Homework 1 Solutions

Due: Monday Sept. 16, 2024 at 10pm

This assignment has a total of 23 pts possible. Your score out of 20 will noted and scaled to 5 points (maximum of 5).

In my solutions I will try to highlight the complete answer to a question using yellow. I may also include extra information to explain an answer or provide hints for similar questions we encounter in the future which I will highlight in orange.

Question 1 – Conceptual Questions: (2pts each)

Part A What does it mean to say two variables are *associated* with each other?

Two variables are associated if one variable gives us information about another.

Part B What does it mean to say two variables are *independent* of each other?

Two variables are independent if there is no relationship / association between them.

Part C What does the distribution of a variable tell us?

The distribution tells us how frequently certainly values of a variable show up in our data.

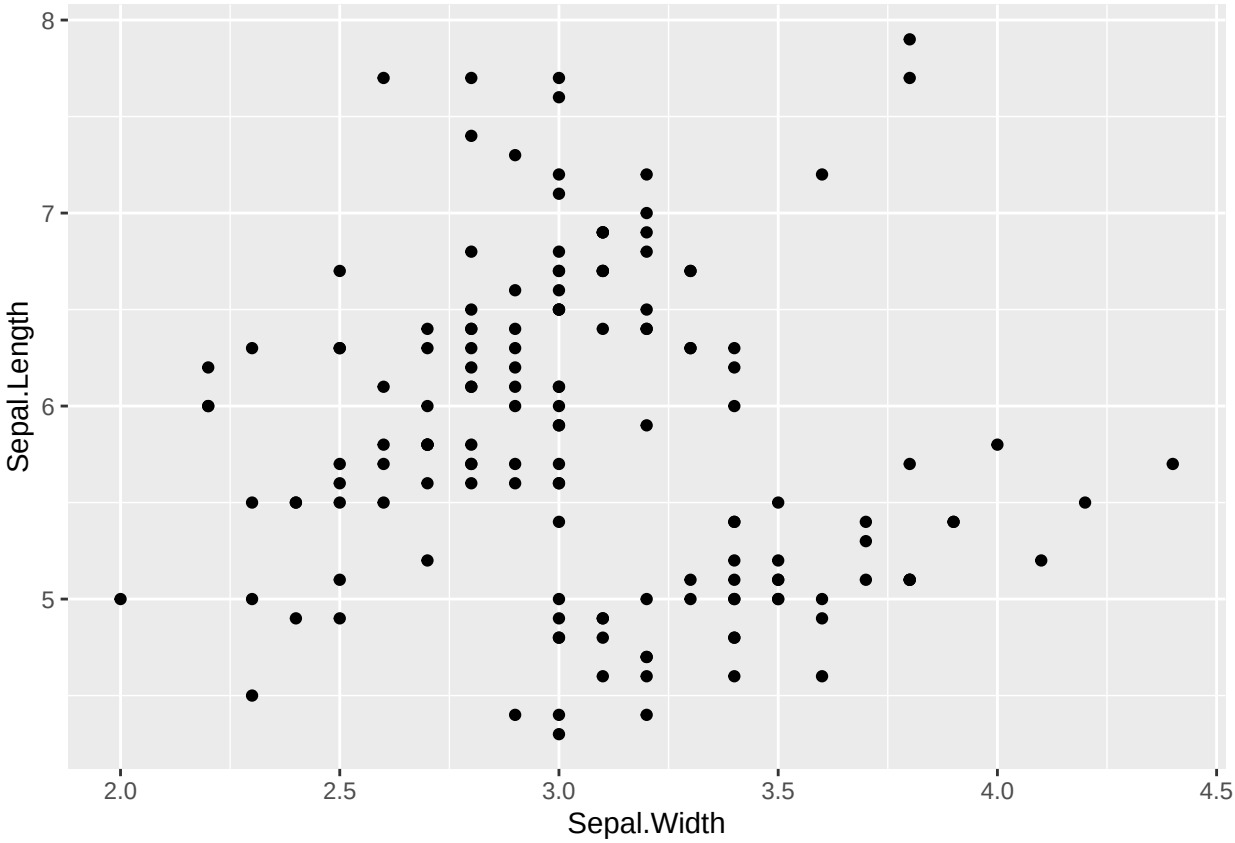
Question 2 For this question, we will be using the `iris` dataset, giving the measurements, in centimeters, of the variables for sepal and petal length and width. You can read more on the dataset here.

