Lecture 2
Describing and Visualizing Distributions

Review

Statisticians use to **data** to answer questions about **populations**

A population is the set of **ALL** observations of interest

Our data is usually a subset of **observations** from the population called a **sample**

The way in which we collect our data is called the **sampling design**

Warm Up

In elections, television networks use exit polling (interviewing voters after they leave the voting booth) to declare the winner well before all votes are counted. In the 2010 California gubernatorial election between candidates Jerry Brown (D) and Meg Whitman (R), an exit poll projected Brown to be the winner early into election night. Specifically, the network responsible for the poll interviewed 3,889 voters at the booth and determined that 53.1% favored the democratic candidate.

What is the statistical question?

Who will win the California Race for Governor

What is the population? What is the population size N?

The population is all eligible voters in the state of California

What is the sample ? What is the sample size n?

All voters interviewed at the election booth; 3,889 people

What is the statistic being calculated?

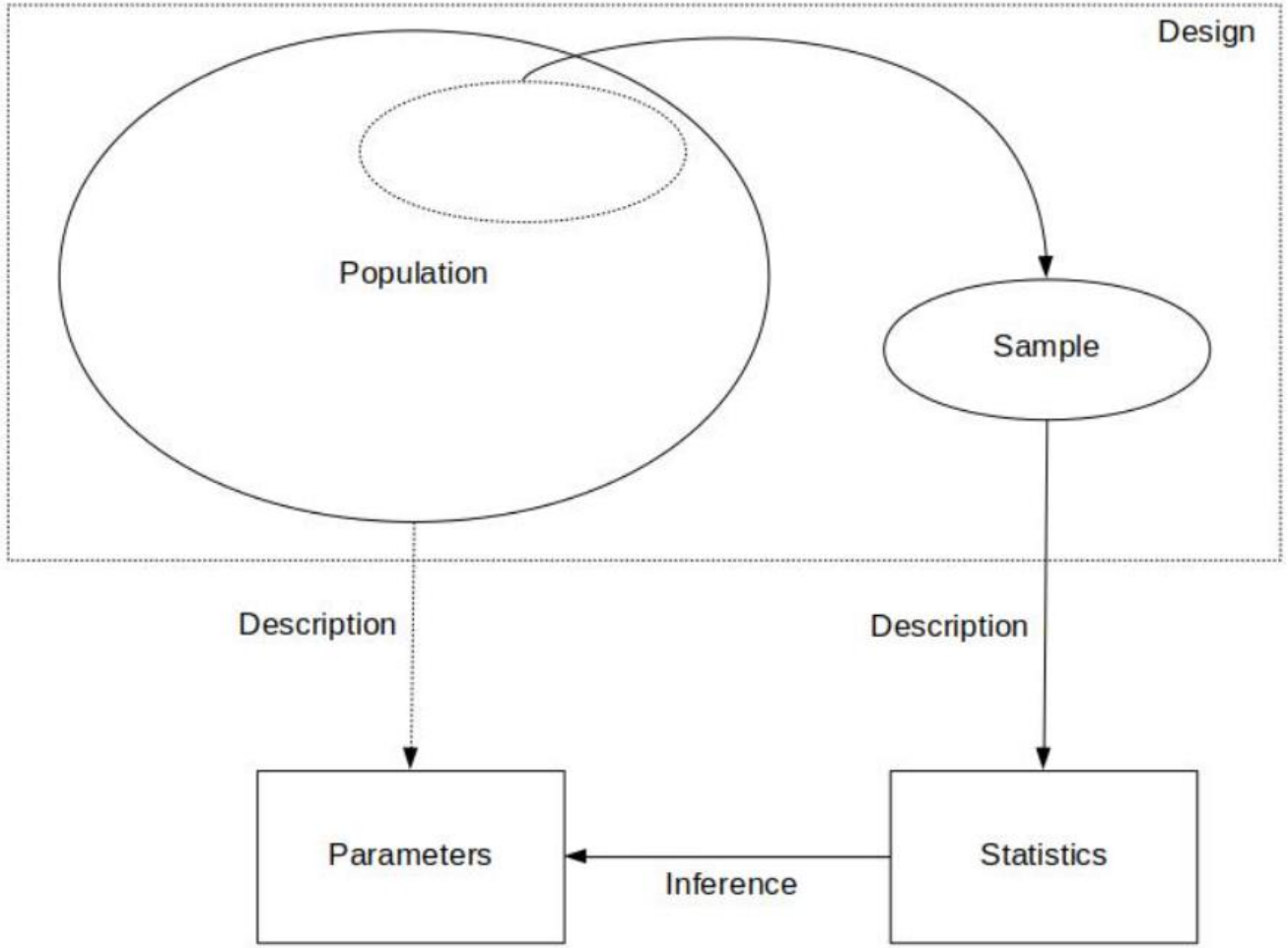
The proportion of 3,889 voters casting their vote for Democrat Jerry Brown

