Week 5: Data Communication and Ethics

DSC 365: Introduction to Data Science

2024-09-17

Graphical Perception

To visually display data, information is encoded into a graph. The viewer then visually decodes this information, known as graphical perception, to gain knowledge. A successful graphic allows the user to perform graphical perception accurately and efficiently

In previous weeks we have discussed how to choose the right plot to visualize the data. This week we will focus on the other two principles:

- Keep it simple
- Show the data clearly

To achieve these two goals, you need to make use of several optional functions in the ggplot as well as understand what need to be include in your writing.

Keep it Simple

A general guideline is to use simplicity in the design.

- Minimize the number of unique symbols to not overload the user's memory.
- A sample of graphics found that graphics with the best overall ratings had fewer features on average.
- Beauty with the addition of unnecessary features does not always equal good content as attention is limited

Show the Data Clearly: Preattentive Features

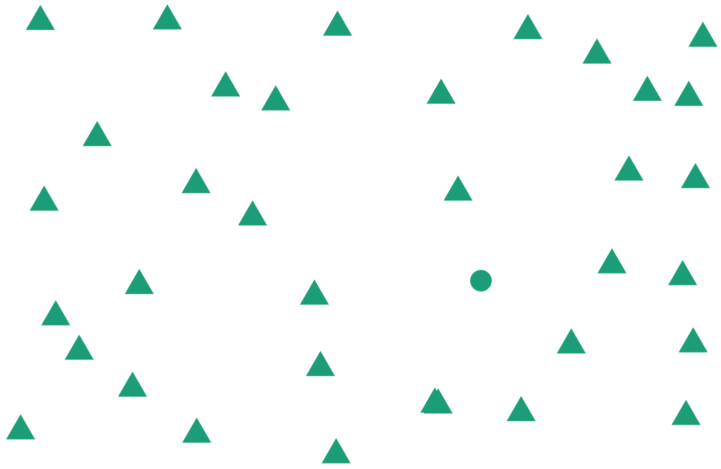
Pre-Attentive Features are things that “jump out” in less than 500 ms

- Color, shape, angle, movement, spatial localization

There is a hierarchy of features + Color is stronger than shape

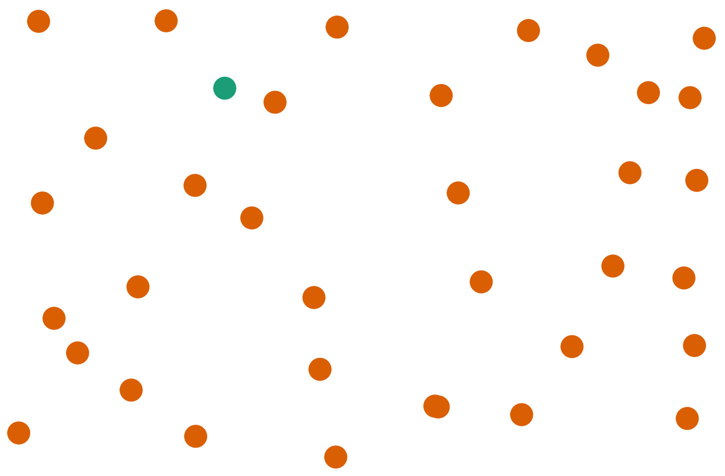
Reduces the amount of work users of your graph have to do when they view it

Find the Target - How quickly did you spot the circle?

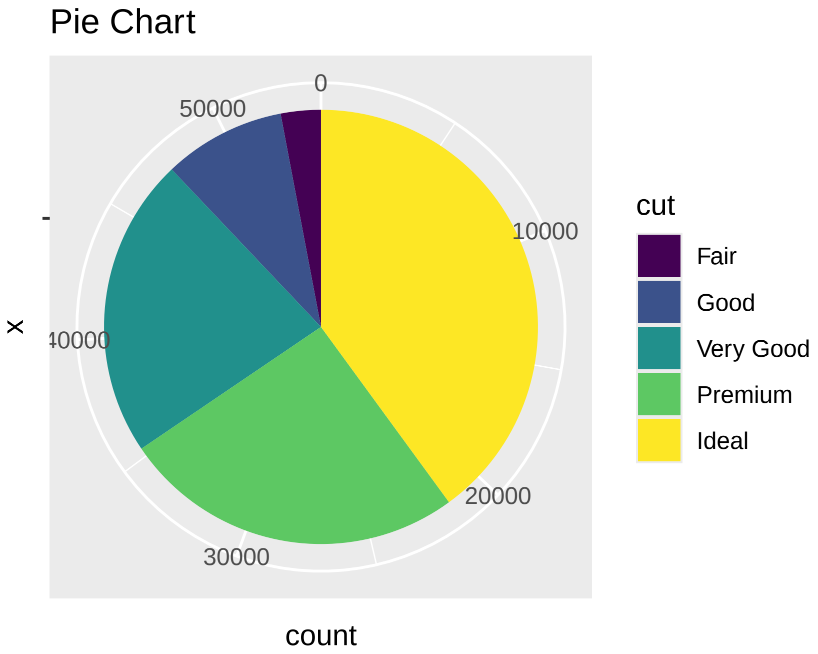


```r
data$shape <- factor(c(rep(2, 25), 1, rep(2, 10)))

ggplot(data, aes(x, y)) + geom_point(aes(colour = shape), size = 5, shape = I(19)) + theme
```

