Lecture 2.4 – Basic Data Analysis
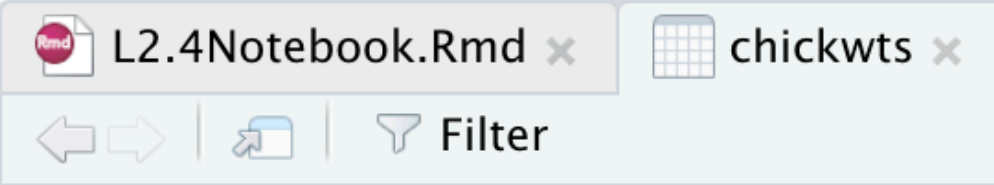
Learning Objectives:

3.4. Execute qualitative and quantitative data analyses.

How Can You Make Sense of Data?

- There is a huge amount of data that exists on nearly everything these days, and that amount is increasing every day.

- It's only useful if you have the skills to make sense of it, to extract meaning, and to communicate that meaning to others!

- In this lecture, we'll describe some tools that can help you understand patterns in data.

An Example Data Set

- The `chickwts` data set is the results of a feeding experiment in baby chickens.

- Chicks were fed one of six diets and their weights (in grams) were measured after a certain amount of time.

- The data set consists of two columns:

  ▸ `weight`: weight of chicks at the end of the experiment (in grams)
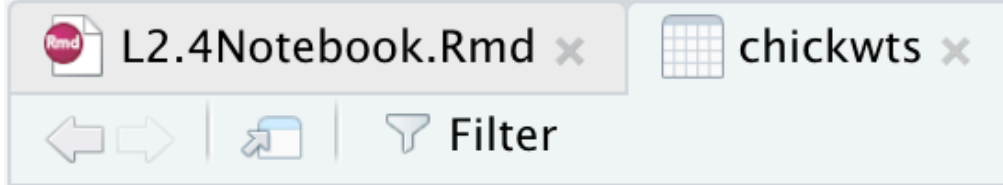
  ▸ `feed`: The diet they were fed.

	L2.4Notebook.Rmd ×		chickwts ×
	⇦⇨ ⤴ ▽ Filter		

	weight	feed
1	179	horsebean
2	160	horsebean
3	136	horsebean
4	227	horsebean
5	217	horsebean
6	168	horsebean
7	108	horsebean
8	124	horsebean
9	143	horsebean

How can we initially describe the data?

- Description of independent versus dependent variables. (What is suppose to be influencing what?)

- Independent variable: `feed`

- Dependent variable: `weight`

L2.4Notebook.Rmd ×	chickwts ×

Filter

	weight	feed
1	179	horsebean
2	160	horsebean
3	136	horsebean
4	227	horsebean
5	217	horsebean
6	168	horsebean
7	108	horsebean
8	124	horsebean
9	143	horsebean

How can we initially describe the data?

- Variable types are an important consideration to understanding data because they define what sort of visualizations and statistics you are able to perform.

| Variable Type	Data Class in R	Example
Continuous	numeric	speed of cars on a track
Discrete	integer	number of birds in a yard
Categorical	factor	color of sea stars (purple, orange, or brown)

How can we initially describe the data?

