Lecture 3
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**

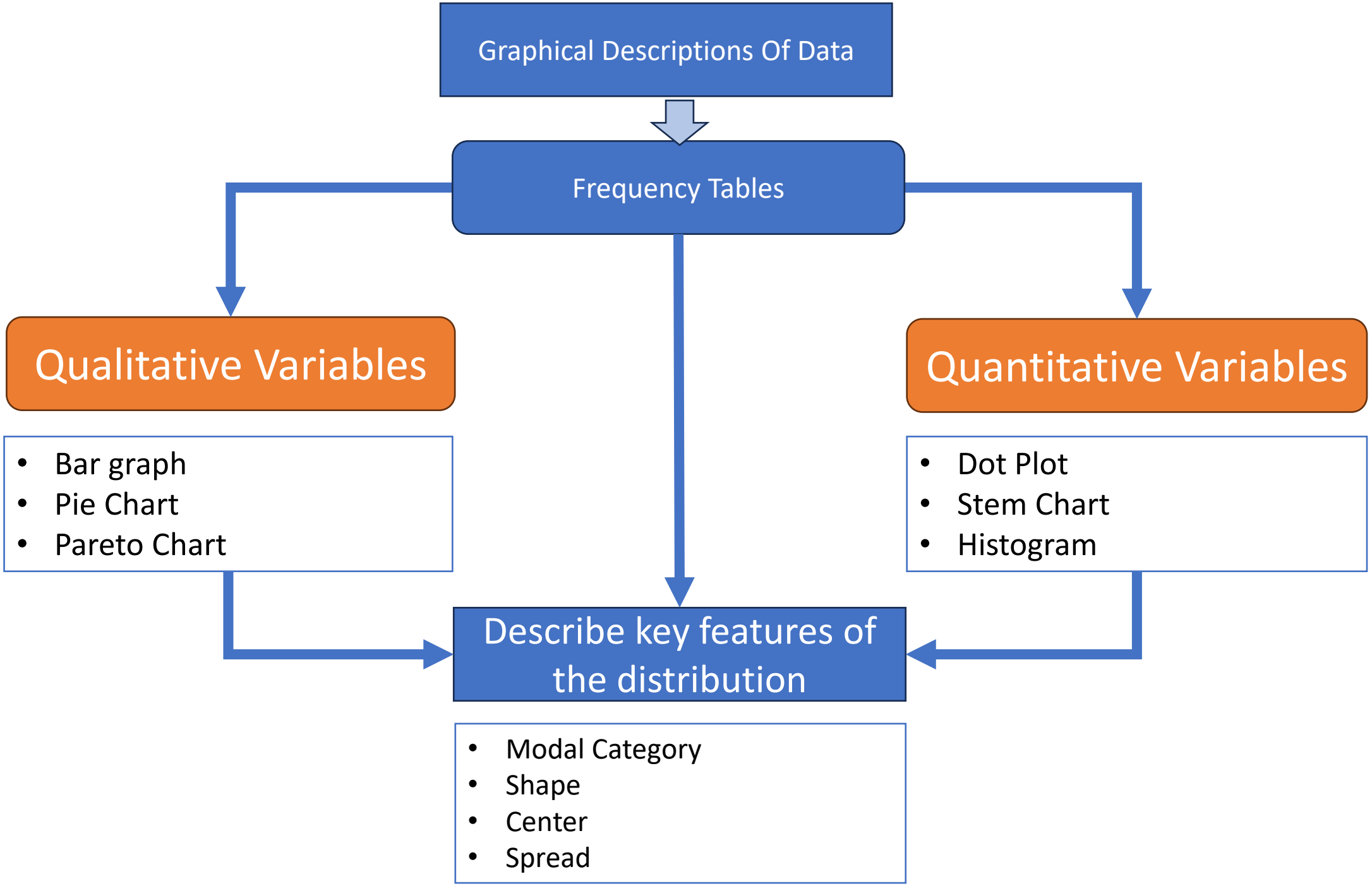
The way in which we collect our data is called the **sampling design**

Review

- A natural first step of statistical description is to look at graphical summaries of the observations for our variables

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

- A **frequency table** is a tabular descriptions of the distribution of a variable – it can be applied to either quantitative or qualitative variables


Graphical Descriptions Of Data
Frequency Tables
Qualitative Variables
Quantitative Variables
• Bar graph
• Pie Chart
• Pareto Chart
• Dot Plot
• Stem Chart
• Histogram
Describe key features of the distribution
• Modal Category
• Shape
• Center
• Spread

