BINF 5003: Data Mining, Modeling, and Biostatistics

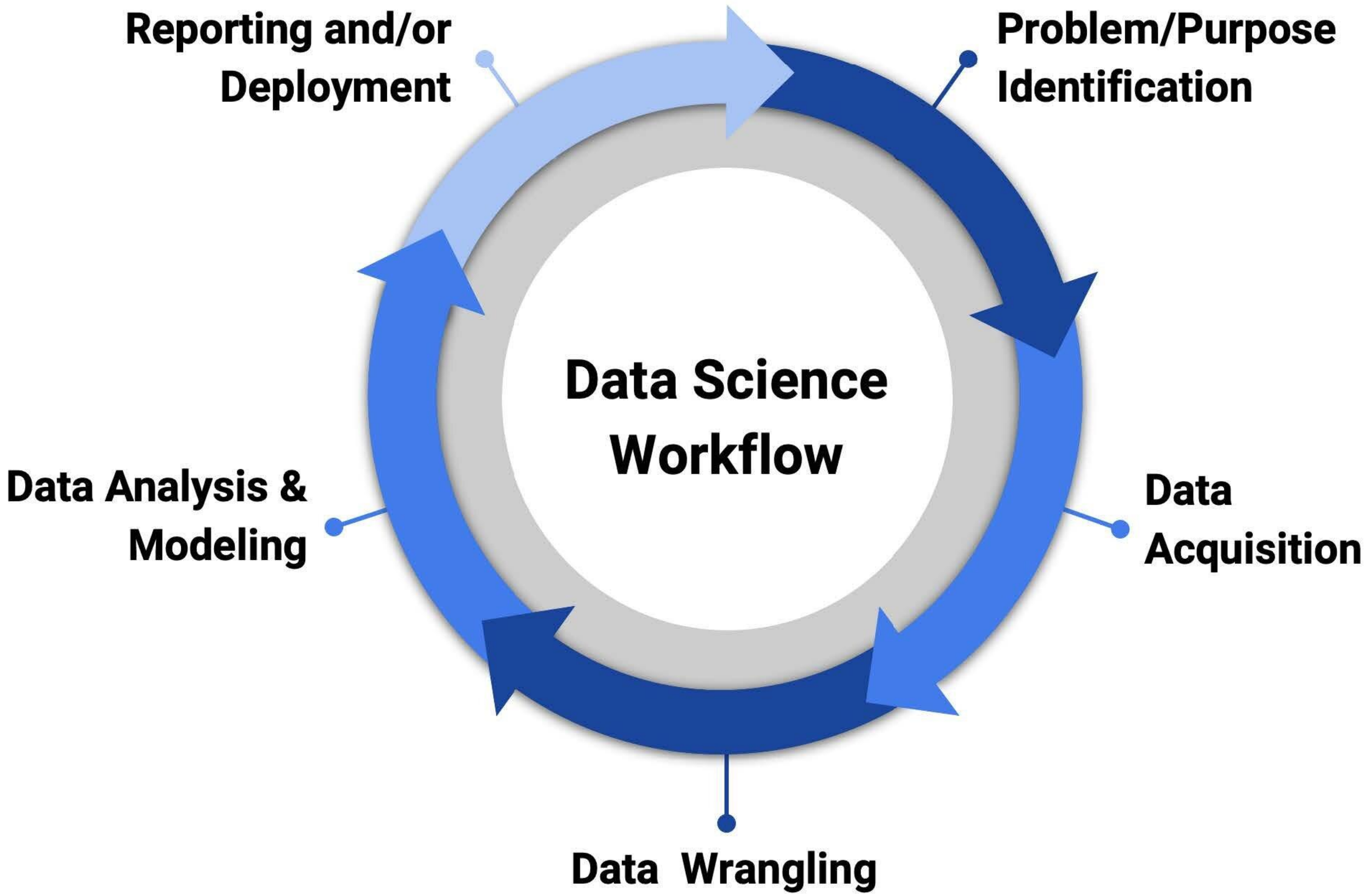
Week 4

Module 3 – Visualization

Overview

- Numeric data and distribution patterns
- 1D and 2D plots
- Plotting with base R and tidyverse functions

Take a step back



Visualizing data

- Continuous
    - Large/unlimited range of values
    - E.g., gene expression values

- Categorical/discrete
    - Limited range of values
    - e.g., clinical diagnosis, benign vs malignant

- Text and logical can be visualized too, although using different methods
    - E.g., Word clouds

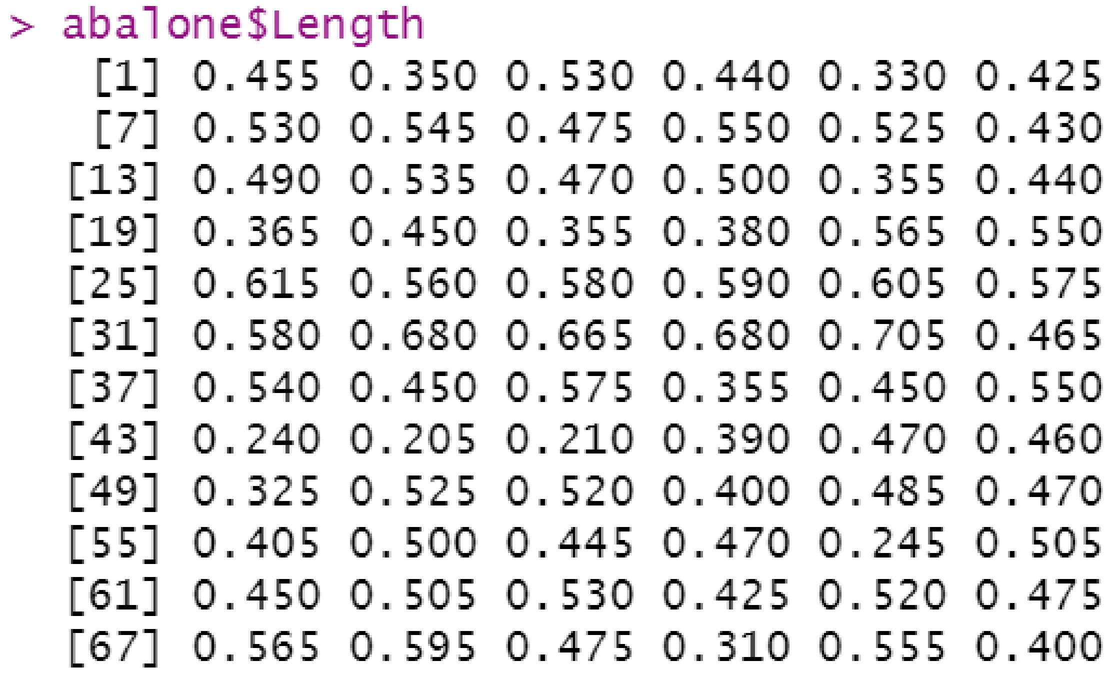
Easier to see patterns in visualization

- Difficult to interpret a list of numbers
  - Generally, the list needs to be sorted
  - Too many individual values to relate each of them

