Visualizing Data

Grinnell College

September 6, 2024

Nathan Friedrichsen

Goals for Class Today

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

1. use appropriate graphs to display and describe quantitative and categorical data
2. describe the **distribution** of a variable
3. make graphs that describe the relationship between 2 or more variables

Motivation

Let's look at the Tips data seen previously. Here are 20 observations out of 244 regarding the tips given to one waiter over the course of several months in one restaurant.

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

Do more customers come to the restaurant on certain days?

Hard to tell by looking at table

Motivation

Data collection has made remarkable progress in the last decades, giving us a greater quantity of data than most could ever dream of. However, just looking at tables of data is not very useful.

Better approaches:

1. **Data Visualization** displaying data in ways that make patterns more noticeable
2. **Numerical Summaries** calculating numbers that tell us about certain aspects of the data

Data Visualization

Previously, I made a big deal out of determining if a variable is categorical or quantitative

Why? Because the type of variable is going to determine how we make graphs to display and describe the data

The types graphs we make also will be determined by *how many* different variables we are working with

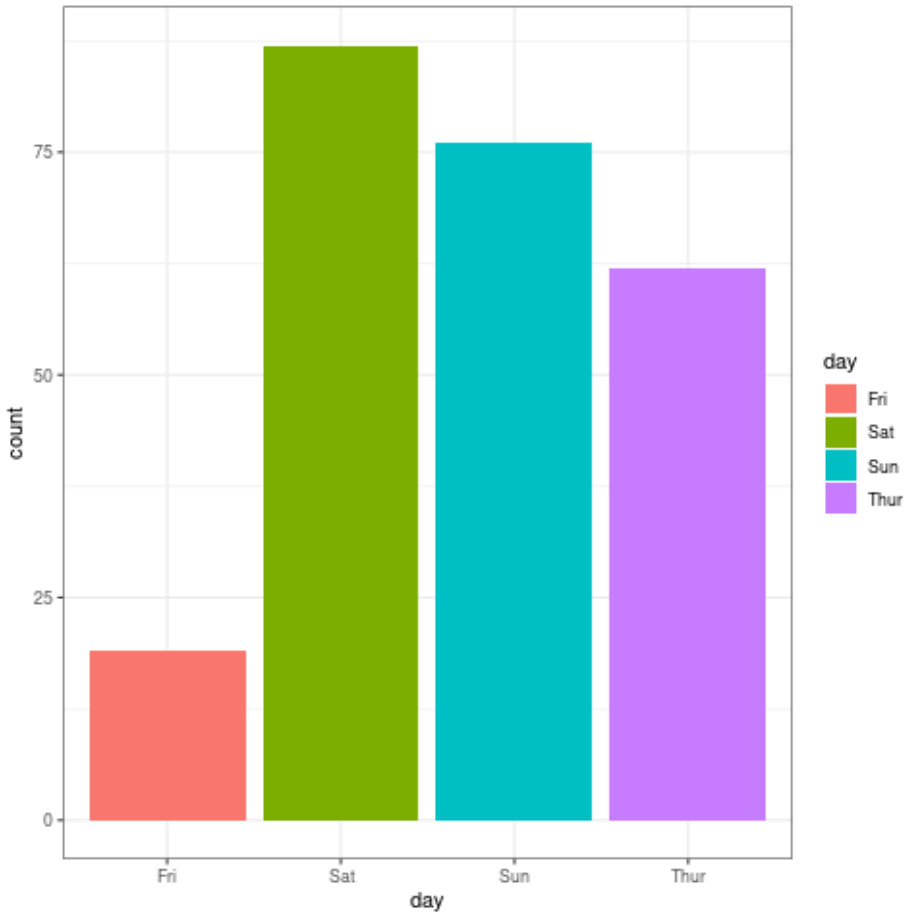
Distribution

In order to better understand patterns in our data, we will often combine graphics with short descriptions of what we see. A term we will use often:

The **distribution** of a variable refers to how frequently certain values of that variable show up in our data