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

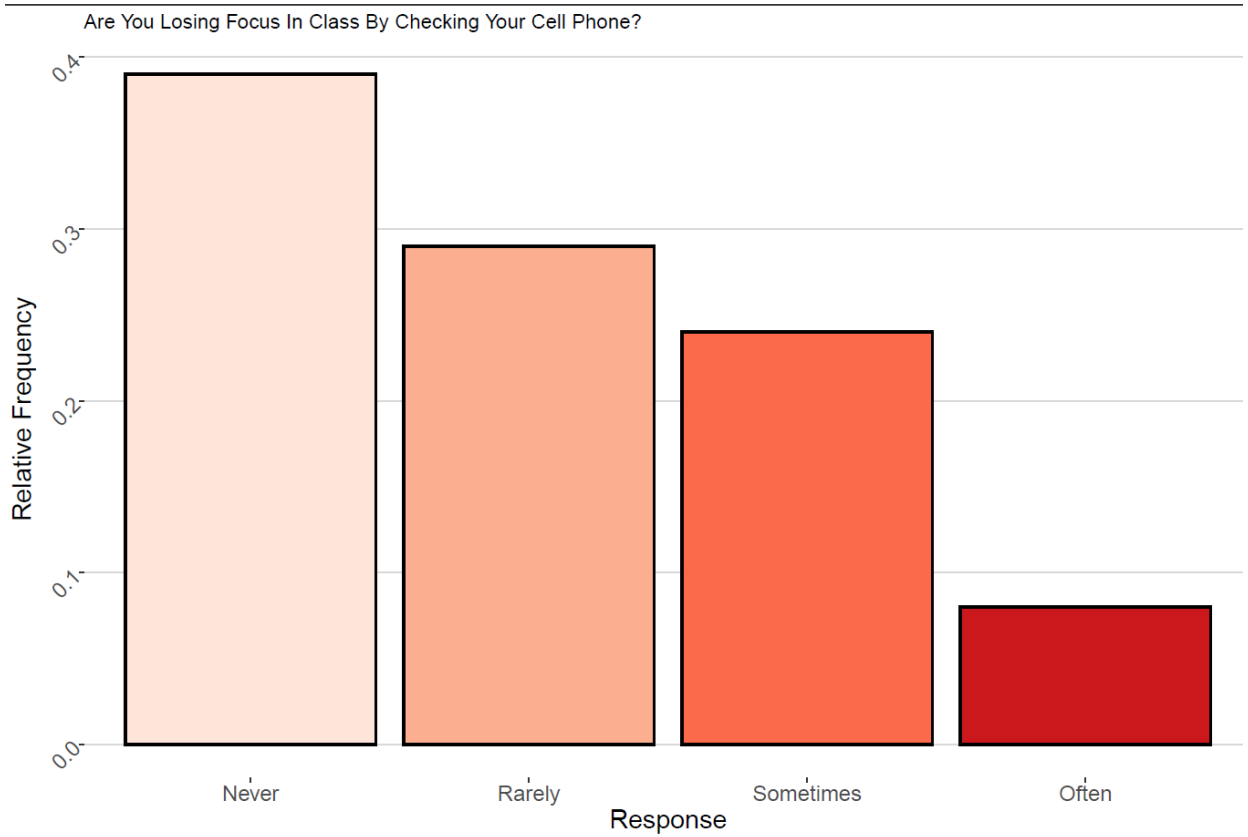
Waiting Time (Min)	Frequency	Relative Frequency	Cumulative Relative Frequency
< 50	21	0.077	0.077
50 - 60	56	0.206	0.283
60 - 70	26	0.096	0.379
70 - 80	77	0.283	0.662
80 - 90	80	0.294	0.956
> 90	12	0.044	1

