STA 141 Worksheet 2

Richard McCormick

September 12, 2023

Due Date: Thursday, September 21, 2023 before 11:00am.

Instructions

Worksheets must be turned in as a PDF file through Canvas. The worksheet is worth a total of **15 points**, which is 3 percent of your overall grade.

Exercises

Before you start! You will likely need to install the `tidyverse` package. To do this type `install.packages("tidyverse")` into the console window in RStudio.

Once a package is installed on your computer, you need to add it to your library any time that you want to use it. The following code block will be on all of your worksheets from now on. It will load the packages you need into your library but it won't appear on the final pdf.

Exercise 1

(a) Load the `diabetes` dataset that is available on canvas.

```r
diabetes <- read.csv( 'diabetes.csv' )
```

(b) Use the structure function `str()` to see the variable types of each column.

```r
str( diabetes )
```

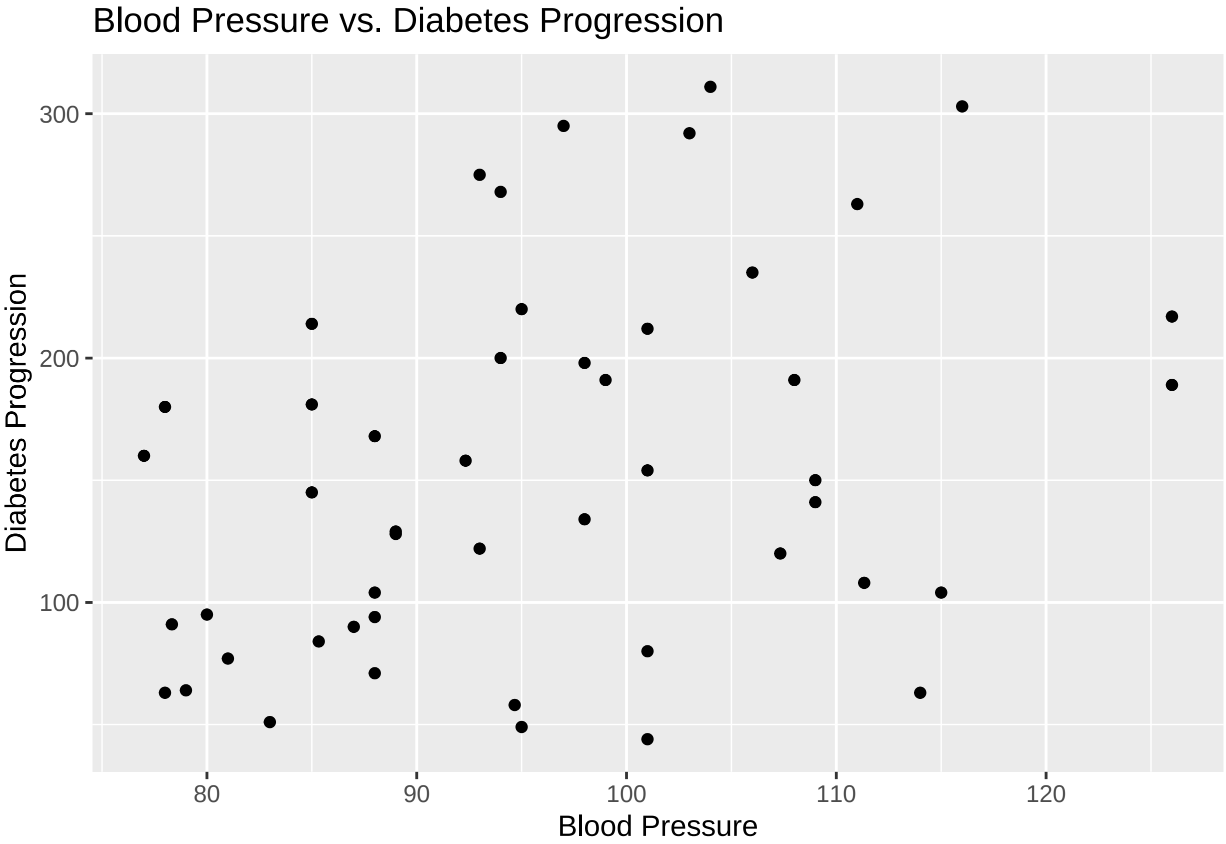
```
## 'data.frame':    49 obs. of  11 variables:
##  $ AGE: int  66 25 32 46 61 64 65 60 53 60 ...
##  $ SEX: int  2 2 1 1 1 2 2 2 1 2 ...
##  $ BMI: num  26.2 24.3 30.5 24.9 25.8 27.3 30.2 23.4 22 27.5 ...
##  $ BP : num  114 95 89 115 98 109 98 88 94 106 ...
##  $ S1 : int  255 162 182 198 235 186 219 153 175 229 ...
##  $ S2 : num  185 98.6 110.6 129.6 125.8 ...
##  $ S3 : int  56 54 56 54 76 38 40 58 59 51 ...
##  $ S4 : num  4.55 3 3 4 3 5 5 3 3 4 ...
##  $ S5 : num  4.25 3.85 4.34 4.28 5.11 ...
##  $ S6 : int  92 87 89 103 82 99 84 95 98 91 ...
##  $ Y  : int  63 49 129 104 134 150 198 104 200 235 ...
```