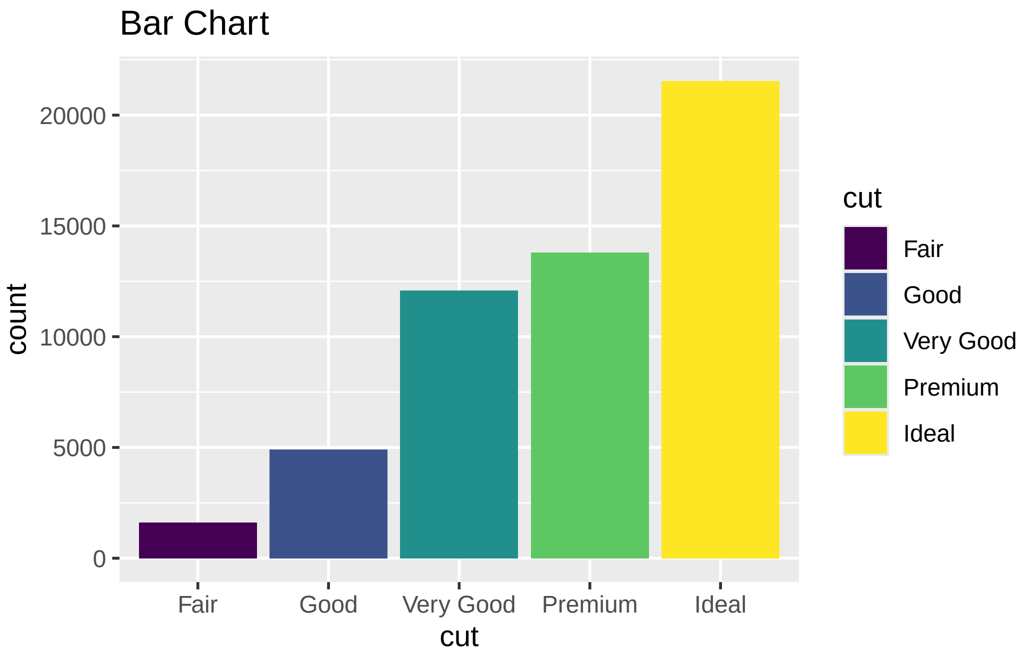


Show Data Clearly: Pie vs Bar Chart



Anything that can be put into a pie chart can also be put into a bar chart

- Research has shown that people have more difficulty with angle measurements than length measurements, so always prefer bar chart over pie chart.



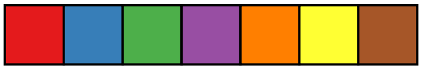
Show Data Clearly: Color

- Hue: shade of color (red, orange, yellow…)
- Intensity: amount of color
- Both color and hue are pre-attentive. Bigger contrast corresponds to faster detection.
- Use color to your advantage
- When choosing color schemes, we will want mappings from data to color that are not just numerically but also perceptually uniform
    - Avoid rainbow color gradients
- Distinguish between sequential scales and categorical scales
- Be conscious of what color means
    - Leverage common association

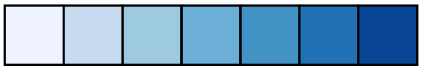
Our eyes are optimized for perceiving the yellow/green region of the color spectrum

Gradients

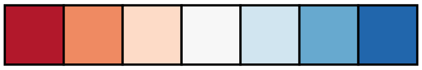
Qualitative schemes: no more than 7 colors



Quantitative schemes: use color gradient with only one hue for positive values



Quantitative schemes: use color gradient with two hues for positive and negative values. Gradient should go through a light, neutral color (white)



Color Blindness

Not everyone perceives color in the same way. Some individuals have colorblindness or color deficiencies.

You can take a test designed to screen for colorblindness here.

Suggestions: + Design for a black-and-white photocopier + Use a monochromatic color gradient scheme where possible. + Suggested 2-color gradient: blue/purple - white - orange (safe for most types of colorblindness) + Utilize double encoding: use color and another aesthetic (line type, shape) + Avoid any scheme that uses green-yellow-red signaling if you have a target audience that may include colorblind people. + Can use this website to help pick palettes as well + The `viridis package` (from last week's lab) contains a set of color scales designed to span as wide a palette as possible, making it easier to see differences in your data and are also perceptually uniform.

Example: Hair Color

The data `HairEyeColor` provides the information of hair colors and eye colors in a statistics class. The data is recorded in a three-way table. The first step is to convert the data into a column-wise data frame.

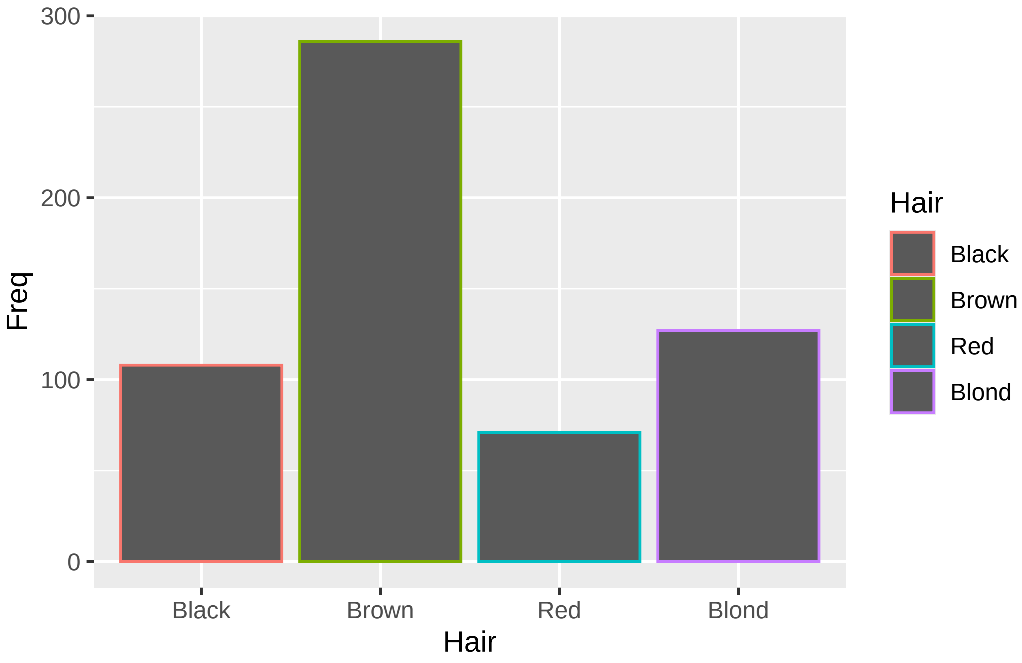
```
# A tibble: 4 x 2
  Hair   Freq
  <fct> <dbl>
1 Black   108
```

```
2 Brown   286
3 Red      71
4 Blond   127
```

