada") %>%
  mutate(loggdp = log10(gdpPercap)) %>% #Log scale gdpPercap
  ggplot(aes(lifeExp, loggdp)) +
  geom_point() +
  theme_bw() +
  xlab("Canada's life expectancy") +
  ylab("GDP per capita")
```



2 Pick one categorical variable and one quantitative variable to explore. Answer the following questions in whichever way you think is appropriate: What are possible values (or range, whichever is appropriate) of each variable? What values are typical? What's the spread? What's the distribution? Etc., tailored to the variable at hand. Feel free to use summary stats, tables, figures.

Categorical: country

```
levels(gapminder$country) #Possible values for countries
```

```
##   [1] "Afghanistan"              "Albania"
##   [3] "Algeria"                  "Angola"
##   [5] "Argentina"                "Australia"
##   [7] "Austria"                  "Bahrain"
##   [9] "Bangladesh"               "Belgium"
##  [11] "Benin"                    "Bolivia"
##  [13] "Bosnia and Herzegovina"   "Botswana"
##  [15] "Brazil"                   "Bulgaria"
##  [17] "Burkina Faso"             "Burundi"
##  [19] "Cambodia"                 "Cameroon"
##  [21] "Canada"                   "Central African Republic"
##  [23] "Chad"                     "Chile"
##  [25] "China"                    "Colombia"
##  [27] "Comoros"                  "Congo, Dem. Rep."
##  [29] "Congo, Rep."              "Costa Rica"
##  [31] "Cote d'Ivoire"            "Croatia"
##  [33] "Cuba"                     "Czech Republic"
##  [35] "Denmark"                  "Djibouti"
```

```
##  [37] "Dominican Republic"       "Ecuador"
##  [39] "Egypt"                    "El Salvador"
##  [41] "Equatorial Guinea"        "Eritrea"
##  [43] "Ethiopia"                 "Finland"
##  [45] "France"                   "Gabon"
##  [47] "Gambia"                   "Germany"
##  [49] "Ghana"                    "Greece"
##  [51] "Guatemala"                "Guinea"
##  [53] "Guinea-Bissau"            "Haiti"
##  [55] "Honduras"                 "Hong Kong, China"
##  [57] "Hungary"                  "Iceland"
##  [59] "India"                    "Indonesia"
##  [61] "Iran"                     "Iraq"
##  [63] "Ireland"                  "Israel"
##  [65] "Italy"                    "Jamaica"
##  [67] "Japan"                    "Jordan"
##  [69] "Kenya"                    "Korea, Dem. Rep."