(c) Change the `SEX` variable to be categorical.

```r
diabetes$SEX <- as.factor( diabetes$SEX )
```

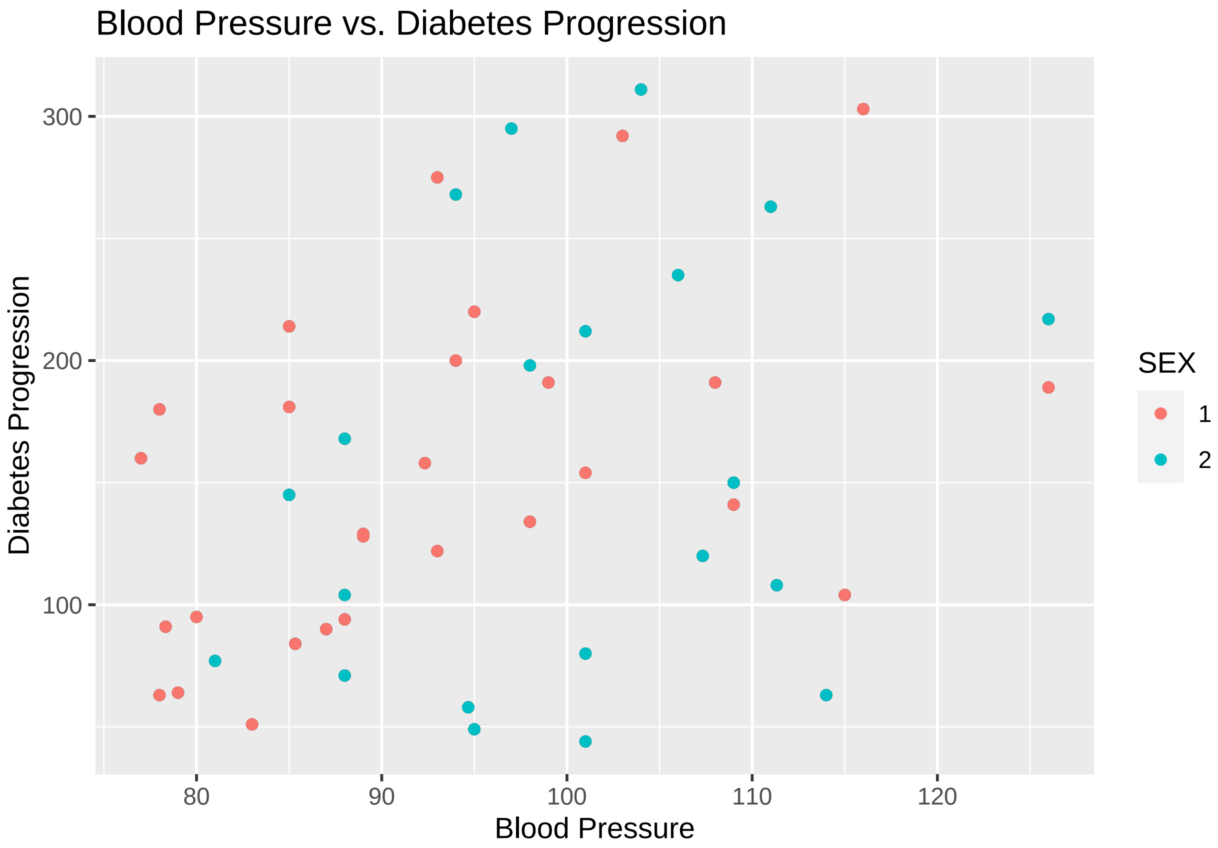
(d) Use ggplot to create a points plot of blood pressure (BP) against Diabetes Progression (Y).

```
diabetes_plot <- ggplot( data=diabetes, aes( x=BP, y=Y ) ) +
  geom_point() +
  labs( x="Blood Pressure",
        y="Diabetes Progression",
        title="Blood Pressure vs. Diabetes Progression" )

diabetes_plot
```



(d) Repeat your points plot of blood pressure (BP) against Diabetes Progression (Y) but map the sex variable (SEX) to the color of the points.

```
diabetes_plot <- diabetes_plot +
  geom_point( aes( color=SEX ) )

diabetes_plot
```

