Visualizing Data

Grinnell College

January 26, 2026

Goals for Class Today

We are going to learn how to do the following today:

1. Common graphics
    - Univariate Graphs (one variable graphs)
        - Bar chart (categorical)
        - Histogram (numeric)
    - Multivariate Graphs (two or more graphs)
        - Bar charts
        - Scatterplots
2. Critique of graphics
3. Describe the **distribution** of a variable
4. Make graphs that describe the relationship between 2 or more variables

Motivation: Tips Data

First 20 observations of the tips given to a waiter over the course of several months in a restaurant. Do more customers come to the restaurant on certain days?

Total Bill	Tip	Sex	Smoker	Day	Time	Size
13.42	1.58	Male	Yes	Fri	Lunch	2
16.27	2.50	Female	Yes	Fri	Lunch	2
10.09	2.00	Female	Yes	Fri	Lunch	2
20.45	3.00	Male	No	Sat	Dinner	4
13.28	2.72	Male	No	Sat	Dinner	2
22.12	2.88	Female	Yes	Sat	Dinner	2
24.01	2.00	Male	Yes	Sat	Dinner	4
15.69	3.00	Male	Yes	Sat	Dinner	3
11.61	3.39	Male	No	Sat	Dinner	2
10.77	1.47	Male	No	Sat	Dinner	2
15.53	3.00	Male	Yes	Sat	Dinner	2
10.07	1.25	Male	No	Sat	Dinner	2
12.60	1.00	Male	Yes	Sat	Dinner	2
32.83	1.17	Male	Yes	Sat	Dinner	2
35.83	4.67	Female	No	Sat	Dinner	3
29.03	5.92	Male	No	Sat	Dinner	3
27.18	2.00	Female	Yes	Sat	Dinner	2
22.67	2.00	Male	Yes	Sat	Dinner	2
17.82	1.75	Male	No	Sat	Dinner	2
18.78	3.00	Female	No	Thur	Dinner	2

We need something more than a data file to answer that

Claim

Most statistical analysis can be done graphically; the math is high-end filler

- If there is a relationship, it's usually visual
- If there isn't a relationship, it's usually visual
- Trick is finding the right graphics
- If I don't see the data somewhere in the paper I'm suspicious
  - And yes, it's easy to lie with graphics
- I'm basing this off of my time doing consulting work during grad school

Data Visualization

Why do we graph data?

- It (hopefully) allows us to interpret data...
  - quickly and
  - easily

Which graph is chosen is driven by

- the type of data
- the number of variables and
- what we are trying to convey (context)

Context

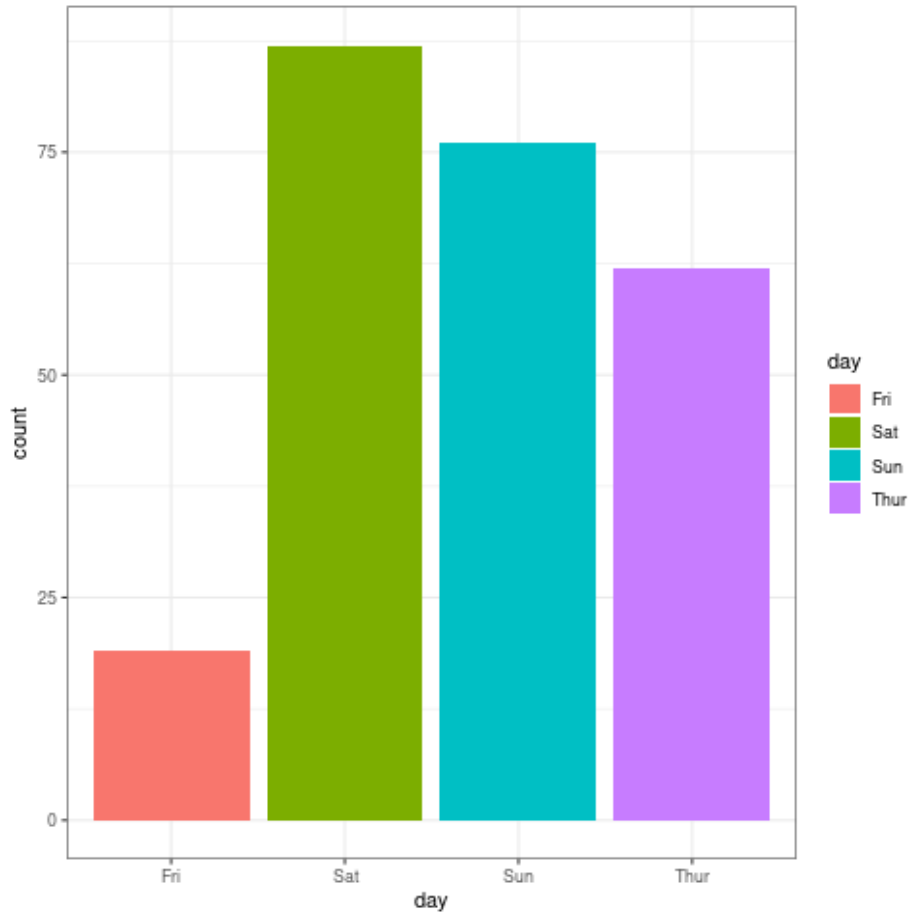
Context gives us the goal of what we want to convey from our data

- Eg If we are interested in the fastest 100m dashes in the Olympics over time we could plot the shortest time for each year
  - three numeric variables (all run times, shortest run times, and years)
  - The context lets us ignore "all run times"
  - More parsimonious graphs

- Often, for a single variable, we are interested in the **distribution**
  - The distribution of a variable is the frequency certain values occur
  - Eg the distribution of of sizes of penguin colonies
  - Heavily tied to probability

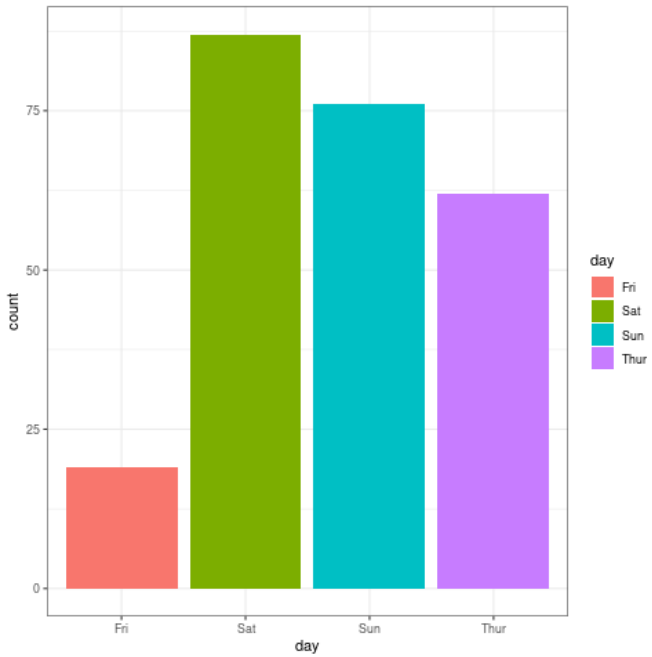
One Categorical Variable: Graph

When we have one categorical variable, a **bar chart** is often used to tally the frequencies (counts) of that categorical variable



One Categorical Variable: Distribution

To describe the distribution of a categorical variable we need to talk about....