##  [71] "Korea, Rep."              "Kuwait"
##  [73] "Lebanon"                  "Lesotho"
##  [75] "Liberia"                  "Libya"
##  [77] "Madagascar"               "Malawi"
##  [79] "Malaysia"                 "Mali"
##  [81] "Mauritania"               "Mauritius"
##  [83] "Mexico"                   "Mongolia"
##  [85] "Montenegro"               "Morocco"
##  [87] "Mozambique"               "Myanmar"
##  [89] "Namibia"                  "Nepal"
##  [91] "Netherlands"              "New Zealand"
##  [93] "Nicaragua"                "Niger"
##  [95] "Nigeria"                  "Norway"
##  [97] "Oman"                     "Pakistan"
##  [99] "Panama"                   "Paraguay"
## [101] "Peru"                     "Philippines"
## [103] "Poland"                   "Portugal"
## [105] "Puerto Rico"              "Reunion"
## [107] "Romania"                  "Rwanda"
## [109] "Sao Tome and Principe"    "Saudi Arabia"
## [111] "Senegal"                  "Serbia"
## [113] "Sierra Leone"             "Singapore"
## [115] "Slovak Republic"          "Slovenia"
## [117] "Somalia"                  "South Africa"
## [119] "Spain"                    "Sri Lanka"
## [121] "Sudan"                    "Swaziland"
## [123] "Sweden"                   "Switzerland"
## [125] "Syria"                    "Taiwan"
## [127] "Tanzania"                 "Thailand"
## [129] "Togo"                     "Trinidad and Tobago"
## [131] "Tunisia"                  "Turkey"
## [133] "Uganda"                   "United Kingdom"
## [135] "United States"            "Uruguay"
## [137] "Venezuela"                "Vietnam"
## [139] "West Bank and Gaza"       "Yemen, Rep."
## [141] "Zambia"                   "Zimbabwe"
```

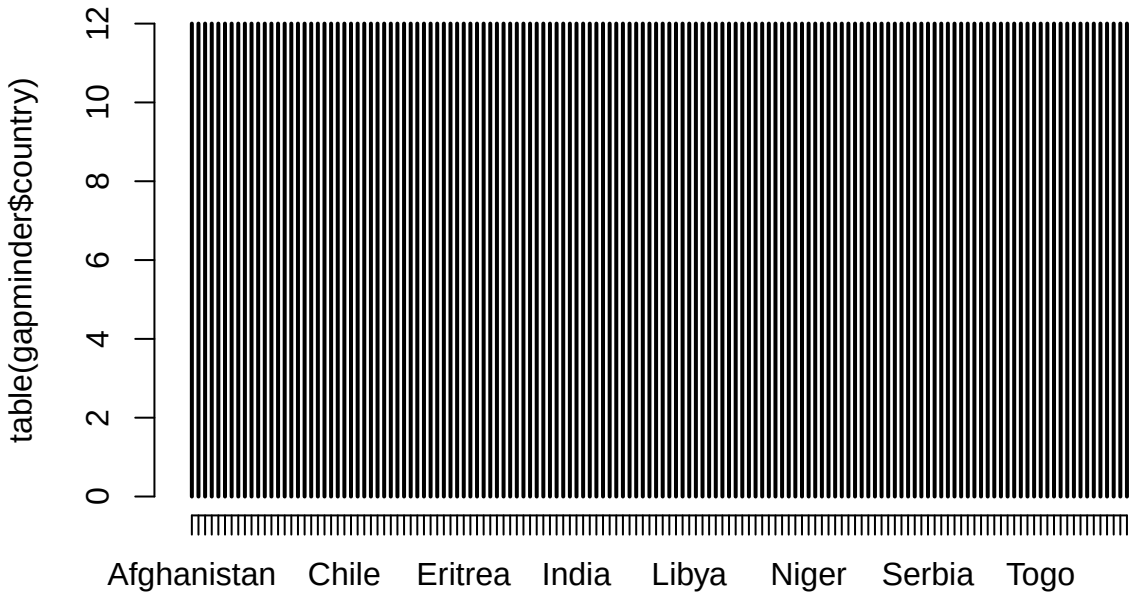
```r
nlevels(gapminder$country) #Total number of possible countries
```