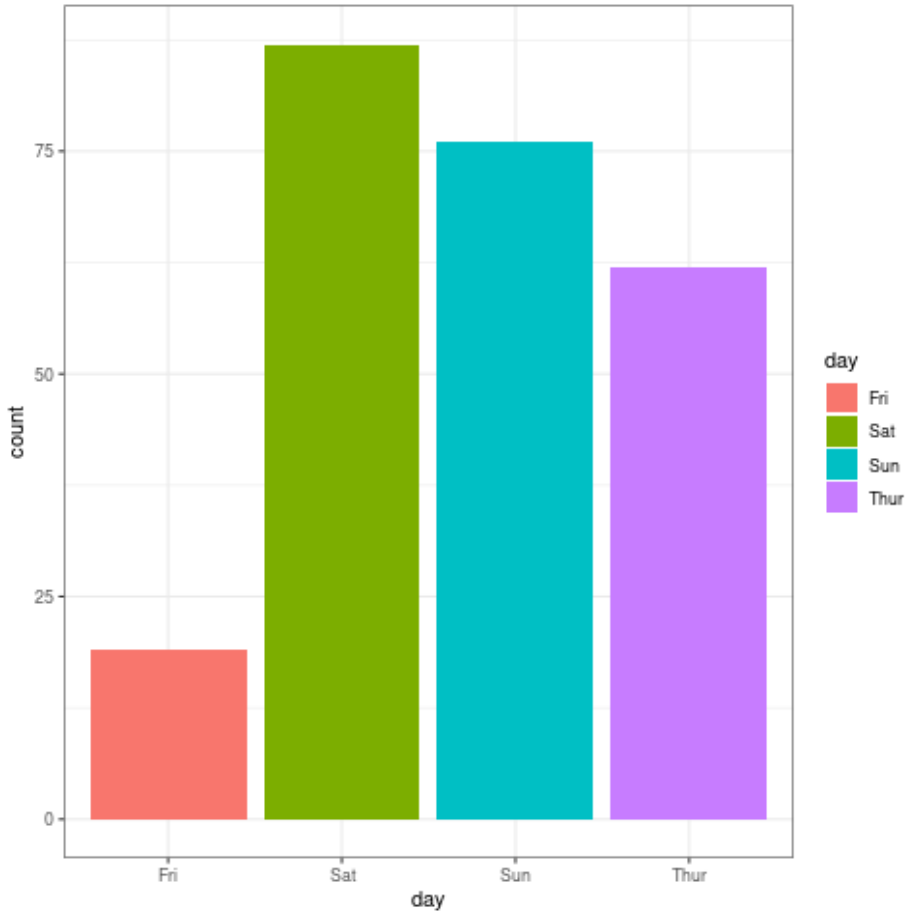
One Categorical Variable

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



One Categorical Variable

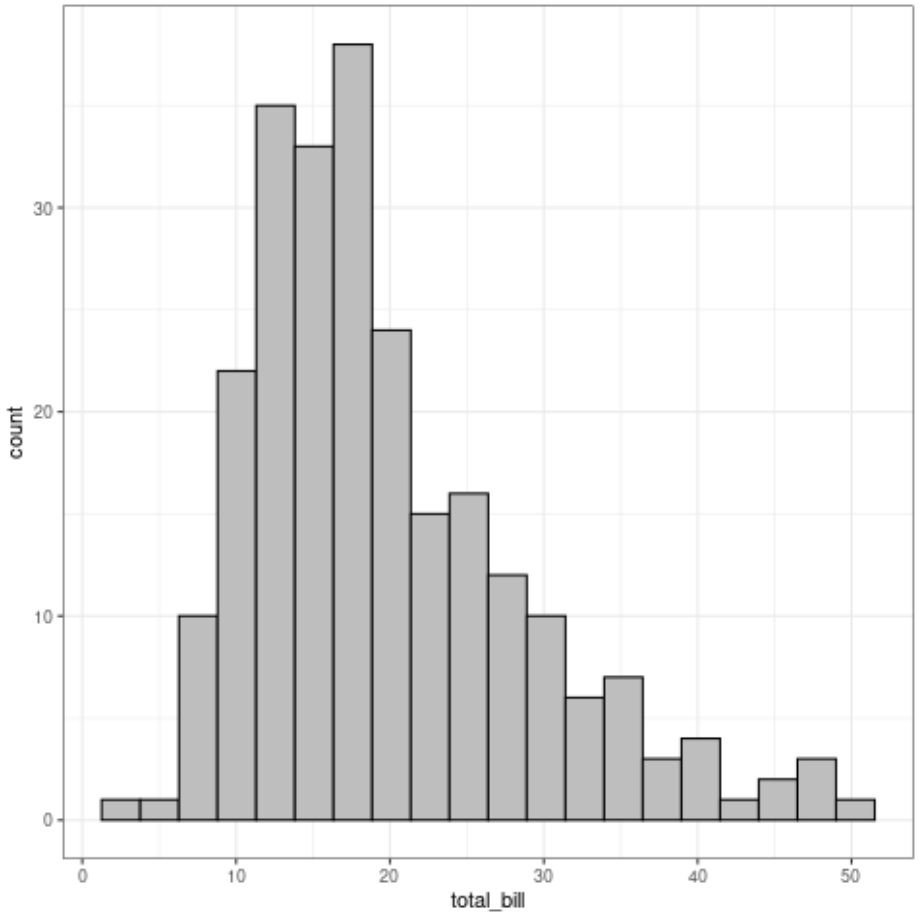
To describe the distribution of a categorical variable, we just need to talk about how likely each category is, and mention the most and least likely categories (helpful to include supporting values)



Distribution of customers?

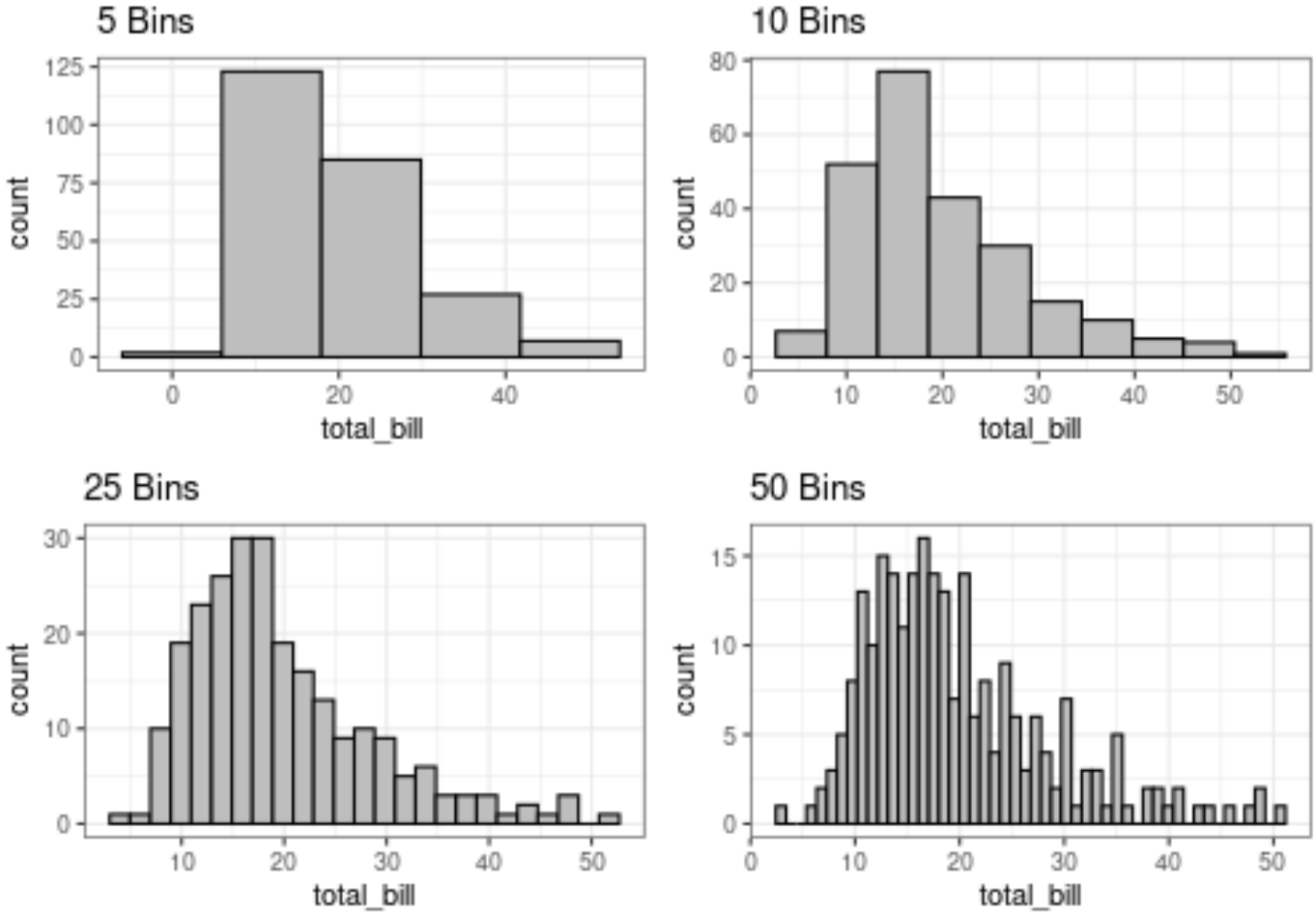
One Quantitative Variable

For quantitative variables, a **histogram** is often used to show the distribution of values. Histograms group numeric values into equally spaced intervals known as bins, then display the frequencies/counts (or % or proportion) of data in each bin:

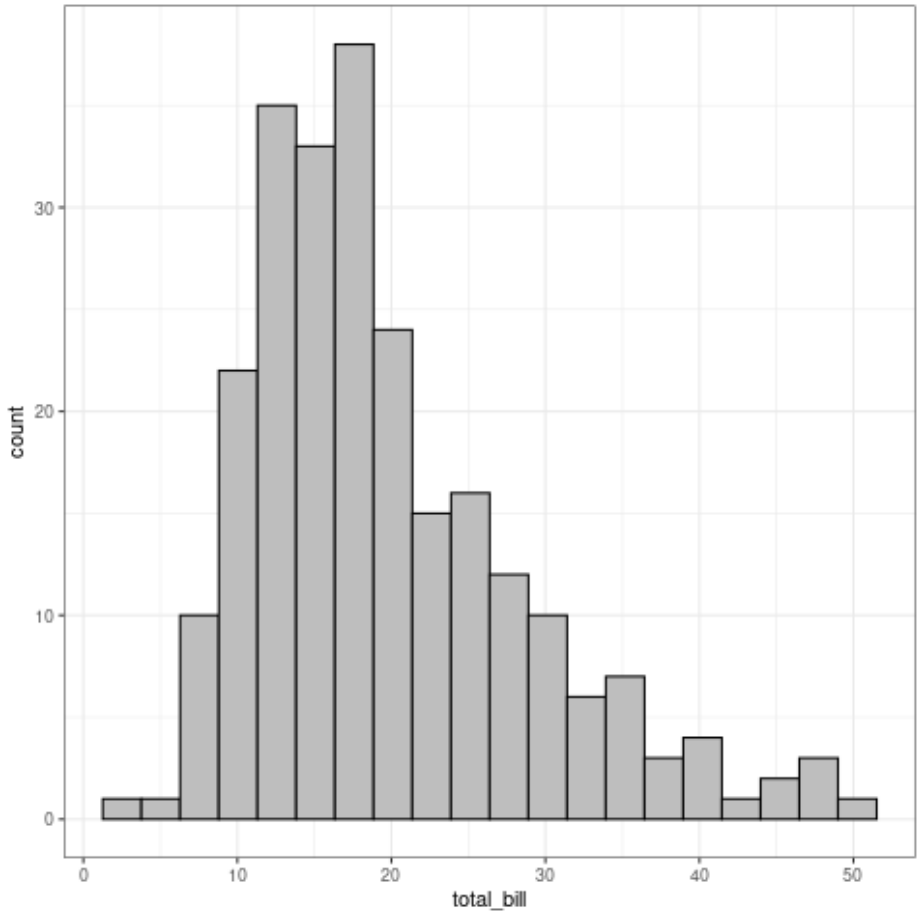


Histogram Bin Width

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



One Quantitative Variable



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?)

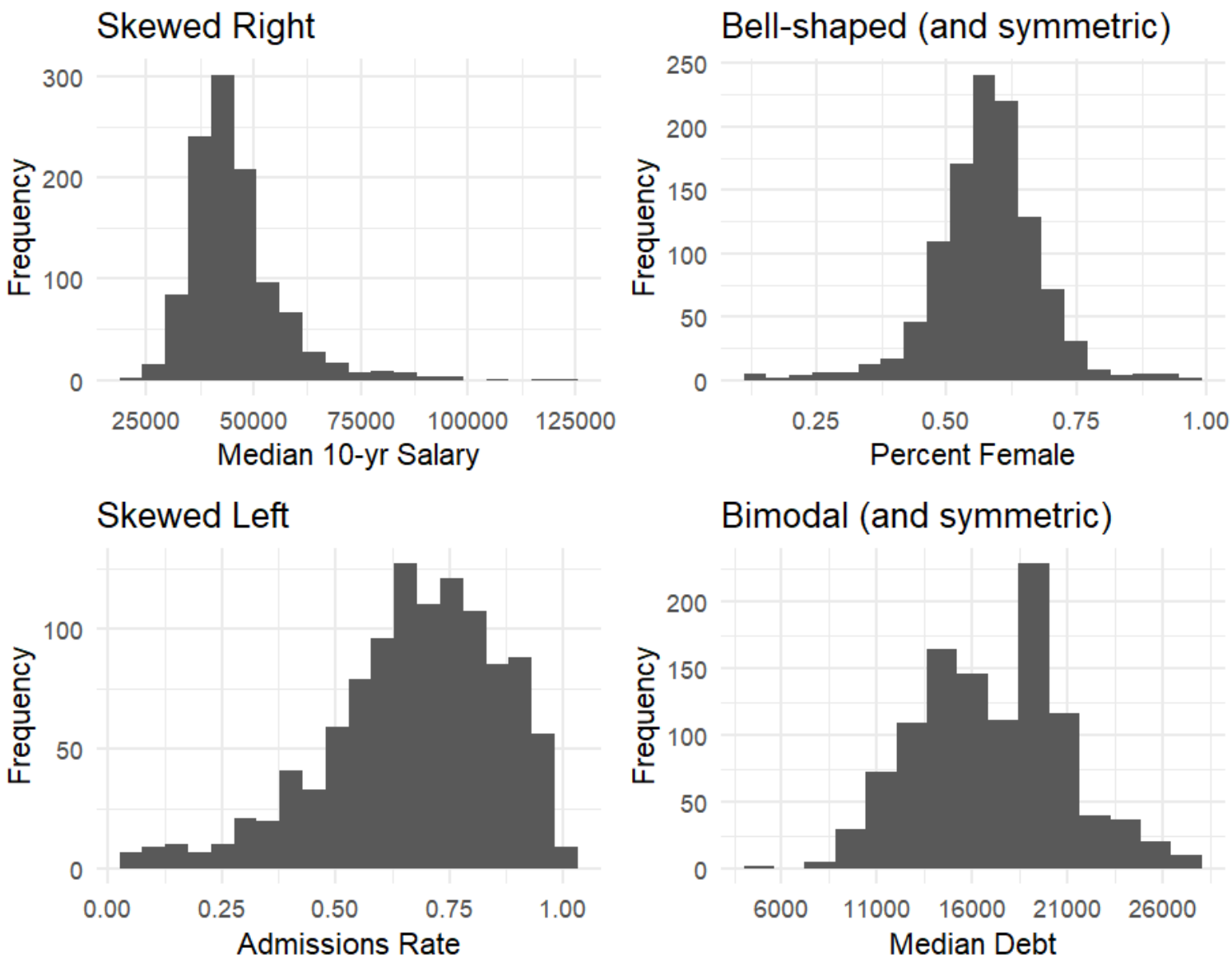
One Quantitative Variable - Distribution

Describing the **distribution** of a quant. variable is more nuanced than for a cat. variable

We need to mention all of the following things:

- **Shape** - is the distribution symmetric, skewed, bell-shaped, bimodal?
- **Center** - where does the data bunch up (approx. 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?