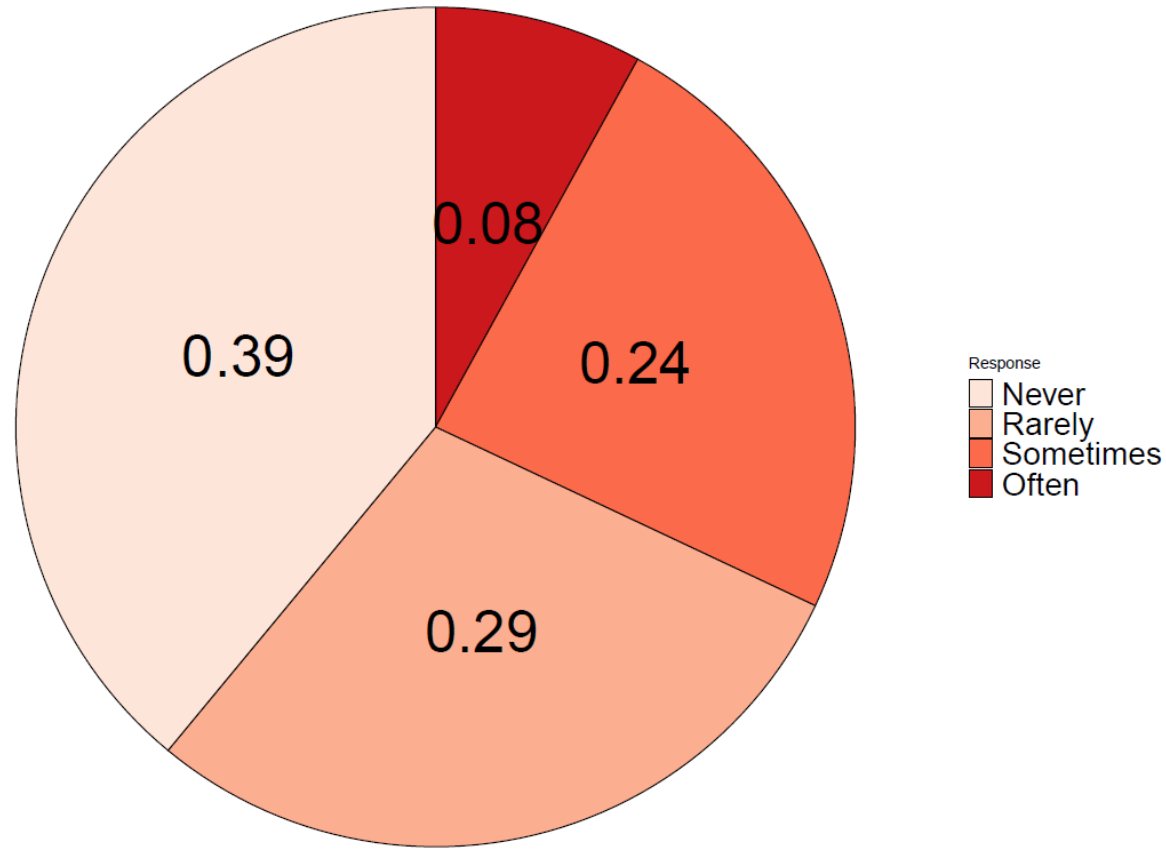
Visualizing Distributions of Categorical Data