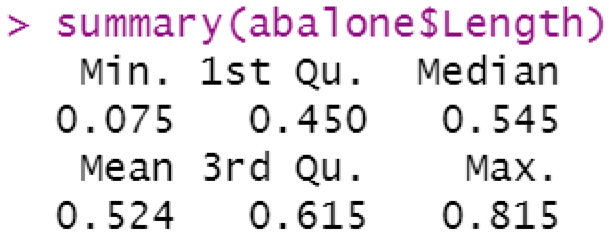
- Summaries are limited to key statistics

- Plots show every data point in relation to each other
  - Can represent data with multiple dimensions

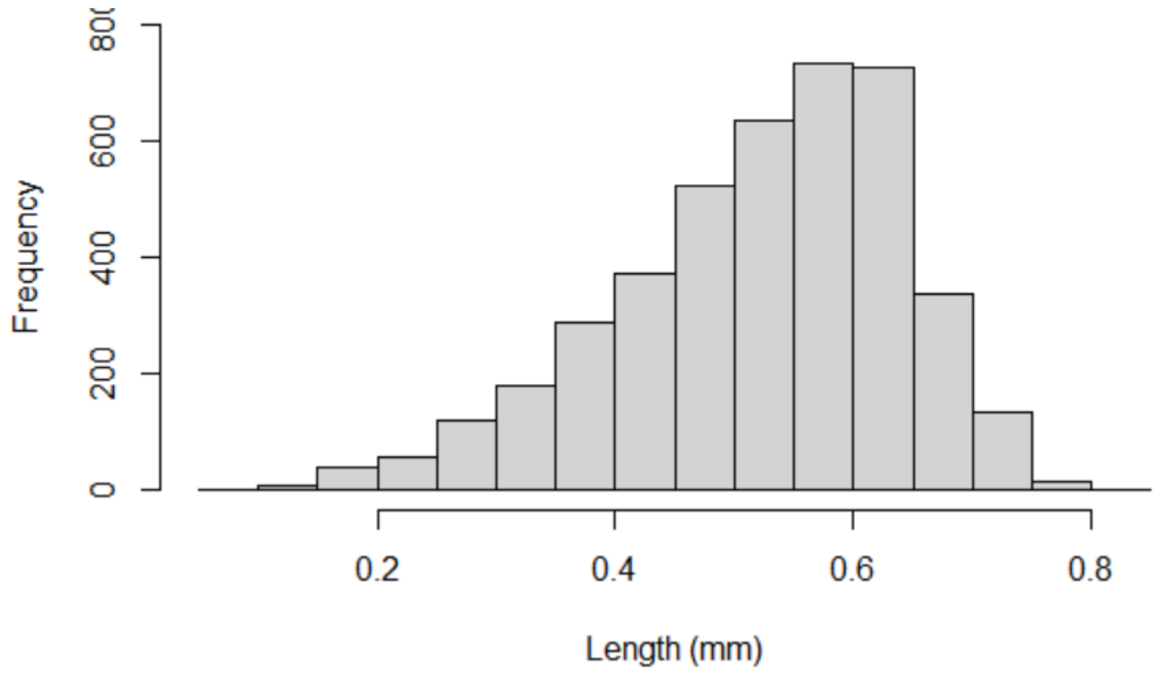
What is the distribution of abalone length?



```
> abalone$Length
  [1] 0.455 0.350 0.530 0.440 0.330 0.425
  [7] 0.530 0.545 0.475 0.550 0.525 0.430
 [13] 0.490 0.535 0.470 0.500 0.355 0.440
 [19] 0.365 0.450 0.355 0.380 0.565 0.550
 [25] 0.615 0.560 0.580 0.590 0.605 0.575
 [31] 0.580 0.680 0.665 0.680 0.705 0.465
 [37] 0.540 0.450 0.575 0.355 0.450 0.550
 [43] 0.240 0.205 0.210 0.390 0.470 0.460
 [49] 0.325 0.525 0.520 0.400 0.485 0.470
 [55] 0.405 0.500 0.445 0.470 0.245 0.505
 [61] 0.450 0.505 0.530 0.425 0.520 0.475
 [67] 0.565 0.595 0.475 0.310 0.555 0.400
```



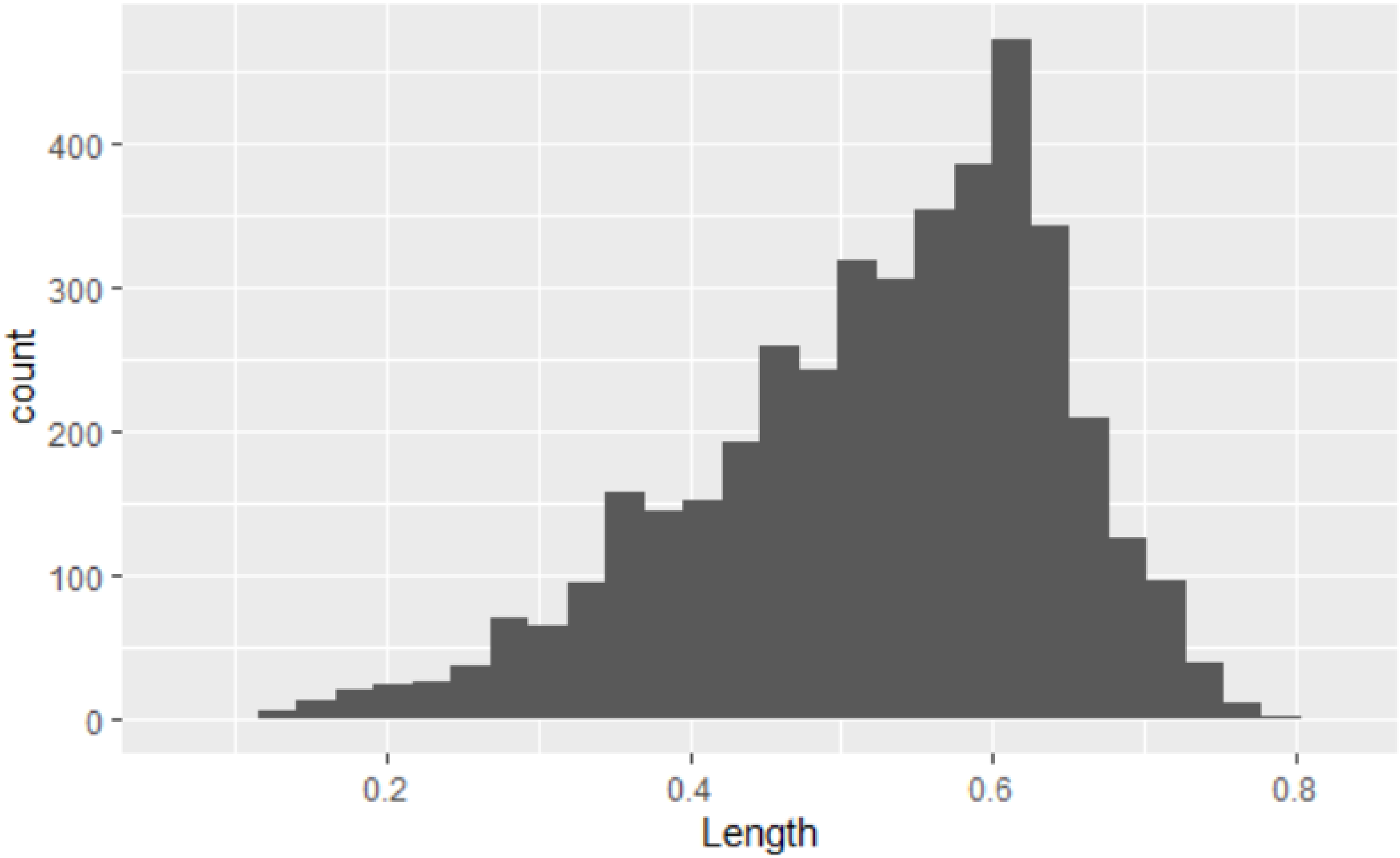
```
> summary(abalone$Length)
   Min. 1st Qu.  Median 
  0.075   0.450   0.545 
   Mean 3rd Qu.    Max. 
  0.524   0.615   0.815 
```