Recap:

1. **Design** – the goal/statistical question we want to answer and how we plan to obtain our data

2. **Description** – a preliminary exploration and summary of the data

3. **Inference** – using statistics to make decisions or predictions about the data



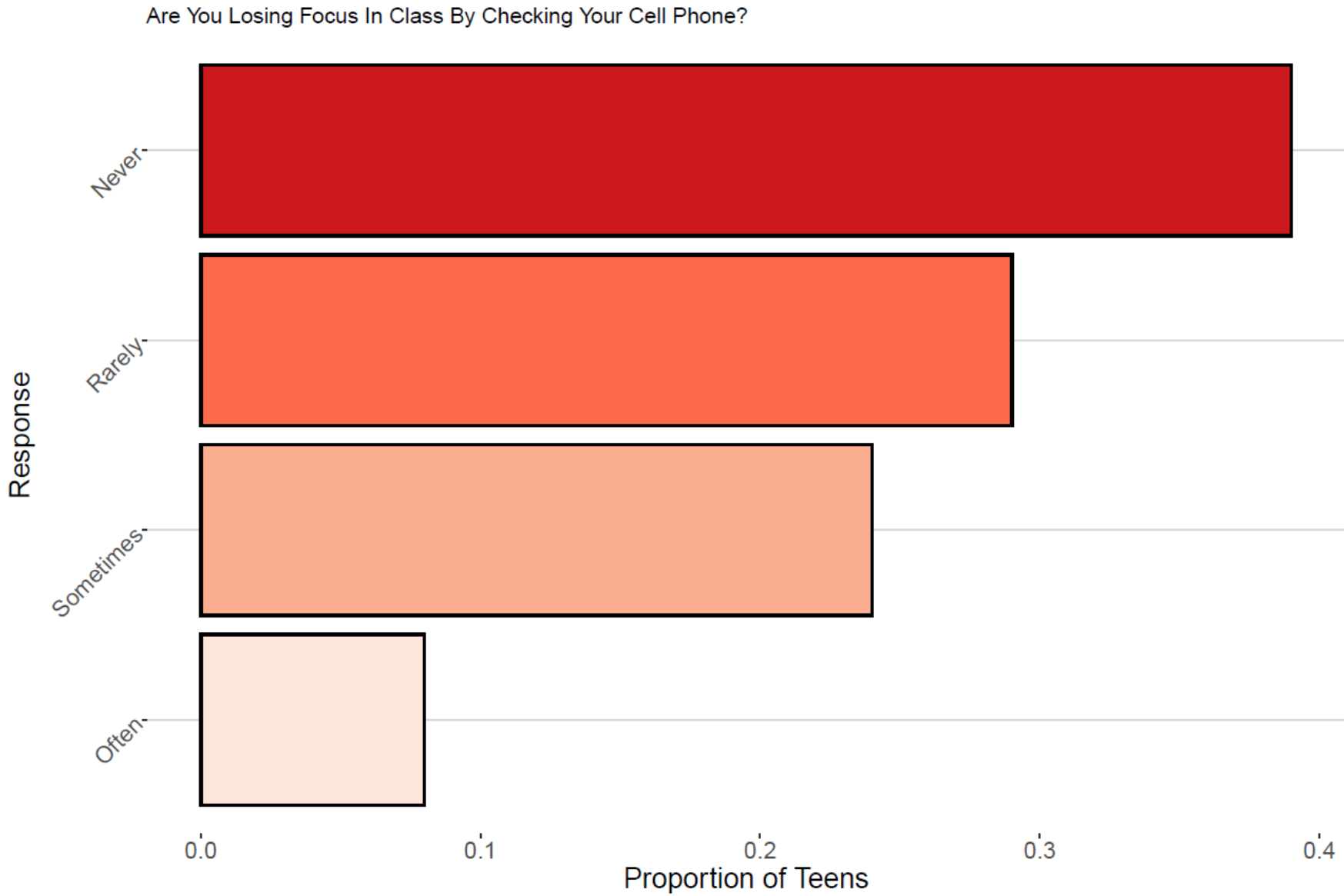
Descriptive statistics

- Methods for summarizing, visualizing and characterizing data
- The data can be from samples OR populations
- Goal: simplify the data without distorting or losing information
- Summaries are easier to understand

