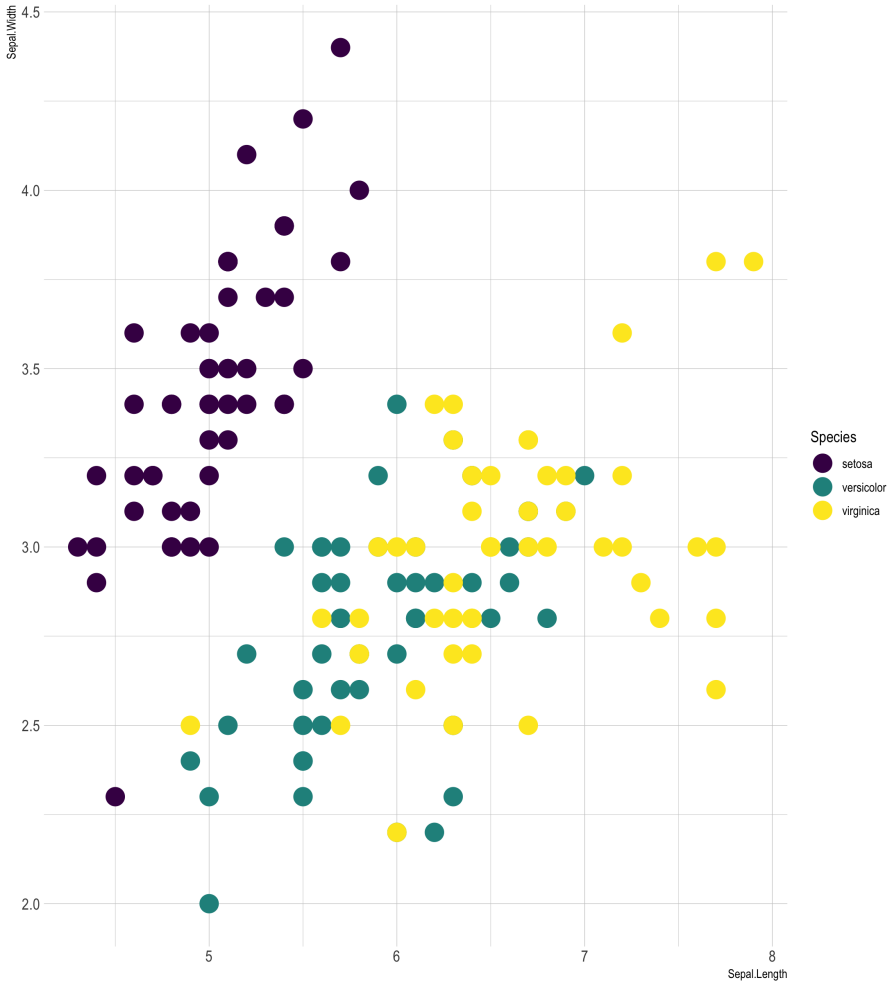
Due Date: February 17th, 2022
Problems due: 1
Download data set from: https://www.isibang.ac.in/~athreya/incovid19/dataopen/Master.csv

1. (From the course archives of Computational Biology) Examine the structure of the inbuilt dataset `iris` . How many observations and variables are in the dataset?

   (a) Using `ggplot` produce the following scatter plot:



   (b) Use the `dplyr` package to do the following computations.
      i. Create a new data frame `iris1` that contains only the species virginica and versicolor with sepal lengths longer than 6 cm and sepal widths longer than 2.5 cm. How many observations and variables are in the dataset?
      ii. Now, create a `iris2` data frame from `iris1` that contains only the columns for `Species, Sepal.Length,` and `Sepal.Width.` How many observations and variables are in the dataset?

iii. Create an `iris3` data frame from `iris2` that orders the observations from largest to smallest sepal length. Show the first 6 rows of this dataset.

iv. Create an `iris4` data frame from `iris3` that creates a column with a sepal area (length * width) value for each observation. How many observations and variables are in the dataset?

v. Create `iris5` that calculates the average sepal length, the average sepal width, and the sample size of the entire `iris4` data frame and print `iris5`.

vi. Finally, create `iris6` that calculates the average sepal length, the average sepal width, and the sample size for each species of in the `iris4` data frame and print `iris6`.

vii. In these exercises, you have successively modified different versions of the data frame `iris1 iris1 iris3 iris4 iris5 iris6`. At each stage, the output data frame from one operation serves as the input fro the next. A more efficient way to do this is to use the pipe operator %>% from the tidyr package. Rework all of your previous statements into an extended piping operation that uses `iris` as the input and generates `iris6` as the output.

2. From Philip Spector's page on dates and times https://www.stat.berkeley.edu/~s133/dates.html read about dates and times in `R`.

   (a) What is the significance of January 1, 1970 ?

   (b) What is the difference between `as.Date` and `POSIXlt` ?

3. From the book R for Data Science read how to use the `lubridate` package and work with dates and times. Do the following Exercises.

   (a) Exercise 16.2.4

   (b) Exercise 16.3.4

   (c) Exercise 16.4.5