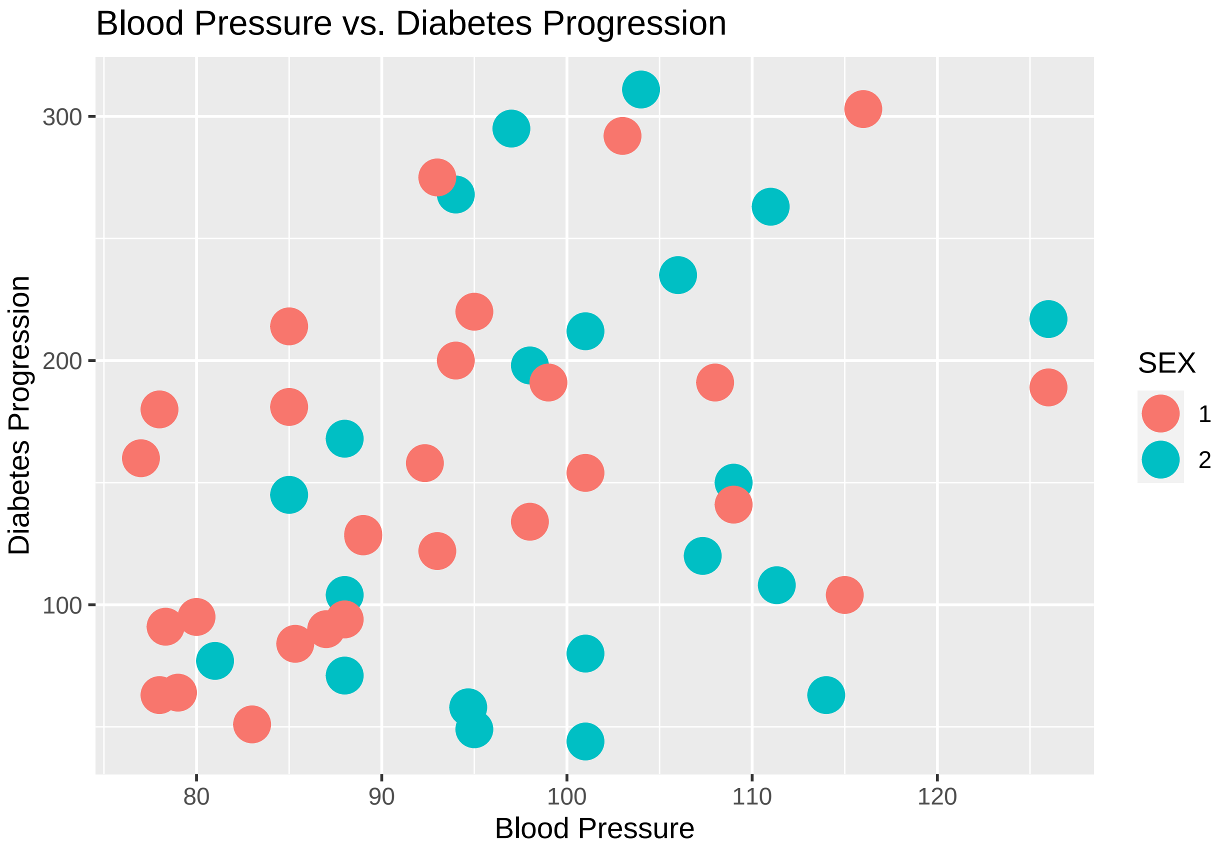


(e) Is the color in (d) an example of hue or saturation?

The color is an example of hue, not saturation. This is because we changed sex to be a categorical variable, which saturation cannot map to.

(f) Repeat your points plot from (d) but change the size of the points to size 6. (Hint: as we are not mapping the size aesthetic to a variable, it does not go in the aesthetics!)

```
diabetes_plot <- diabetes_plot +
  geom_point( aes( color=SEX ), size=6 )

diabetes_plot
```



Exercise 2

(a) Load the fish dataset that is available on canvas.

```r
fish <- read.csv( 'Fish-1.csv', fileEncoding="UTF-8-BOM" )
```

(b) Run the structure command in the following code block using whatever you called your dataset in part (a). After looking at the output decide if there are any variables that need to be cast to categorial variables (factors). If so, do this in the second code block.

```r
str( fish )
```