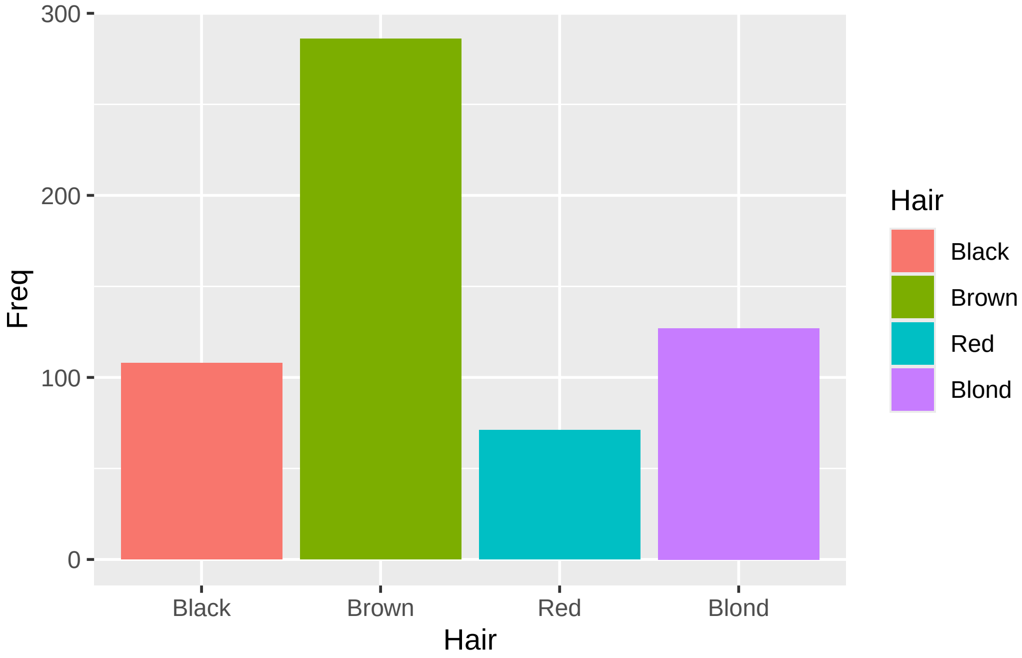
Suppose now we want to see the distribution of hair color in this class. What kind of the plot we should use?

What is the difference between the two chunks of code below?

```r
hairData %>%
  ggplot(aes(x=Hair, y=Freq)) +
  geom_col(aes(color = Hair))
```

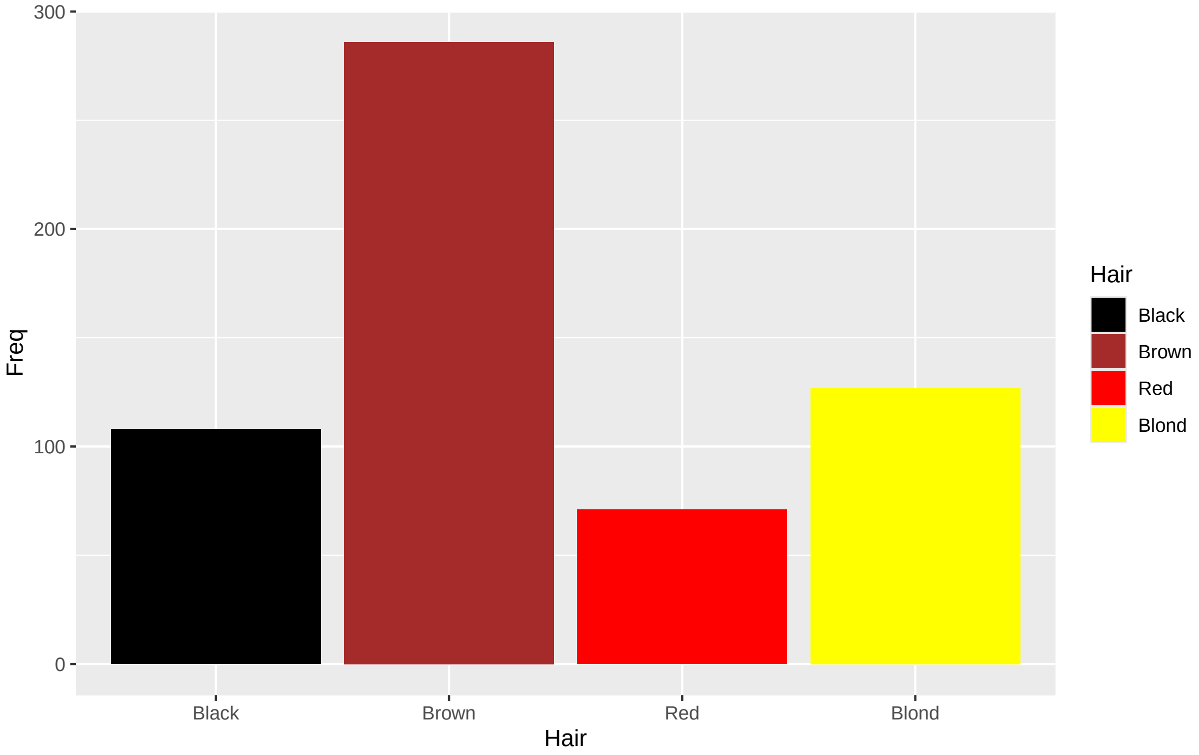


```r
hairData %>%
  ggplot(aes(x=Hair, y=Freq)) +
  geom_col(aes(fill = Hair))
```



It is a little weird that the color in the visualization does not match the color. Let's try to define the color by ourselves.