```
## [1] 142
```

```r
head(table(gapminder$country)) #How many time a country appears in the data
```

```
##
## Afghanistan     Albania     Algeria      Angola   Argentina   Australia
##          12          12          12          12          12          12
```

```r
#Plot the spread/distribution of the number of times a country appears
plot(table(gapminder$country))
```



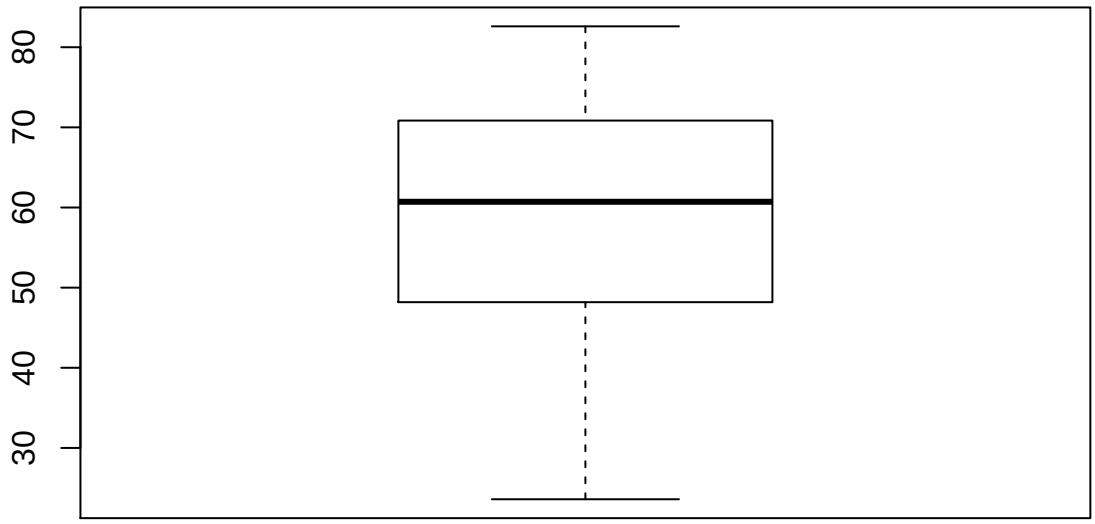
There are 142 different countries that are expressed with the levels function. Each country appears 12 times uniformly.

Quantitative: lifeExp

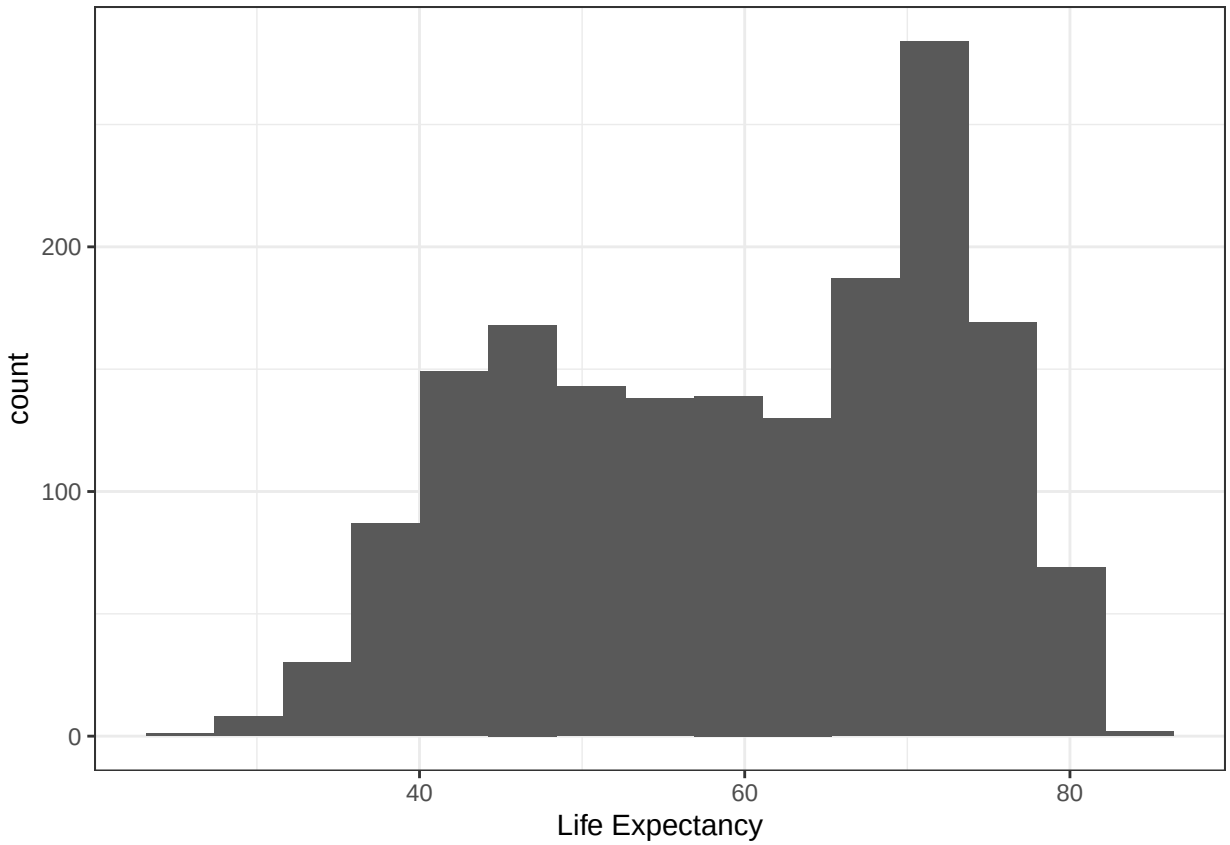
```r
summary(gapminder$lifeExp) #Summary statistics
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   23.60   48.20   60.71   59.47   70.85   82.60
```