Ex. Are teens distracted by their cell phones? A study conducted by the Pew Research center surveyed 743 U.S teenagers ages 13-17 to understand how cell phones in class impacted their ability to concentrate. Each student was asked to rate the impact of cell phone use on their concentration based on a Likert scale

Student	Age	Response
1	13	Never
2	13	Sometimes
3	15	Never
4	17	Often
⋮	⋮	⋮
743	16	Rarely

Possible Responses: Never, Rarely, Sometimes, Often


Are You Losing Focus In Class By Checking Your Cell Phone?
Response
Never
Rarely
Sometimes
Often
0.0
0.1
0.2
0.3
0.4
Proportion of Teens

Features of Distributions

- A **distribution** of a variable gives (a) the values that occur and (b) how often each value occurs

Features of A Distribution

Categorical Variables:

- **Modal category** – the category with the highest frequency

Quantitative Variables:

- **Shape** – do observations cluster into certain areas, are values spread out or more densely packed together?
- **Center** – where is the middle point on the distribution, where does a typical value fall?
- **Variability** – how tightly to observations cluster around the center of the distribution

Displaying Distributions: Frequency Tables

- **Frequency table –** a table listing the distinct values of variable together with the number of observations of each value
- **Frequency** – the number of times a value occurs
- **Relative Frequency** – the proportion of observations that assume a given value

$$\text{RF} = \frac{f}{n},$$

- **Cumulative Relative Frequency** - the proportion of observations equal to or less than a given value (more on this later)

Ex. Are teens distracted by their cell phones?

Response	Frequency	Relative Frequency	Cumulative Relative Frequency
Never	289	0.39	0.39
Rarely	216	0.29	0.68
Sometimes	178	0.24	0.92
Often	60	0.08	1.00

Ex. Mendels Pea Plants - In one of Gregor Mendel's classic studies, he bred 8023 pea plants and observed the color of the pea pods.