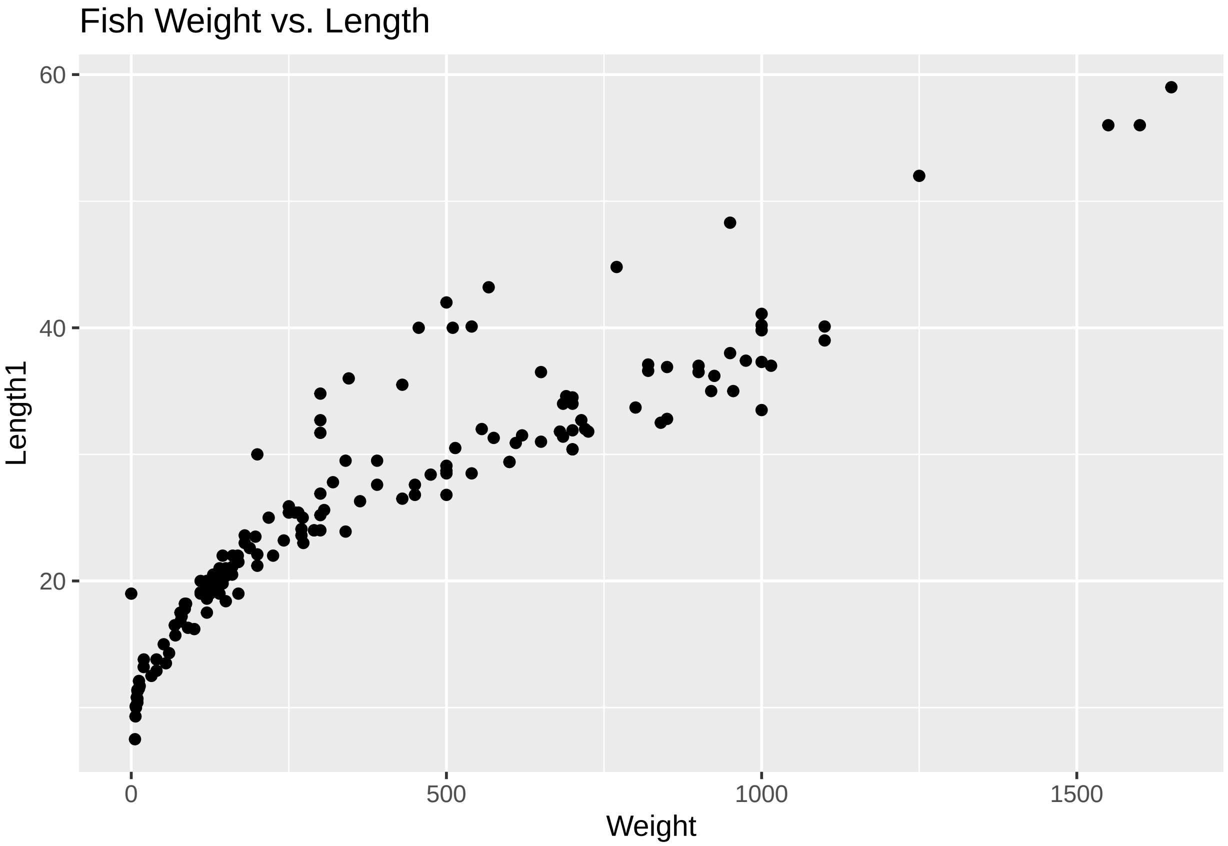
```
## 'data.frame':    159 obs. of  7 variables:
##  $ Species: chr  "Bream" "Bream" "Bream" "Bream" ...
##  $ Weight : num  242 290 340 363 430 450 500 390 450 500 ...
##  $ Length1: num  23.2 24 23.9 26.3 26.5 26.8 26.8 27.6 27.6 28.5 ...
##  $ Length2: num  25.4 26.3 26.5 29 29 29.7 29.7 30 30 30.7 ...
##  $ Length3: num  30 31.2 31.1 33.5 34 34.7 34.5 35 35.1 36.2 ...
##  $ Height : num  11.5 12.5 12.4 12.7 12.4 ...
##  $ Width  : num  4.02 4.31 4.7 4.46 5.13 ...
```

```r
fish$Species <- as.factor( fish$Species )
str( fish )
```

```
## 'data.frame':    159 obs. of  7 variables:
##  $ Species: Factor w/ 7 levels "Bream","Parkki",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Weight : num  242 290 340 363 430 450 500 390 450 500 ...
##  $ Length1: num  23.2 24 23.9 26.3 26.5 26.8 26.8 27.6 27.6 28.5 ...
##  $ Length2: num  25.4 26.3 26.5 29 29 29.7 29.7 30 30 30.7 ...
##  $ Length3: num  30 31.2 31.1 33.5 34 34.7 34.5 35 35.1 36.2 ...
##  $ Height : num  11.5 12.5 12.4 12.7 12.4 ...
##  $ Width  : num  4.02 4.31 4.7 4.46 5.13 ...
```

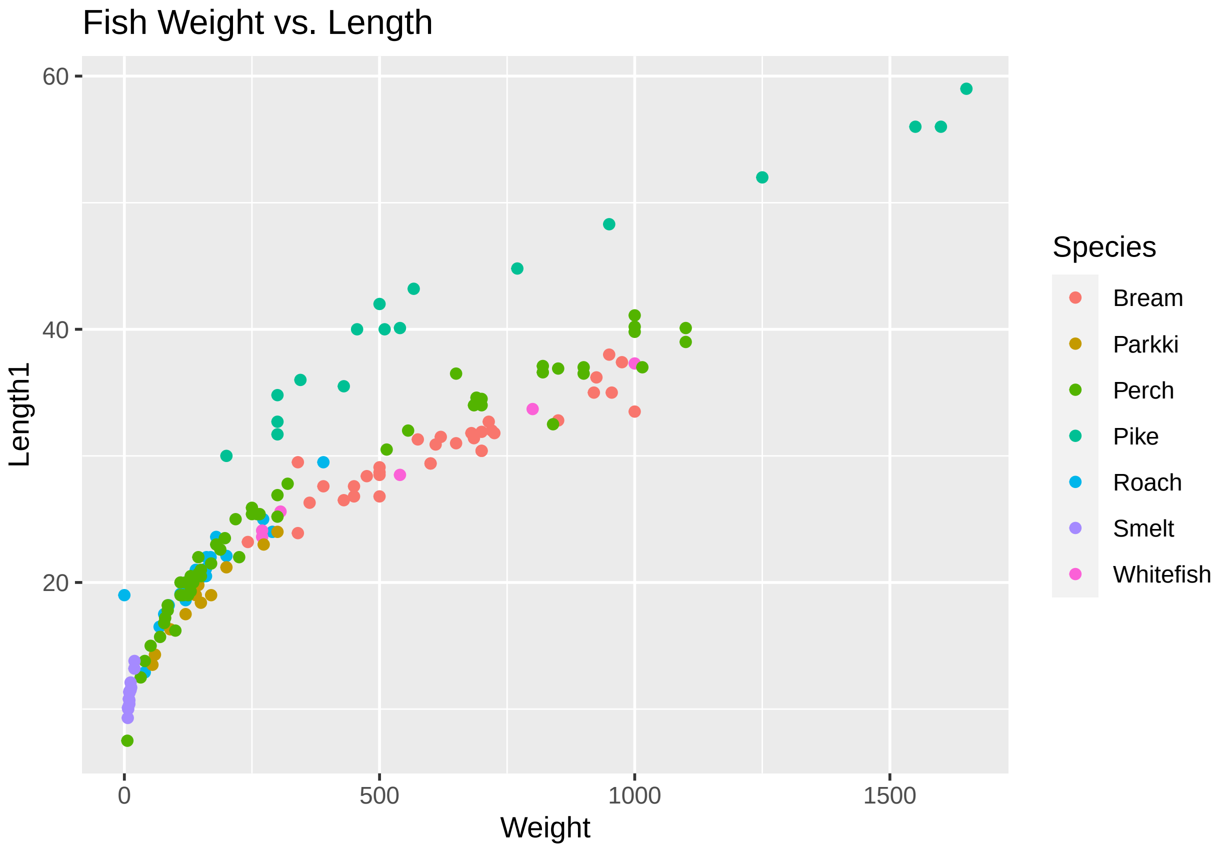
(c) Create a points plot of the fish's weights against their lengths (using the `Length1` variable).

```
fish_plot <- ggplot( data=fish ) +
  geom_point( aes( x=Weight, y=Length1 ) ) +
  labs( title="Fish Weight vs. Length" )

fish_plot
```