```r
boxplot(gapminder$lifeExp) #Visual illustration of the summary statistics
```



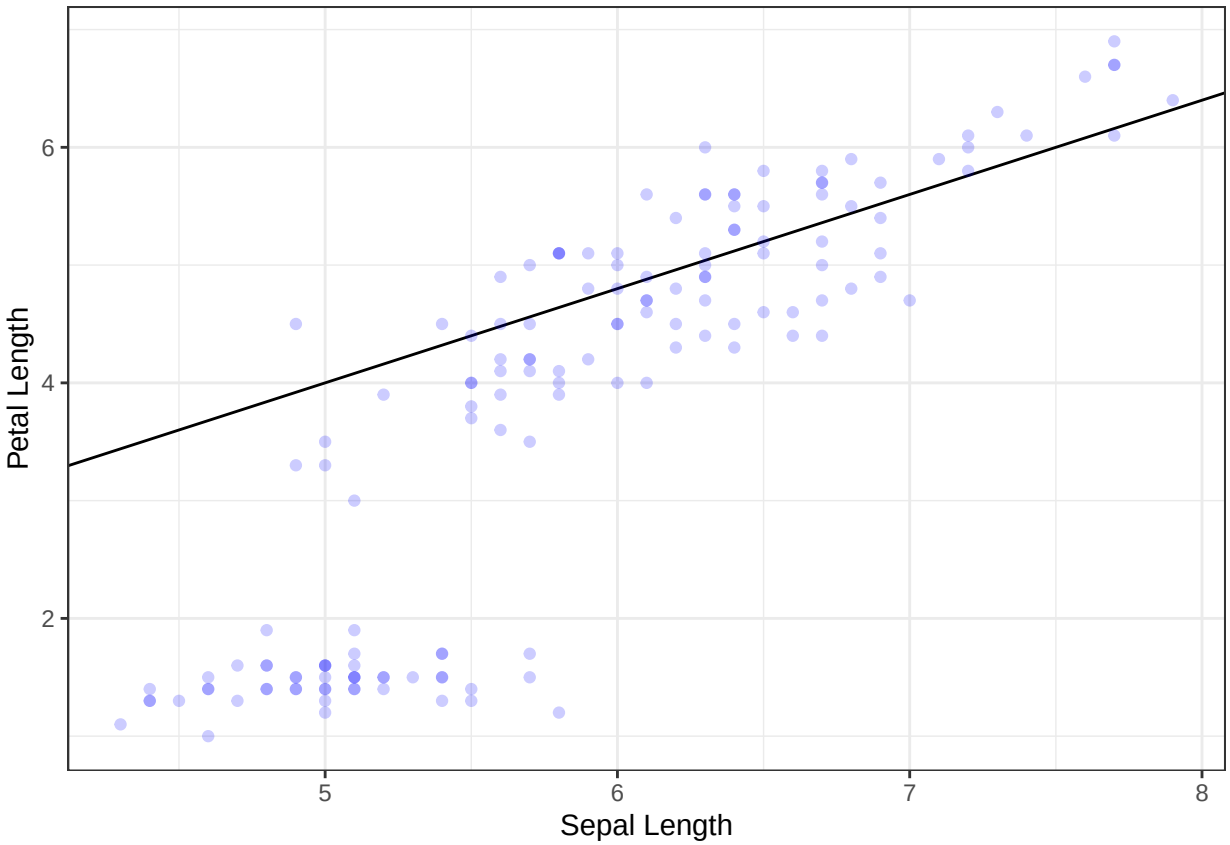