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

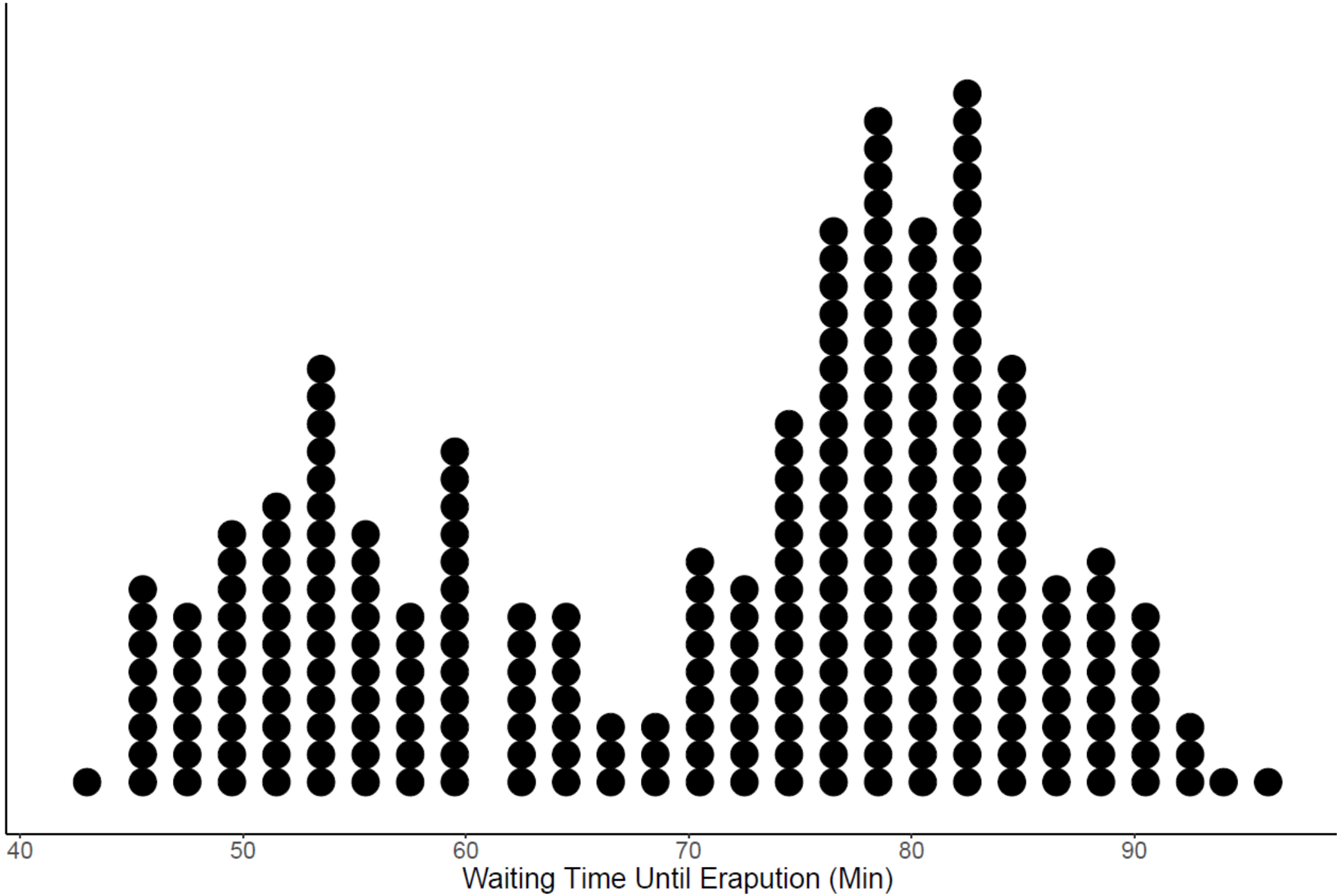

0.08
0.39
0.24
0.29
Response
Never
Rarely
Sometimes
Often

Visualizing Distributions: Quantitative Variables

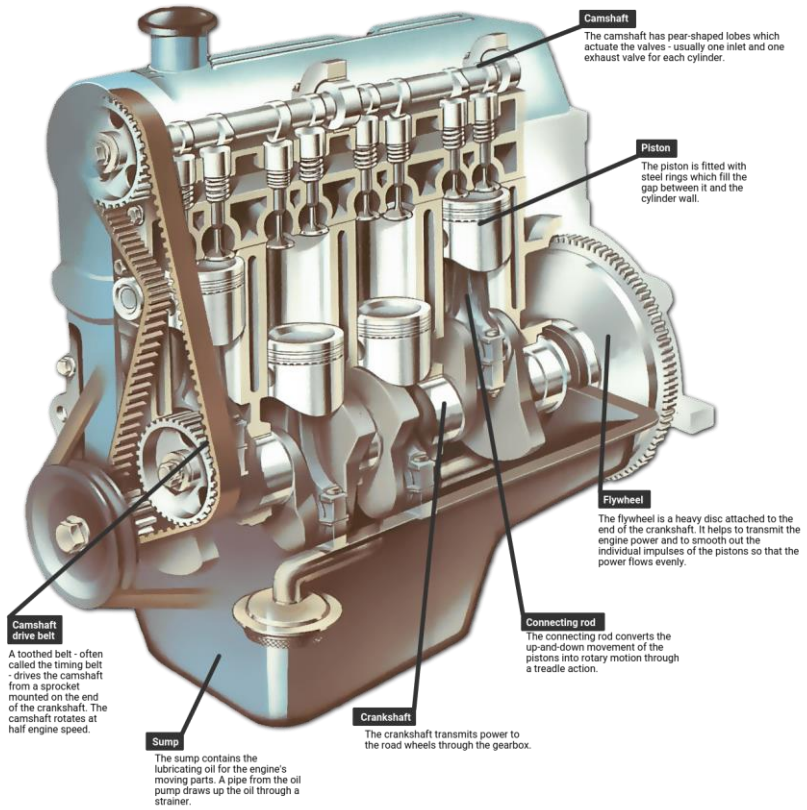
- **Dot plots –** shows a dot for each observation placed above the value for that observation