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

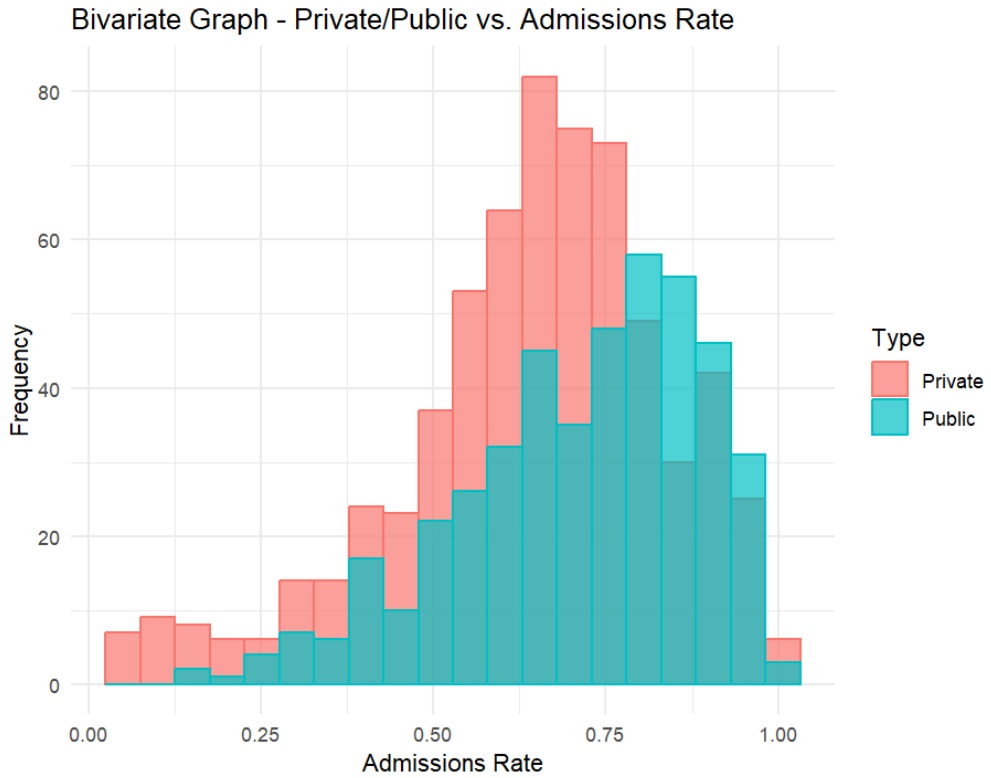
In a histogram, we identify outliers by looking for gaps in the bins

Bivariate Graphs

Up until now we have only looked at graphs that displayed one variable at a time. These are often called **univariate** graphs

Bivariate graphs show the relationship between two variables

- type of graph we use still depends on whether the variables are categorical or quantitative



Association

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

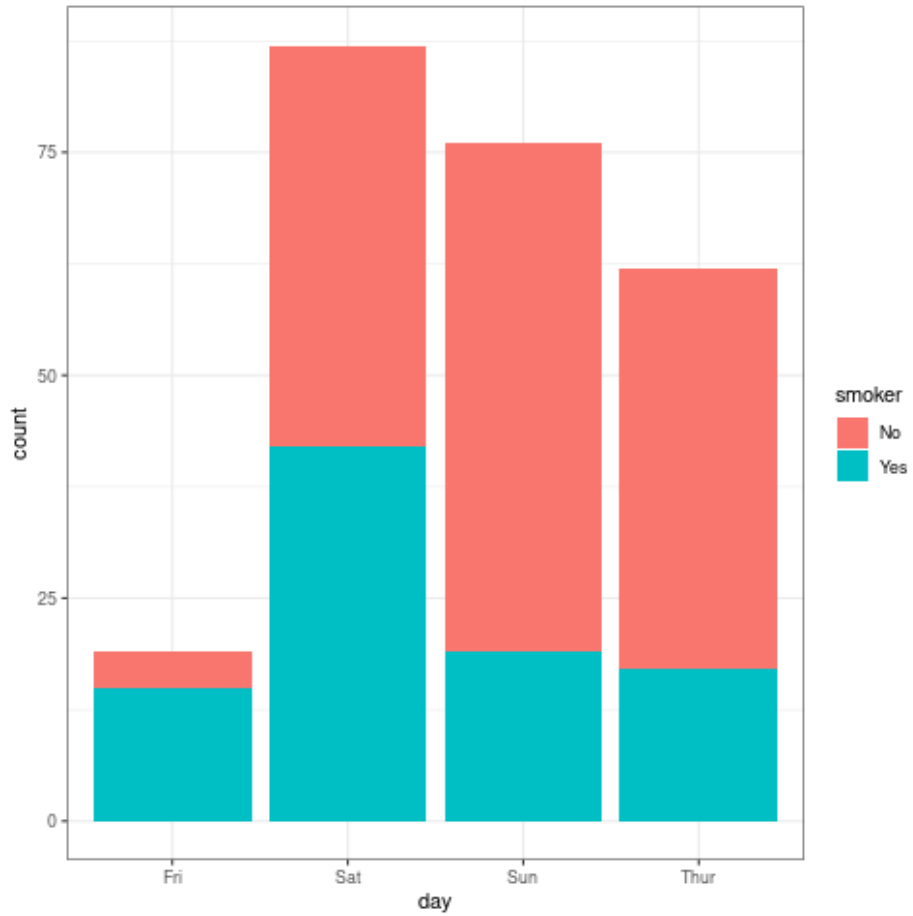
- When there seems to be some connection between two variables (knowing about one variable tells us about the other), we say they are **associated**.
- If there does not seem to be a relationship between the variables, we say they are **independent**.