- how likely each category is
- the most and least likely category
- any interesting patterns
  - WARNING: The order of a bar chart is sometimes meaningless so it can't be read left to right
- numbers help

When should we use pie charts?

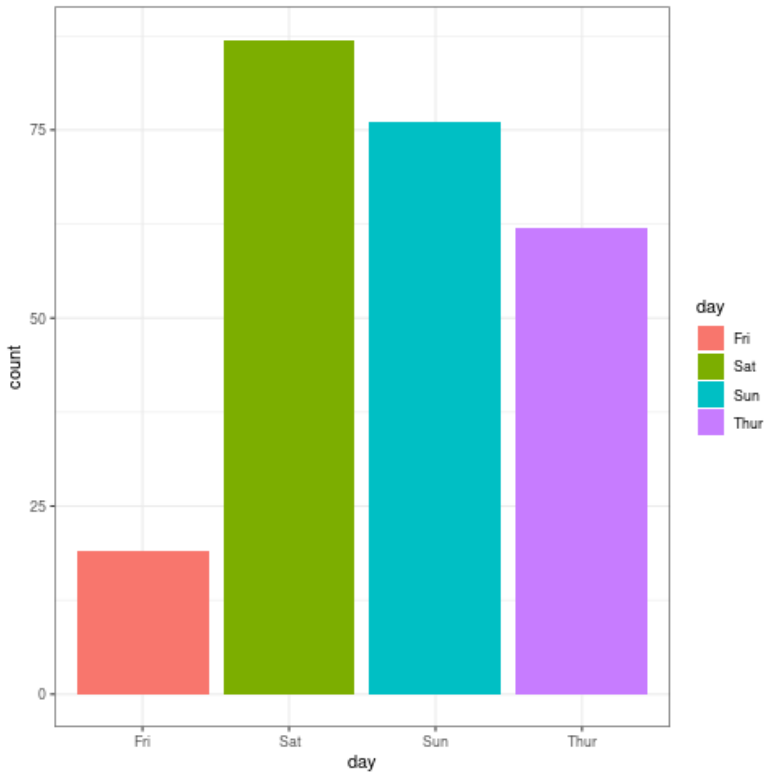
When do we use pie charts?

.....never?

- Humans are bad are reading angles/pie slices
- Bar charts can convey the same info, easier
- Dimensions of the graph are weird (bar chart with polar coordinates?)
- No strong advantage to pie charts in my opinion

Counterpoint: Use a table

As a counterpoint to my emphasis on graphics the bar chart (with few categories) has such a poor data-ink ratio that a small table would suffice



- **data-ink ratio**: The relevant amounts of ink/pixels used to convey information vs the amount of ink/pixels used (roughly)
- High data ink ratio is parsimonious and spartan (in a good way)
- Low data ink ratio takes up space and doesn't add much

Counterpoint: Use a table

As a counterpoint to my emphasis on graphics the bar chart (with few categories) has such a poor data-ink ratio that a small table would suffice

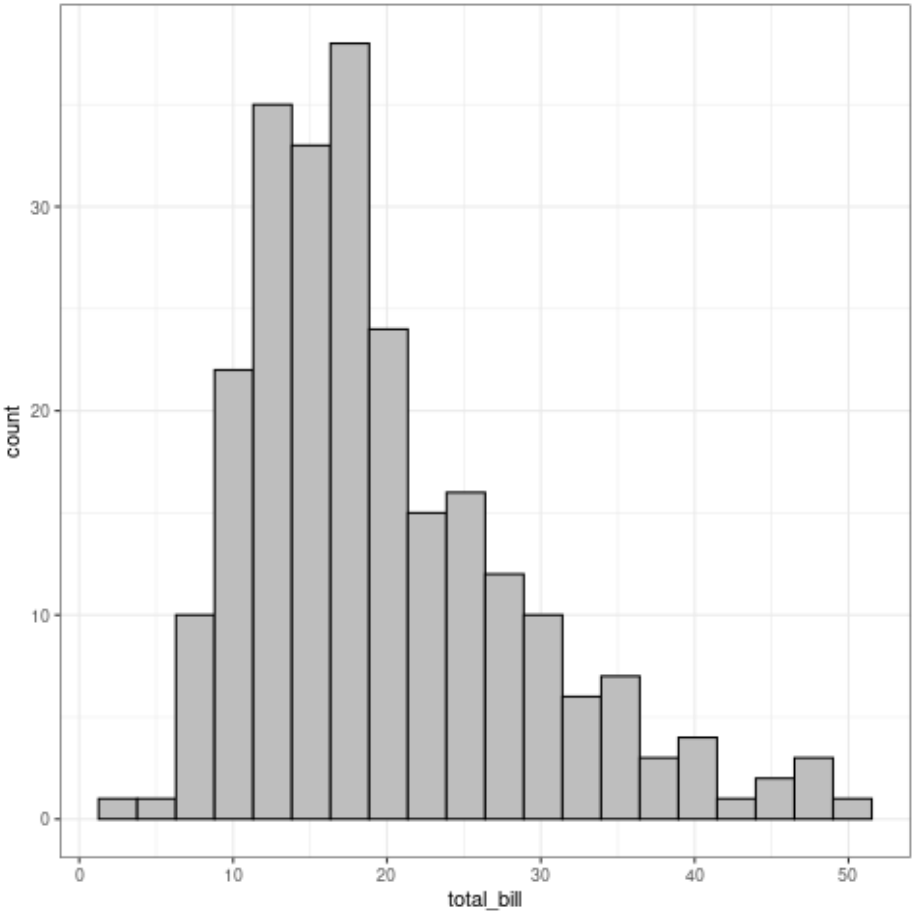
- **data-ink ratio**: The relevant amounts of ink/pixels used to convey information vs the amount of ink/pixels used (roughly)
- High data ink ratio is parsimonious and spartan (in a good way)
- Low data ink ratio takes up space and doesn't add much