```
#Plot the histogram to see which values are typical and understand the spread/distribution
gapminder %>%
  ggplot(aes(lifeExp)) +
  geom_histogram(bins=15) +
  theme_bw() +
  xlab("Life Expectancy")
```



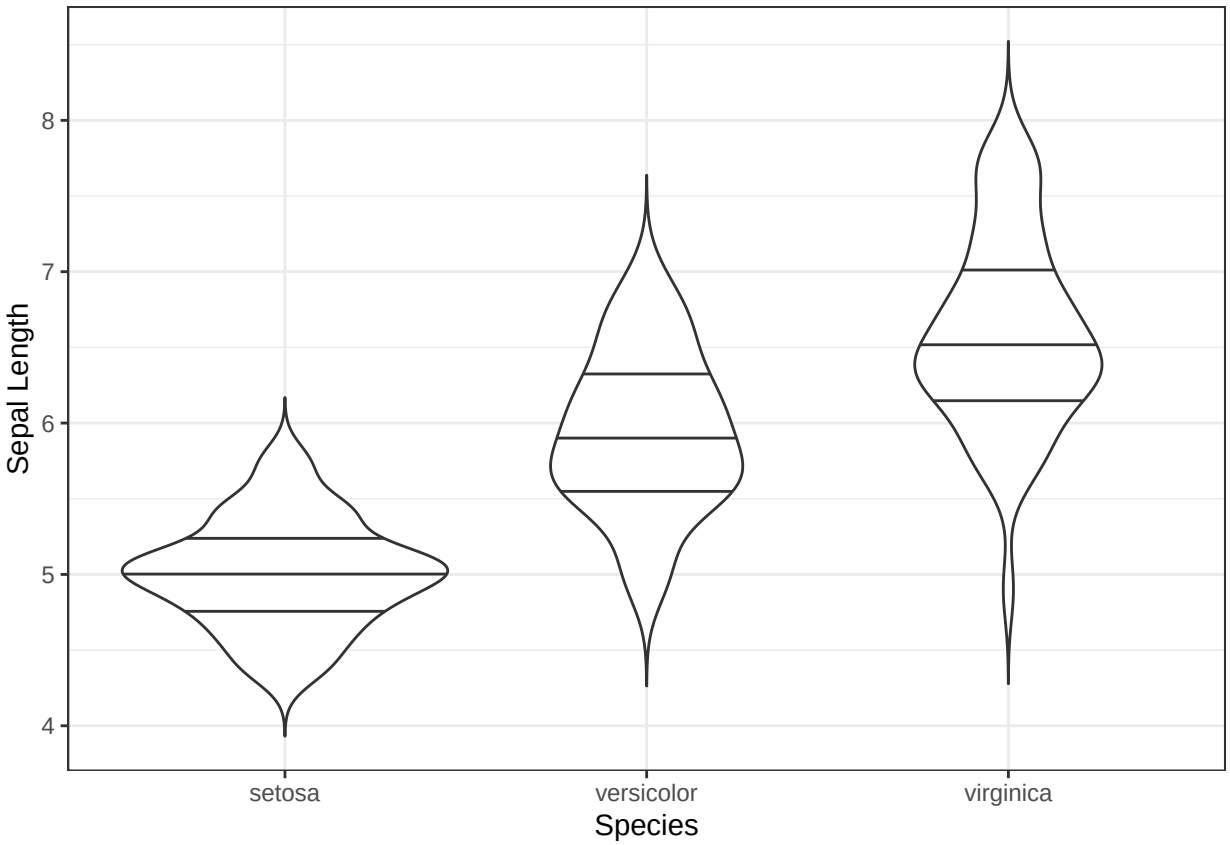
lifeExp variable is between 23.60 and 82.60 with a mean of 60.71 for all data points. According to the histogram of lifeExp variable which shows its distribution, it is more likely that this variable is between 70 and 75.

3 Make two plots that have some value to them. That is, plots that someone might actually consider making for an analysis: A scatterplot of two quantitative variables, one other plot besides a scatterplot.

```
data <- datasets::iris

# Scatter plot
ggplot(data) +
  geom_abline(intercept = 0, slope = 0.8) +
  geom_point(aes(x = Sepal.Length, y = Petal.Length), alpha = 0.2, color = "blue") +
  theme_bw() +
  xlab("Sepal Length") +
  ylab("Petal Length")
```




```
# The other plot

data %>%
  ggplot(aes(Species, Sepal.Length)) +
  geom_violin(draw_quantiles = c(.25, .5, .75), trim=FALSE) +
  theme_bw() +
  xlab("Species") +
  ylab("Sepal Length")
```



Optional Question

The code results in half of the data available for both Afghanistan and Rwanda. Out of 12 data points for each contry, only 6 of them were captured with this filtering. Hence, they did not succeed in getting all the data for Rwanda and Afghanistan. The code can be fixed with two different ways.

```
knitr::kable(filter(gapminder, country == c("Rwanda", "Afghanistan")))
```

country	continent	year	lifeExp	pop	gdpPercap
Afghanistan	Asia	1957	30.332	9240934	820.8530
Afghanistan	Asia	1967	34.020	11537966	836.1971
Afghanistan	Asia	1977	38.438	14880372	786.1134
Afghanistan	Asia	1987	40.822	13867957	852.3959
Afghanistan	Asia	1997	41.763	22227415	635.3414
Afghanistan	Asia	2007	43.828	31889923	974.5803
Rwanda	Africa	1952	40.000	2534927	493.3239
Rwanda	Africa	1962	43.000	3051242	597.4731
Rwanda	Africa	1972	44.600	3992121	590.5807
Rwanda	Africa	1982	46.218	5507565	881.5706
Rwanda	Africa	1992	23.599	7290203	737.0686
Rwanda	Africa	2002	43.413	7852401	785.6538

```
#First correction
DT::datatable(filter(gapminder, country %in% c("Rwanda", "Afghanistan")))
```

```r
#or

#Second correction
DT::datatable(filter(gapminder, country == "Rwanda" | country == "Afghanistan"))
```