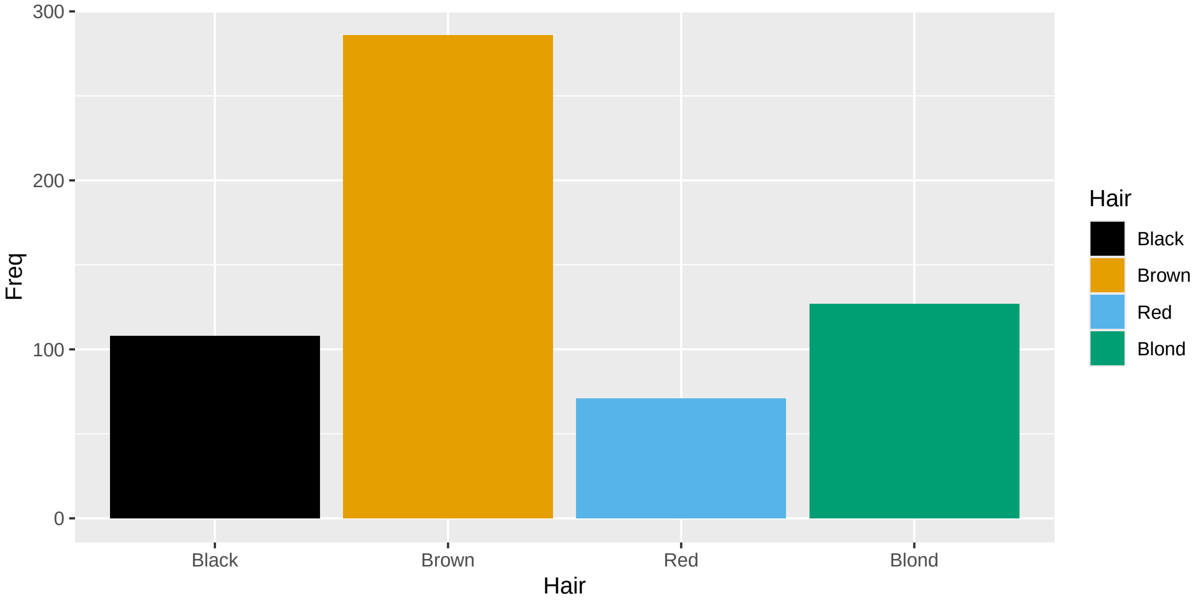
```
ggplot(hairData, aes(x = Hair, y = Freq)) +
  geom_col(aes(fill = Hair)) +
  scale_fill_manual(breaks = c("Black", "Brown", "Red", "Blond"), #<<
                    values=c("black", "brown", "red", "yellow")) #<<
```



You may have realized that in the previous figure, the color red and brown are pretty close. For an extreme case, what if you have a reader who is color-blind? There are a lot of research on which palette then to use.

```
cbPalette <- c("#000000","#E69F00","#56B4E9","#009E73","#F0E442")

ggplot(hairData, aes(x = Hair, y = Freq)) +
  geom_col(aes(fill = Hair)) +
  scale_fill_manual(values=cbPalette)
```

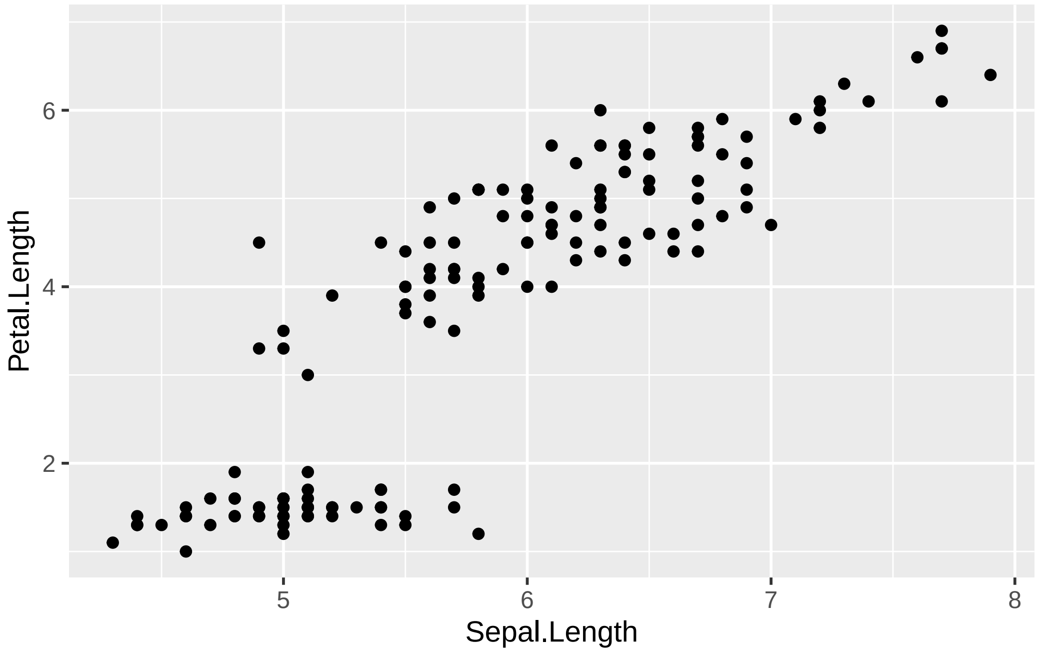


Try it for yourself!!