Fri	Sat	Sun	Thur
19	87	76	62

Table: Total Days Worked

One Numeric Variable: Graph

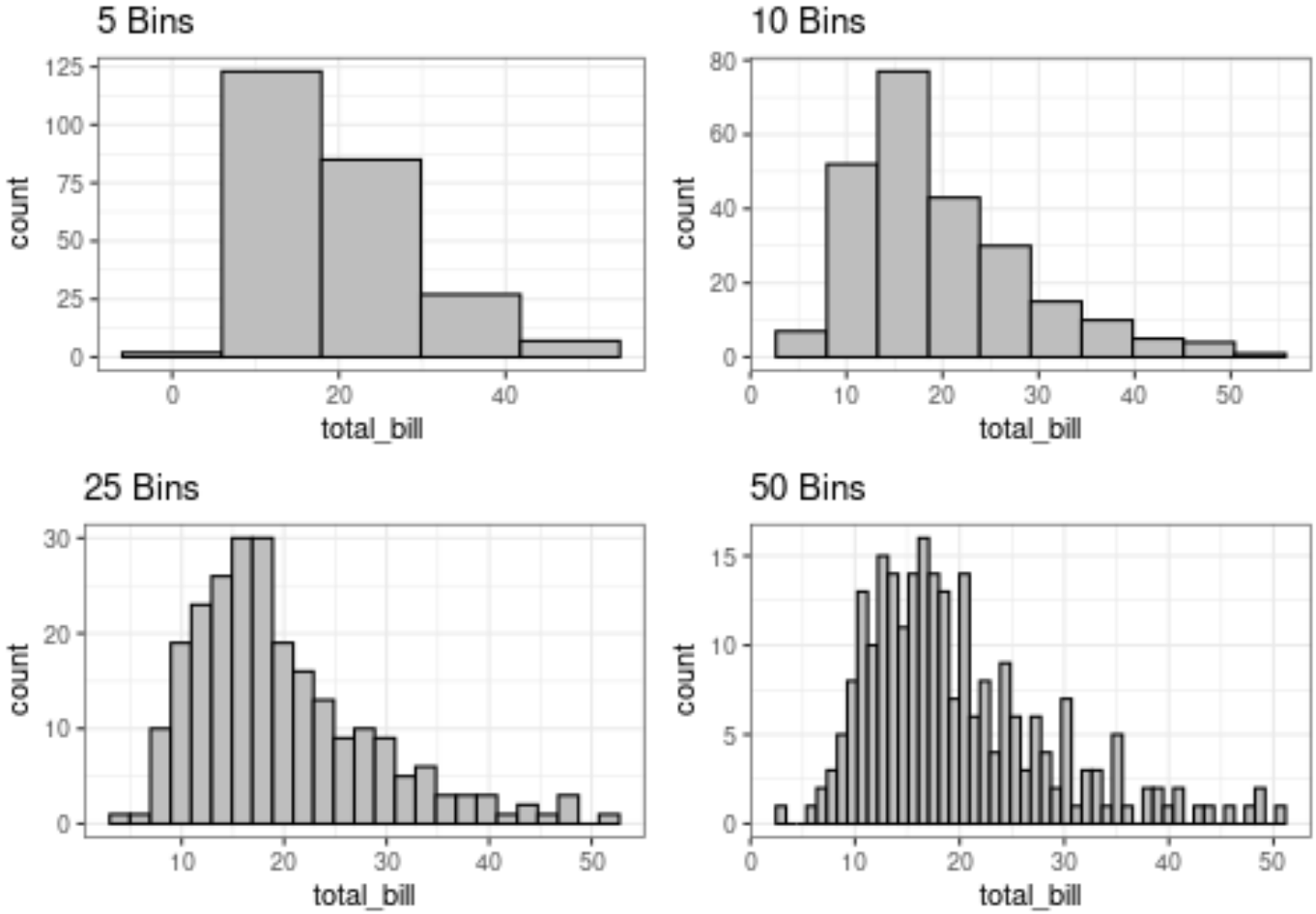
For a numeric variable, we use **histogram** to show the distribution.



- Groups data into equally spaced intervals/bins
- Each bin has either the count/frequence displayed or the proportion/percent displayed
- Why favor one over the other?

Histogram Bin Width

Using wider/narrower bin width can drastically change the histogram
- too wide: can't tell exactly where data points are
- too narrow: overly detailed and hard to read

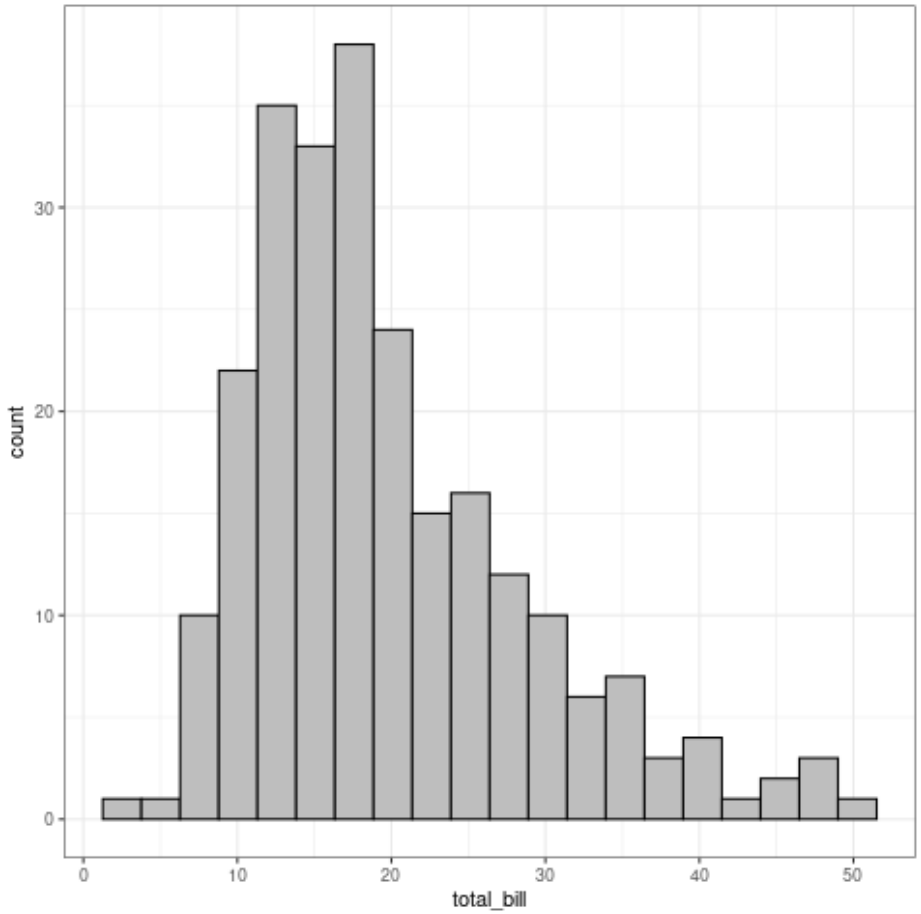


Grey Area: How many bins should be used?

- Several formulaic choices.....
  - run ?hist and follow the R documentation to find the three ways it (natively) calculates the breaks (bins) and their references
- It's impossible to say with any certainty
  - Sometimes graphs need a lot of little bins to see "wave" like patterns
  - Other times the forest gets lost in the trees with a lot of little bins
- My very rough rule: More than 10, less than 50

NOTE: It should NOT matter how many bins you make because we shouldn't be coming to different conclusions based on an arbitrary (??) decision....that indicates big problems. So don't get caught up picking exact bin sizes.

