Quantitative Variables – Part 2

Distributions, Associations, and Scatterplots

Grinnell College

Outline

What we are going to cover today:

1. More on mean vs. median
2. associations
3. scatterplots

Which measures to use for Center and Spread?

In general we will prefer to use **moment statistics** (mean and standard deviation) if we can, but there are certain situations where the mean and standard deviation are not good measures of center and spread

The *shape* of the distribution, as well as whether we have *outliers* will determine whether we use **order statistics** (median and IQR) or **moment statistics** (mean and standard deviation) to describe the center and spread

Which measures to use for Center and Spread?

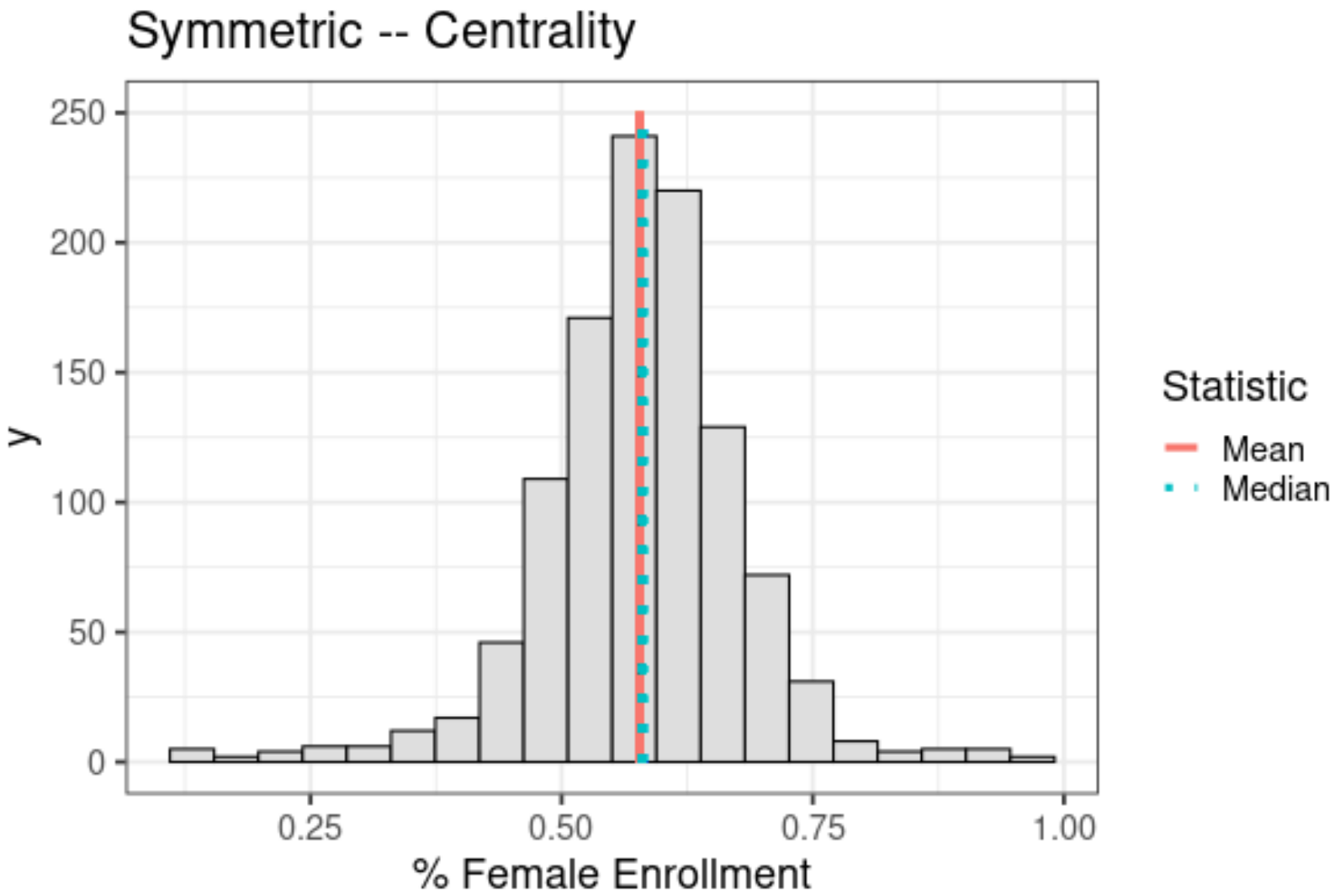
Order statistics are robust, moment statistics are not robust.

- A skewed distribution can affect the mean and std. dev. a lot
  - skew $\rightarrow$ mean & std. dev. not good measures of center & spread
- Outliers can affect the mean and std. dev. a lot
  - outliers $\rightarrow$ mean & std. dev. not good measures of center & spread

Summary:

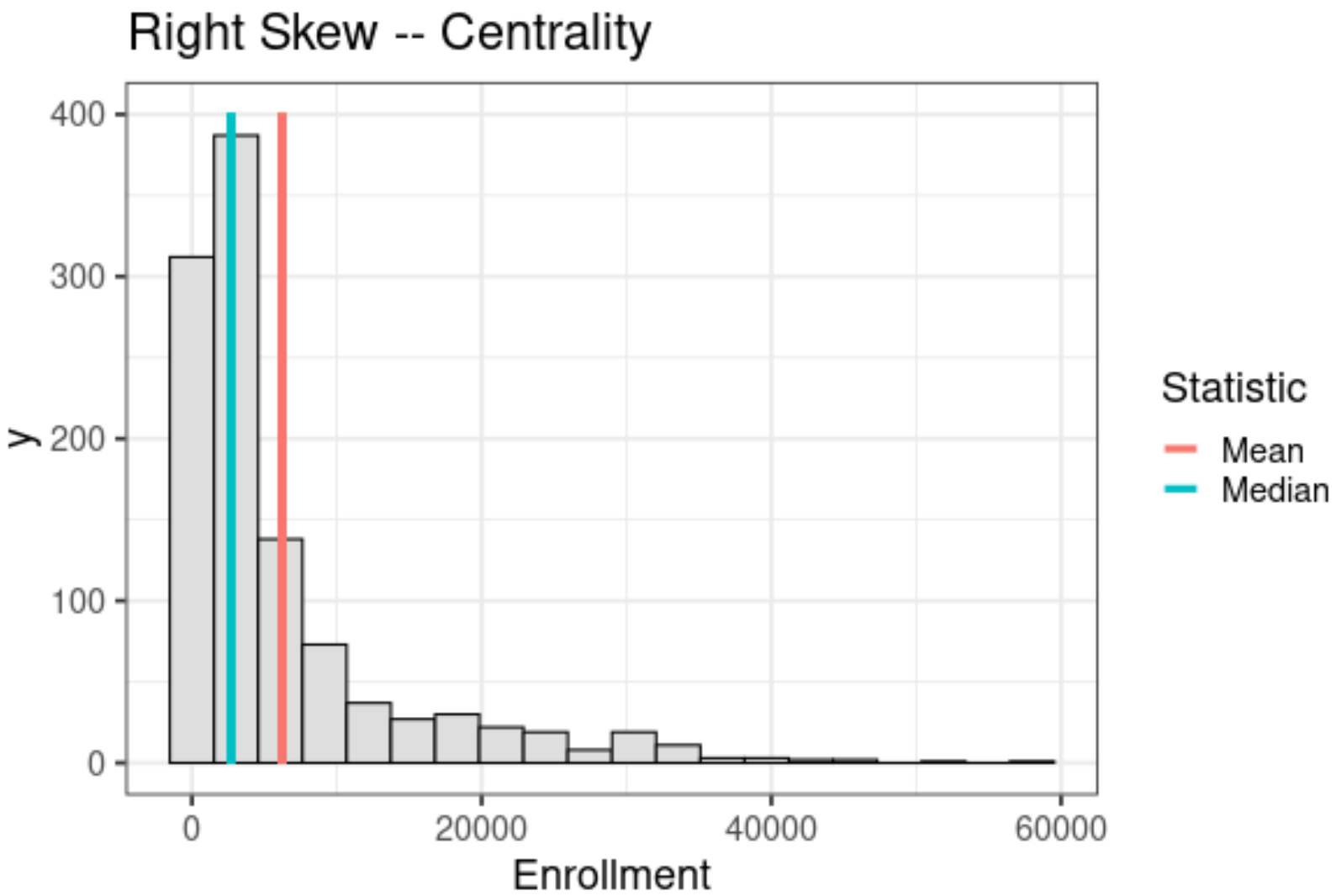
Symmetric shape with no outliers $\rightarrow$ mean and std. dev.
Skewed shape or outliers (or both) $\rightarrow$ median and IQR
Multimodal $\rightarrow$ median and IQR