Now try by yourself (or with a group) with another type of figure.

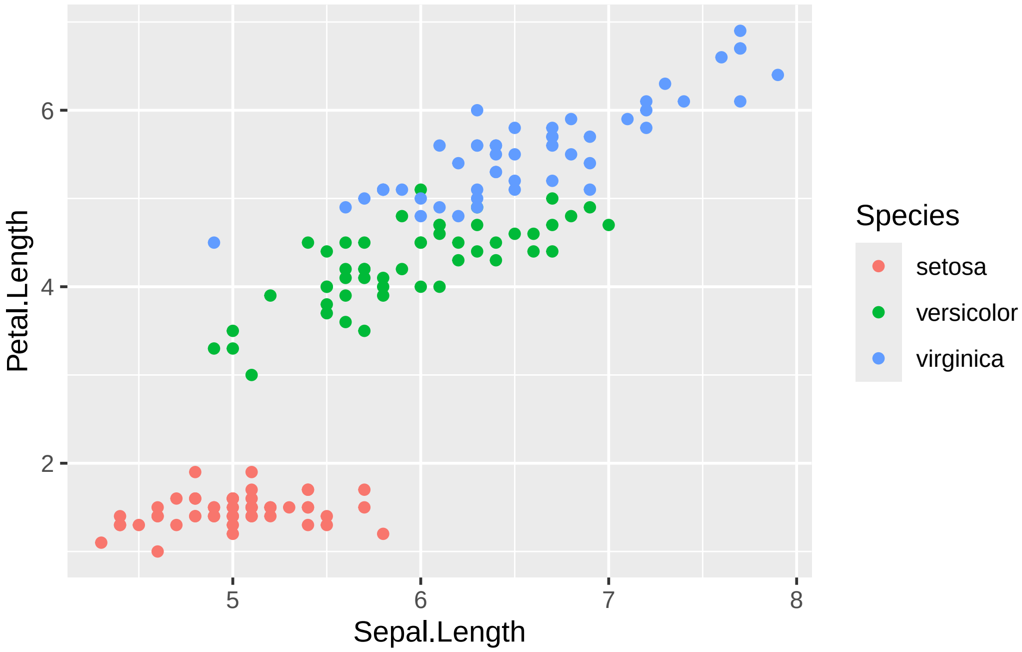
(a). Use the iris data below. Plot the figure to show the relation between Sepal.Length and Petal.Length.

```
iris %>%
  ggplot(aes(x = Sepal.Length, y = Petal.Length))+
  geom_point()
```



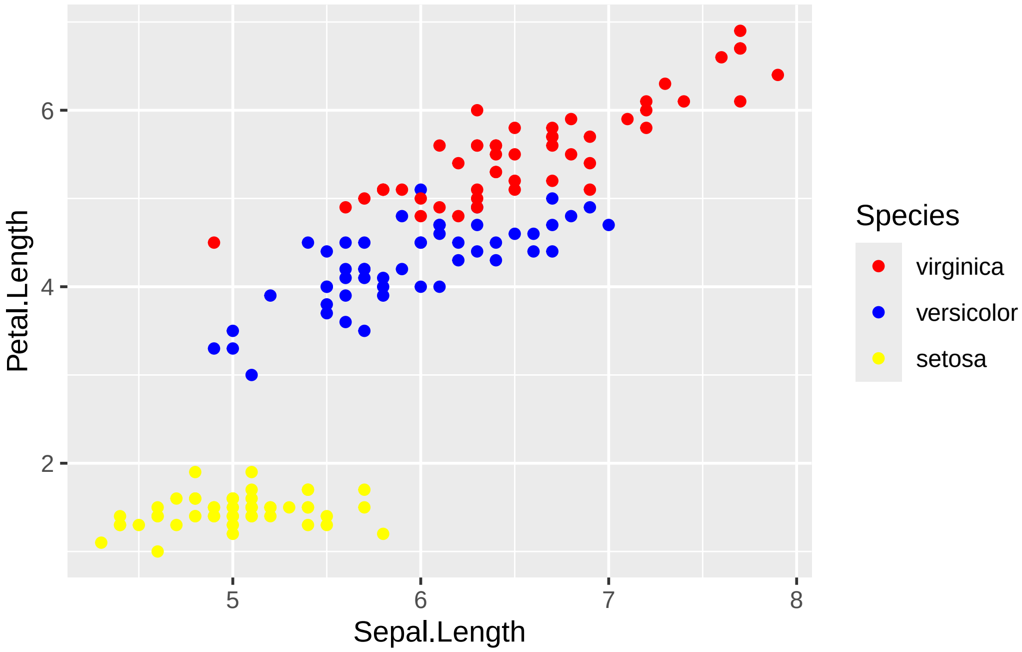
(b). Use colors to distinguish the species. Which function you have used? fill or color?

```
iris %>%
  ggplot(aes(x = Sepal.Length, y = Petal.Length))+
  geom_point(aes(color = Species))
```



(c). Use red to represents `virginica`, blue to represents `versicolor`, and yellow to represents `setosa`.