- **Continuous**: variables that exist on a continuous scale. Things reported with decimal places are often continuous. Continuous data can exist at any point in the variable's range. Continuous data are often represented as numerical data class in R.
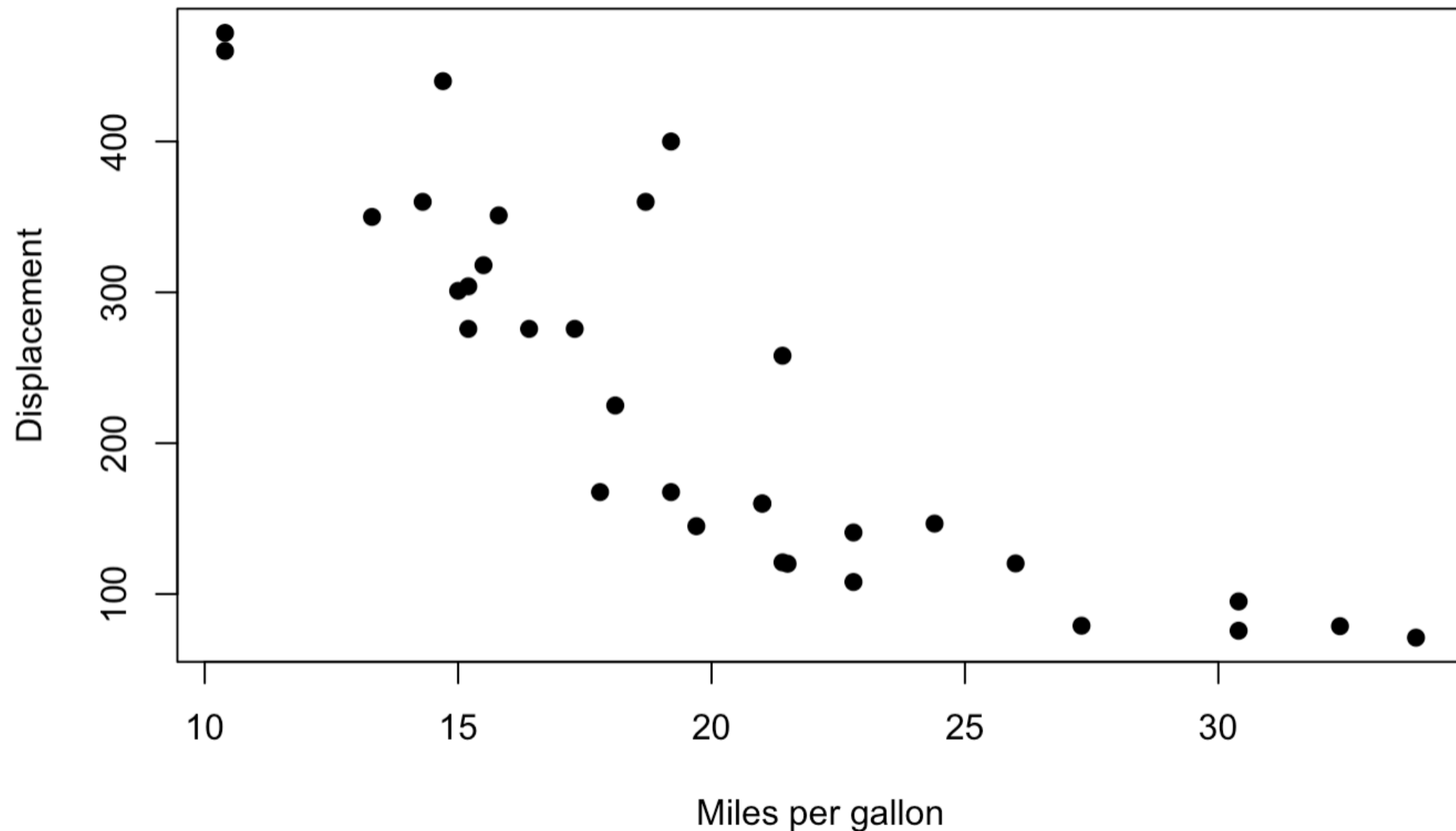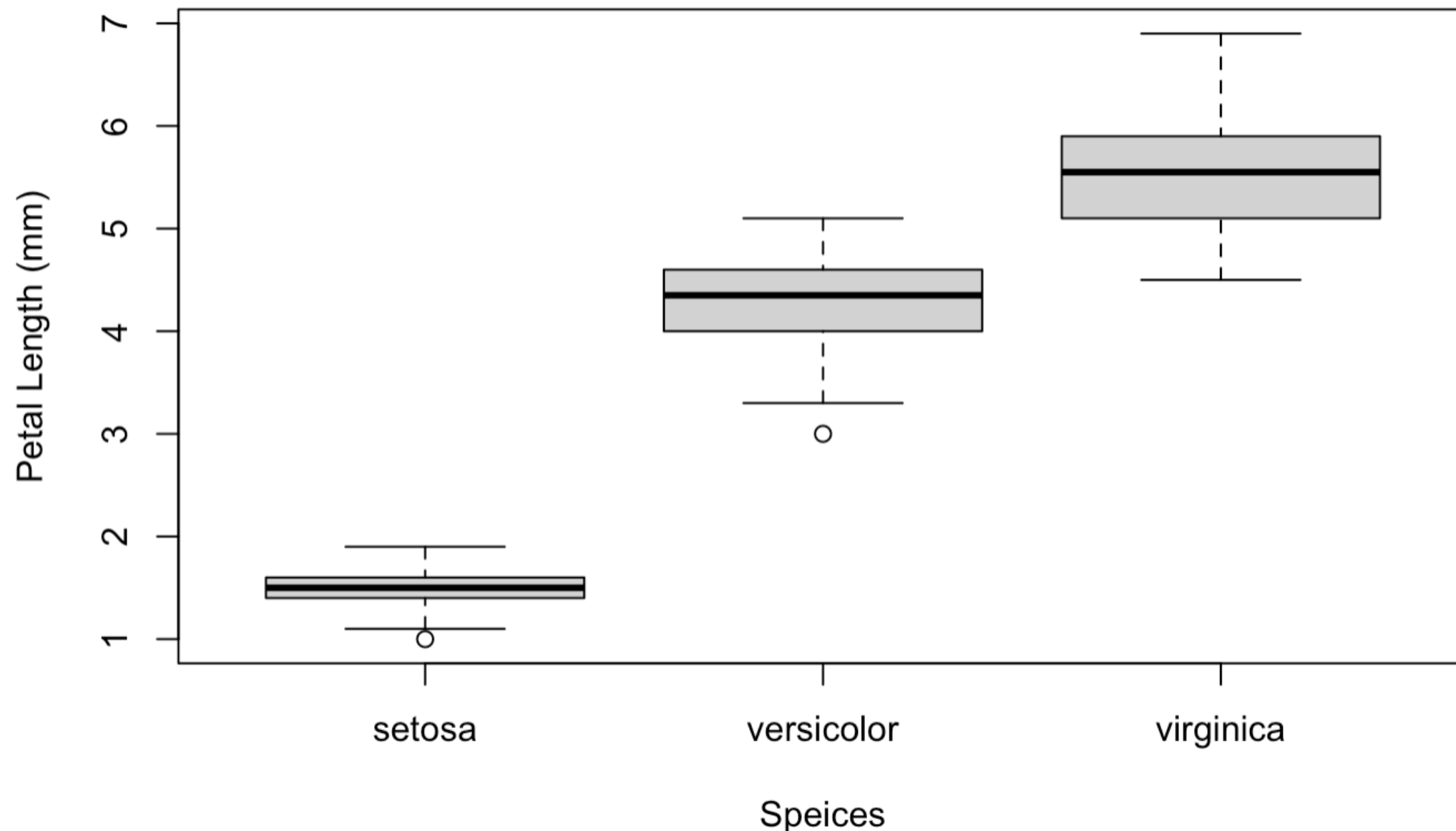
How can we initially describe the data?