Comparing Mean with Median



Symmetric distributions: mean $\approx$ median (generally)

Comparing Mean with Median

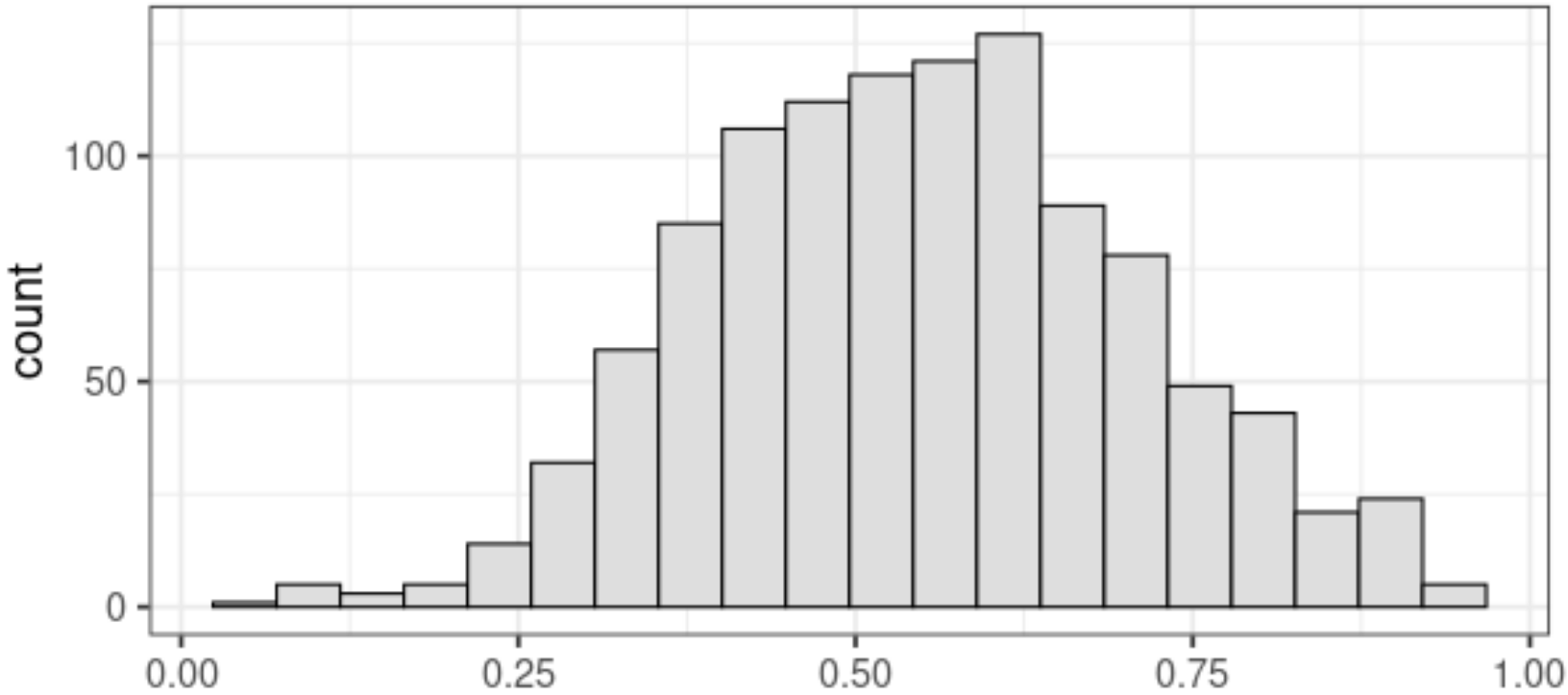


right skew: mean $>$ median left skew: mean $<$ median

Practice

1. Determine the shape of the distribution. Will mean and median be similar?

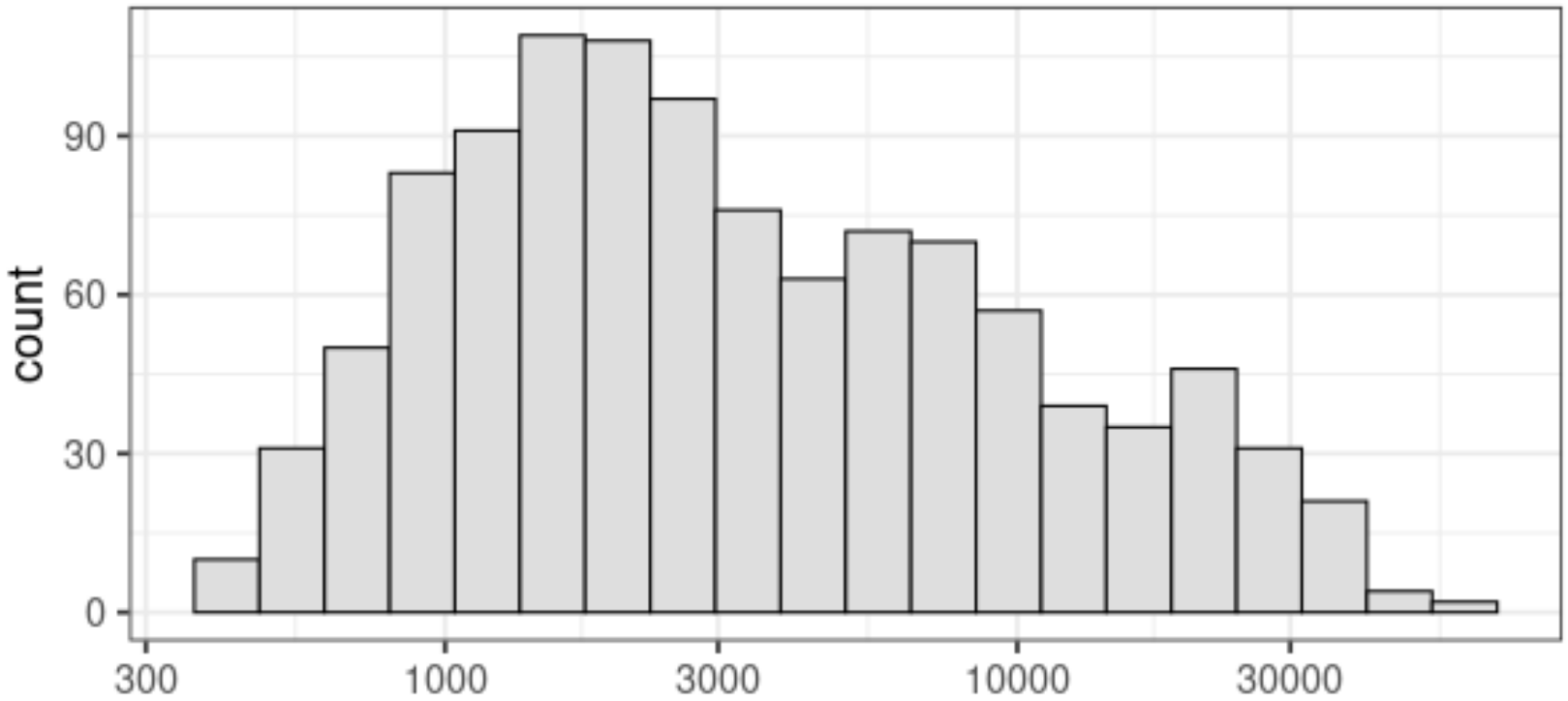
2. Should we use mean and standard deviation or median and IQR