When discussing an association between two variables we'll sometimes designate an explanatory variable (suspected cause) and a response variable (suspected effect)

NOTE: this does not always mean that one variable is causing a change in the other

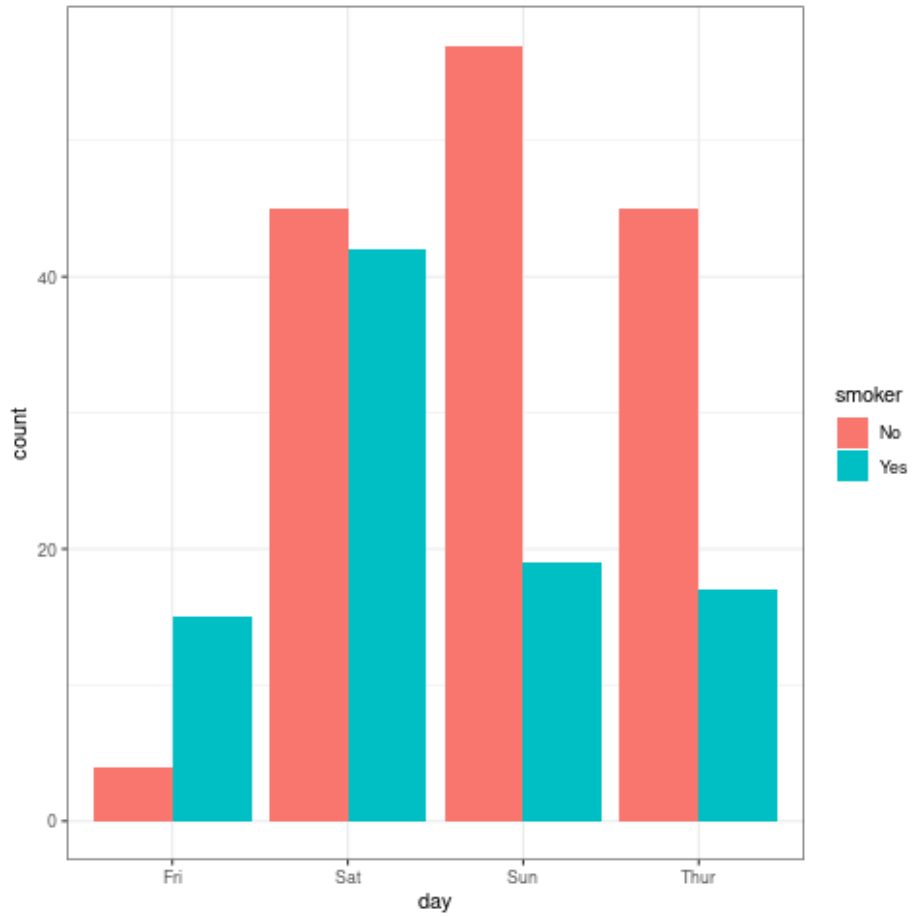
Categorical + Categorical → Stacked Bar

The first type of bivariate bar chart is known as a **stacked bar chart**, which allows us to break down one variable in terms of another. Here, we consider if any smokers were included in the party



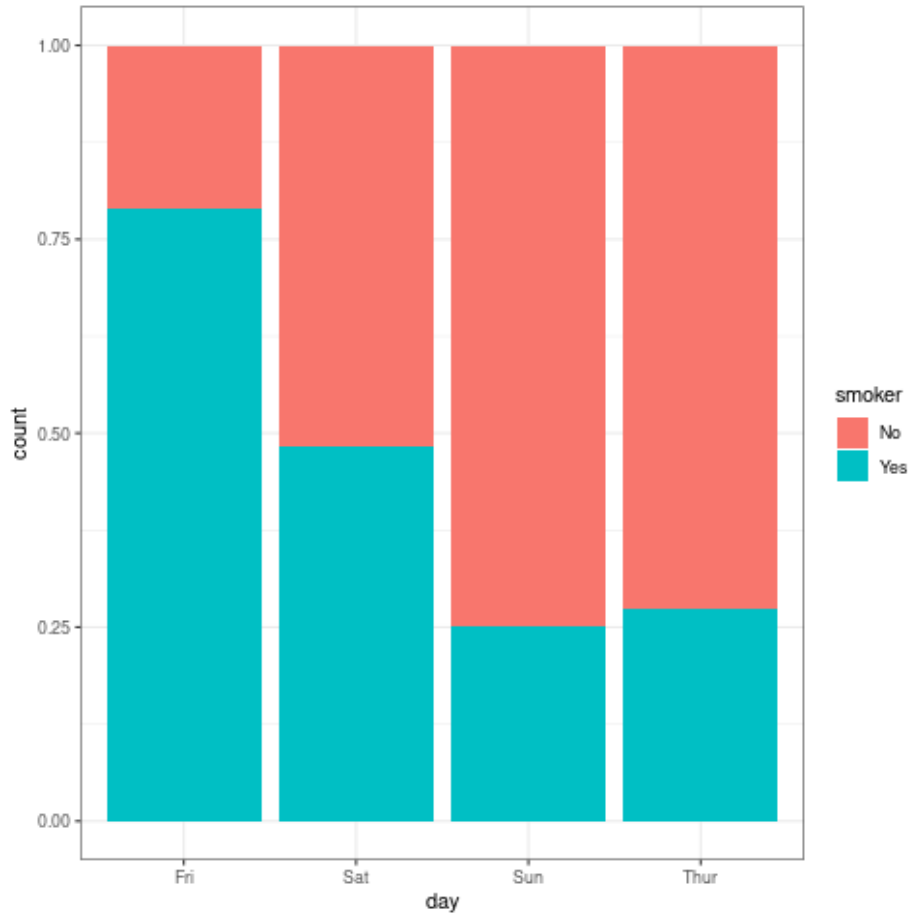
Categorical + Categorical → Dodge Bar

The second type of bivariate bar chart is known as a **dodged bar chart**, which presents both variables alongside one another. This makes comparing within groups much simpler