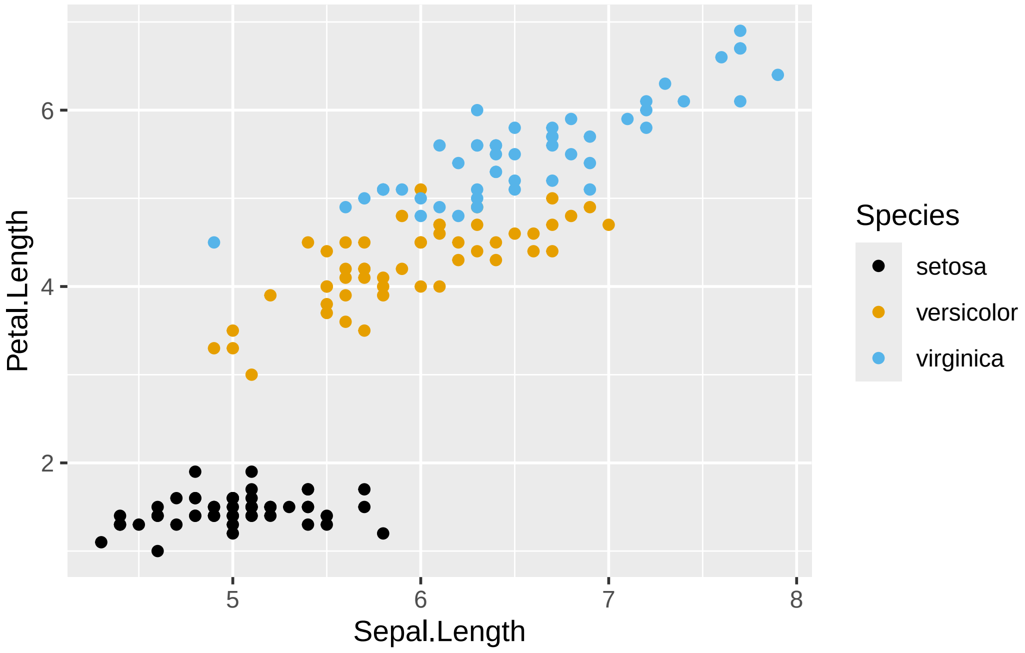
```
iris %>%
  ggplot(aes(x = Sepal.Length, y = Petal.Length))+
  geom_point(aes(color = Species)) +
  scale_color_manual(breaks = c("virginica", "versicolor", "setosa"),
                     values=c("red", "blue", "yellow"))
```



(d). Use the color blind friendly color to distinguish three different species

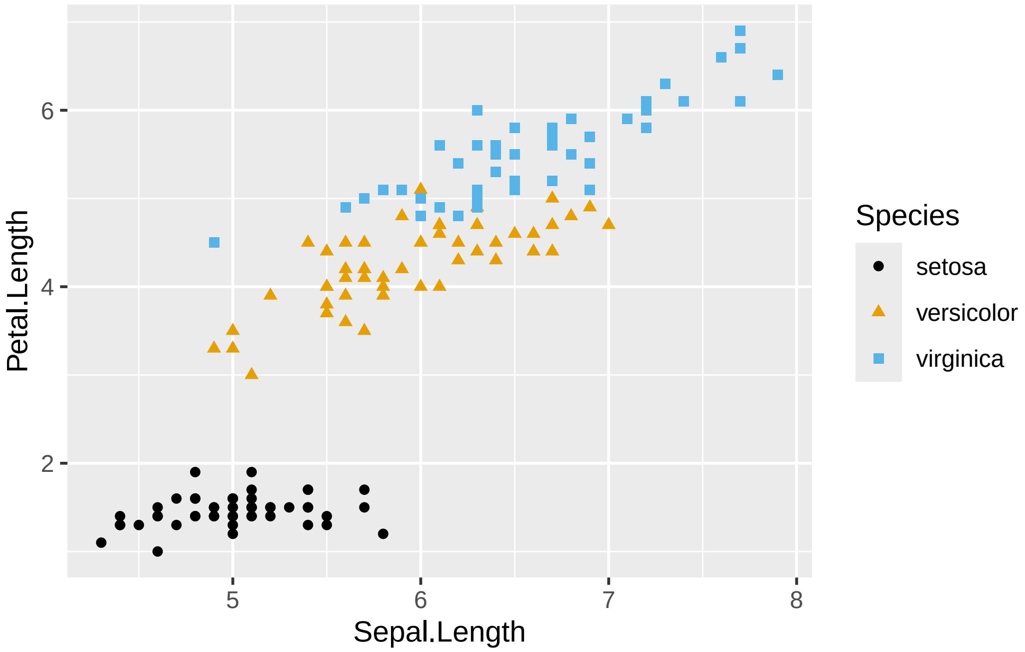
```
cbPalette <- c("#000000","#E69F00","#56B4E9","#009E73","#F0E442")

iris %>%
  ggplot(aes(x = Sepal.Length, y = Petal.Length))+
  geom_point(aes(color = Species)) +
  scale_color_manual(values=cbPalette)
```



(e). Besides of the color, use different shape of the points to represents the difference (Hint: Use Google or the Help Documentation).

```
iris %>%
  ggplot(aes(x = Sepal.Length, y = Petal.Length))+
  geom_point(aes(color = Species, shape = Species)) +
  scale_color_manual(values=cbPalette)
```



Data Misrepresentation

Common Ways to Misreprsent Data

- Scales!
- Omitting Data
- Cherry picking the data
- Misleading pie chart/Using the wrong chart type

What's Wrong with These Plots?

Identify the issues with the plots, and talk about how you would fix them?

Election Maps

Visualzing Uncertainty

- Uncertainty is fundamental to data analysis and models

- Effectively communicating uncertainty in visualization is hard and is still being researched.
- Can be important to include, but don’t present the. uncertainty information, without describing in detail what it means.