Practice

1. Determine the shape of the distribution. Will the mean or median be larger?

2. Should we use mean and standard deviation or median and IQR

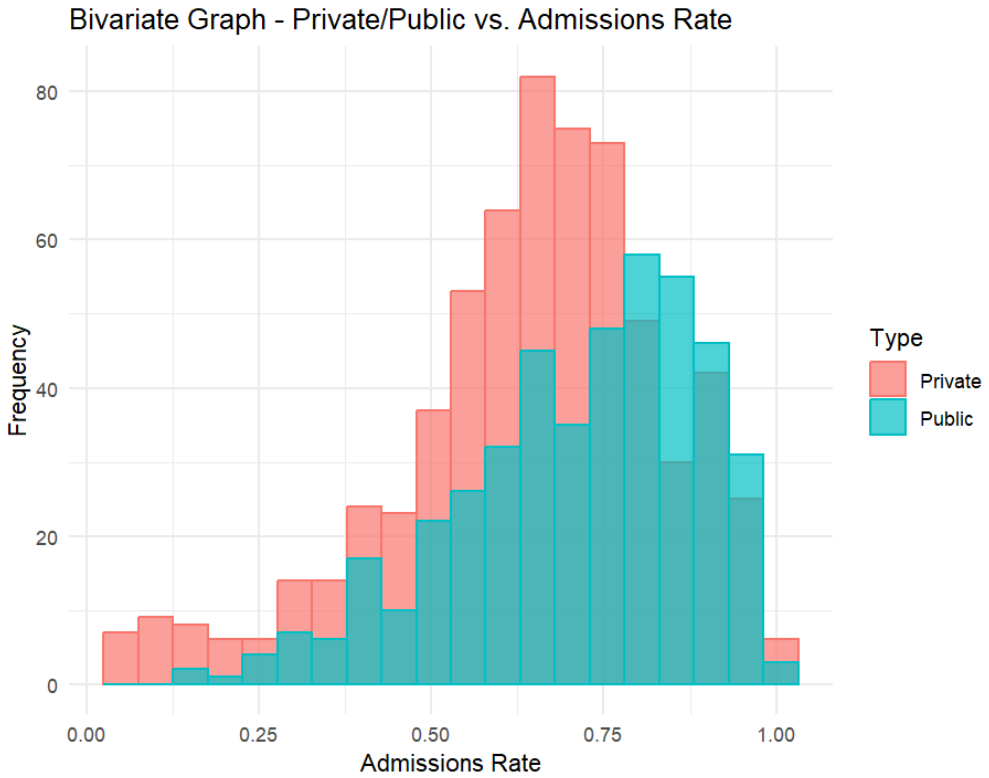


Bivariate Graphs

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



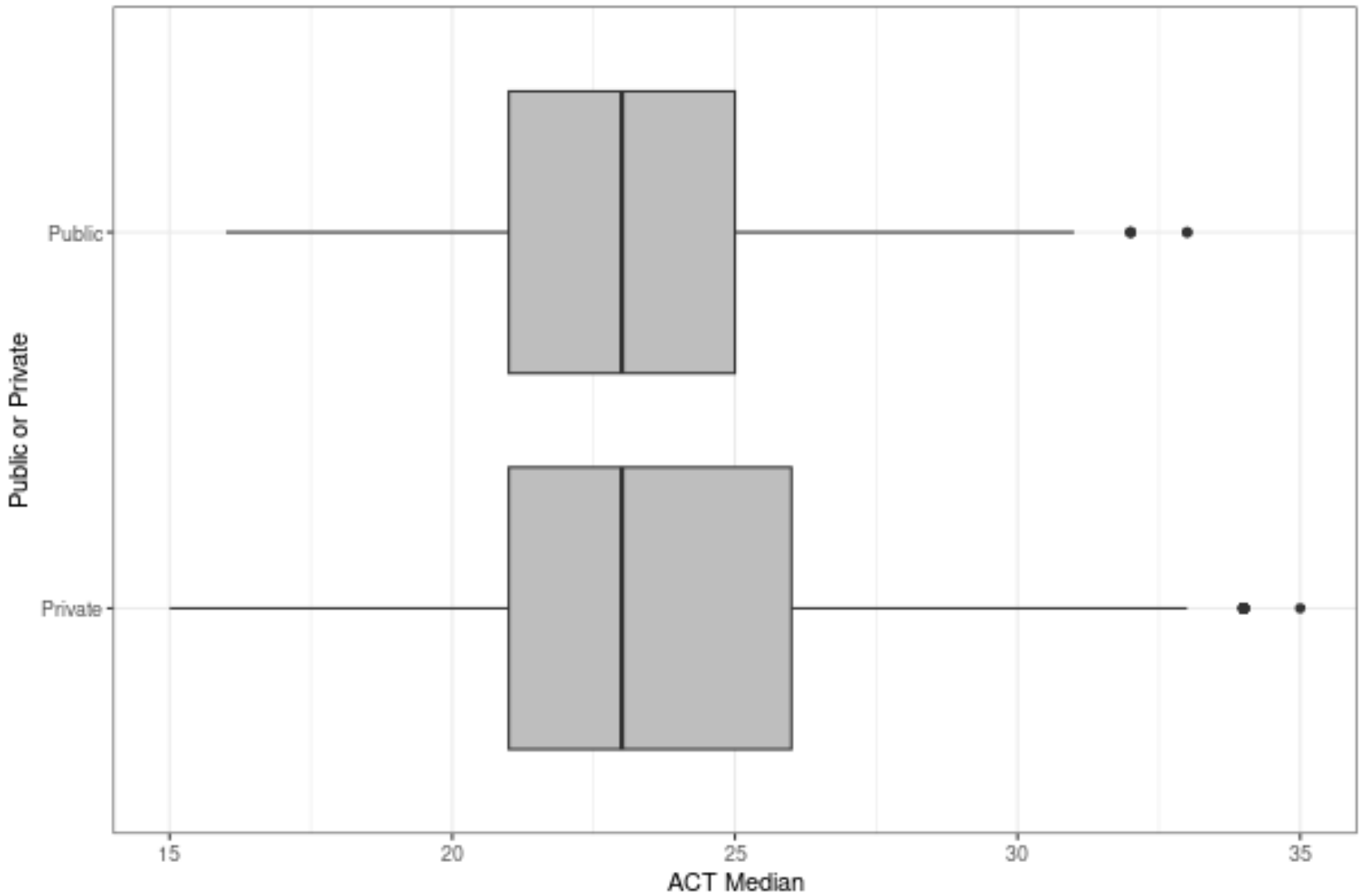
Association

It is going to be very common for us to try to see if there is a relationship between two (or more) variables

We will say there is an **association** between variables if knowing something about one tells us about the other. This means they are related to each other in some way.