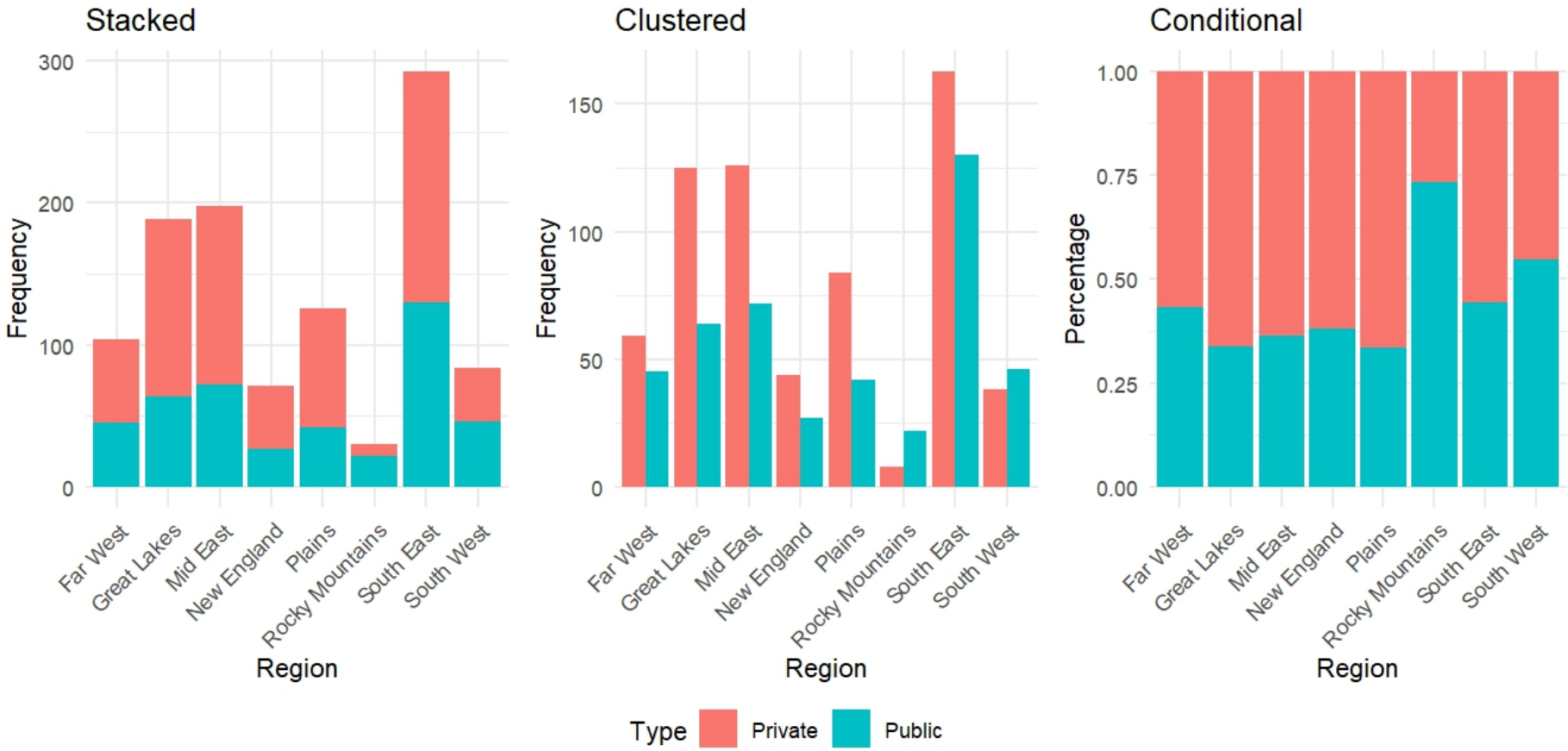
Categorical + Categorical → Filled Bar

The last type of bivariate bar chart is known as a **filled bar chart**, offering proportions. Although we lose absolute counts, we can now see relative frequencies within each group