Pea Color	Frequency	Relative Frequency
Yellow	6022	0.751
Green	2001	0.249

Try it out: Computing Frequencies

- Roll a six-sided die $n = 10$ times and record the number rolled each time

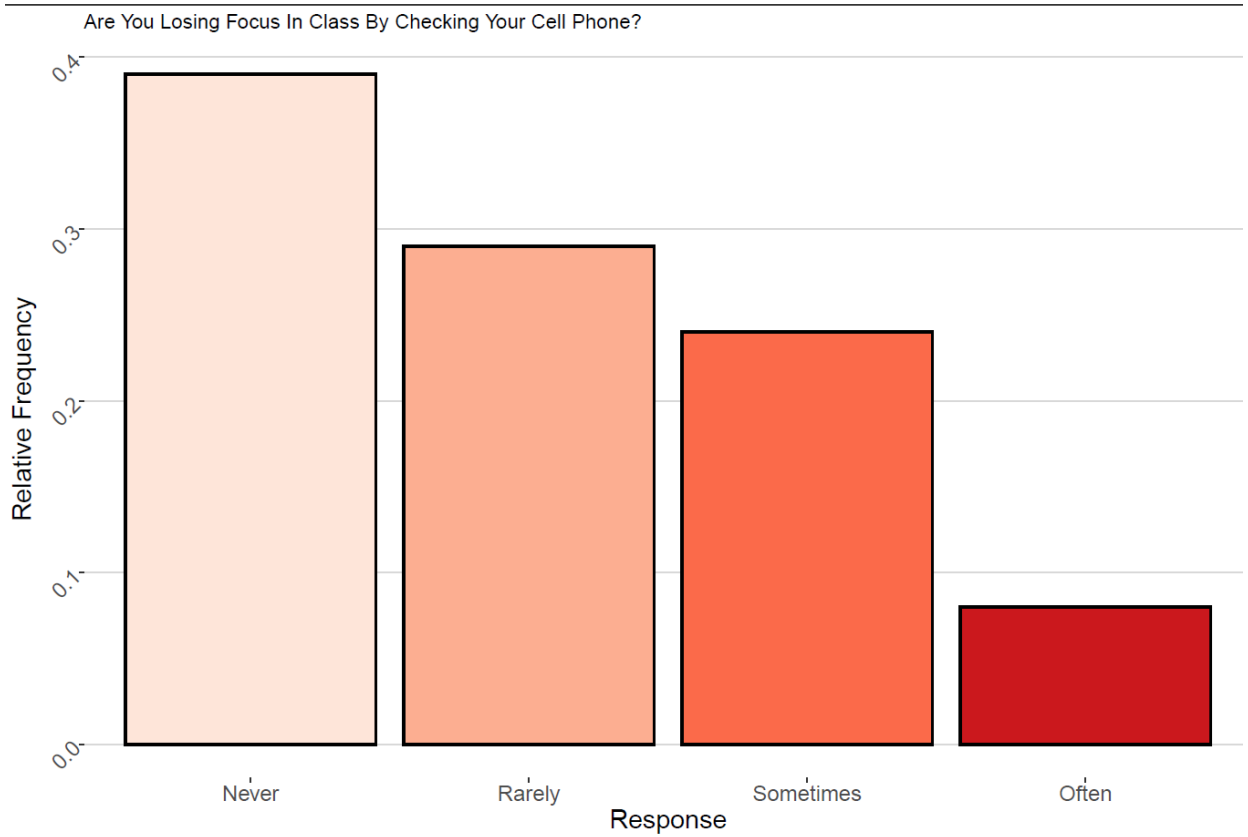
- Data = 1,2,3,3,4,4,4,5,6,6

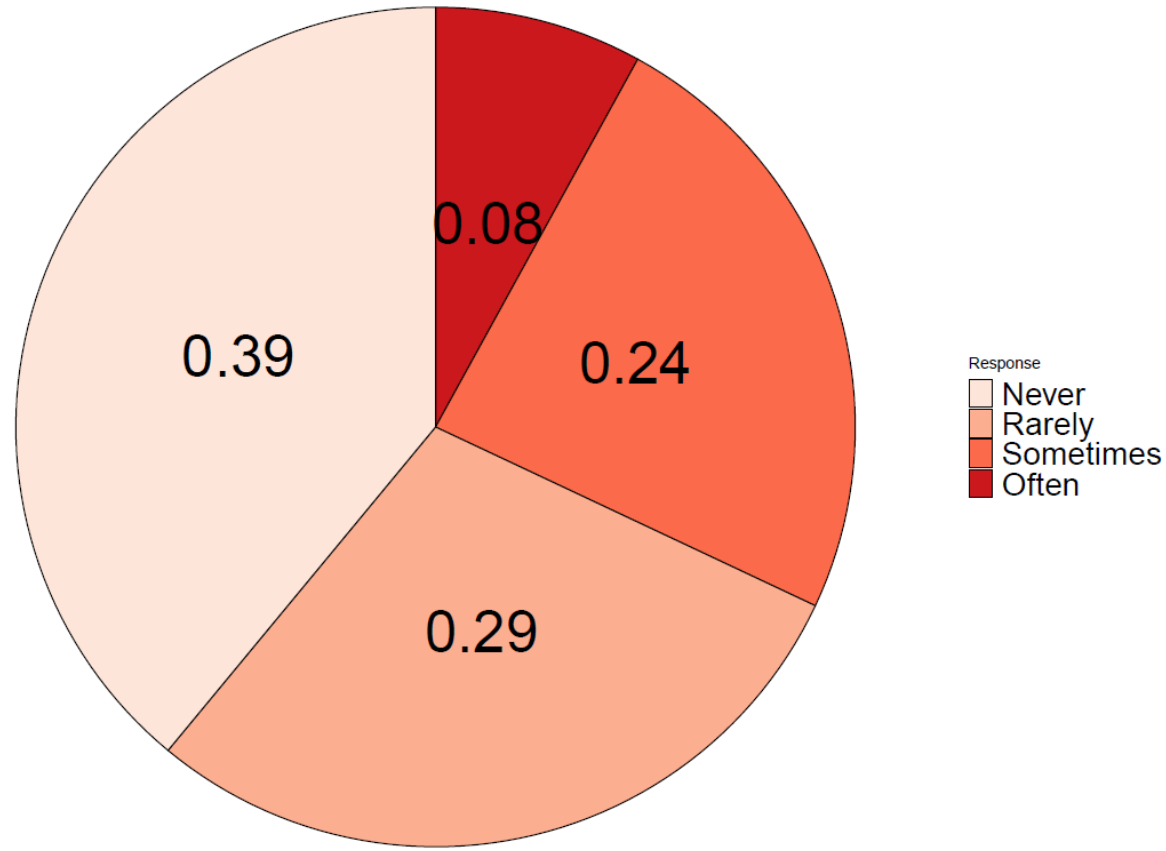
Visualizing Distributions of Categorical Data

Pie Charts - a circle divided into ‘slices’ corresponding to each category. The size of a slice shows the proportion of observations in a category

Bar graph – displays a vertical bar for each category. The height of the bar shows the percentages of observations in the category

Pareto Chart - a bar chart with the categories ordered by decreasing frequency


Are You Losing Focus In Class By Checking Your Cell Phone?
Relative Frequency
0.4
0.3
0.2
0.1
0.0
Never
Rarely
Sometimes
Often
Response