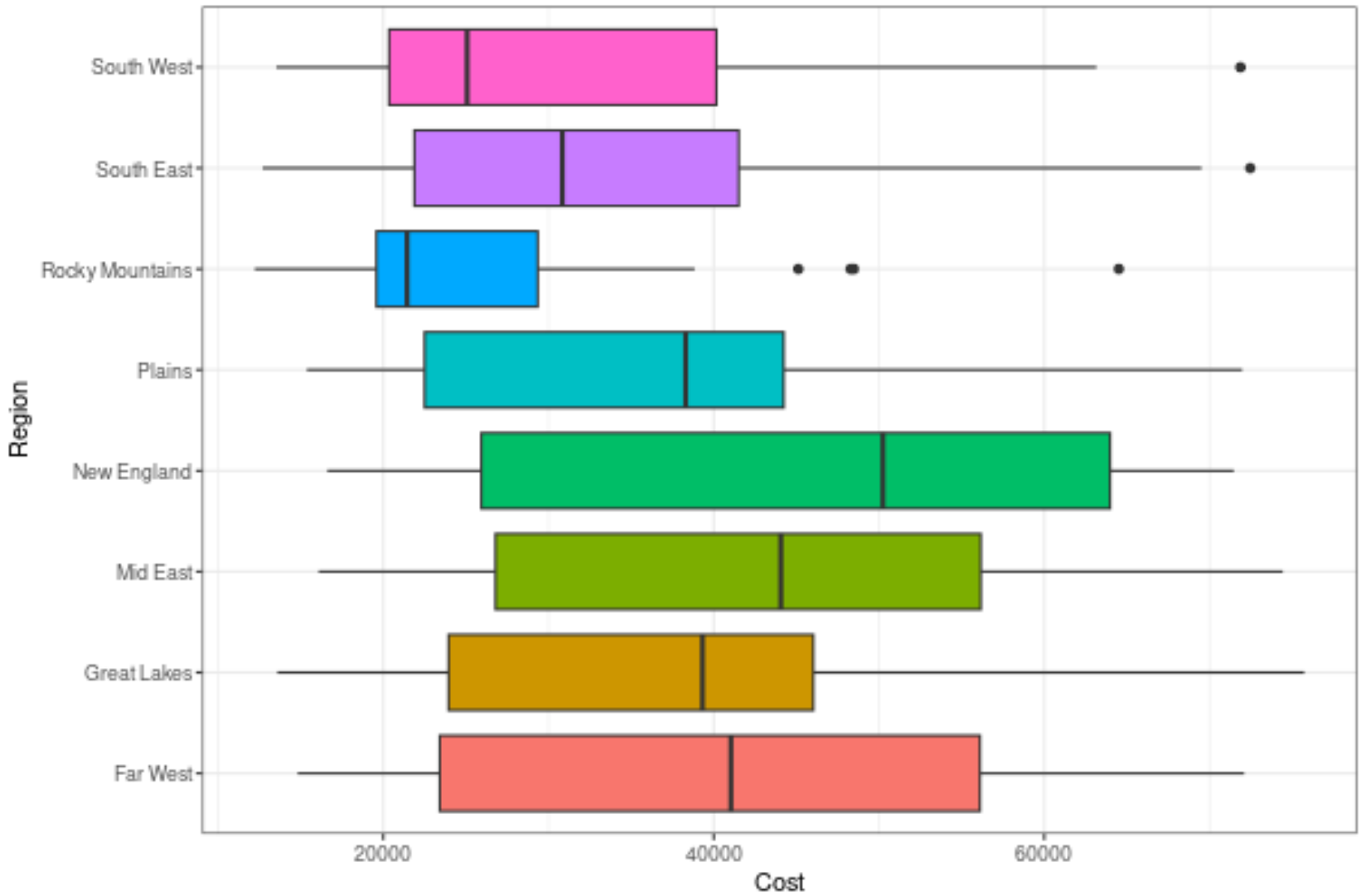
We will say two variables are **independent** if they are not associated. This means the variables don't affect each other much or at all.

Quantitative + Categorical $\rightarrow$ Side-by-side Box plots



Association?

Quantitative + Categorical → Side-by-side Box plots

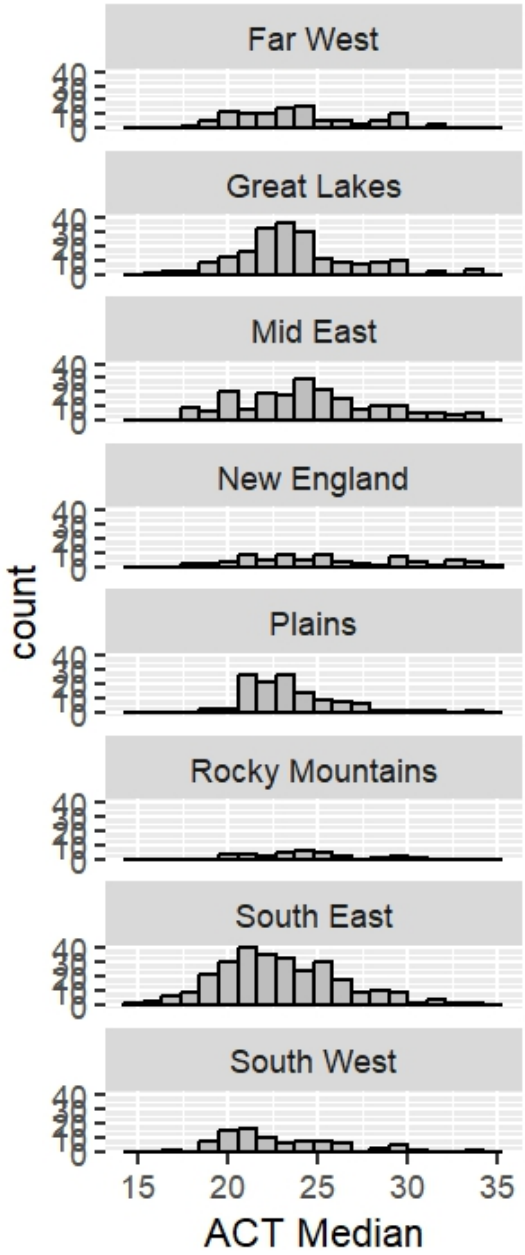


Association?