- **Discrete**: variables that are numbers but can only be pulled from specific numbers (and no where in between). Discrete data can be represented as integers or factors data classes in R (other classes are also possible).
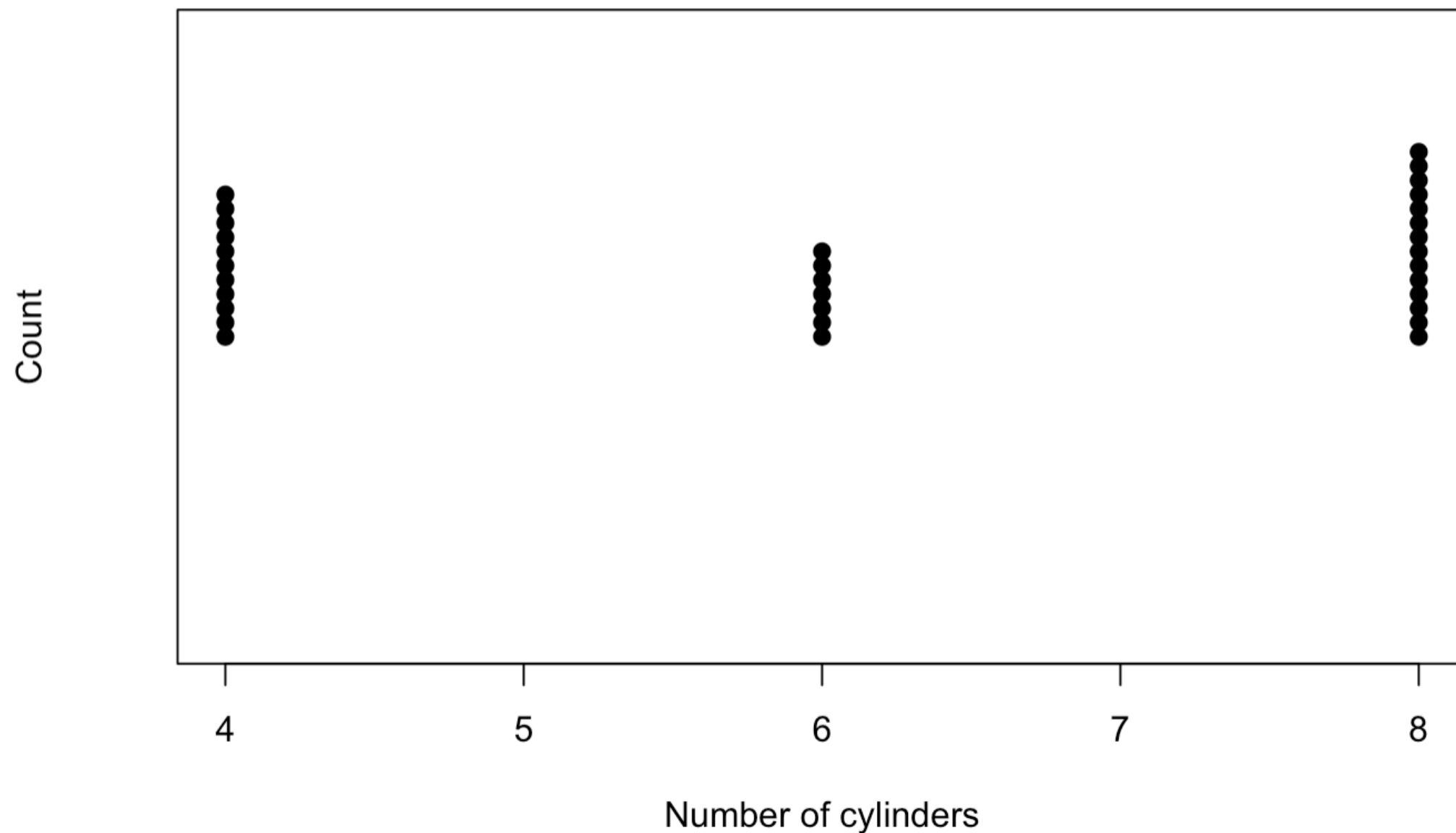
How can we initially describe the data?