1D plots

- Histogram

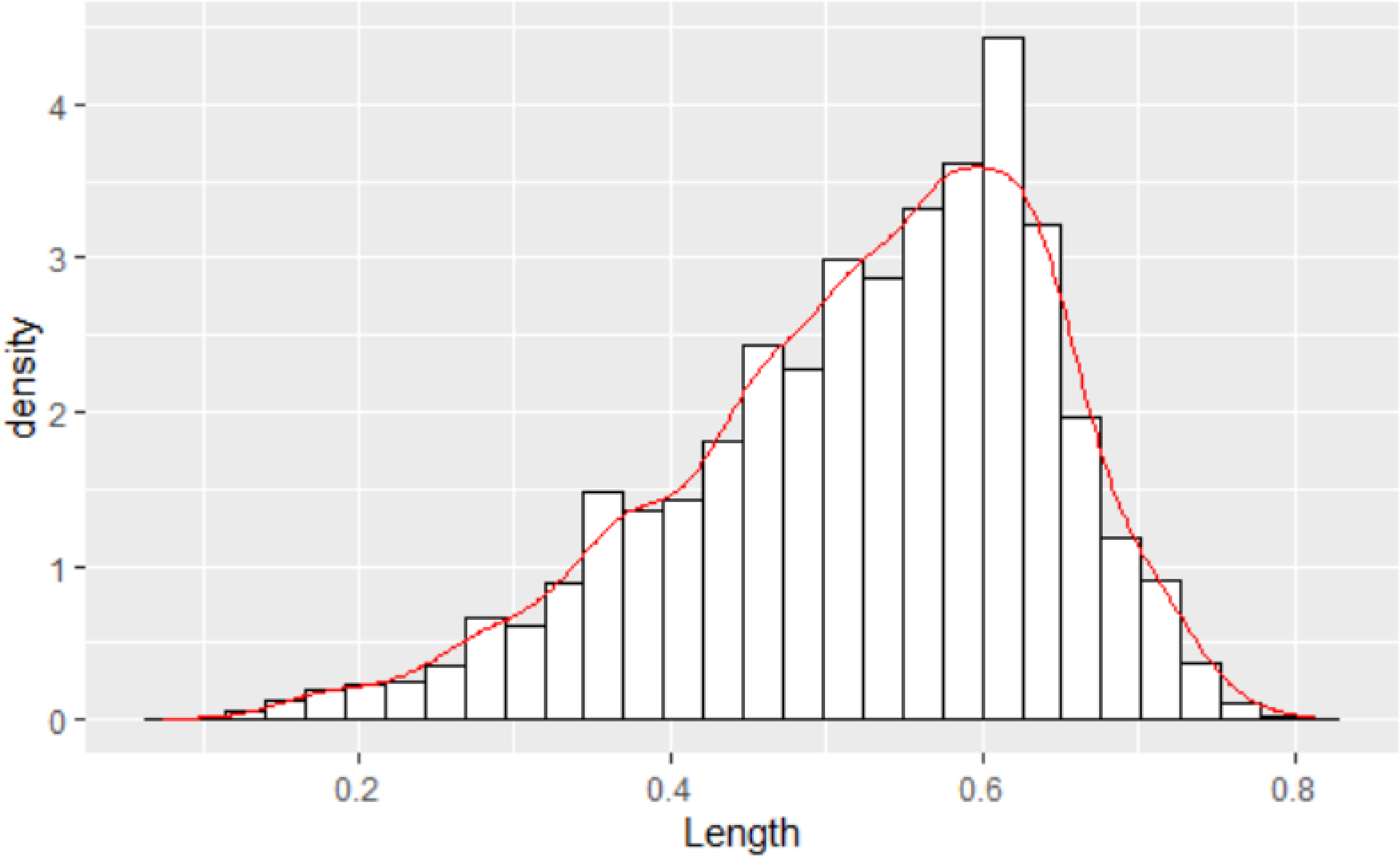
- 1 continuous variable



1D plots

- Histogram + Density

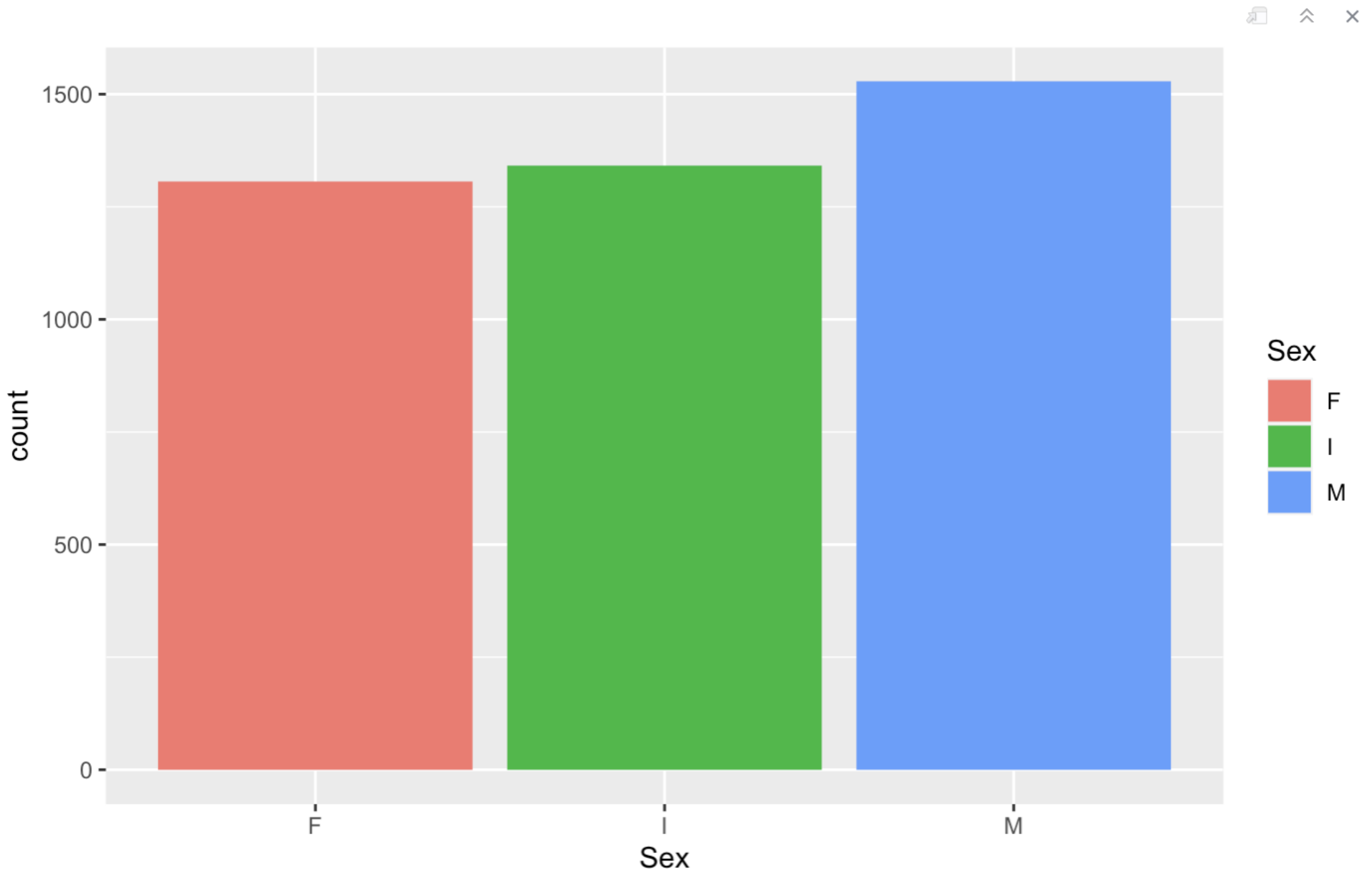
- 1 continuous variable



1D plots

- Bar plot