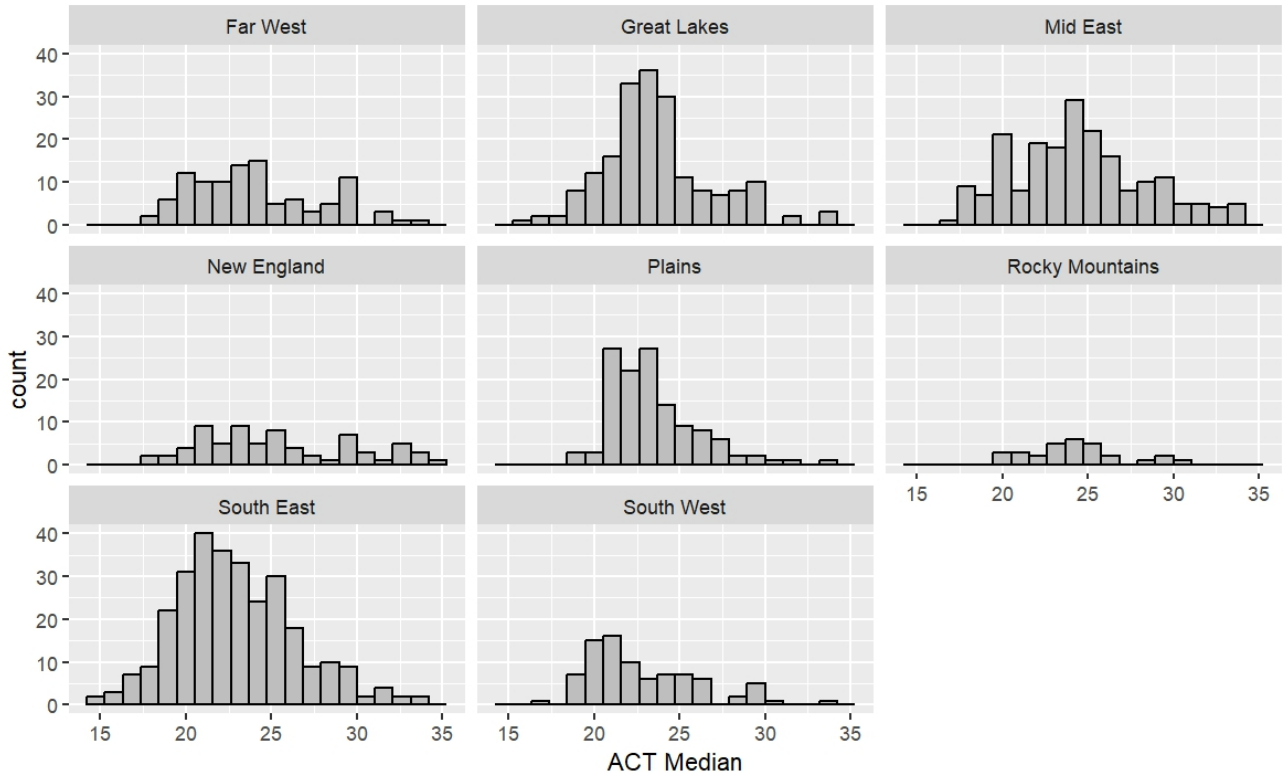
Quantitative + Categorical → Stacked Histograms (bad)

Instead of doing side-by-side box plots, you may ask why we couldn't do side-by-side (stacked) histograms

Technically we can, they just get difficult to read and compare



Quantitative + Categorical → Grid of Histograms (ok-ish)

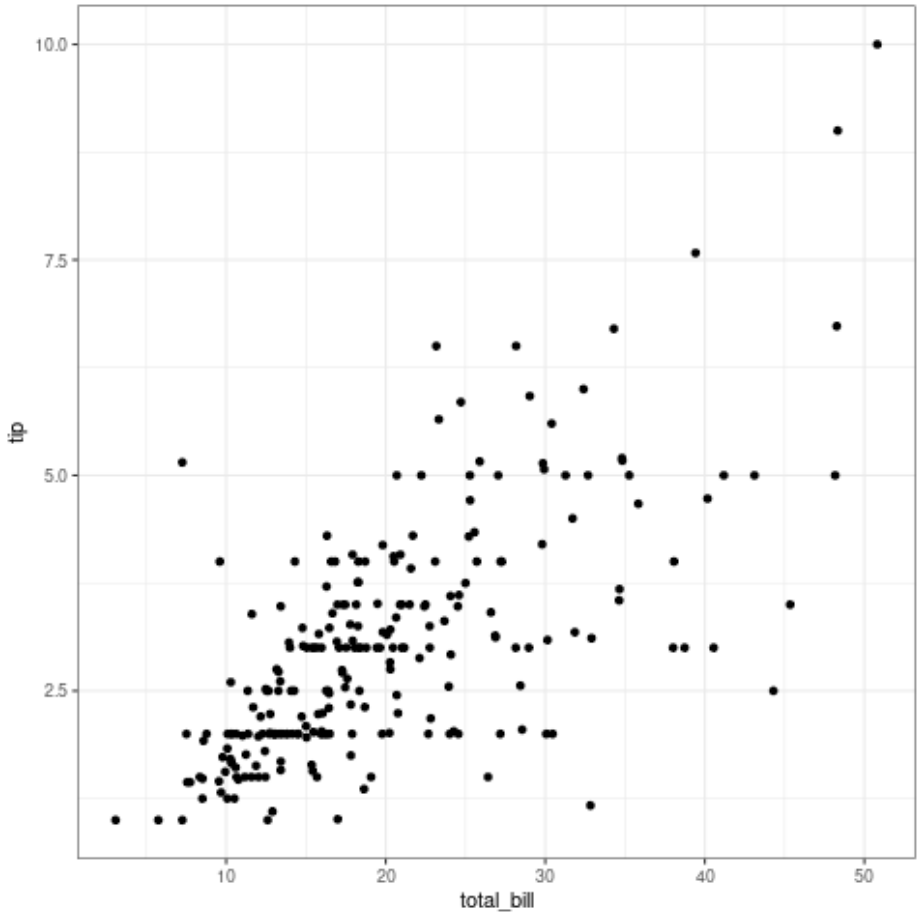


This does not have a special name that I know of... but is another way to display many histograms.

- easier to read the individual histograms
- still harder to compare each group than if we had just used box plots

Quantitative + Quantitative → Scatterplots

Visual summaries investigating the relationship between two quantitative variables are often presented with a **scatterplot**. Each dot corresponds to values for one observation in the data set.