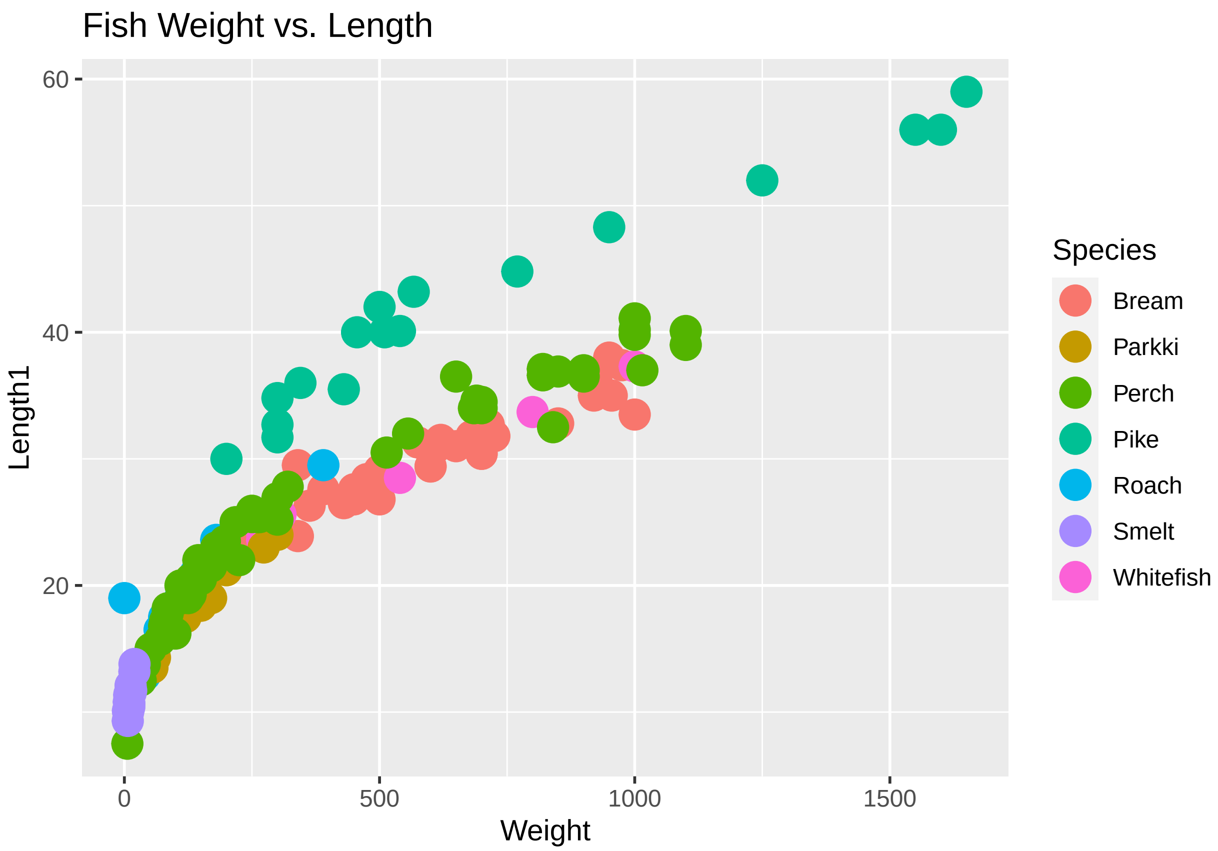
(c) Create the same points plot as in part (b) but now map the Species variable to color.

```
fish_plot <- ggplot( data=fish ) +
  geom_point( aes( x=Weight, y=Length1, color=Species ) ) +
  labs( title="Fish Weight vs. Length" )


fish_plot
```