Steps to construct a dot plot

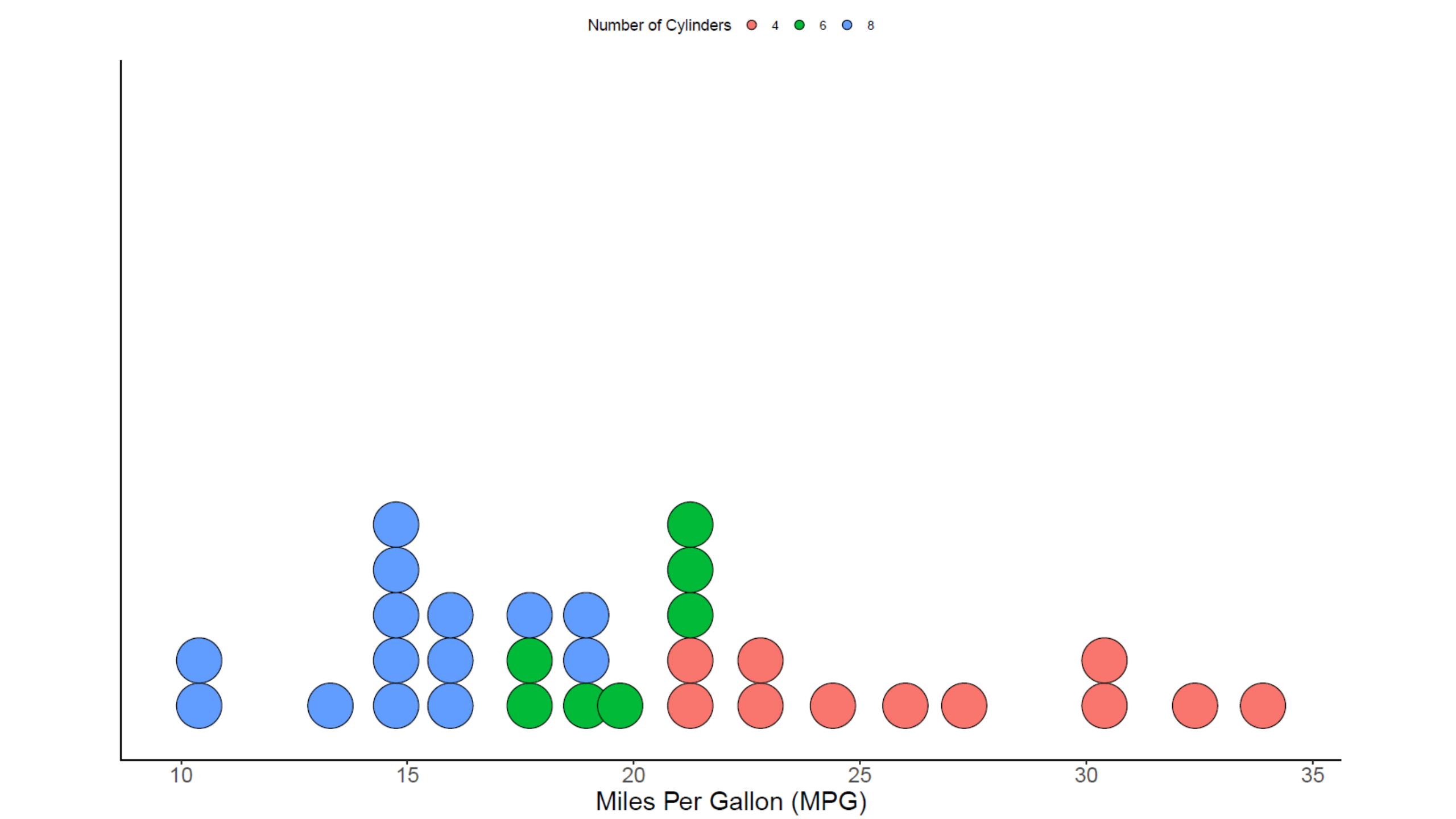
1. **Draw a horizontal line and mark the line with regular values of the variable**
2. **For each observation, place a dot above its value on the number line**
   - Works best with quantitative discrete data
   - Doesn't work well if the variable is continuous and takes on many distinct values…
   -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