0.08
0.39
0.24
0.29
Response
Never
Rarely
Sometimes
Often

Frequency Tables for Continuous Variables

- The number of possible values is usually very large
- Convert continuous values into discrete groups (sometimes called bins):

Steps:

1. Divide the range of the variable into a set of non-overlapping intervals
2. Count the number of values that fall into each interval

Example: Old Faithful Eruption Times

Observation	Eruption Time	Waiting Time
1	3.600	79
2	1.800	54
3	3.333	74
4	2.283	62
5	4.533	85
6	2.883	55
7	4.700	88
8	3.600	85
9	1.950	51
10	4.350	85
11	1.833	54
⋮	⋮	
272	4.467	74

Old Faithful Eruption Times: Frequency Table

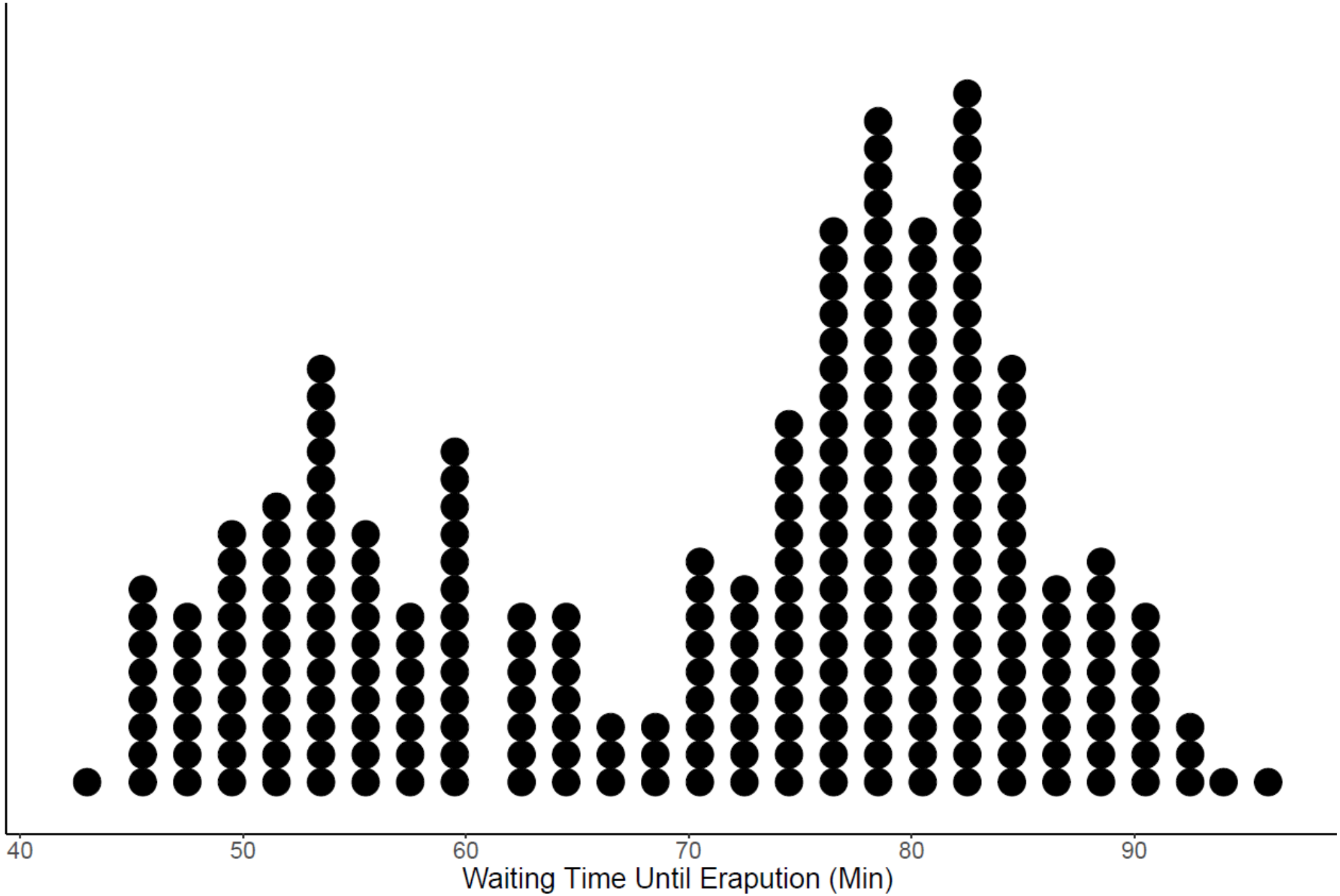
Waiting Time (Min)	Frequency	Relative Frequency	Cumulative Relative Frequency
< 50	21	0.077	0.077
50 - 60	56	0.206	0.283
60 - 70	26	0.096	0.379
70 - 80	77	0.283	0.662
80 - 90	80	0.294	0.956
> 90	12	0.044	1