(d) Create the same points plot as in part (c) but now change the size of the points to 5.

```
fish_plot <- ggplot( data=fish ) +
  geom_point( aes( x=Weight, y=Length1, color=Species ), size=5 ) +
  labs( title="Fish Weight vs. Length" )

fish_plot
```



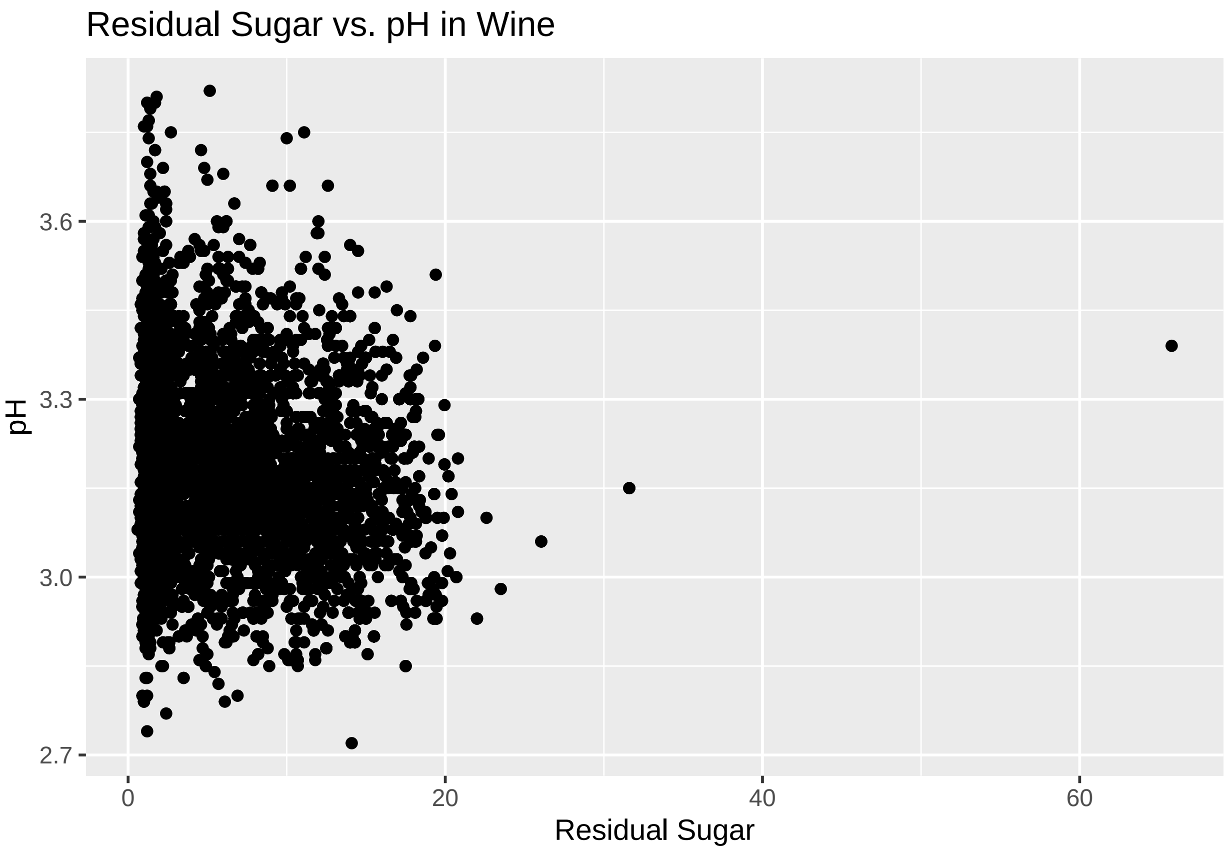
Exercise 3

(a) Load the `wine` dataset that is available on canvas.

```
wine <- read.csv( 'wine.csv' )
```