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

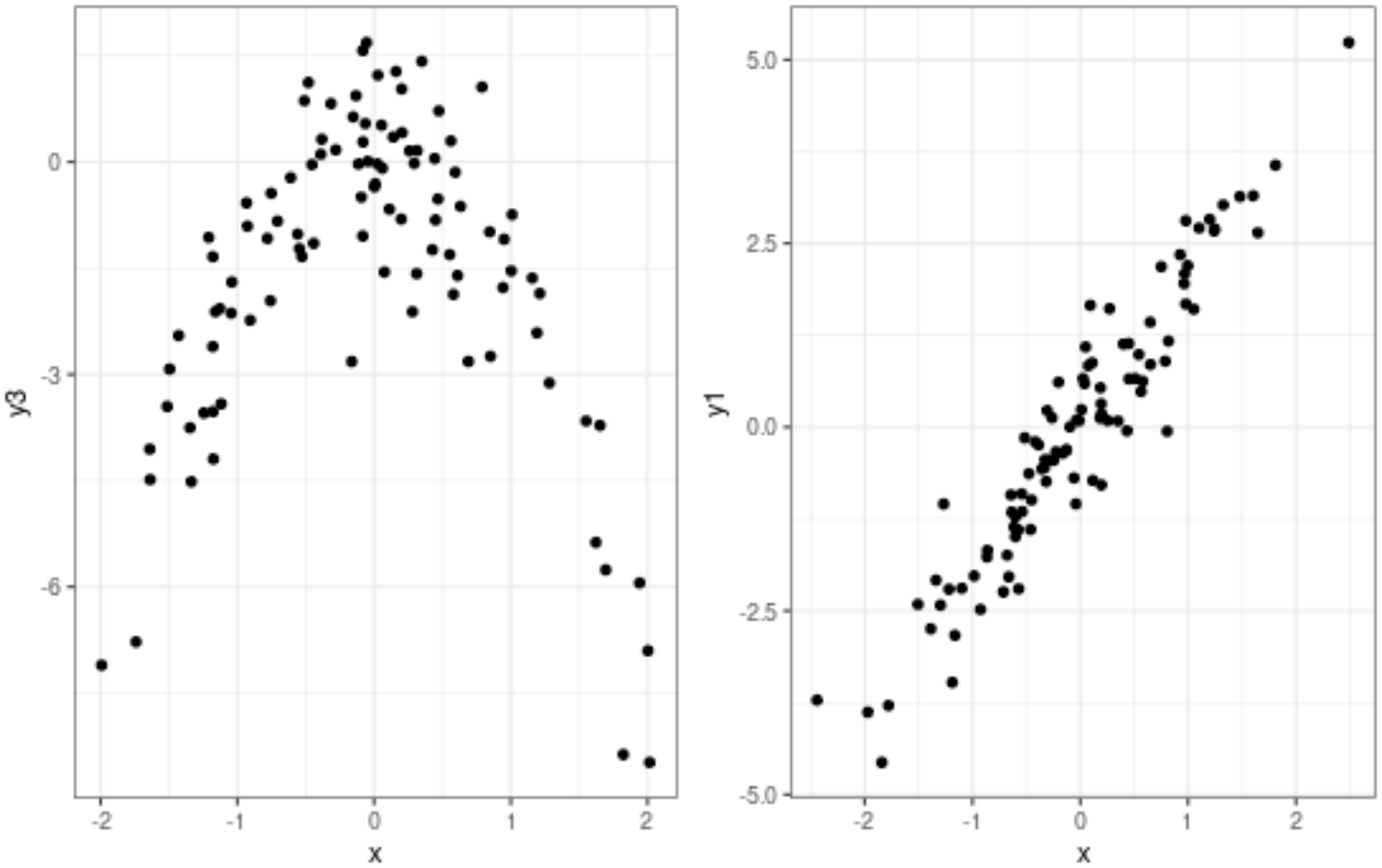
Describing a Scatterplot

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

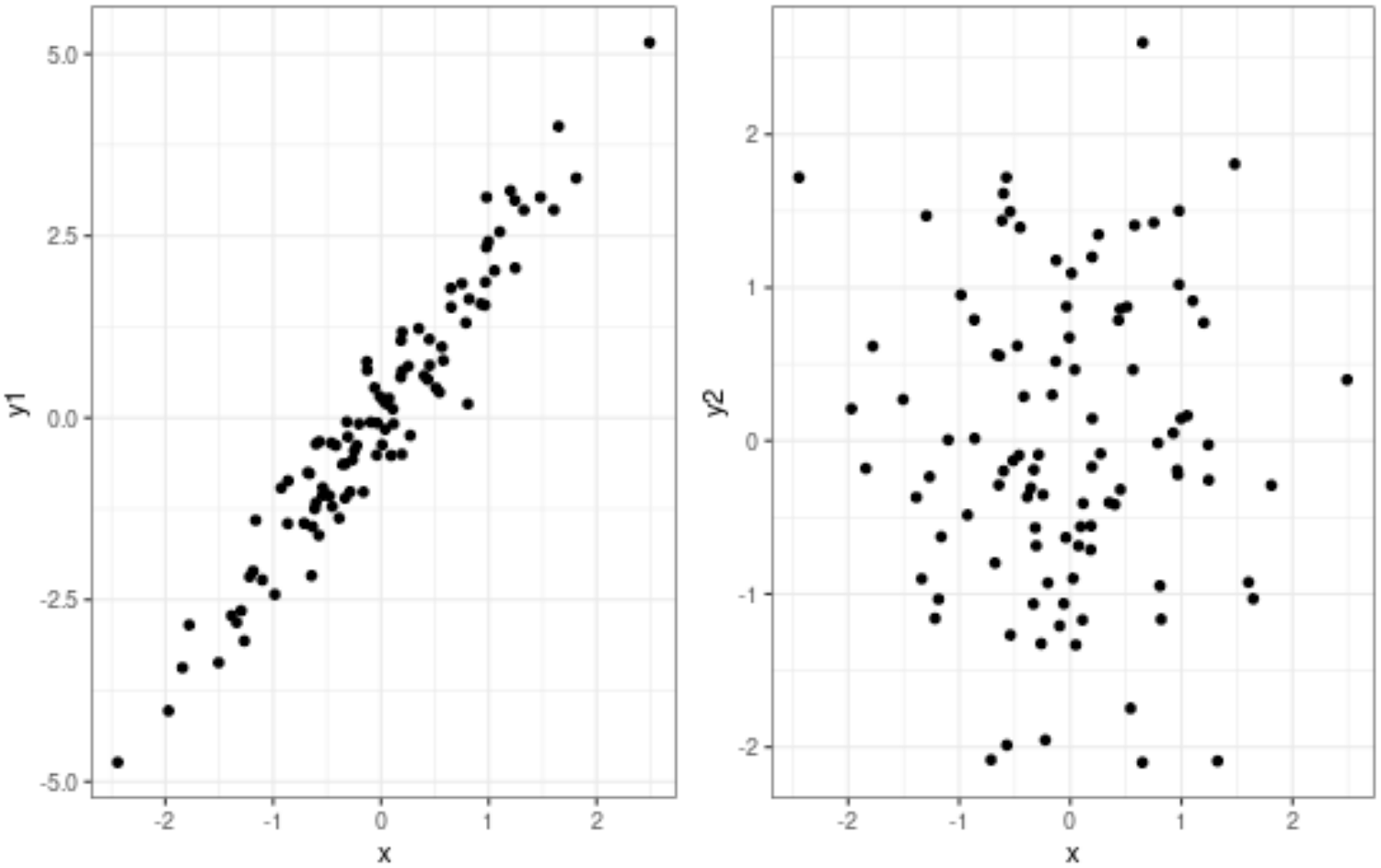
- **Form**: what type of pattern exists (linear / non-linear / curved)
- **Direction**: how the values of one variable relate to the values of the other variable (positive / negative)

Note: later in the course we will come back to scatterplots and talk about how strongly the data fits the pattern

Forms of Quantitative Relationships



Forms of Quantitative Relationships



More on Associations

Just because there is an association between two variables does **not** always mean that one variable is causing a change in another.