Writing About Your Figure

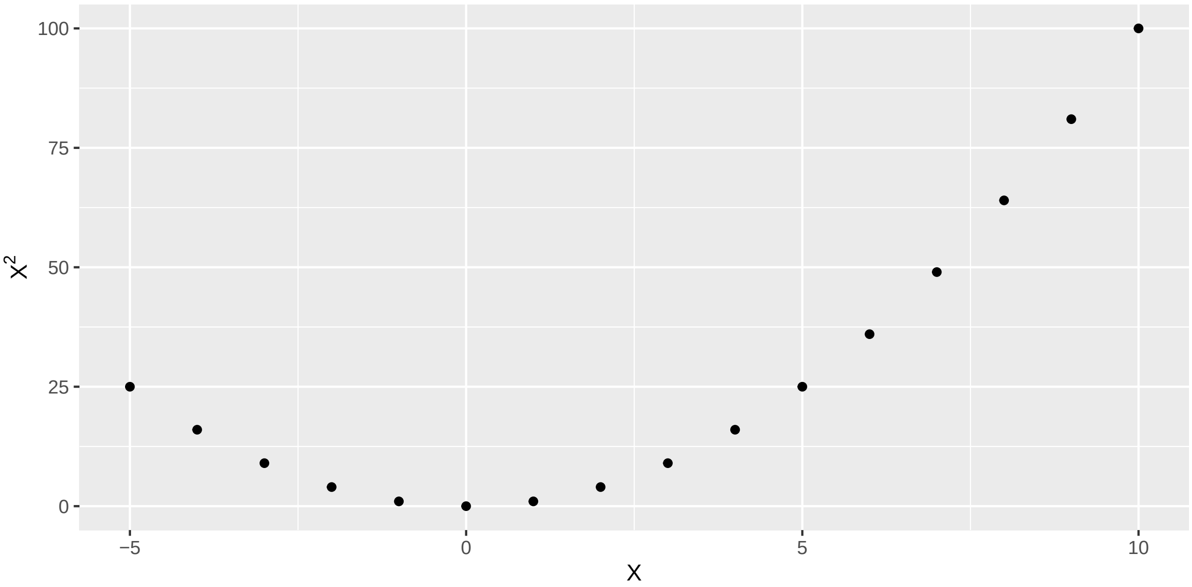
Your visualizations should include:

- Clearly labeled axes
- Clear and descriptive captions
- Readable and easily understandable legends (if applicable)

Labels

(a). Use plain English
(b). If there is mathematically symbols in your label, use `expression`:

```r
x = -5:10
y = x^2
xy = data.frame(x,y)
ggplot(xy, aes(x = x, y=y)) +
  geom_point() +
  xlab("X") + ylab(expression(X^2))#<<
```

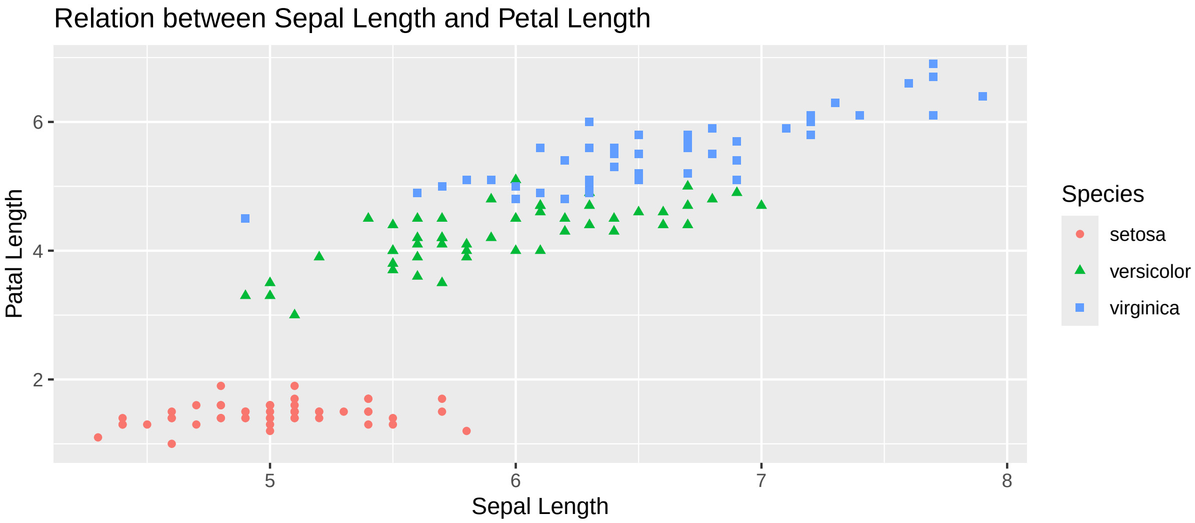


Captions

There are two ways of writing a caption:

1. In-Figure Style

```
ggplot(iris, aes(x=Sepal.Length, y=Petal.Length, group=Species)) +
  geom_point(aes(color = Species, shape = Species)) +
  xlab("Sepal Length") +
  ylab("Patal Length") +
  ggtitle("Relation between Sepal Length and Petal Length") #<<
```



2. Formal report style: use descriptive caption

a). It is not included in the figure. Usually placed below the figure. b). Starts with the word “figure” or “table” followed by a number and a colon c). One sentence to clearly explained What and How is being compared d). One sentence to summarize the main conclusion. e). Any other necessary context

What this then looks like:

The caption only shows in the knit version of documents. Why we need a descriptive caption? How to check whether you have written a good descriptive caption?

- The reader can capture your main argument without seeing the figure
- The reader can capture the context of your figure without reading the other analysis in the formal report.

Writing About Figures in Reports

A formal data analysis report requires you to answer the question with a figure and explain the potential reasons behind the figure image. Sometimes you need multiple figures to answer one question and sometimes you can answer multiple questions with one figure.