One Quantitative Variable: Distribution



Here, there is quite a bit more we can examine:

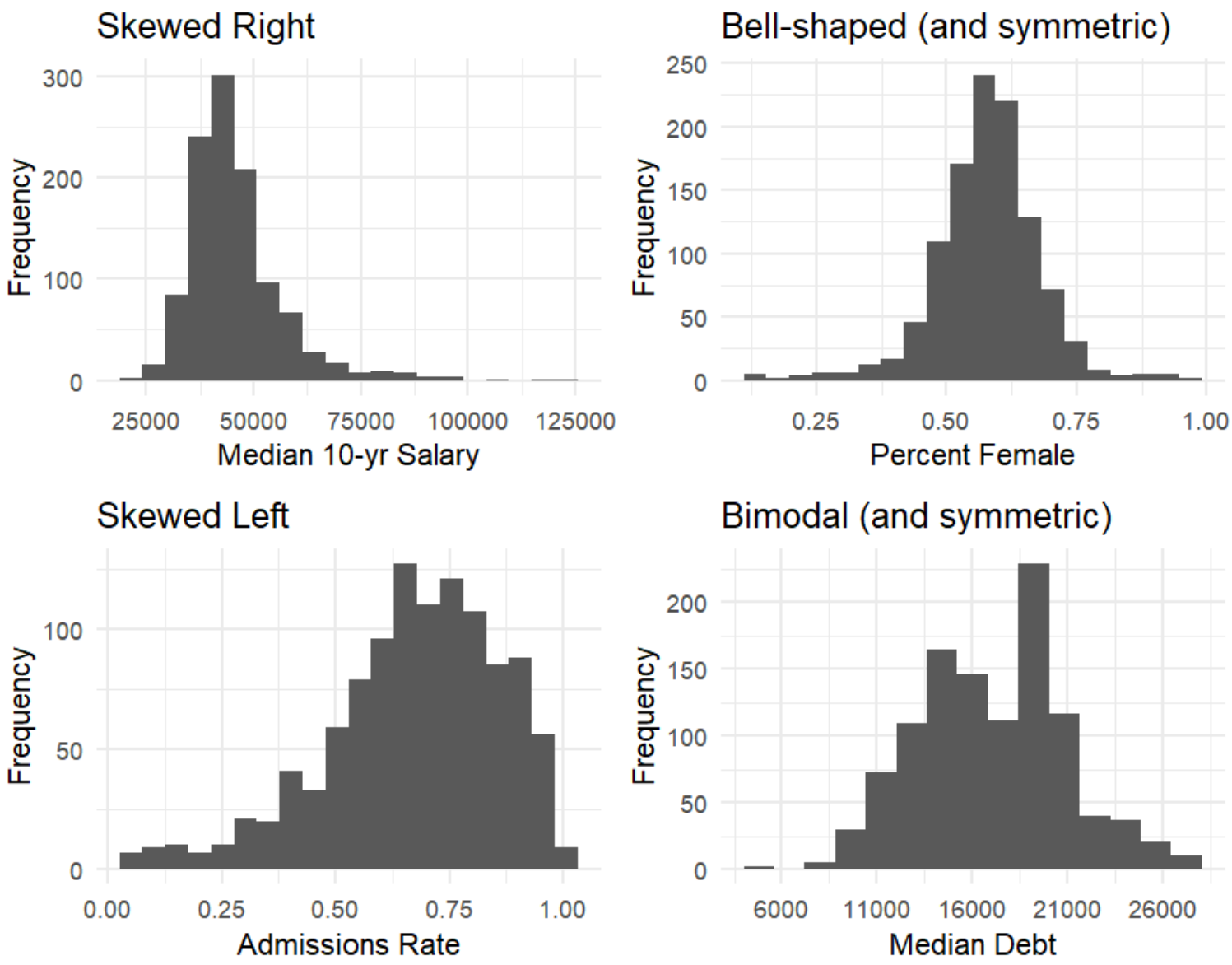
- Where does the "center" appear to be?
- How spread out is this data?
- What about the range of this data?
- Does it appear skewed (more data on one side?)

One Quantitative Variable - Distribution

We need to mention ALL of the following things:

- **Shape** - is the distribution
  - symmetric
  - skewed (mention direction!)
  - uniform (ie flat)
  - bell-shaped (ie one hump + symmetric))
  - bimodal (ie two humps)
- **Center** - where does the data bunch up (eg. mean or median)
- **Spread** - how spread out is the data (ie: range of values)
- **Outliers** - are there values that are much smaller/larger than the rest? Do they not follow the pattern?
  - Even when there isn't outliers, we mention there isn't outliers

Distribution - Shape



Distribution - Center and Spread

- **Center**: typically we use means or medians
- **Spread**: typically we use standard deviation, range, or IQR

We will talk more about how to decide which thing to use for both center and spread in a few days (and how to calculate each)

Outliers

Outliers are data points that look *unusual* in that they either don't follow a pattern that we see in the data or are far away from other points

- If everything is an outlier, nothing is an outlier
- Using histograms, we look for gaps in the bins to identify outliers
  - How does this tie into the note on bin sizes?
- A group of outliers can indicate a **subpopulation**
  - A subpopulation is a subset of the population, usually that share some characteristic (eg Student athletes are a subpopulation of the Grinnell student body)
- A fuzzy point is if the point is far away from other data points but follows the general pattern
  - This will come up during linear regression

Bivariate Graphs

Switiching gears from one variable to two variables but first why?

Association

It is very common for us to try to find a relationship between two (or more) variables

- When there seems to be some connection between two variables (knowing about one variable tells us about the other), we say they are **associated**.
- If there does not seem to be a relationship between the variables, we say they are **independent**.
- Occasionally talk about explanatory variables and response variables (misleading terms!!)

Bivariate Graphs

In your lab groups/neighbors discuss strategies you'd use to plot plot two categorical variables that you are interested in. Eg...

Political Party?	Greenland?
R	Yes
D	No
Ind	Unsure
D	No

Bivariate Graphs

In your lab groups/neighbors discuss strategies you'd use to plot plot two categorical variables that you are interested in. Eg...