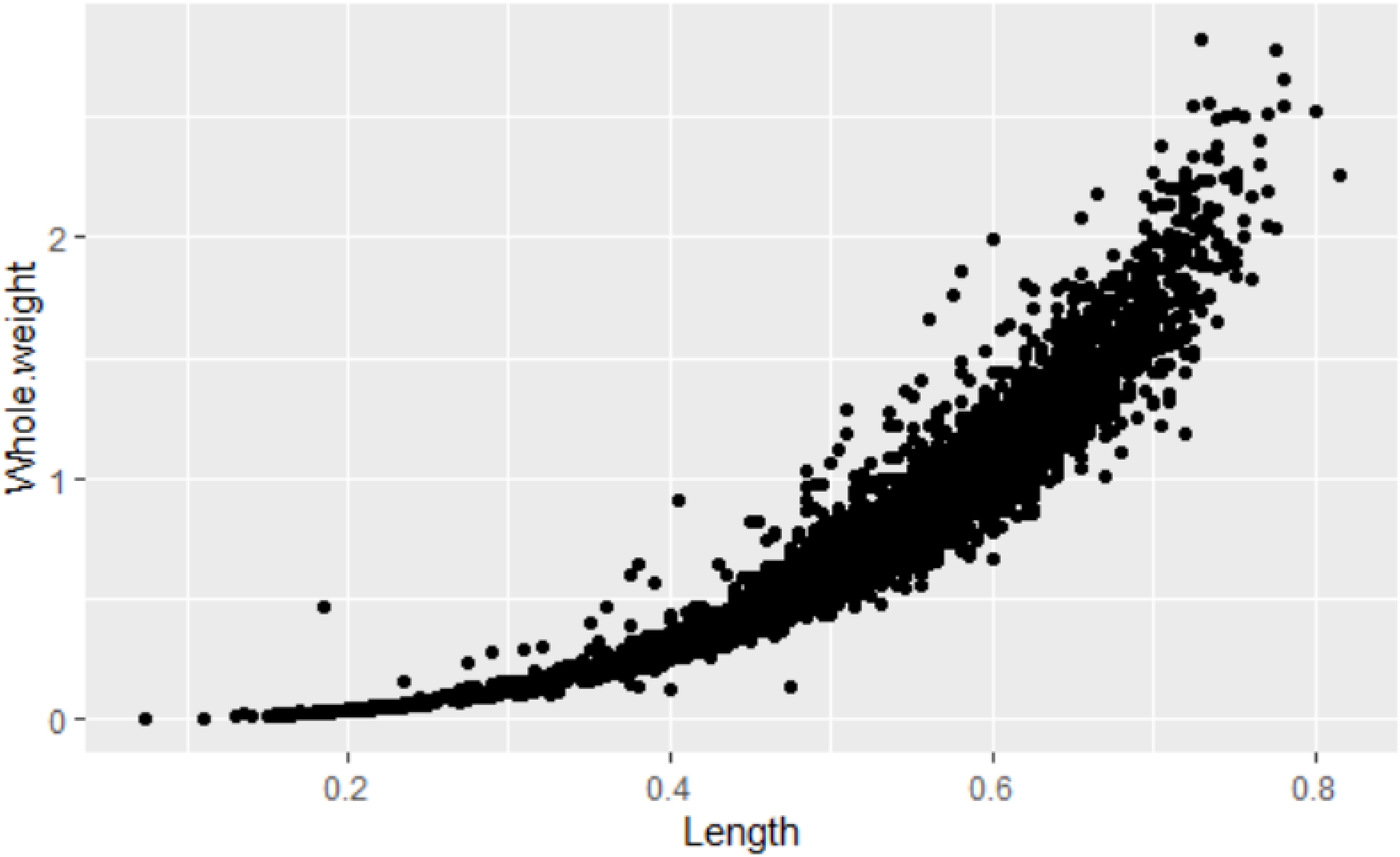
- 1 categorical variable



2D plots

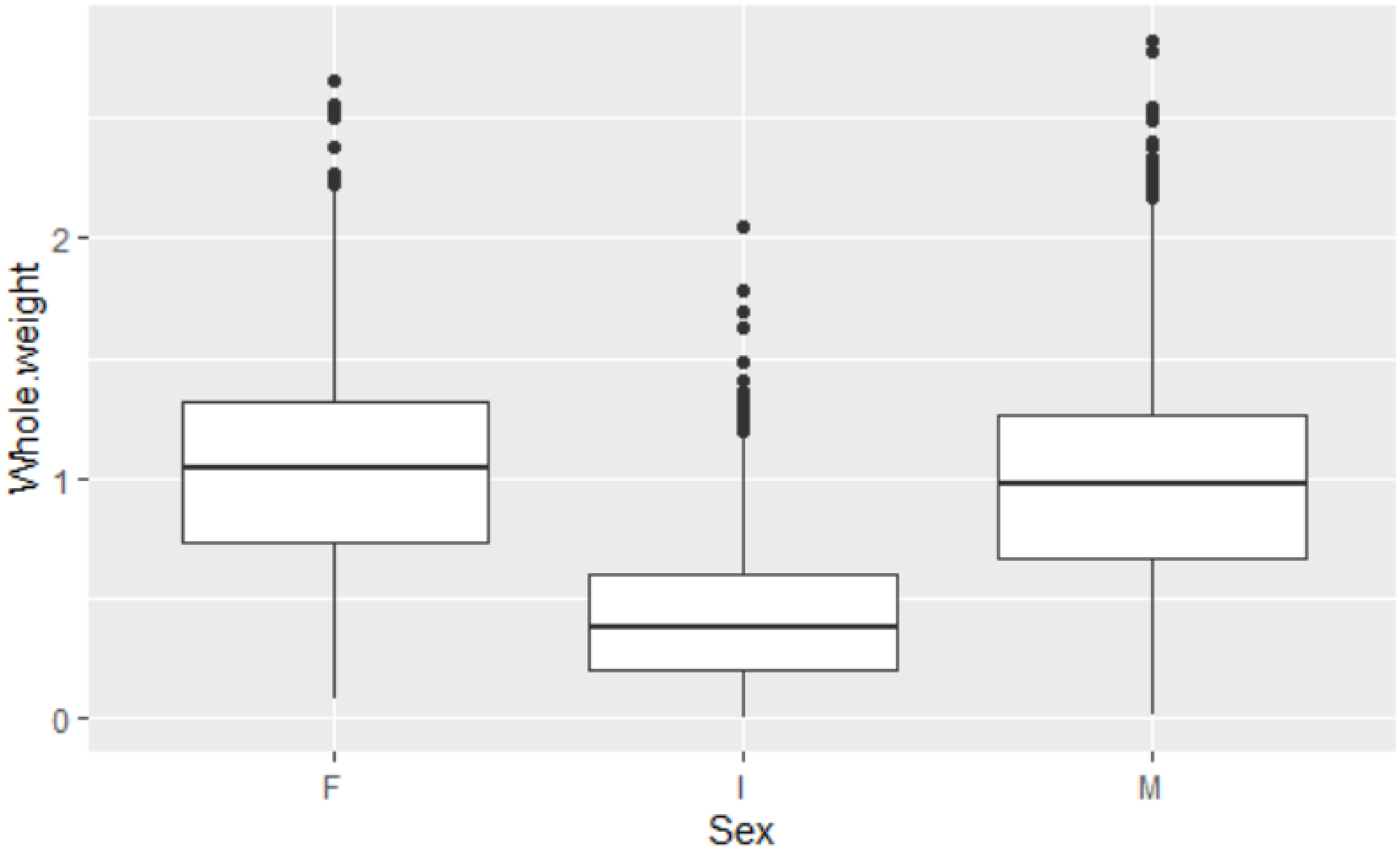
- Point plot

- 2 continuous variables



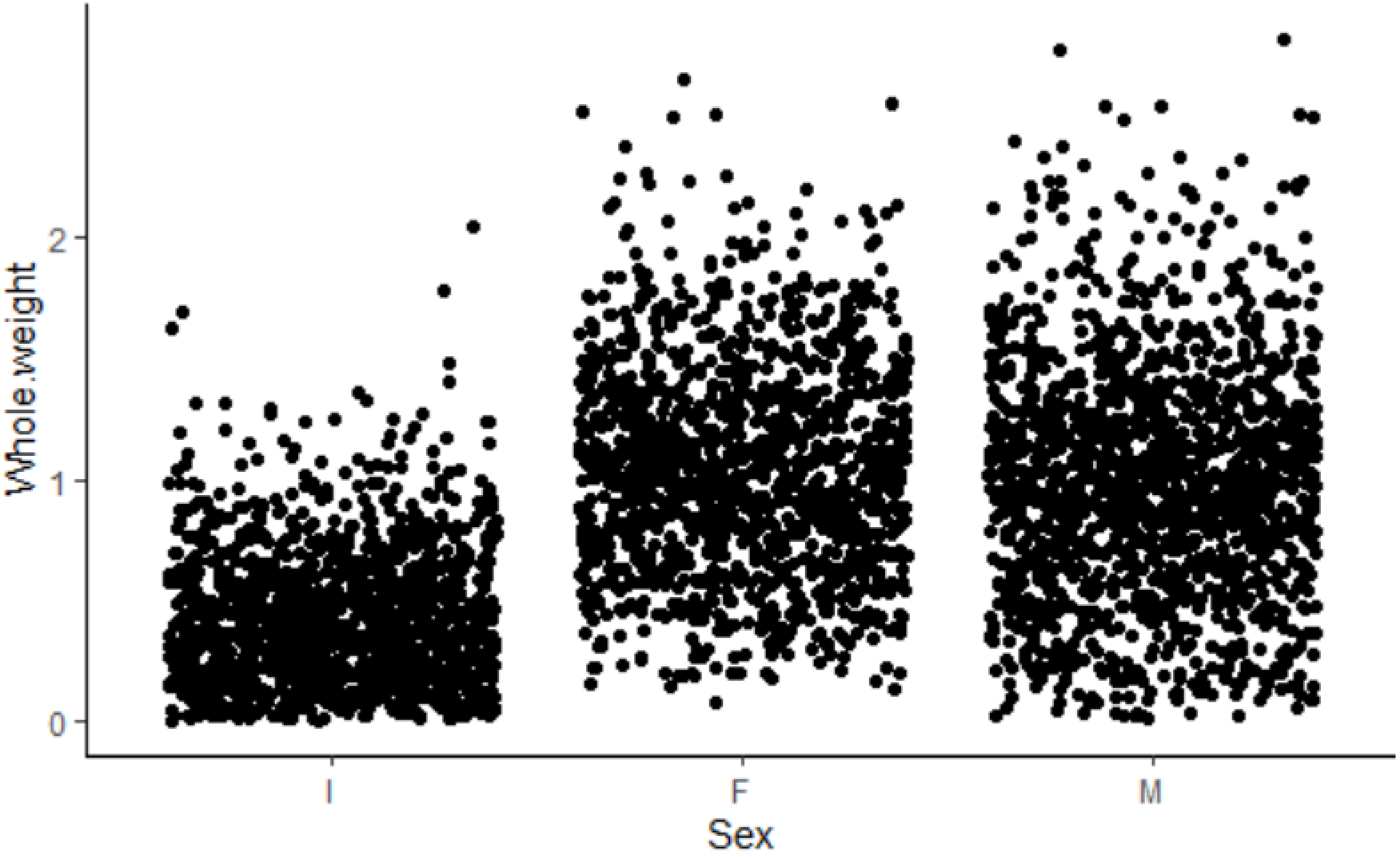
2D plots

- Box plot