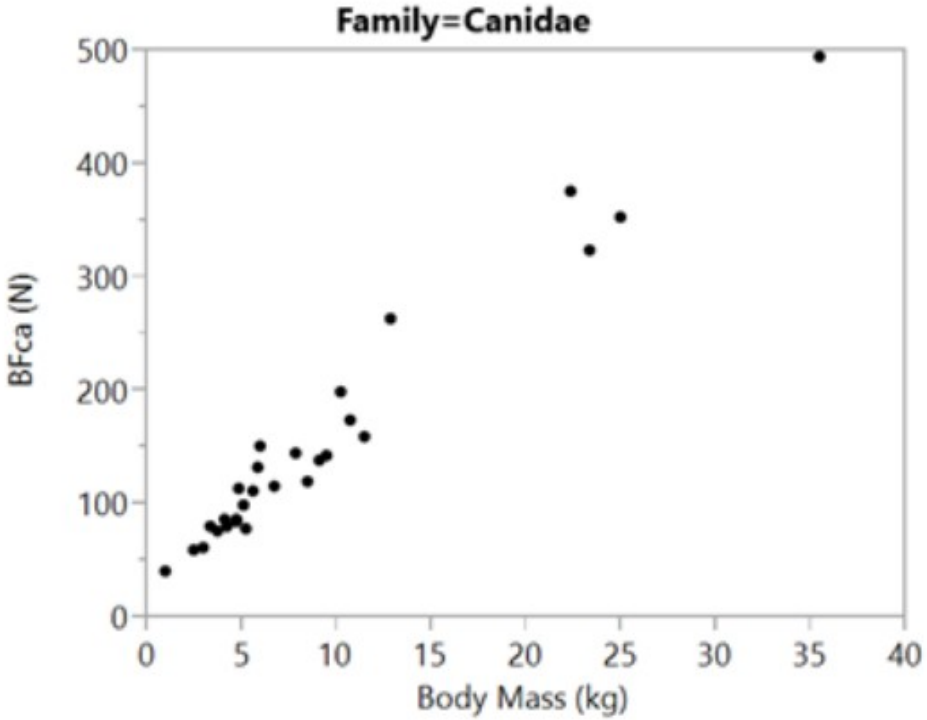
- TVs per household and life expectancy
- ice cream sales and shark attacks

Regardless, sometimes we may make a distinction as to which variable *we may think* affects the other.

The variable we think is causing the change will be called the **explanatory** variable. The variable responding to the other we will call the **response** variable.

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).

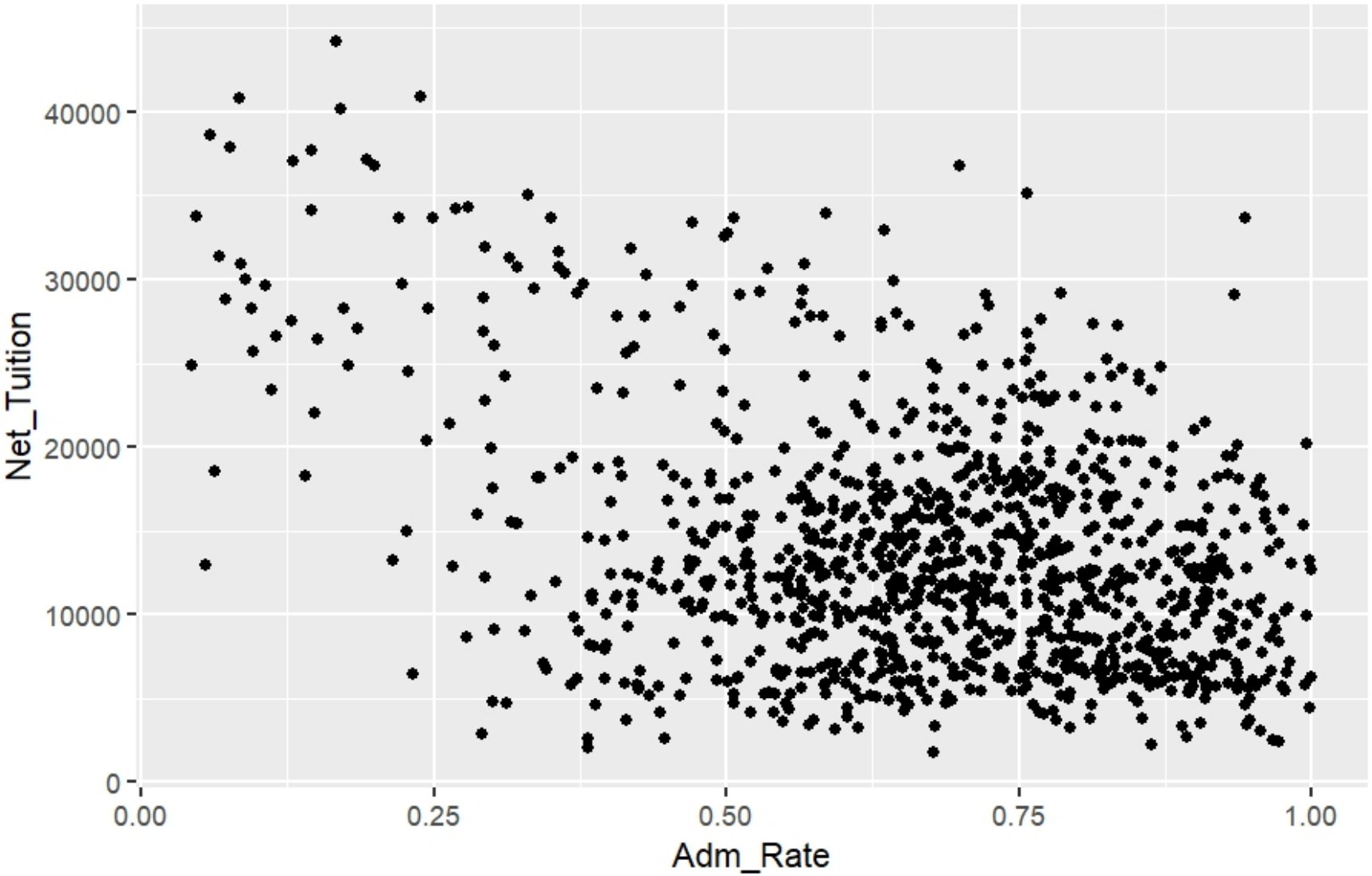


Explanatory:

Response:

Description: The relationship between bite force and body mass for Canidae animals is