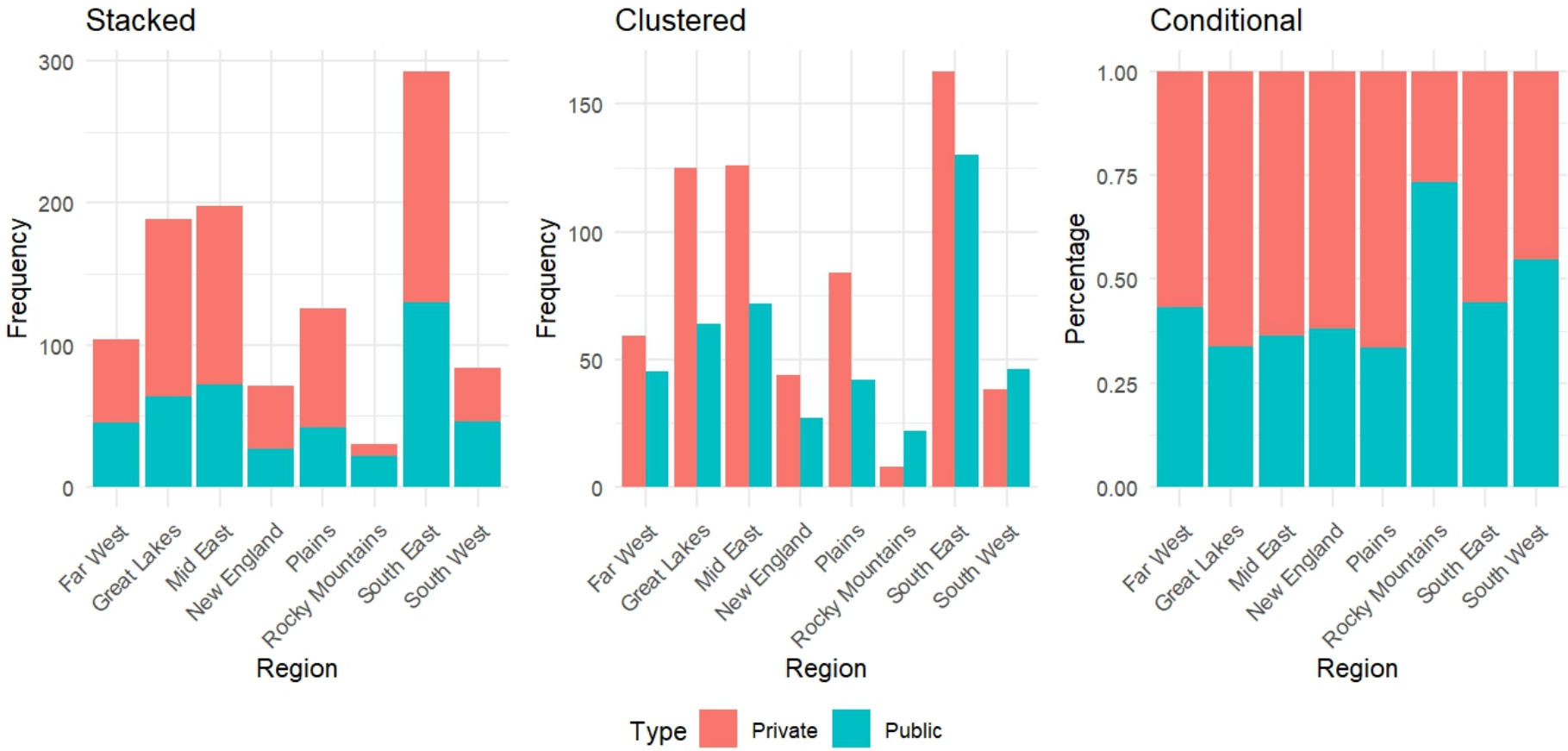
Political Party?	Greenland?
R	Yes
D	No
Ind	Unsure
D	No

How would your answer change if our goal was...

1. Total count of yes's and no's, with party affiliation noted
2. Proportion of yes's within each party
3. Count of yes's in each political party x Greenland combination

Bivariate Bar Charts

Back to the college data. Are the variables "Region" and "Type" associated? Which bar chart is most helpful?



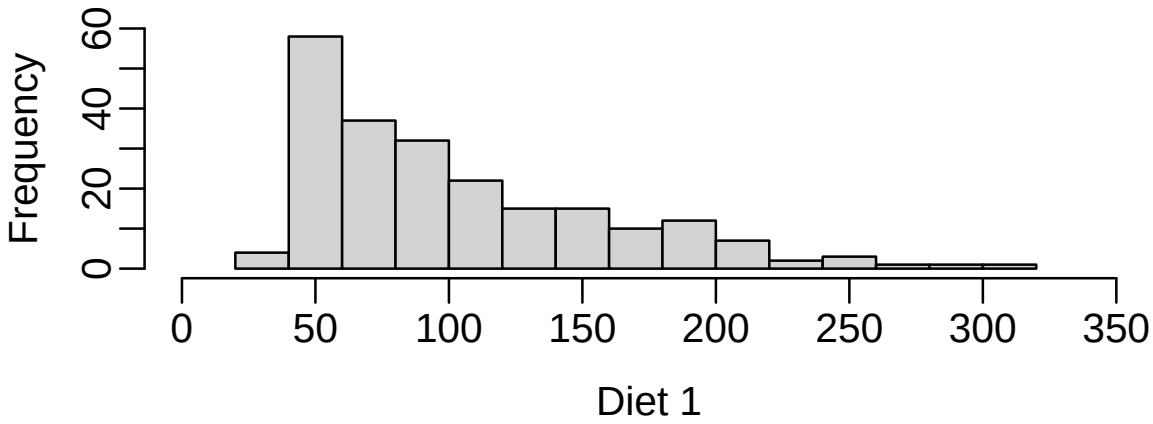
Quick Summary

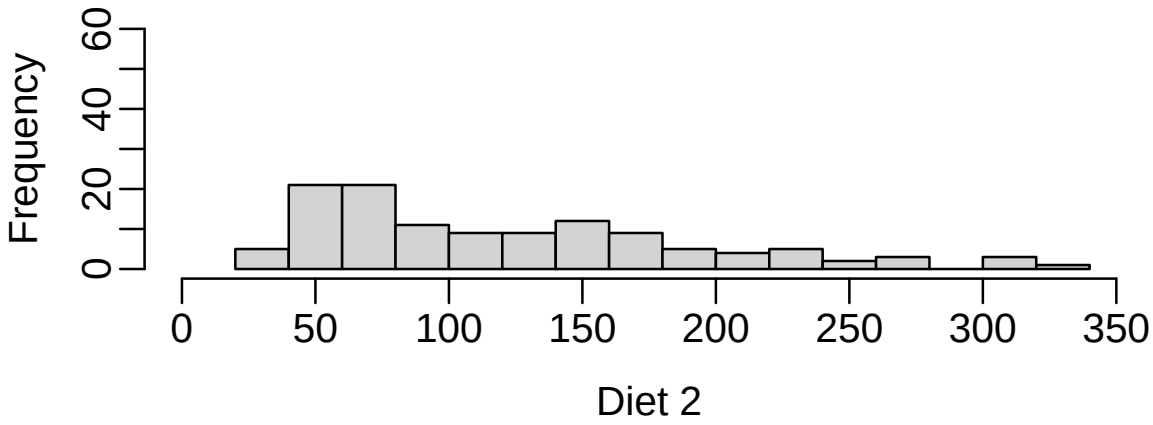
Advantages and Disadvantages to all three

- Stacked bar charts
  - Show the original variable's distribution easily
  - Hard to compare groups within a stack
  - Hard see the distribution for subpopulations
- Clustered (dodged) bar charts
  - Hard to see original distribution
  - Easy to compare groups (via counts)
  - Easy to see distributions for subpopulations
- Conditional (filled) bar chart
  - Impossible to see the original distribution
  - Easy to compare groups (via proportions)
  - Impossible to know subpopulations distributions