- 1 continuous variable x
  1 categorical variable



2D plots

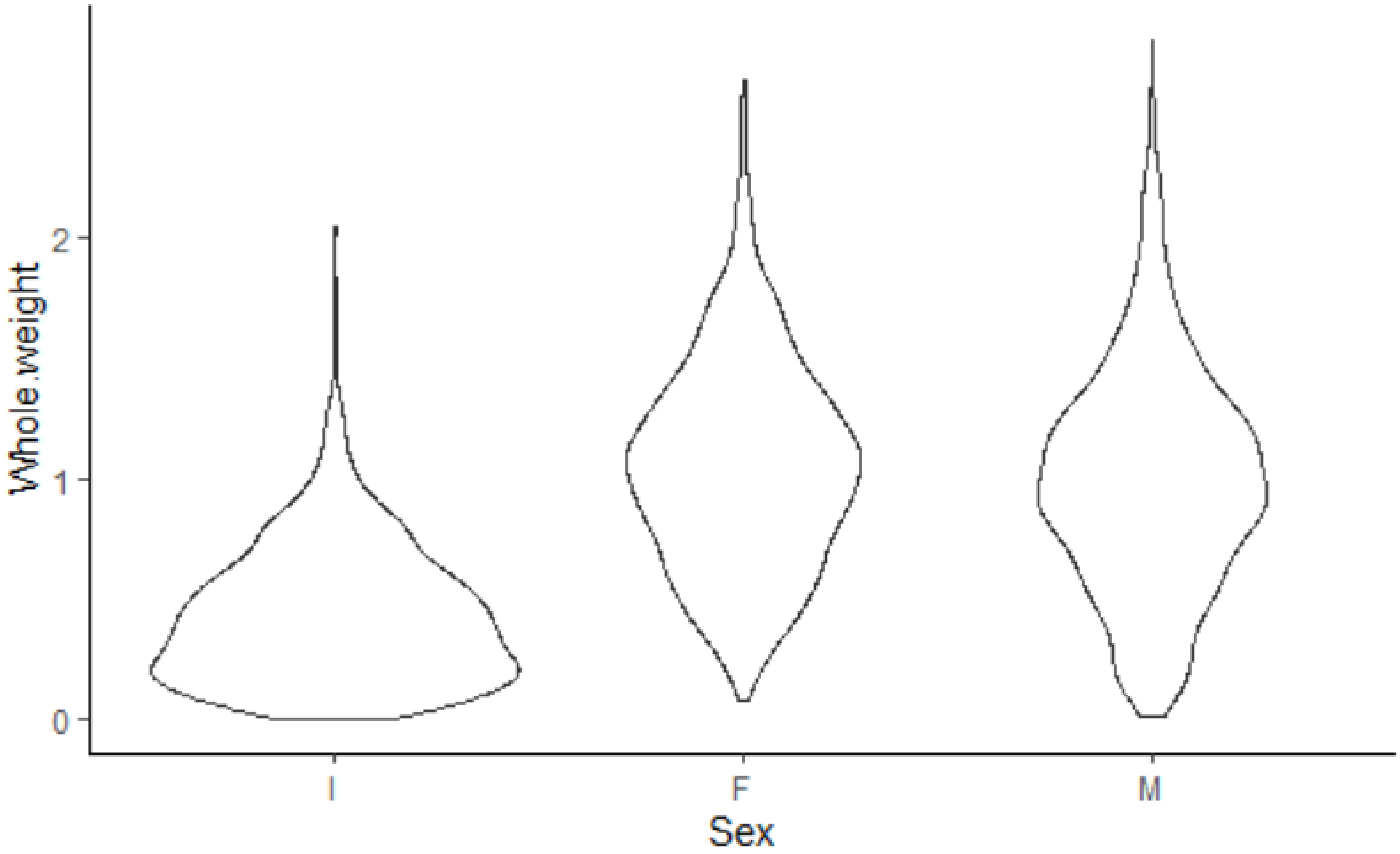
- Jitter plot
- 1 continuous variable x
  1 categorical variable



2D plots

- Violin plot

- 1 continuous variable x