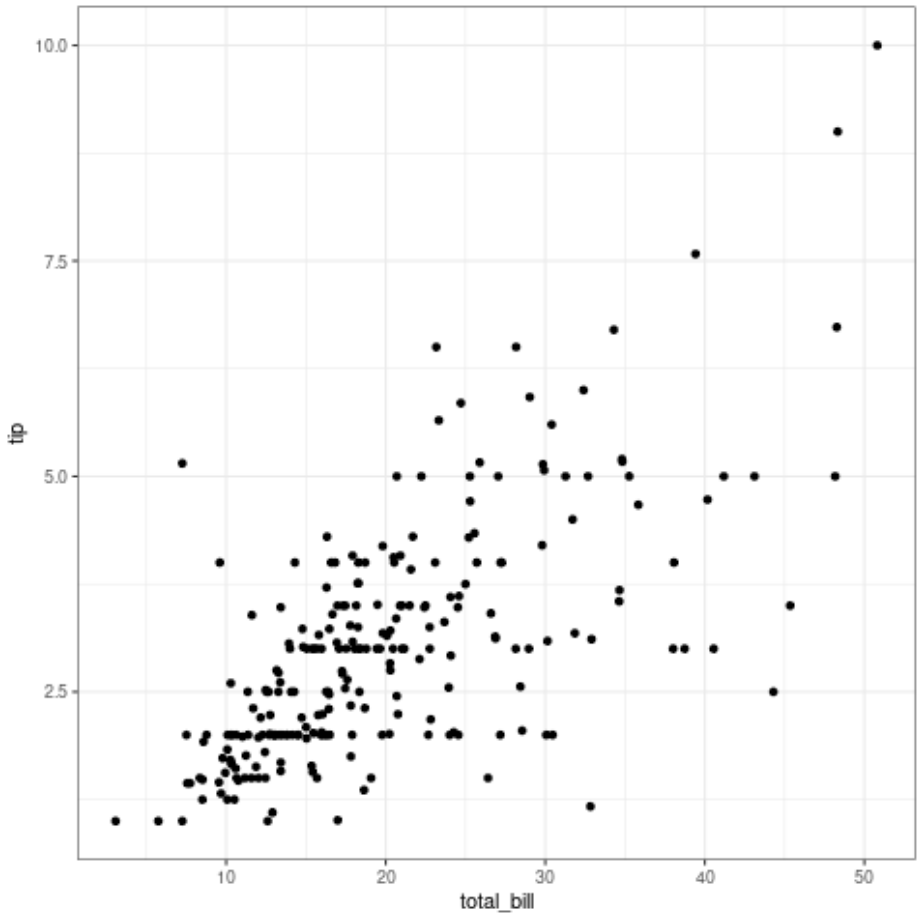
Bivariate Bar Charts

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



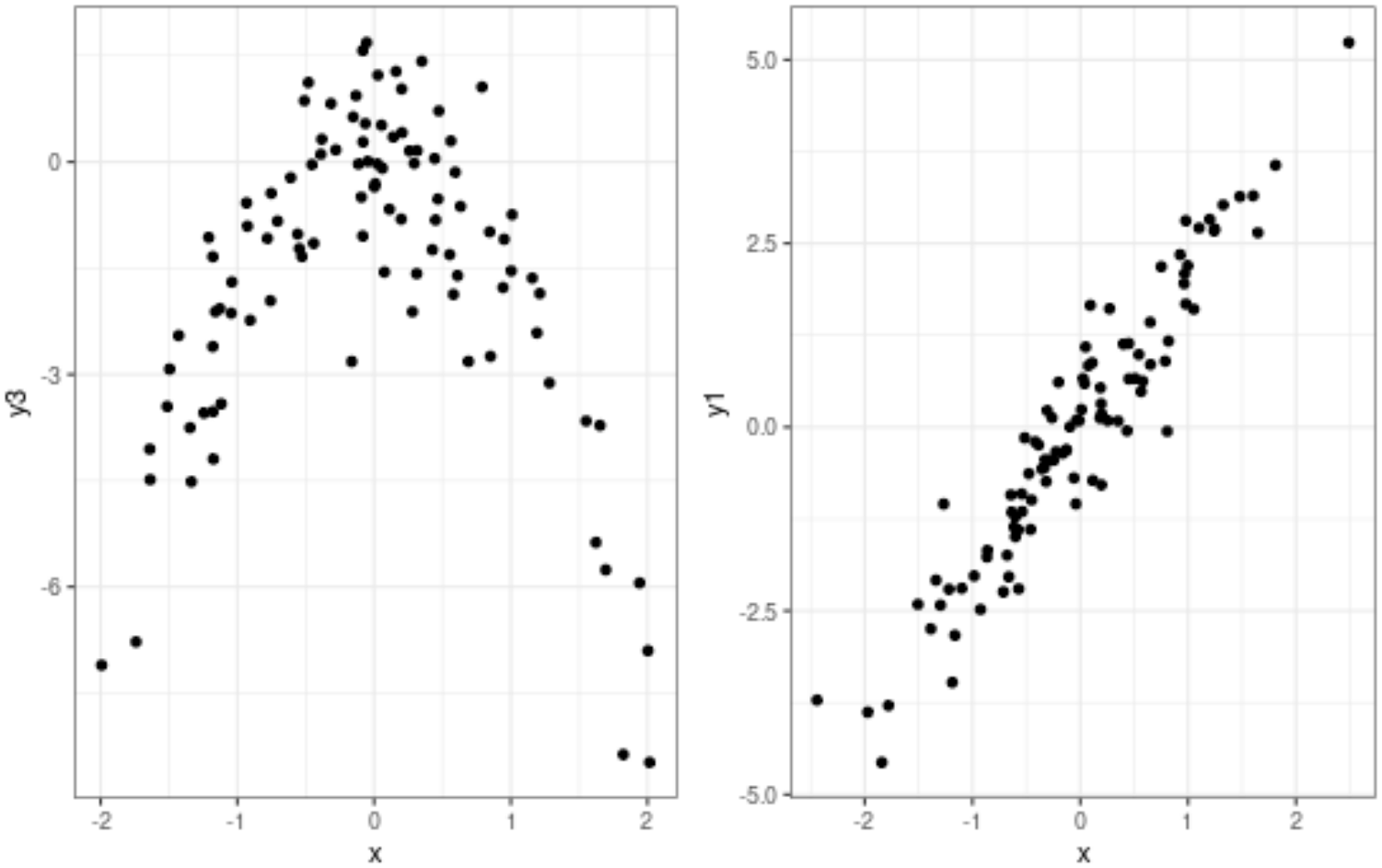
Quantitative + Quantitative → Scatterplots

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

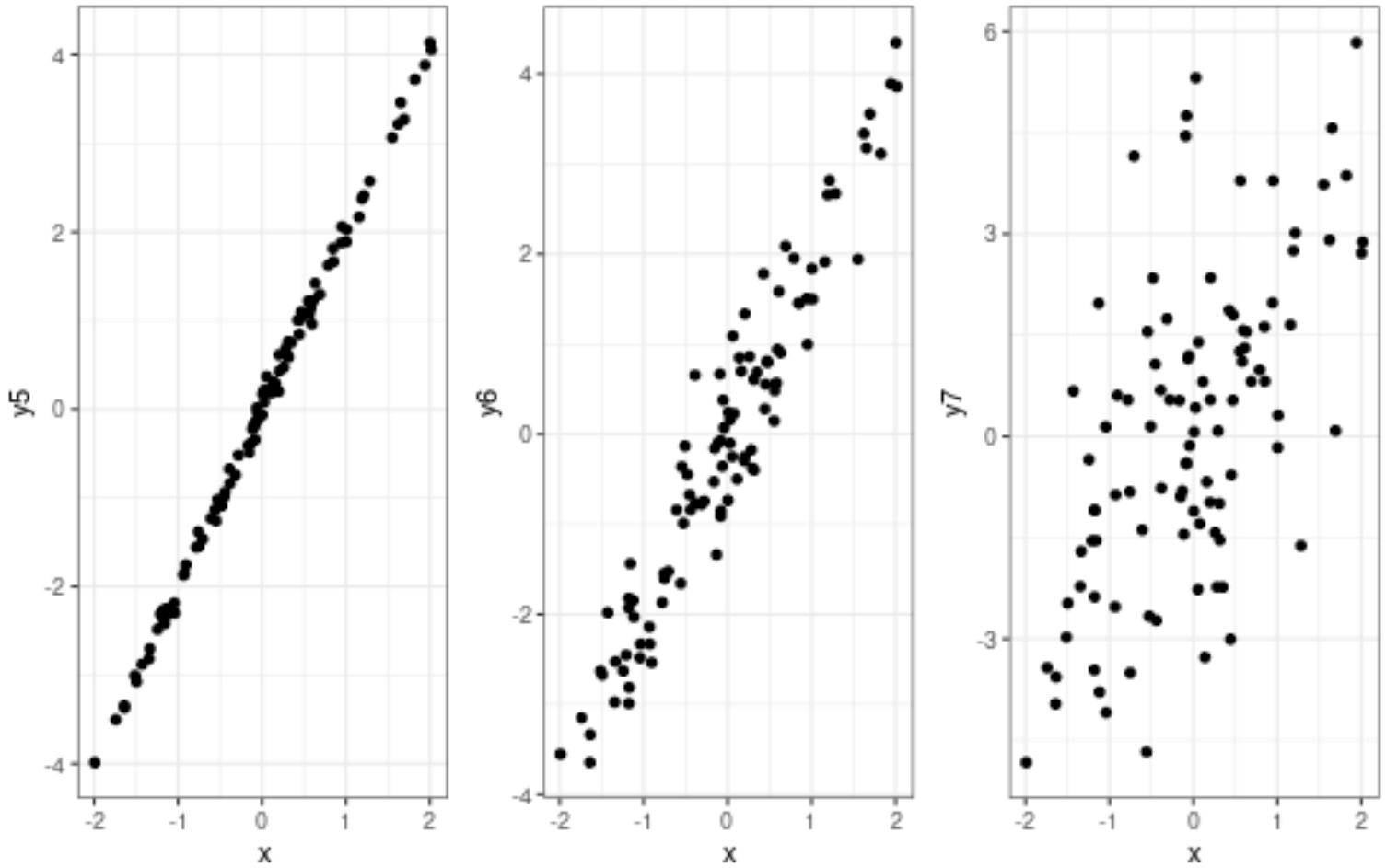


What kind of relationship do we see between the total bill and the tip amount?