Alternative

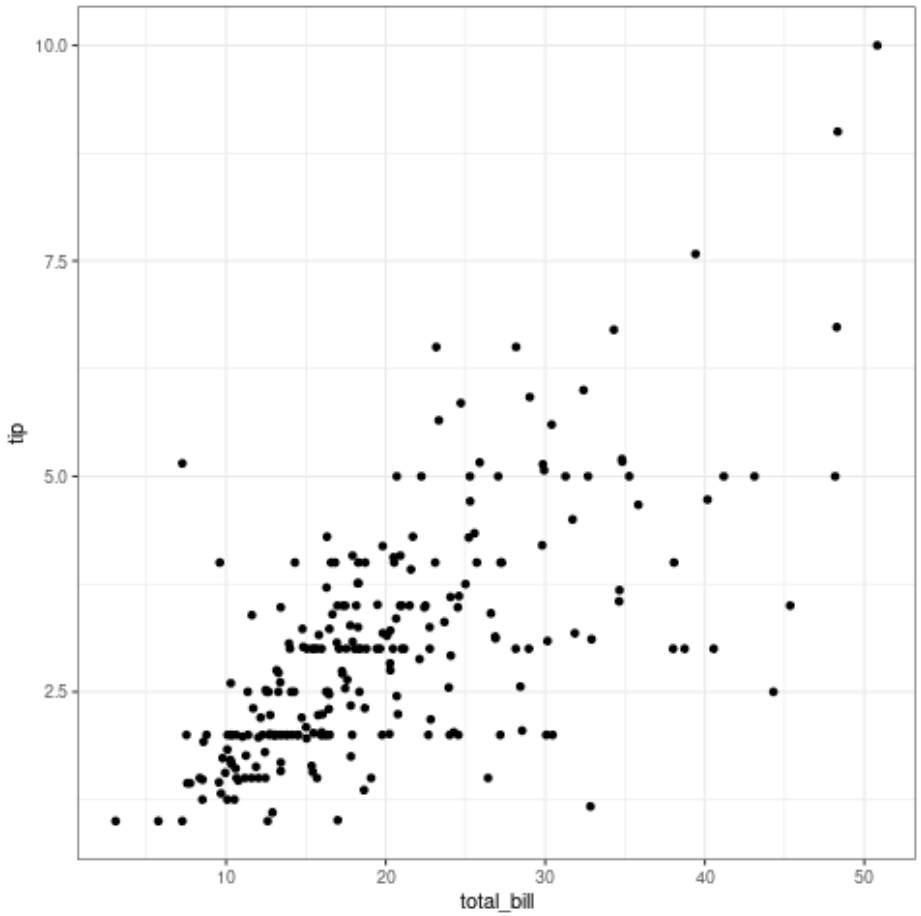
Can also produce two different histograms (KEEP AXIS THE SAME!!)





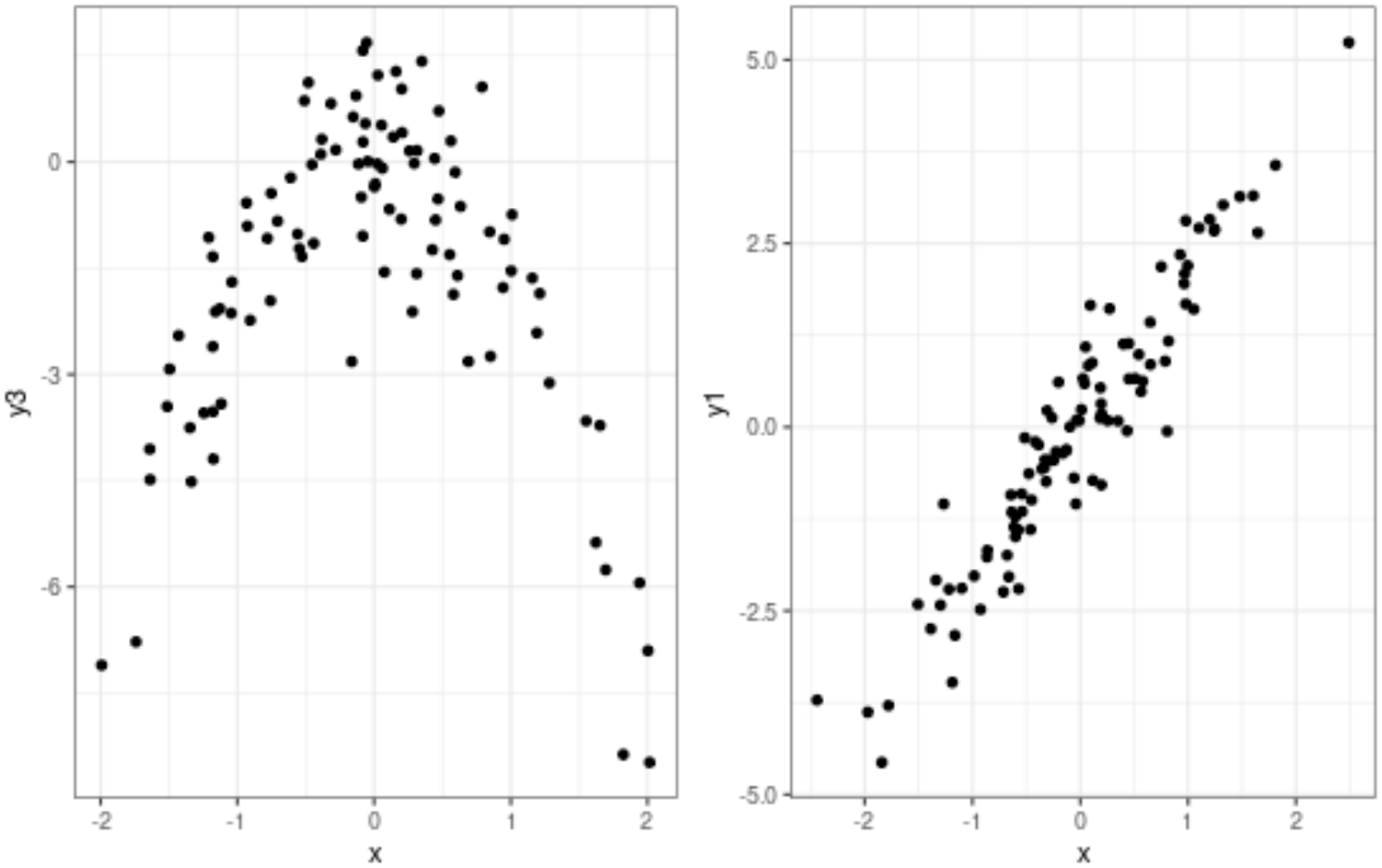
Quantitative + Quantitative → Scatterplots

Visual summaries investigating the relationship between two quantitative variables are often presented with a **scatterplot**