Visualizing Distributions: Quantitative Variables

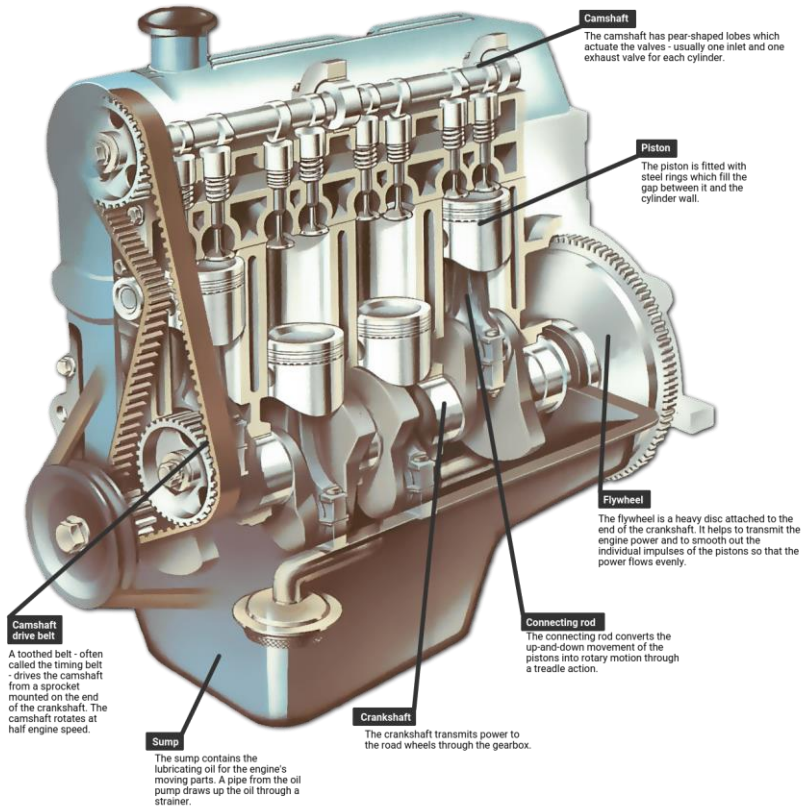
- **Dot plots –** shows a dot for each observation placed above the value for that observation