Use this data to answer the following questions:

- **Part A** How many observations and variables are in the `iris` dataset? In one sentence, briefly describe what constitutes an observation in this data. (2 pts)

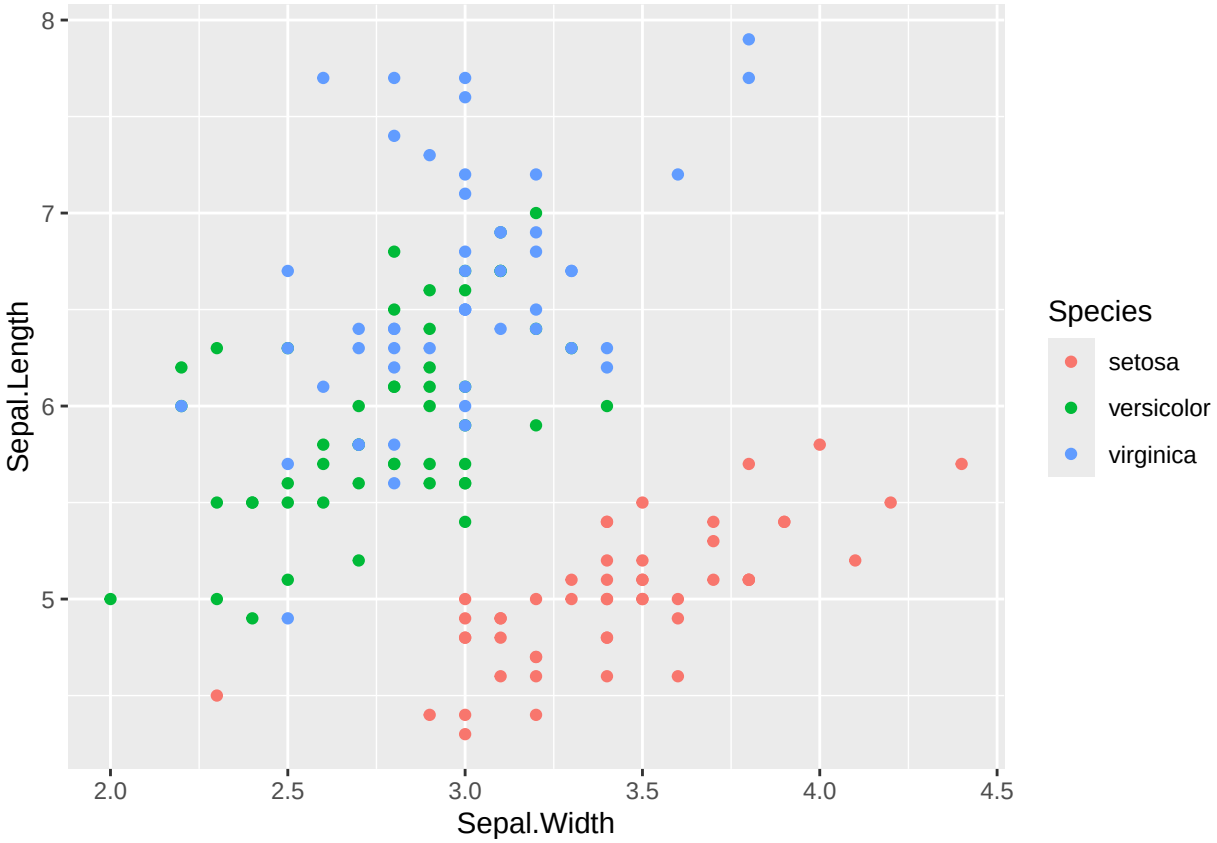
There are 150 observations and 5 variables in this dataset. An observation is an iris flower from the Gaspe Peninsula in 1936.

- **Part B** Use the code below to create the appropriate plot to visualize the relationship between the variables `Sepal.Width` and `Sepal.Length`. Do these two variables appear to be *associated*? If so, comment on the strength of this association. (2 pt)



There is a weak, positive association between `Sepal.Width` and `Sepal.Length`.

- **Part C** Use the code below to create the plot again, this time adding additional information for the variable `Species`. Has anything changed in the association between `Sepal.Width` and `Sepal.Length`? Comment on the **strength**, **form**, and **direction** of any associations you see (1pt)



For each of the three species of iris flower, we can see there is a positive linear relationship. The strengths of these relationships differ slightly. For `virginica` and `versicolor`, the association is moderate. For `setosa` the association is strong.

Chp 4.8, #5

a) We can more easily see the number of participants in each group with the stacked bar chart.

b) We can more easily see the proportion that survived in each group with the proportional bar chart.

c) The proportional bar chart is a better display for this study. It allows us to compare the treatment survival rate to the control group.

Chp 4.8, #6

a) The bottom graph lets us better understand shipping choices of people of different age groups. We can directly compare the age groups in the plot.

b) The top graph lets us better understand age distribution across different types of shipping. We can look at a shipping type and see which age group most/least uses it.

c) The biggest competitor would be USPS. If we look at the 55+ age range (bottom graph), USPS is the biggest shipping method.

d) Fedex should reach out to 55+ to balance their demographics. This group is the smallest demographic for Fedex (top chart).

Chp 5.10, #1

a) Positive association: mammals with longer gestation periods tend to live longer as well. (arguable whether this is linear or non-linear)

b) If we flip the axes, the association will still be positive.

c) Life span and gestation length are **not independent** because we see that there is an association.

Chp 5.10, #2 Graph 1) positive linear association

Graph 2) no association (random scatter of points)

Graph 3) positive non-linear (curved) association

Graph 4) negative linear association