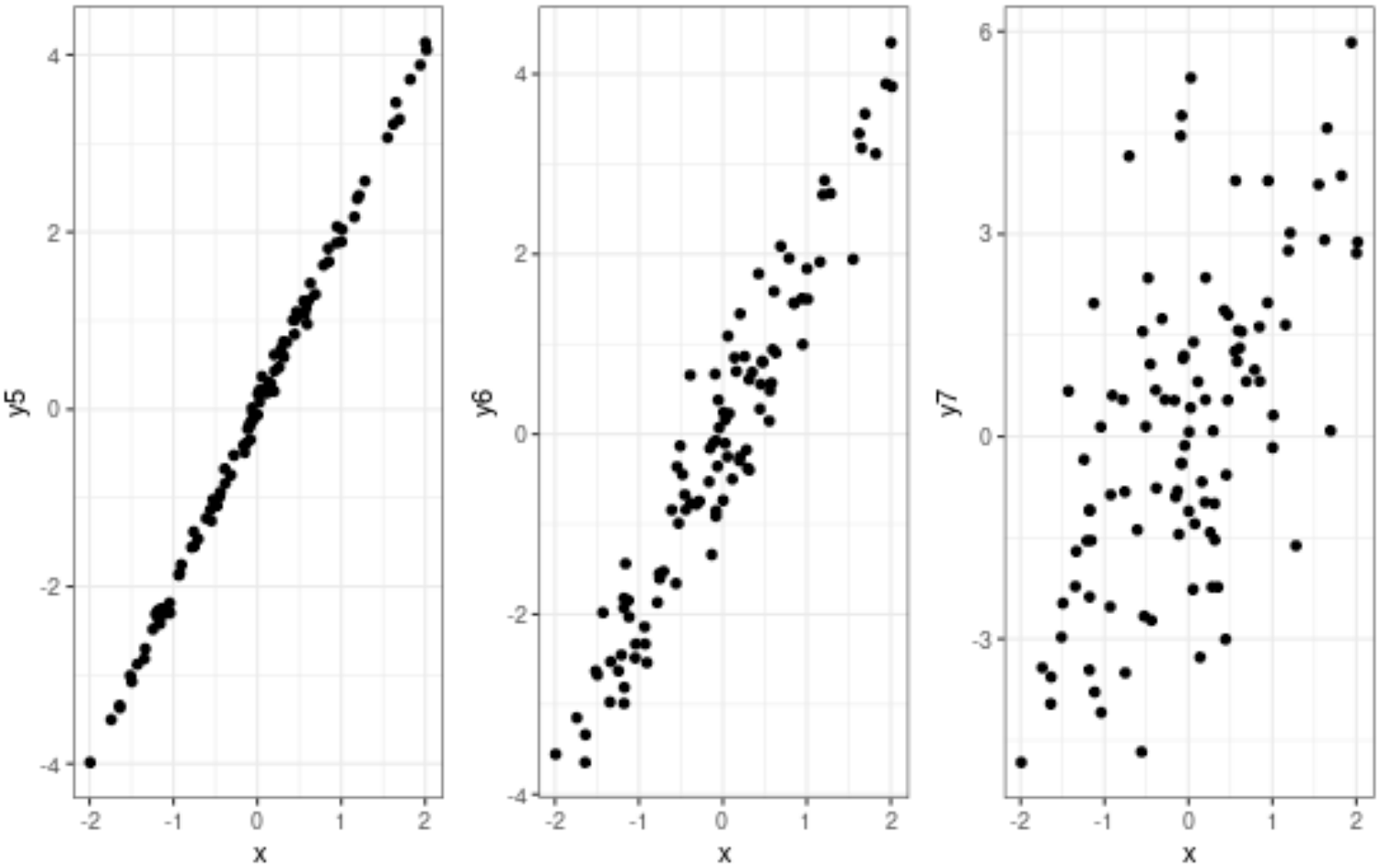


What kind of relationship do we see between the total bill and the tip amount?