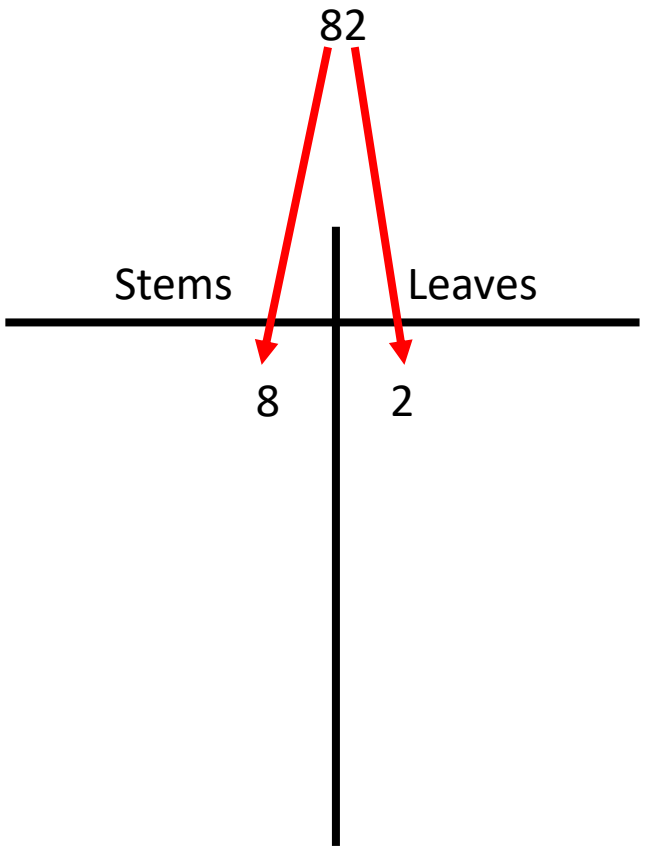
Stem and leaf plot – like a dot plot, a stem and leaf diagram also displays individual observations.

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

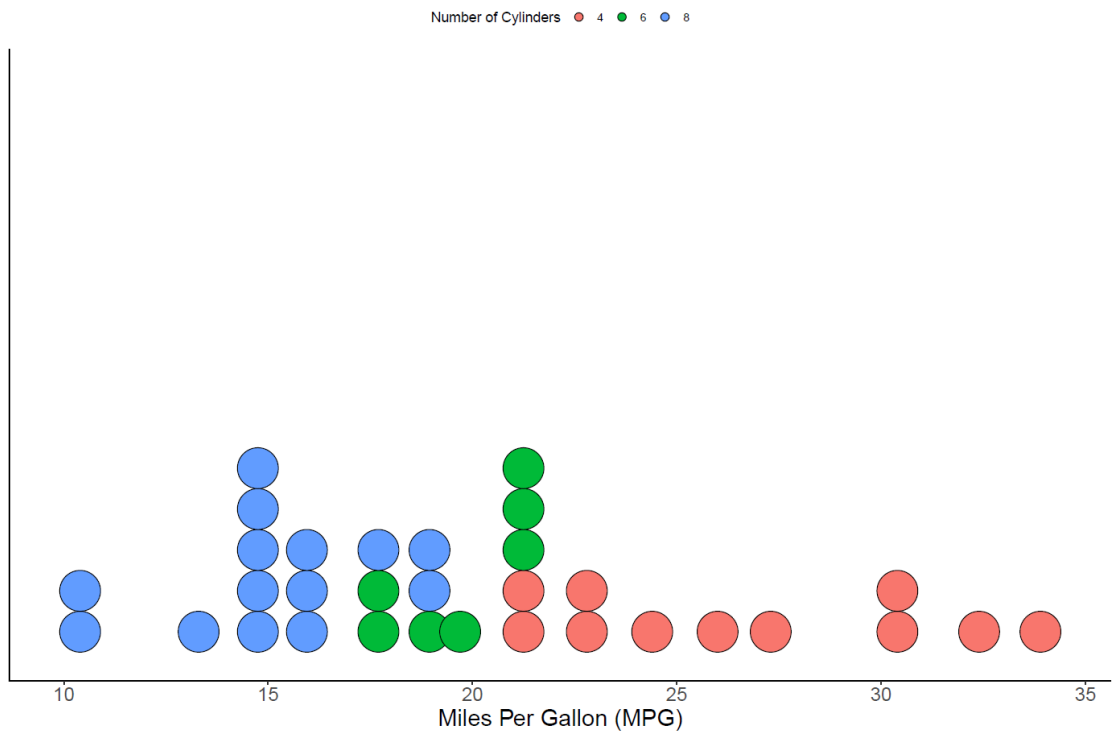
Leaf – the last digit in an observation

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**

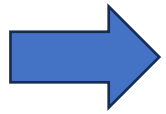


Example: MPG



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

Visualizing Distributions: Quantitative Variables

Stem and leaf plots and **dot plots** are unwieldy for large n

Histogram – uses bars to portray the frequencies or relative frequencies of the possible outcomes for a quantitative variable