  1 categorical variable



Many ways of making plots

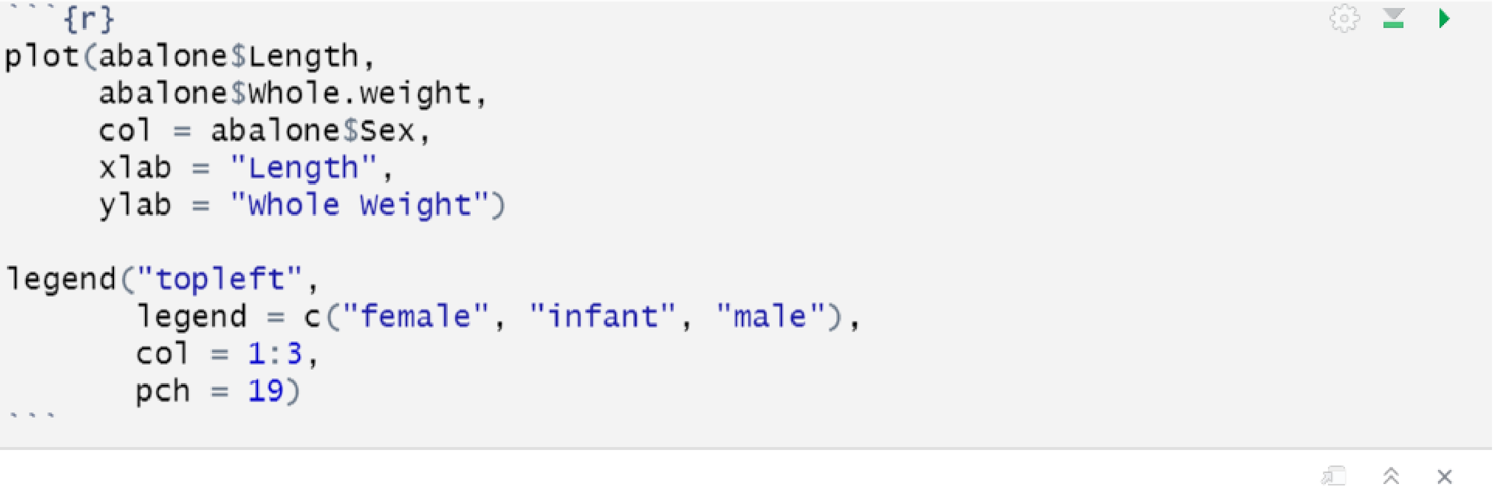
- Base R plots

- ggplot2 package

Base R

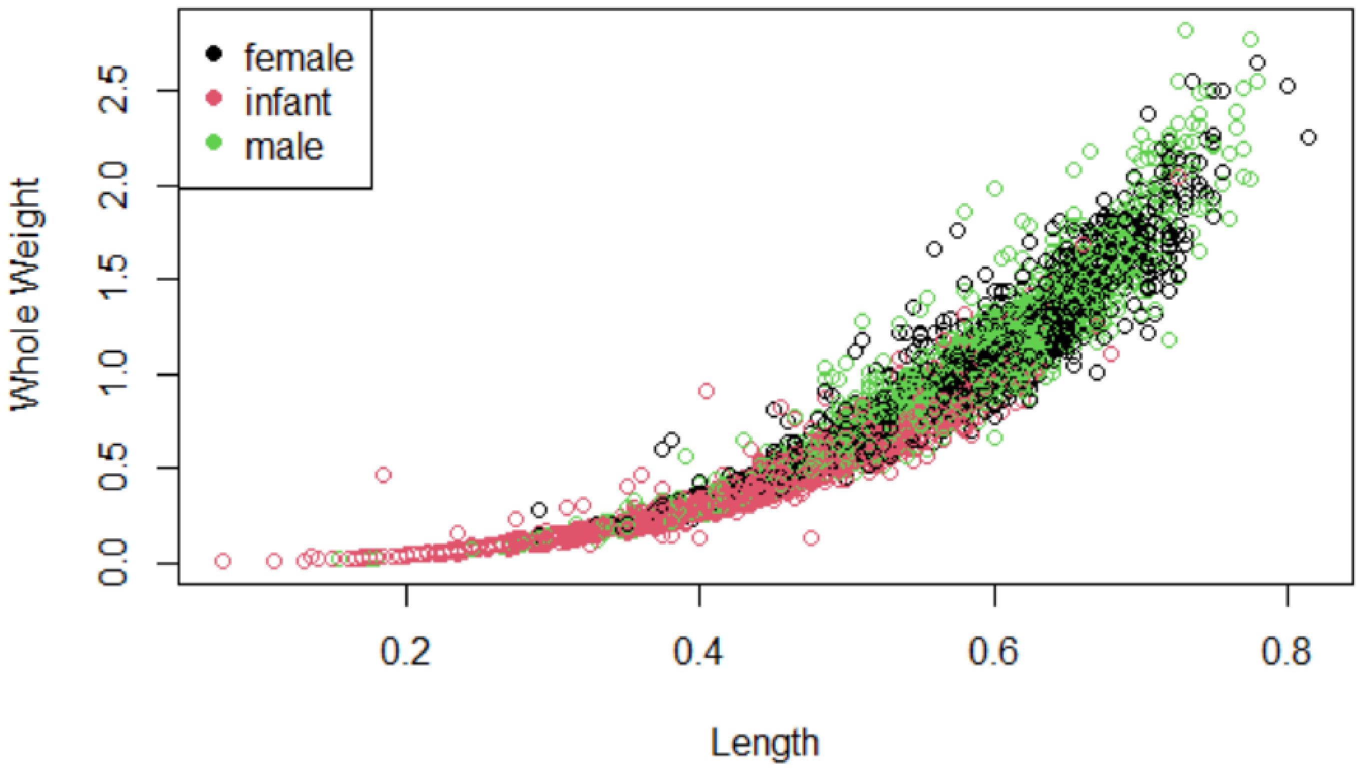
- Plots are built in layers
  - Plot + legend