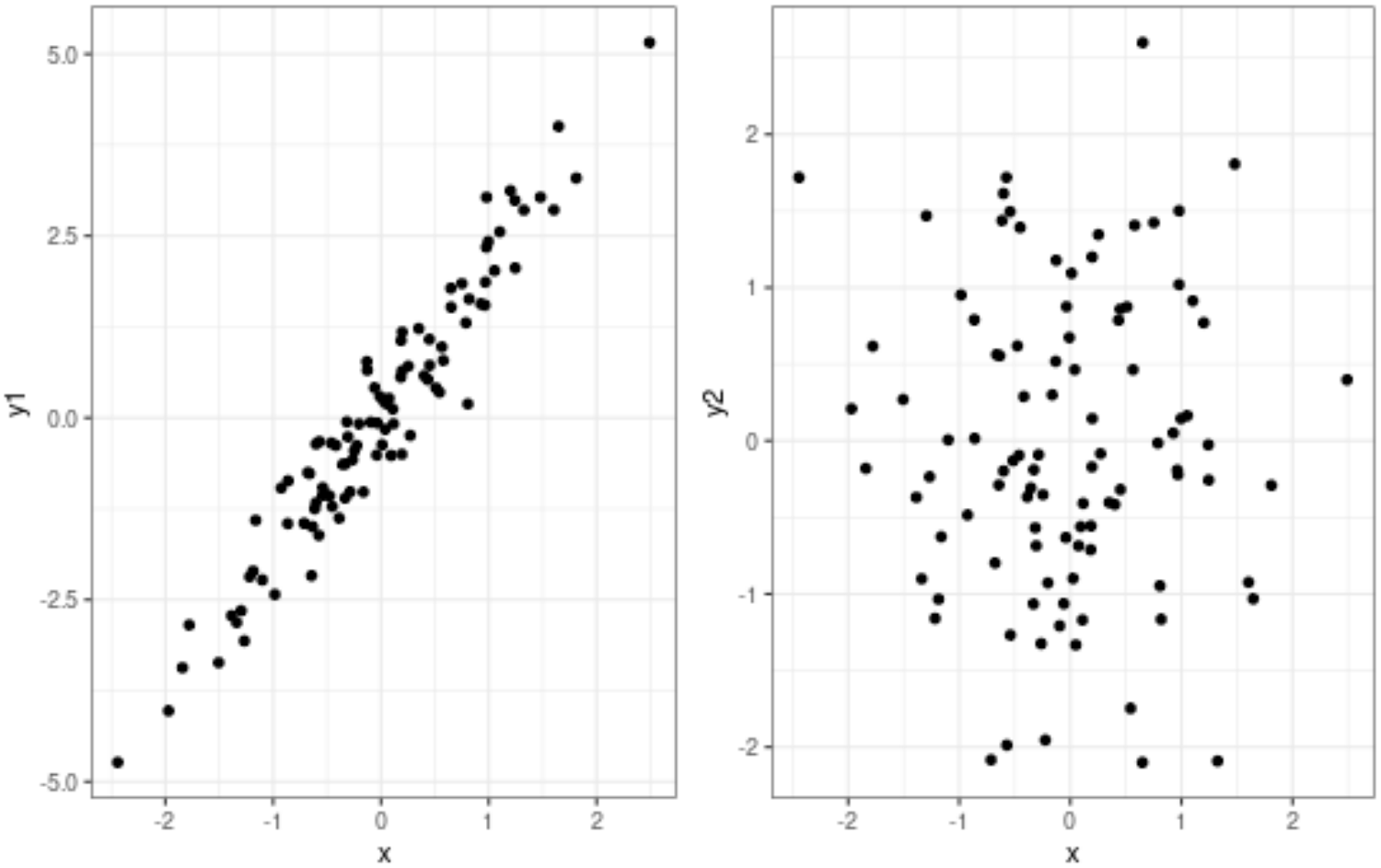
Types of Quantitative Relationships



Types of Quantitative Relationships



Types of Quantitative Relationships



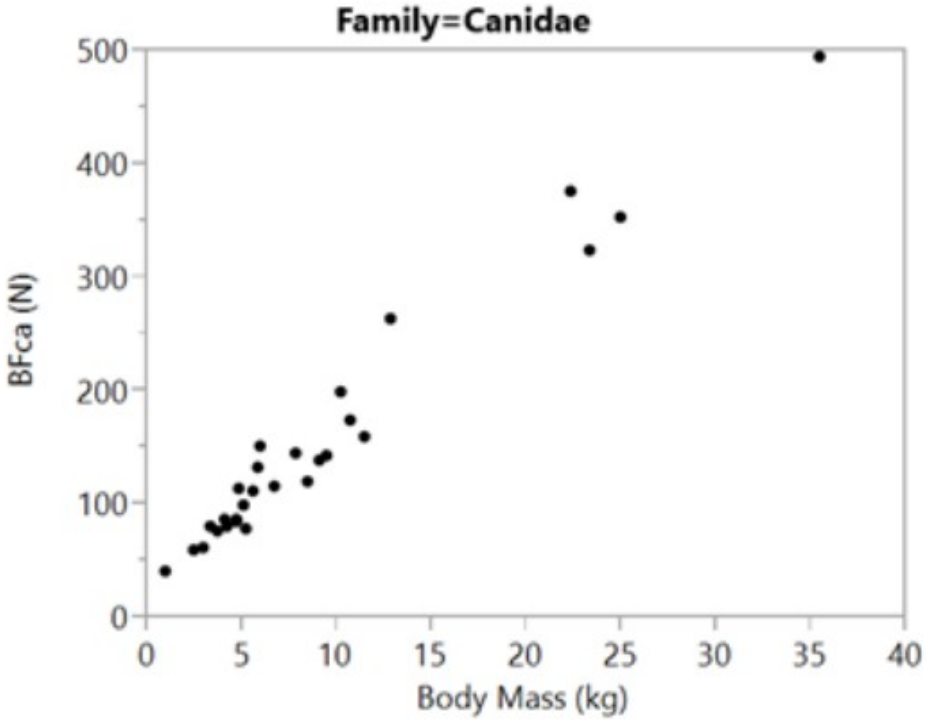
Describing a Scatterplot

To describe the relationship between variables in a scatterplot we need to mention all of the following:

- **Form**: what type of pattern exists (linear/non-linear/curved/cloud)
- **Strength**: how close are the points? (weak/moderate/strong)
- **Direction**: how the values of one variable relate to the values of the other variable (positive/negative)
- **Outliers** are there values that are much smaller/larger than the rest? Do they not follow the pattern?
  - Even if there aren't outliers, mention them

Describing Scatterplots – Example

Canidae is the biological family that contains dogs, wolves, foxes, and similar mammals.
Two variables are bite force (N) and body mass (kg). Which would be the explanatory variable and which would be the response variable?



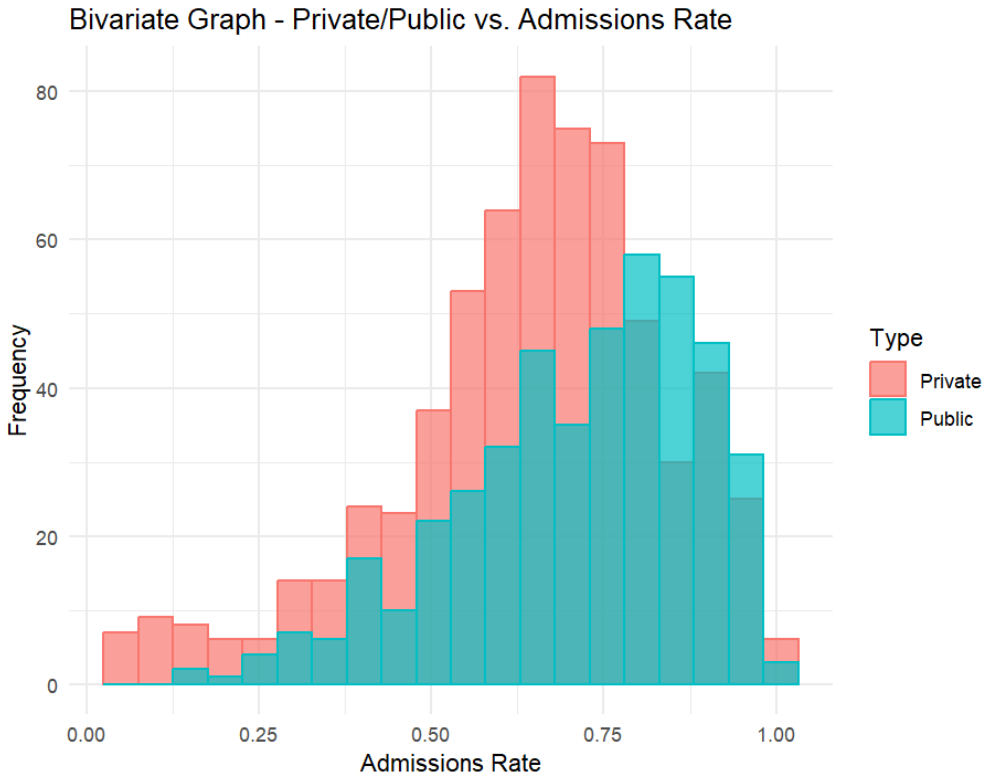
How do we describe the scatterplot?

source: "Bite Forces and Evolutionary Adaptations to Feeding Ecology in Carnivores," by P. Christiansen and S. Wade, Ecology, 88(2), 2007, pp. 347 – 358

Bivariate Graphs

Bivariate graphs show the relationship between two variables and which one we use is still dictated by

- type of variables
- context



The Onion

News

Thousands Of Students Forced To Attend Iowa State After University Sets Acceptance Rate To 140%

Published: March 18, 2019

General Rule

With respect to graphics generally....

- Explanatory variables are on the x-axis
  - Eg Time
- Response variables are on the y-axis
  - Eg Your weight
  - Which together is your weight over time graph
- Often this is intuitive
- Sometimes there is no meaningful ordering
  - length vs width of mussels in the Missouri River

Now and Next Time

- Currently we will finish up the R labs
- Wednesday we will go over more notes on visualization
  - Accessibility in graphics (eg colorblind friendly)
  - Best practices
  - Ggplot2 paradigm
- Ggplot2 Lab