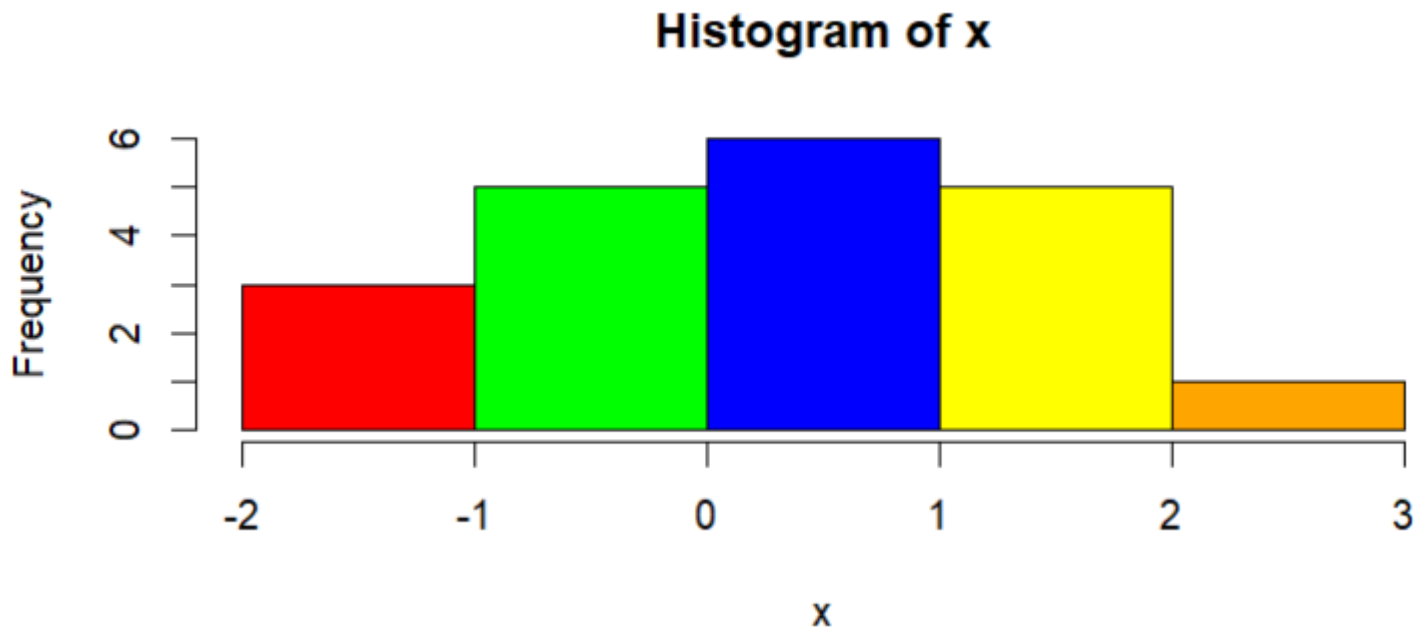
Steps to construct a histogram

1. **Divide the range of the data into intervals of equal width**
2. **Compute the frequency of each interval (i.e construct the frequency table)**
3. **Label the x-axis with the values or endpoints of each interval.**
4. **Draw a bar over each value or interval with height equal to its frequency or relative frequency**

Try it out: Histogram

Consider the following $n = 20$ observations of a continuous variable

Bin 1	Bin 2	Bin 3	Bin 4	Bin 5
x = -1.5, -1.2, -1.0,	-0.8, -0.7, -0.6, -0.1, -0.1,	0.1, 0.1, 0.1, 0.6, 0.6, 0.8,	1.1, 1.2, 1.3, 1.8, 1.9,	2.4



How to choose the number of Bins?