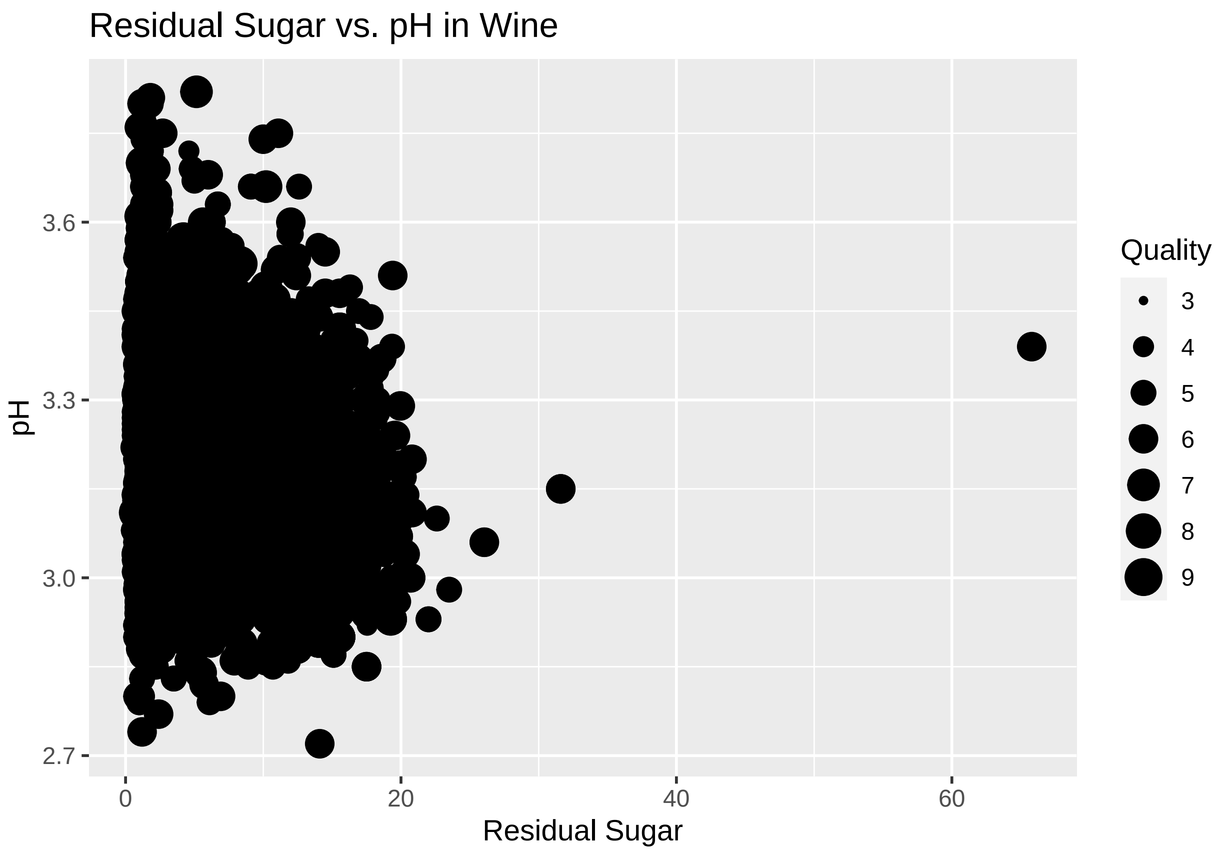
(b) Create a points plot from the wine data using the pH as the response variable and residual sugar as the explanatory variable.

```
wine_plot <- ggplot( data=wine ) +
  geom_point( aes( x=residual.sugar, y=pH ) ) +
  labs( x="Residual Sugar", title="Residual Sugar vs. pH in Wine" )

wine_plot
```



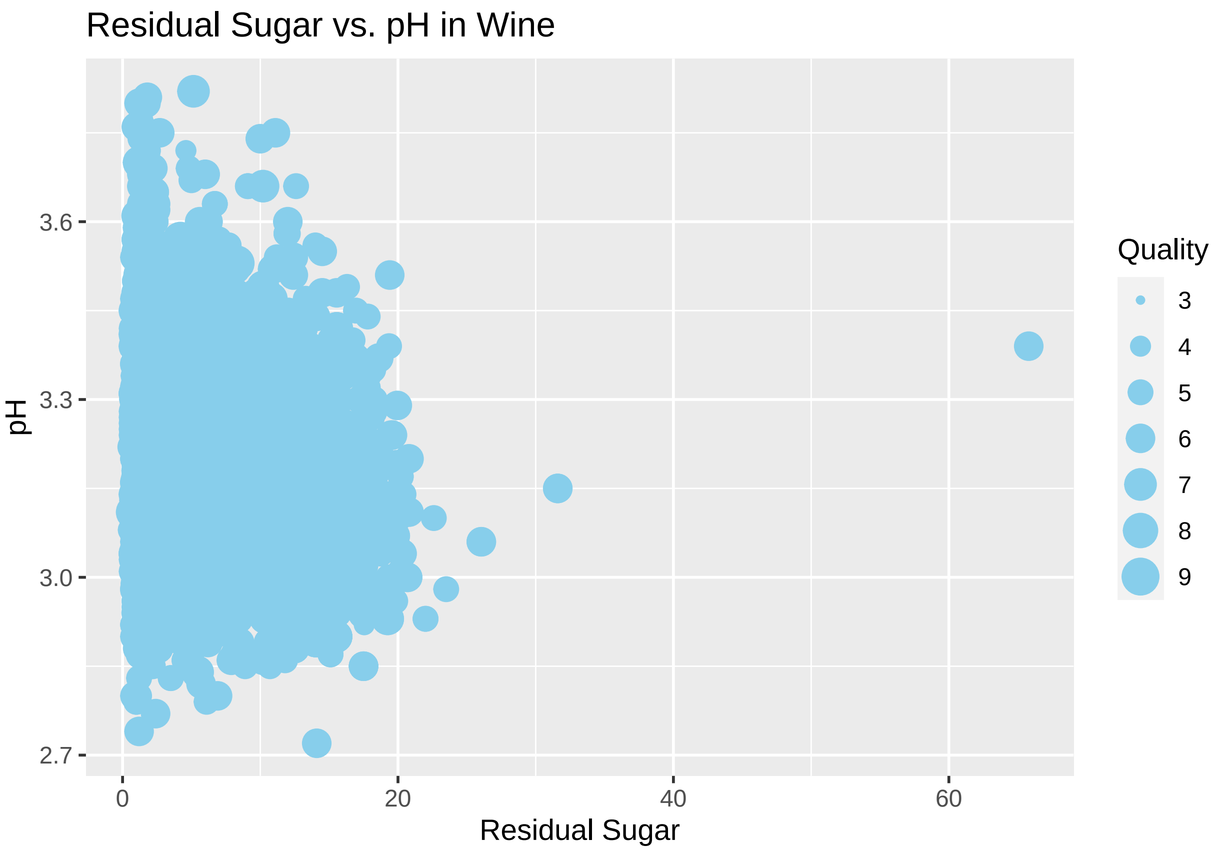
(c) Create the same points plot from (b) but now map the quality variable to the size of the points.

```
wine_plot <- ggplot( data=wine ) +
  geom_point( aes( x=residual.sugar, y=pH, size=quality ) ) +
  labs( x="Residual Sugar",
        title="Residual Sugar vs. pH in Wine",
        size="Quality" )

wine_plot
```