Once you generate one figures, here is what you should do:

- Re-introduce the visualization and main argument
- Cite any statistical evidence or figure characteristics to support your argument
- State secondary argument and other information
- Contrast with the other figures if necessary
- Provide potential reasons and explain the conclusion

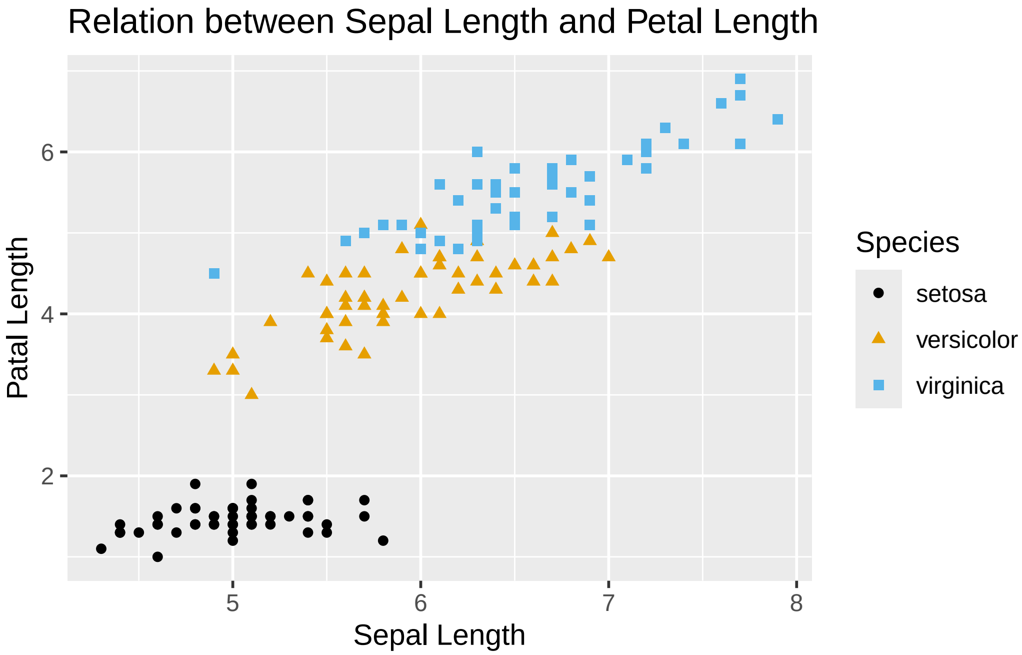


Figure 1: Scatter plot between Iris sepal length and Iris petal length. Different species are indicated by different color and shapes. In general, the length of sepal and petal for Iris have a strong positive linear relationship.

Word Choice

In data writing, it is quite important that you know how to use different word to show your confidence in the statement. This is subjective but here is a list of words you may need, listed from the highest confidence to the lowest:

`demonstrates`

`shows`

`indicates`

`illustrates`

`reveals`

`suggests`

`might indicate`

`may suggest`

`seems to suggest`

Example: Writing About Figures in Reports

Figure 2 indicates that there is no linear relationship between iris sepal width and petal width since the scatter plot did not show a clear linear shape on the distribution of the point.

Meanwhile, the three species does have clear different distributions of the sepal width and petal width and within a species, the the sepal width and petal width may have a positive linear relationship except Setosa, which does not have a clear trend.

The relative size of different iris species is similar to what we have found in Figure 1, where Virginica tends to have the largest size among all three while Setosa is the smallest. One of the potential reasons for the different sizes may be the habitat condition while Virginica tends to grow in the places with enough sunshine but Setosa prefers relatively dark and wet environment.

Data Science Ethics

Ethics in Data Science refers to the responsible and ethical use of the data throughout the entire data lifecycle. This includes the collection, storage, processing, analysis, and interpretation of various data

One of the best selling "statistics" books is called *How to Lie with Statistics* by Darrell Huff (1954) - Shows graphical tools to fool people even with accurate data that are still in use.

As data scientists, we can play a role in shape discourse, so what responsibilities do we have?

ASA Guidelines for Statistical Practice

Revised in 2022: Link

Some highlights:

- Protect and respect the rights and interests of human and animal subjects
  - Data privacy (ex. HIPAA)
  - Consider how your study would impact society, groups, and individuals (Remember the numbers are real people)
- Uses methodology and data that are valid, relevant, and appropriate, without favoritism or prejudice.
  - Ask for help if you don't know how to do something properly
- Promotes reproducibility and replication, whether results are "significant" or not, by sharing data, methods, and documentation to the extent possible.
- Don't only present significant results

Algorithmic Bias

Algorithmic bias occurs when algorithms make decisions that systematically disadvantage certain groups of people

- Biased data -> biased algorithms
  - Ex. Some groups of people may be underrepresented or systematically excluded from data science efforts

Some examples:

- Gender data gap

  - Products, services and strategies are being generalized to women when the research behind them is not based on data involving women

- Facial Recognition software:
  - Joy Buolamwini, a Ghanaian-American graduate student at MIT discovered that the dataset on which many of facial-recognition algorithms are tested contains 78 percent male faces and 84 percent white faces (further reading)

Is this Ethical? Discuss.

In the United States, most students apply for grants or subsidized loans to finance their college education. Part of this process involves filling in a federal government form called the Free Application for Federal Student Aid (FAFSA). The form asks for information about family income and assets. The form also includes a place for listing the universities to which the information is to be sent. The data collected by FAFSA includes confidential financial information (listing the schools eligible to receive the information is effectively giving permission to share the data with them).

It turns out that the order in which the schools are listed carries important information. Students typically apply to several schools, but can attend only one of them. Until recently, admissions offices at some universities used the information as an important part of their models of whether an admitted student will accept admissions. The earlier in a list a school appears, the more likely the student is to attend that school.

Here's the catch from the student's point of view. Some institutions use statistical models to allocate grant aid where it is most likely to help ensure that a student enrolls. For these schools, the more likely a student is deemed to accept admissions, the lower the amount of grant aid they are likely to receive.

Resource