- How to choose the best number of bins is not a straightforward question and there is a lot of literature on the subject

- We can construct our histogram using a specific binwidth w or under a set number of bins k
- $w = \frac{\max x - \min x}{k}$ or $k = \frac{\max x - \min x}{w}$

- Square root method: $k = \text{round}(\sqrt{n})$ (A fairly safe and basic rule of thumb)

- Sturges Rule[1]: $k = \text{round}(\log_2 n) + 1$ (not great for $n < 30$)

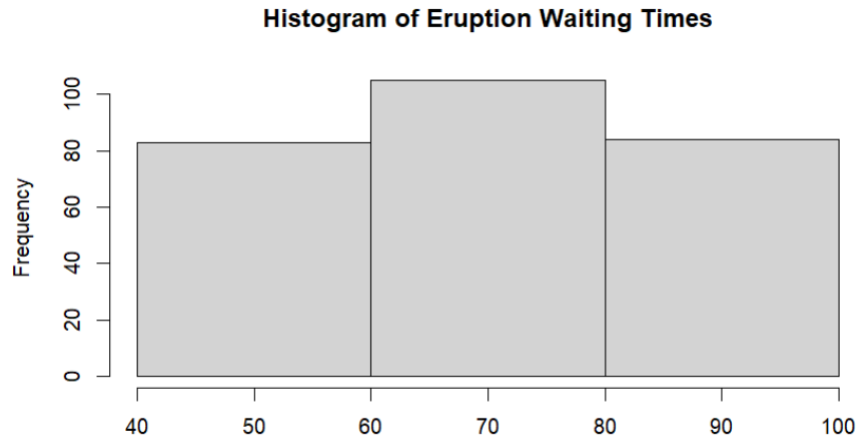
- Rices Rule[2]: $k = 2\sqrt[3]{n}$

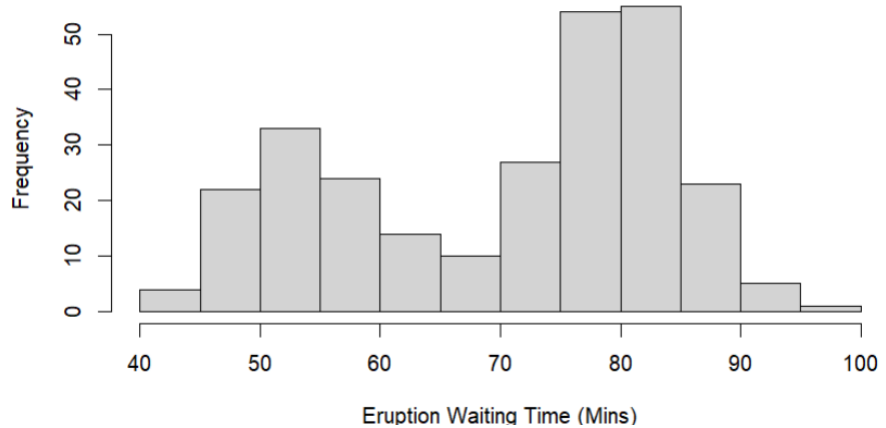
[1] Sturges, Herbert A. "The choice of a class interval." Journal of the american statistical association 21.153 (1926): 65-66.

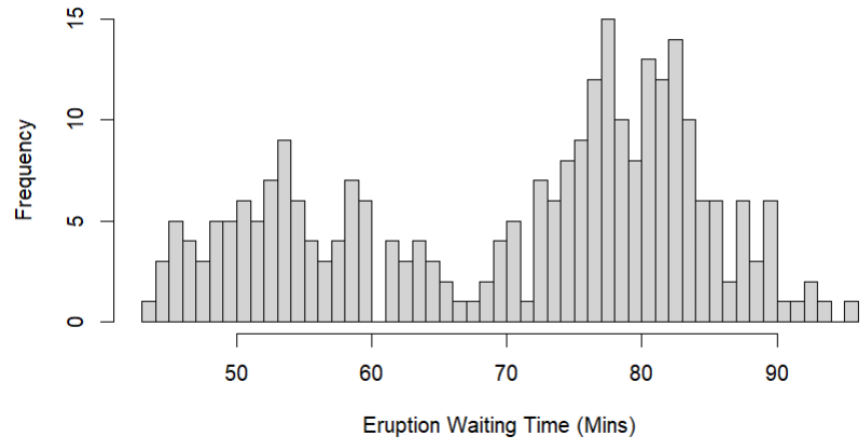
[2] Lane, David. *Online statistics education: A multimedia course of study*. Association for the Advancement of Computing in Education (AACE), 2003. – Chapter 2 “Graphing Distributions

Some tips

- If too few intervals are used, then the graph will be too crude
- If too many intervals are used, graph will contain many short bars and gaps.
  Usually between 5 - 15 intervals are enough.
- Most plotting software will automatically choose the number of bins.
- **ALWAYS** plot the histogram to get an idea about the shape of the distribution of a quantitative variable
- Is the number of observations is small (say n < 50) then it's a good idea to supplement a histogram with a dot plot or stem plot







Example: Old Faithful Eruption Times

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