Steps to construct a dot plot

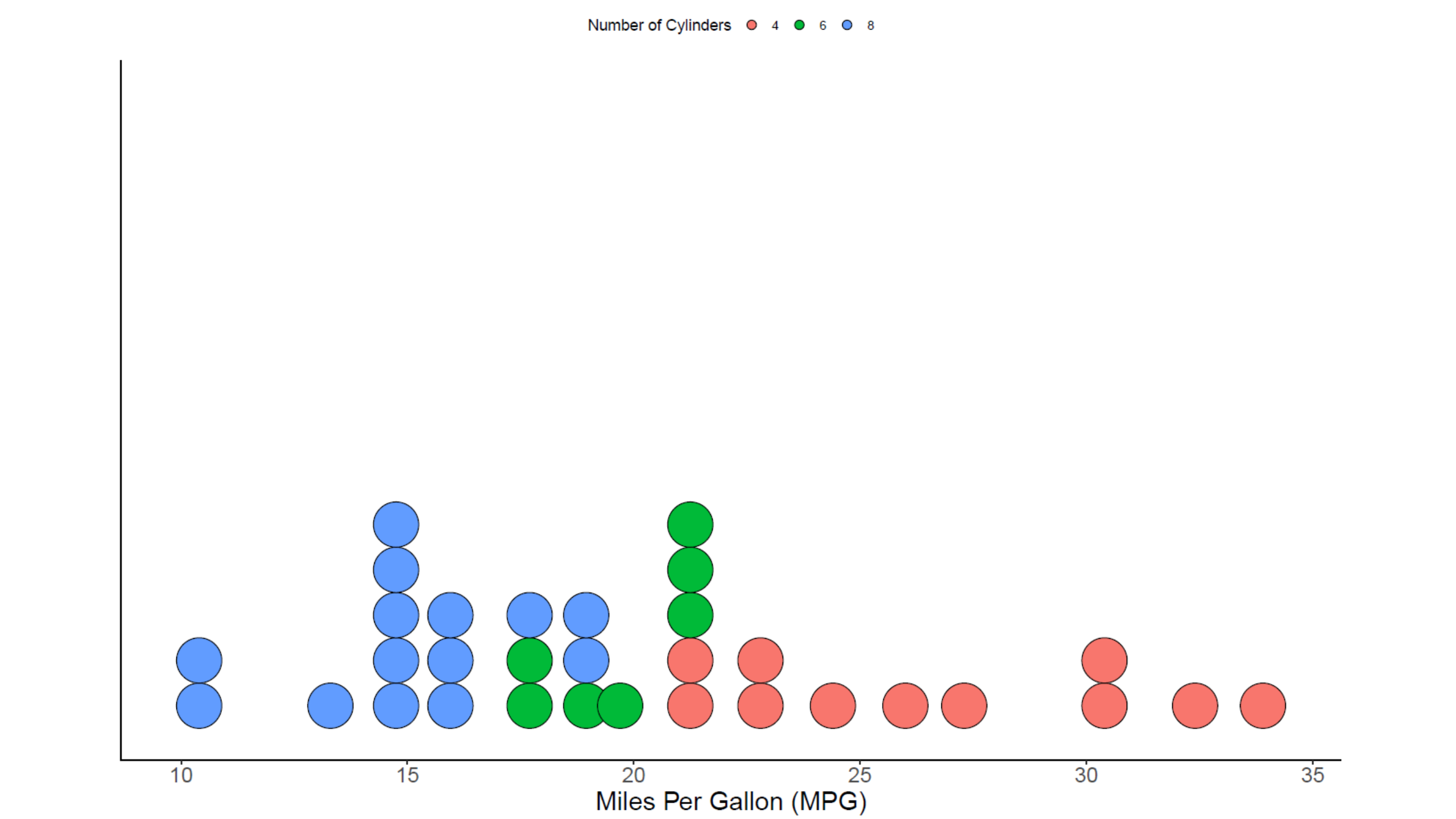
1. **Draw a horizontal line and mark the line with regular values of the variable**
2. **For each observation, place a dot above its value on the number line**
   - Works best with quantitative discrete data
   - Doesn't work well if the variable is continuous and takes on many distinct values…
   - For continuous data, the values may need to be round to the nearest tenth or integer


40
50
60
70
80
90
Waiting Time Until Eraption (Min)

Example: MPG and Engine Cylinders



Observation	MPG	Cylinders	Model
1	21.0	6	Mazda RX4
2	21.0	6	Mazda RX4 Wag
3	22.8	4	Datsun 710
4	21.4	6	Hornet 4 Drive
5	18.7	8	Hornet Sportabout
6	18.1	6	Valiant
⋮	⋮	⋮	⋮
32	21.4	4	Volvo 142E


Number of Cylinders
4
6
8
10
15
20
25
30
35
Miles Per Gallon (MPG)

Visualizing Distributions: Quantitative Variables

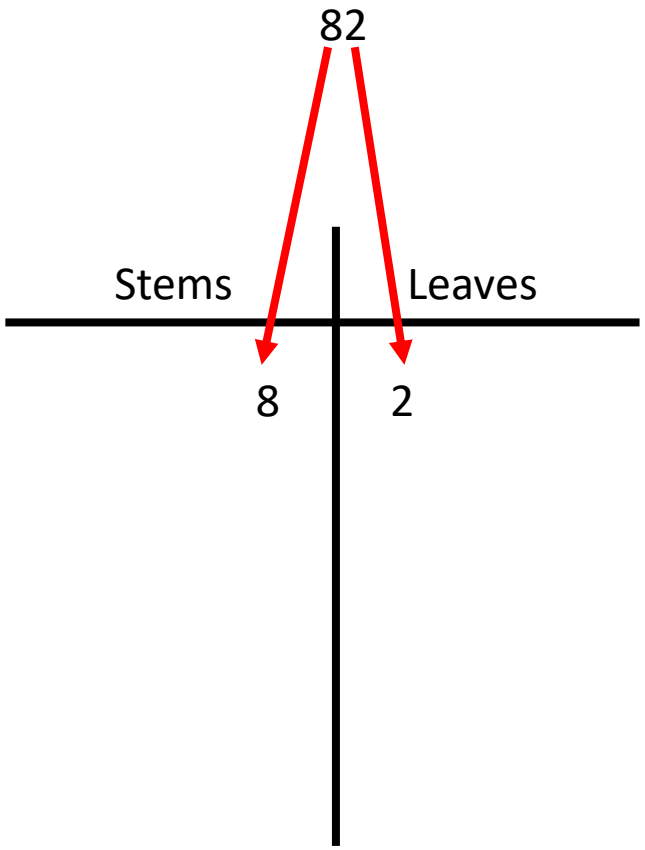
Stem and leaf plot – like a dot plot, a stem and leaf diagram also displays individual observations.