- Plots require the dataset as the first parameter, the rest define the aesthetics of how it is displayed



~~~
```{r}
plot(abalone$Length,
     abalone$Whole.weight,
     col = abalone$Sex,
     xlab = "Length",
     ylab = "Whole Weight")

legend("topleft",
       legend = c("female", "infant", "male"),
       col = 1:3,
       pch = 19)
```
~~~



Base R

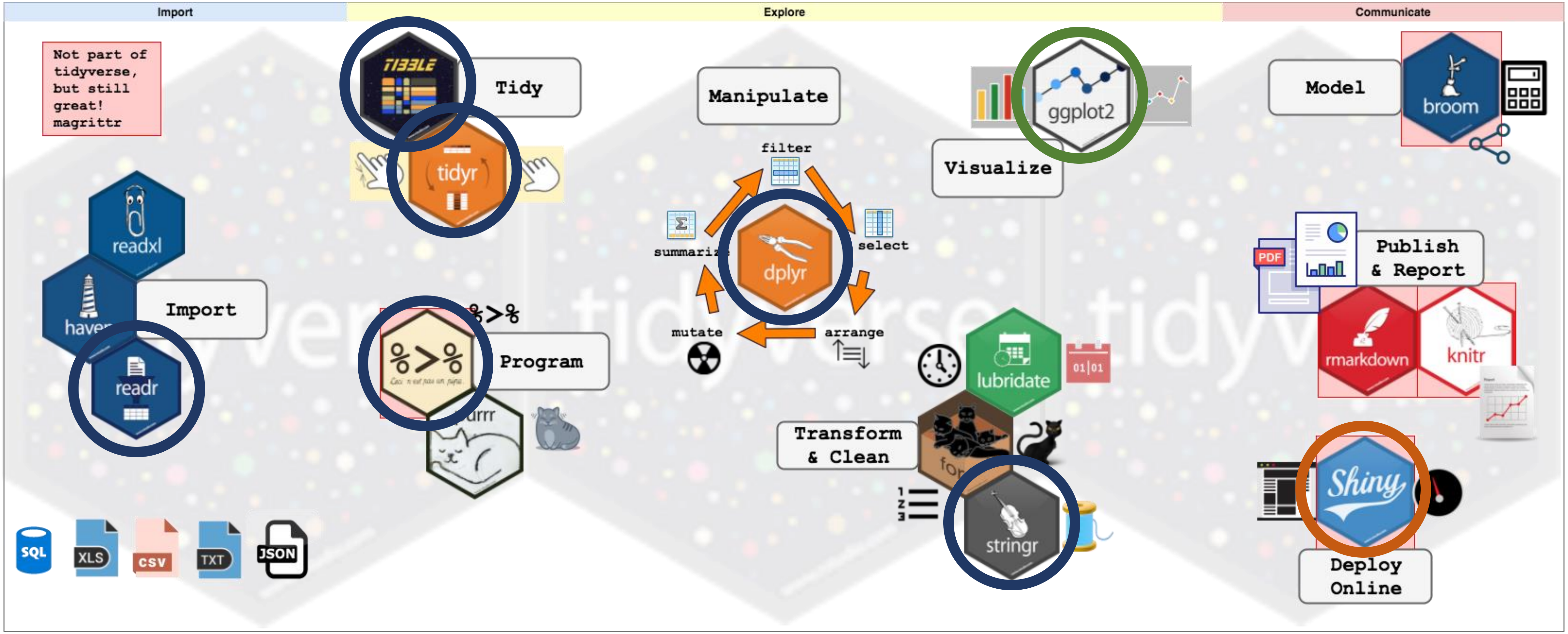
Strengths

- Syntax matches the rest of the base R
- Layers break up the tasks

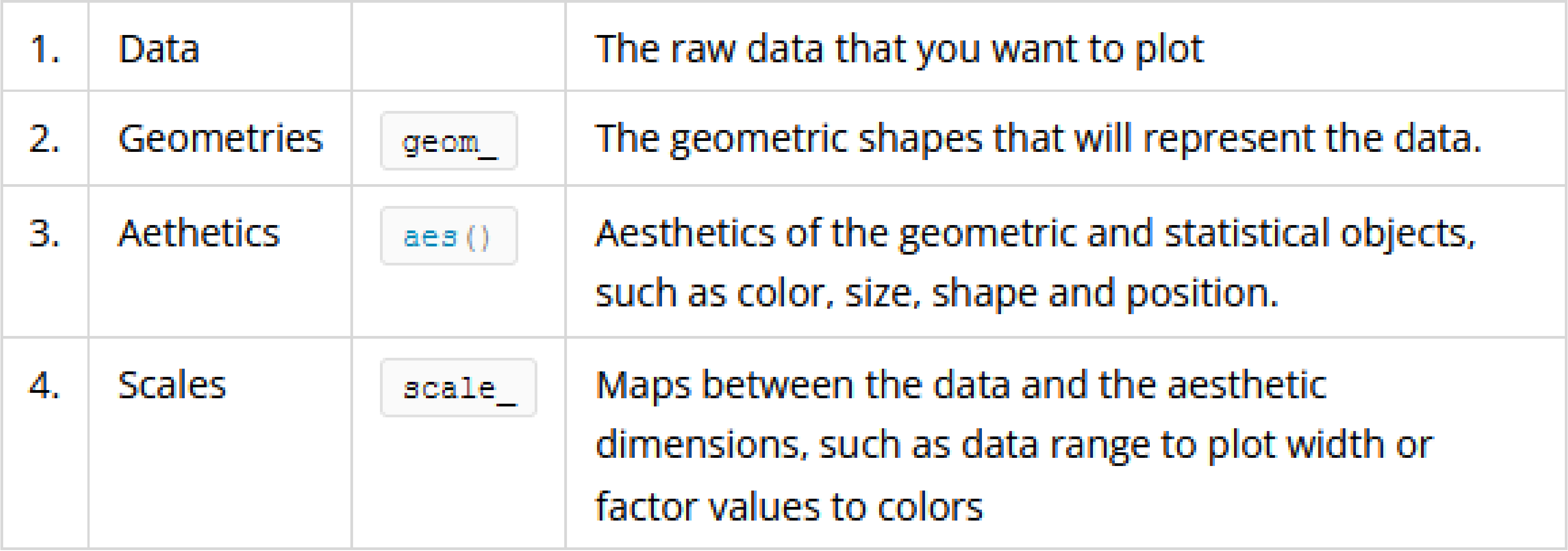
Limitations

- Layers require a base plot to be made
- Each plot type has their own function
  - Need to learn multiple functions
- Not visually appealing