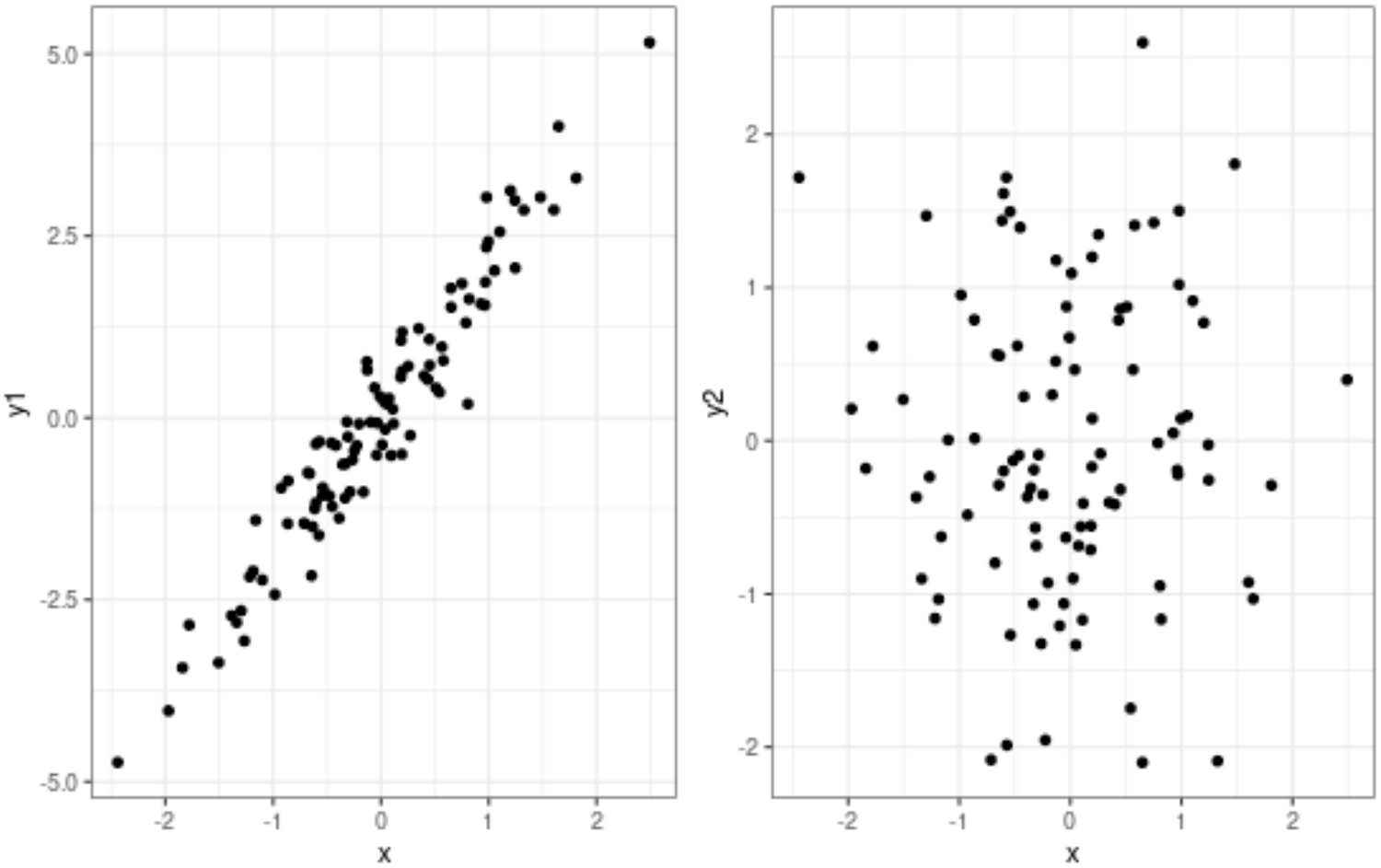
Types of Quantitative Relationships



Types of Quantitative Relationships



Types of Quantitative Relationships



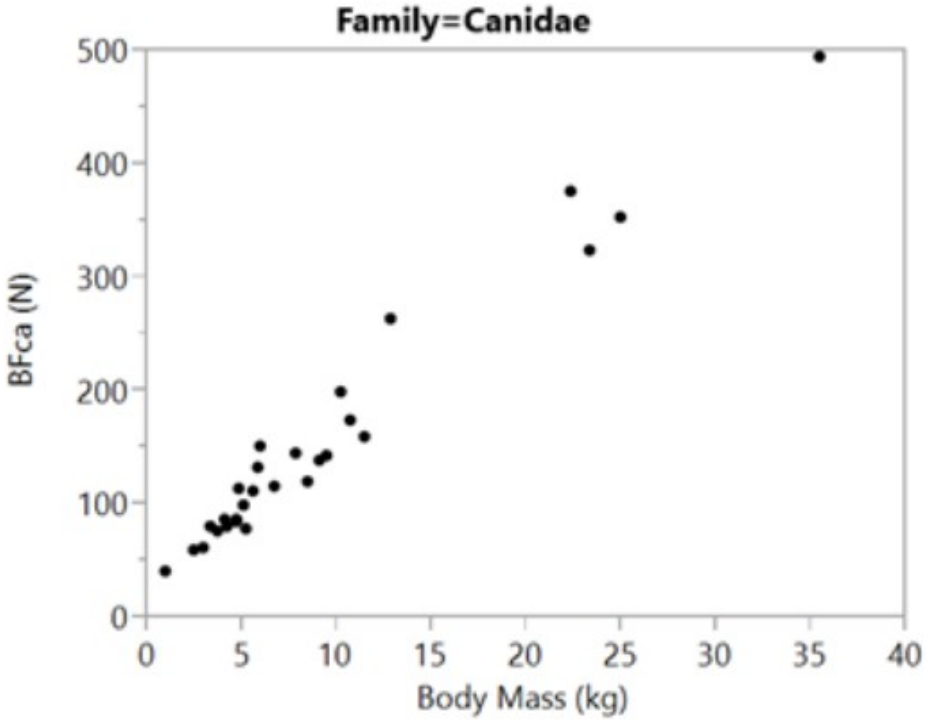
Describing a Scatterplot

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)
- **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**

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

Reflection

We'll take a few minutes to reflect on what we learned. Spend a minute talking with those around you to come up with answers to the following questions:

- Why do we make graphics to display data?
- What is the **distribution** of a variable?
- Why do we care about whether a variable is categorical or quantitative?

Next Time

- One other graph to display quantitative data (boxplot)
- What to do when we have Quantitative + Categorical variables
- Lab for putting this all into practice