Stem – all the digits in an observation except the last digit

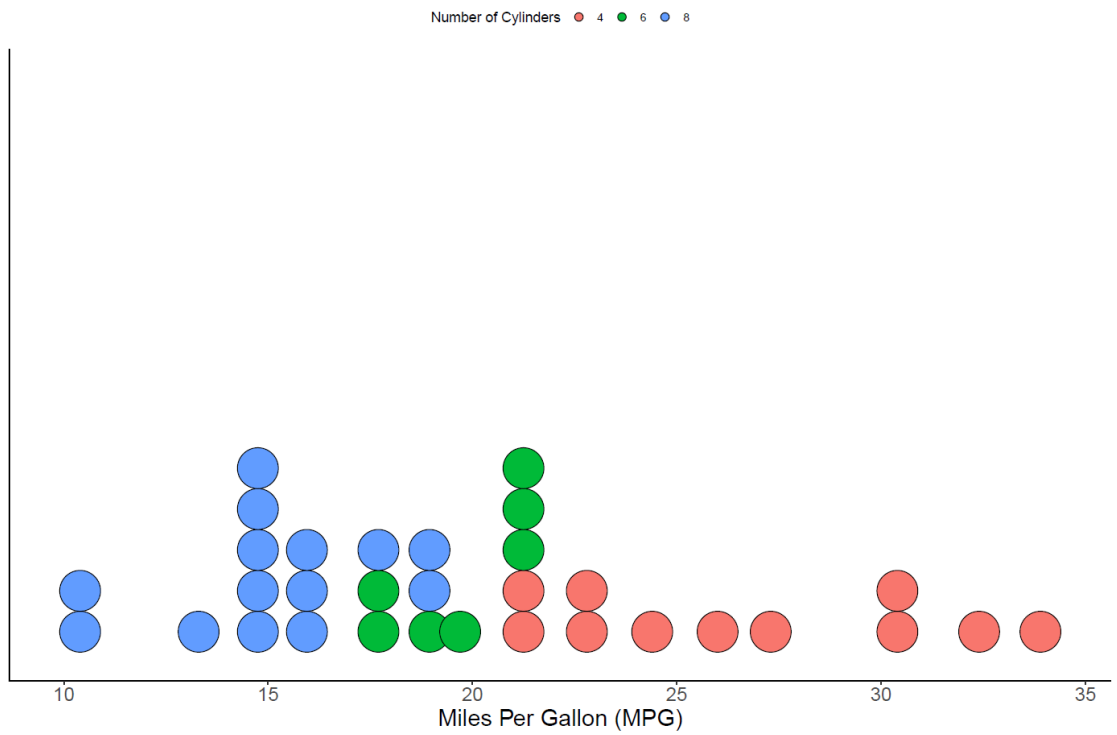
Leaf – the last digit in an observation

Steps to construct a stem and leaf plot

1. **Sort the data in order from smallest to largest.**
2. **Place the stems in a column in increasing order**
3. **Place a vertical line to the right of the stems**
4. **To the right of the vertical line, fill in the leaves that correspond with each stem in increasing order**

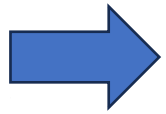


Example: MPG



Observation	MPG	Cylinders	Model
1	21.0	6	Mazda RX4
2	21.0	6	Mazda RX4 Wag
3	22.8	4	Datsun 710
4	21.4	6	Hornet 4 Drive
5	18.7	8	Hornet Sportabout
6	18.1	6	Valiant
⋮	⋮	⋮	⋮
32	21.4	4	Volvo 142E

Stems	Leaves
10	4,4
11	
12	
13	3
14	3,7
15	0,2,2,5,8
16	4
17	3,8
18	1,7
19	2,2,7
20	
21	0,0,4,4,5
22	8,8
23	
24	4
25	
26	0
27	3
28	
29	
30	4,4
31	
32	4
33	9



Stems	Leaves
10	4,4
12	3
14	3,7,0,2,2,5,8
16	4,3,8
18	1,7,2,2,7
20	0,0,4,4,5
22	8,8
24	4
26	0,3
28	
30	4,4
32	4,9

Try it out: Stem and leaf plot

Data = 4.2, 3.8, 4.6, 3.2, 2.7, 8.2, 9.1, 0.2, 1.2, 6.2