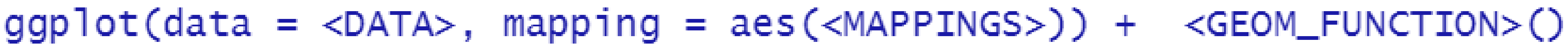
Returning to Tidyverse



One base template for all ggplots



1.	Data		The raw data that you want to plot
2.	Geometries	`geom_`	The geometric shapes that will represent the data.
3.	Aethetics	`aes()`	Aesthetics of the geometric and statistical objects, such as color, size, shape and position.
4.	Scales	`scale_`	Maps between the data and the aesthetic dimensions, such as data range to plot width or factor values to colors

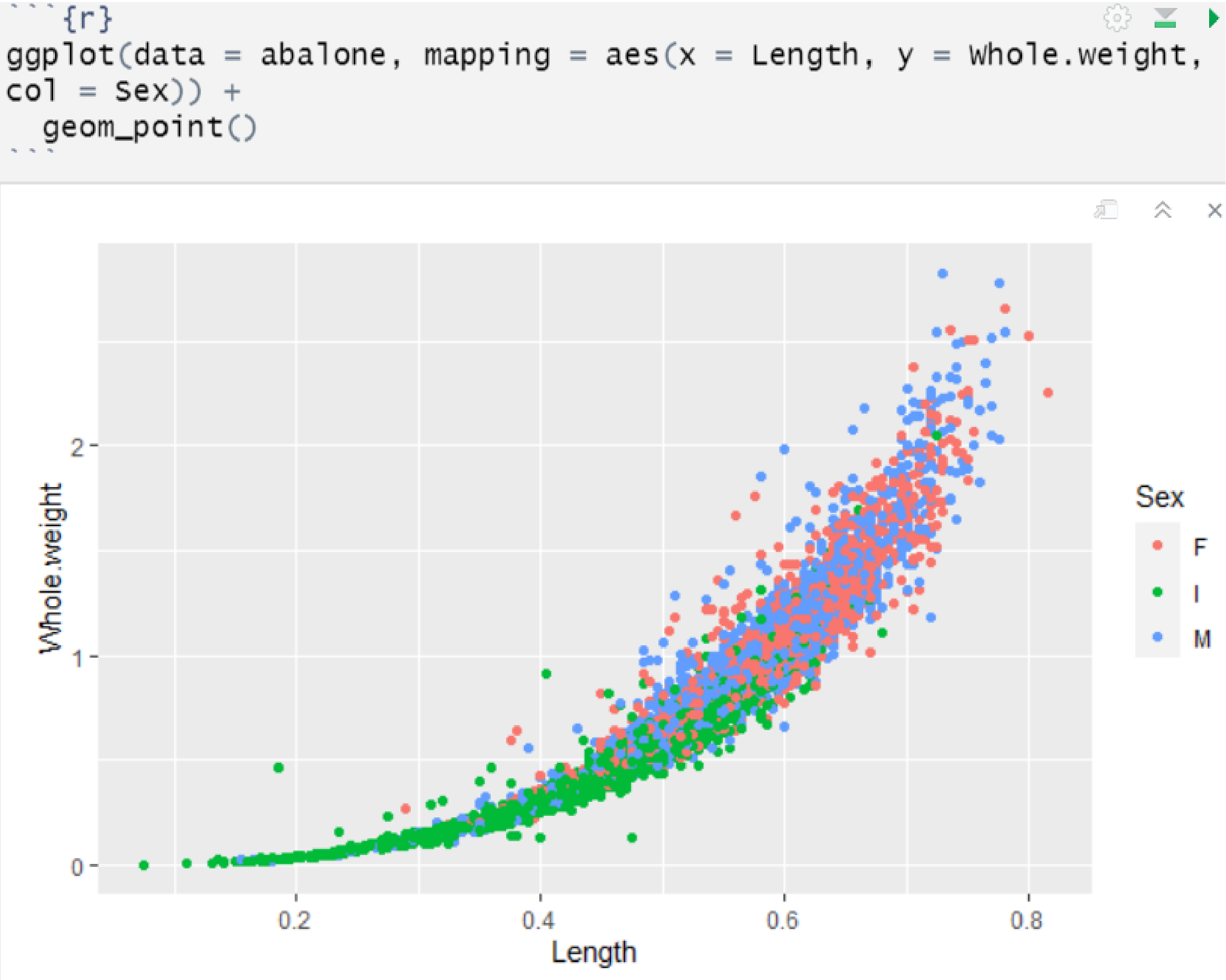


```
ggplot(data = <DATA>, mapping = aes(<MAPPINGS>)) +  <GEOM_FUNCTION>()
```

https://www.r-bloggers.com/2014/03/a-simple-introduction-to-the-graphing-philosophy-of-ggplot2/

ggplot point plot

- Global parameters set in the ggplot() function
- Geometry needs to be specified
- Unified format for all geometries
- Layers connected with `+`
  - Each layer’s aesthetic can be specified individually



~~~
```{r}
ggplot(data = abalone, mapping = aes(x = Length, y = Whole.weight,
col = Sex)) +
  geom_point()
```
~~~

ggplot

Strengths

- One structure for all plots
- Consistent with other tidyverse functions
- Themes and complete graphics system makes plots publication-ready
- Many users

Limitations

- Different syntax compared to base R
- Sometimes may need to wrangle data additionally
- Output types are currently being developed

Wrap up

- Must first understand the structure of the data and the question you are trying to solve in order to select the most appropriate
    - E.g.,. While histograms and bar plots look similar, a single continuous variable should use histograms

- Multiple functions can be used to produce the same/very similar plots.
    - Science publications tend to use ggplot2

Find some inspiration!

- https://r-graph-gallery.com/
- http://www.cookbook-r.com/Graphs/
- https://www.data-to-viz.com/

General check in

- You’re learning many different languages right now – this can be difficult!

- Some concepts will overlap and be reinforced in multiple classes
    - R and Python both have a working directory the respective program is looking at for reading and writing files to your computer

- Other concepts will be different and can make it more difficult to remember
    - The functions for recalling or specifying the working directory are different
    - R indexes from 1, Python indexes from 0