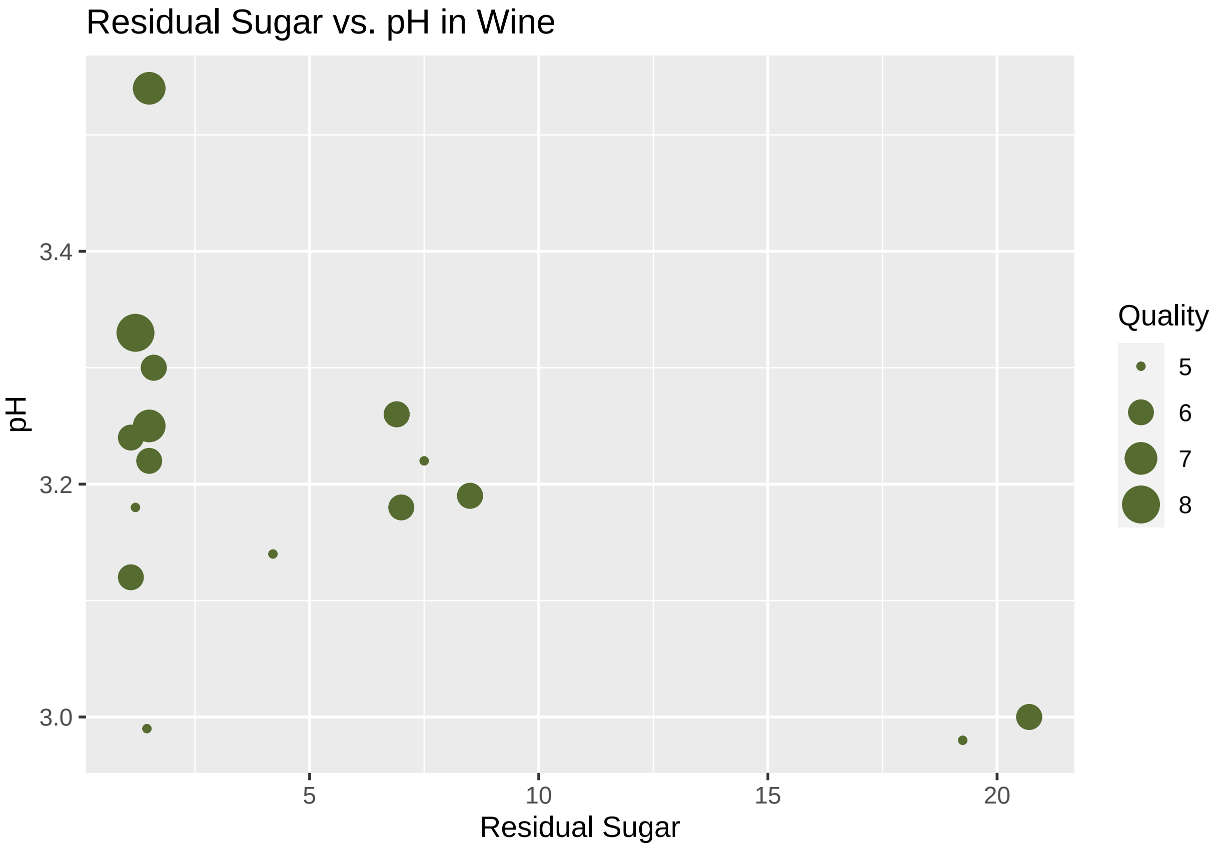
(d) Create the same points plot from (c) but now change the color of all the points to a color of your choice from here: https://www.datanovia.com/en/blog/awesome-list-of-657-r-color-names/

```
wine_plot <- ggplot( data=wine ) +
  geom_point( aes( x=residual.sugar, y=pH, size=quality ), color='skyblue' ) +
  labs( x="Residual Sugar",
        title="Residual Sugar vs. pH in Wine",
        size="Quality" )

wine_plot
```



(e) Create the same plot as in part (d) but this time use only the first 20 rows of the data. You may use the same color or choose another one from the website.

```r
wine_plot <- ggplot( data=wine[1:20,] ) +
  geom_point( aes( x=residual.sugar, y=pH, size=quality ), color='darkolivegreen' ) +
  labs( x="Residual Sugar",
        title="Residual Sugar vs. pH in Wine",
        size="Quality" )

wine_plot
```


Residual Sugar vs. pH in Wine
pH
Residual Sugar
Quality