- **Categorical**: variables which are non-numerical and form discrete groups. Things like color, quality, and species are categorical variables. Categorical data is most often represented in R as data class factor.

How can we initially describe the data?

- *Univariate* versus *multivariate* data: how many numbers do you need to describe the variable?

<table>
<tr><td align="center"><u>Univariate</u></td><td align="center"><u>Multivariate</u></td></tr>
<tr><td align="center">One number
does it</td><td align="center">Needs at least
two numbers</td></tr>
<tr><td align="center">weight
(a number in grams)</td><td align="center">location
(needs one number for
latitude and one for
longitude)</td></tr>
</table>

Check Your Understanding

For each column in `chickwts`, what is the variable type? How could you plot these columns together? (Make the plot.)

Summary Statistics

- Measures of **centrality** are often useful for getting an overview of the data. Each tells you a little different info about the data.

  - **Mean**: arithmetic calculation of the center point of the data set. Calculated with `mean()` in R.

  - **Median**: middle magnitude of a set of observations. Calculated with `median()` in R.

  - **Mode**: most common observation in a set. Calculated with `table()` in R, which will show you the count per observation.

Summary Statistics

- Measures of **spread** describe not only where they are relative to centrality, but also how spread out the data are from each other.

  - **Range**: describes the lowest and highest value in a data set. This can tell you about the absolute spread of the data set. Range is calculated using `range()` in R.

  - **Quantile**: a statistic that tells you the point below which a certain percentage of data lies. Typical quartiles are measured at 0%, 25%, 50%, 75%, and 100%. Quantiles are calculated using `quantile()` in R.

  - **Standard Deviation**: tells you how much values are spread around the mean. Calculated with `sd()` in R.

Summary Statistics

- Pro tips:

    - use `summary()` to quickly retrieve quantiles, mean, median for each column in a data frame!

    - use `tapply()` to apply a function within groupings in a data set!

    - use `IQR()` to calculate the interquartile range for any data.

Check Your Understanding

Calculate the mean and standard deviation of sepal lengths in the iris data set broken down by species. (Hint: use `tapply()`!)

Action Items

1. Complete Assignment 2.3

2. Create a new R project for your Project 1. Write code to load the data set into R.

3. Read Davies Ch. 7.4 and Chang Appendix A