Even More Variables – Scatterplot



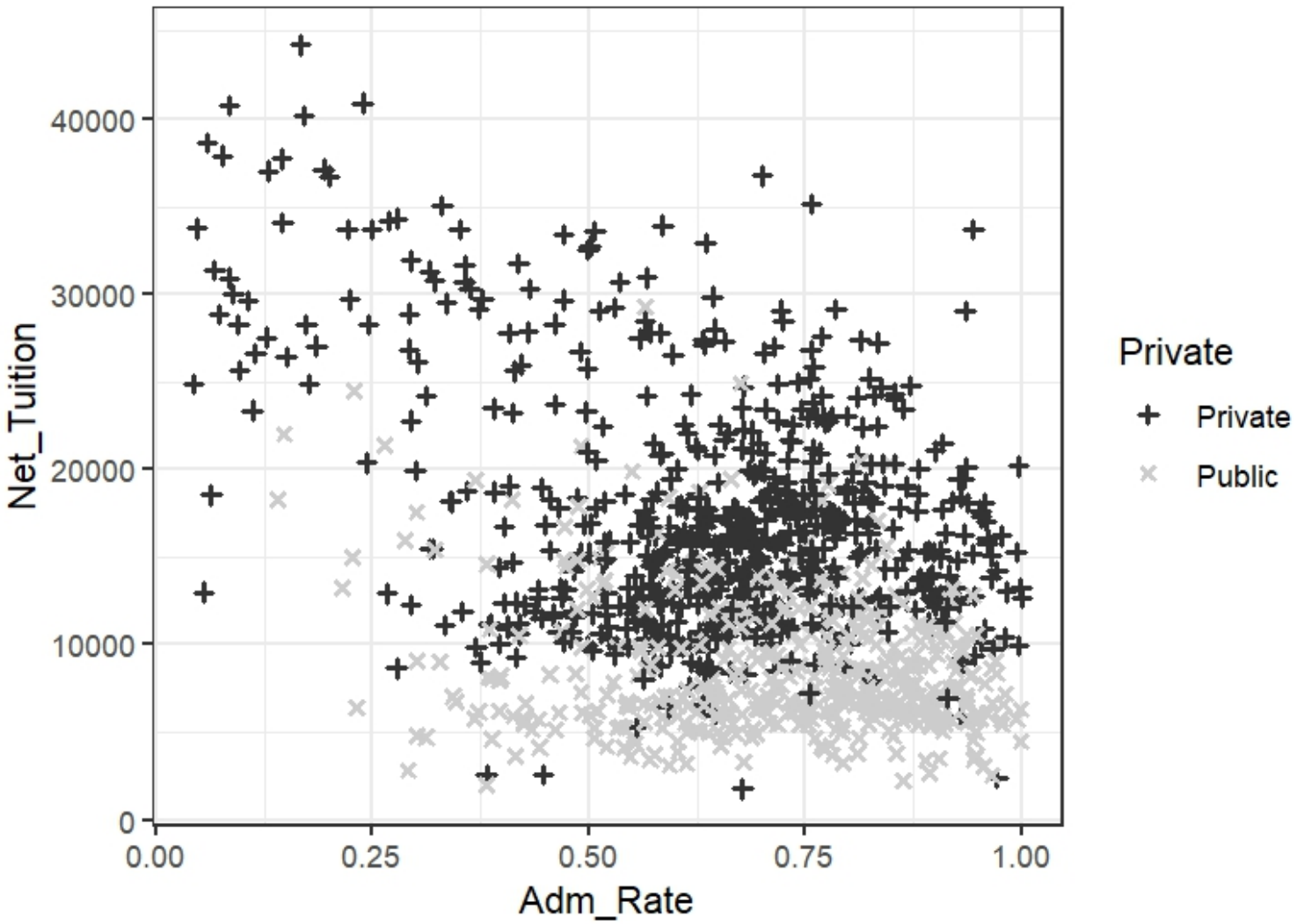
Even More Variables – Scatterplot

Association?

Even More Variables – Scatterplot

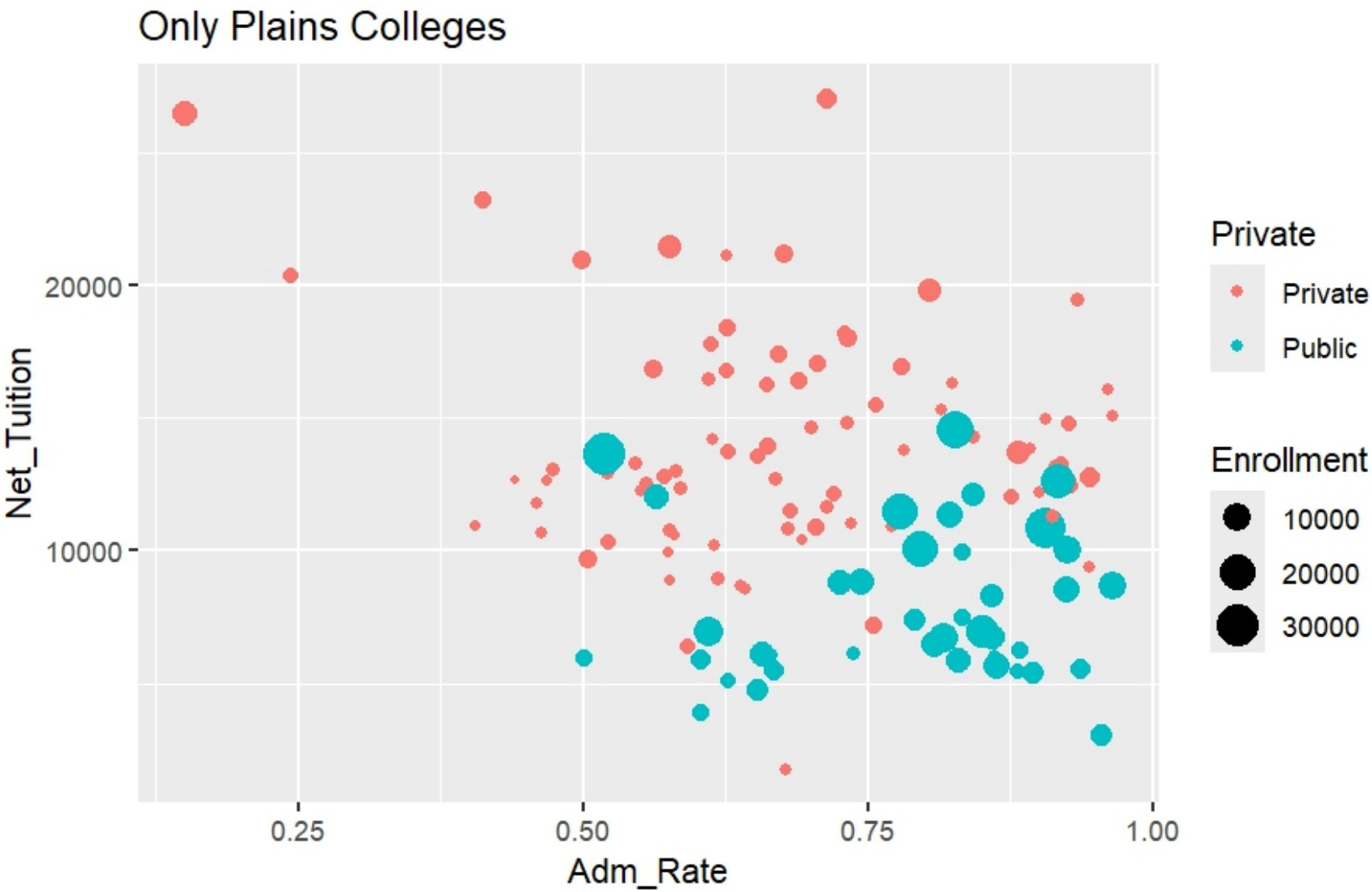
Association?

Even More Variables – Scatterplot



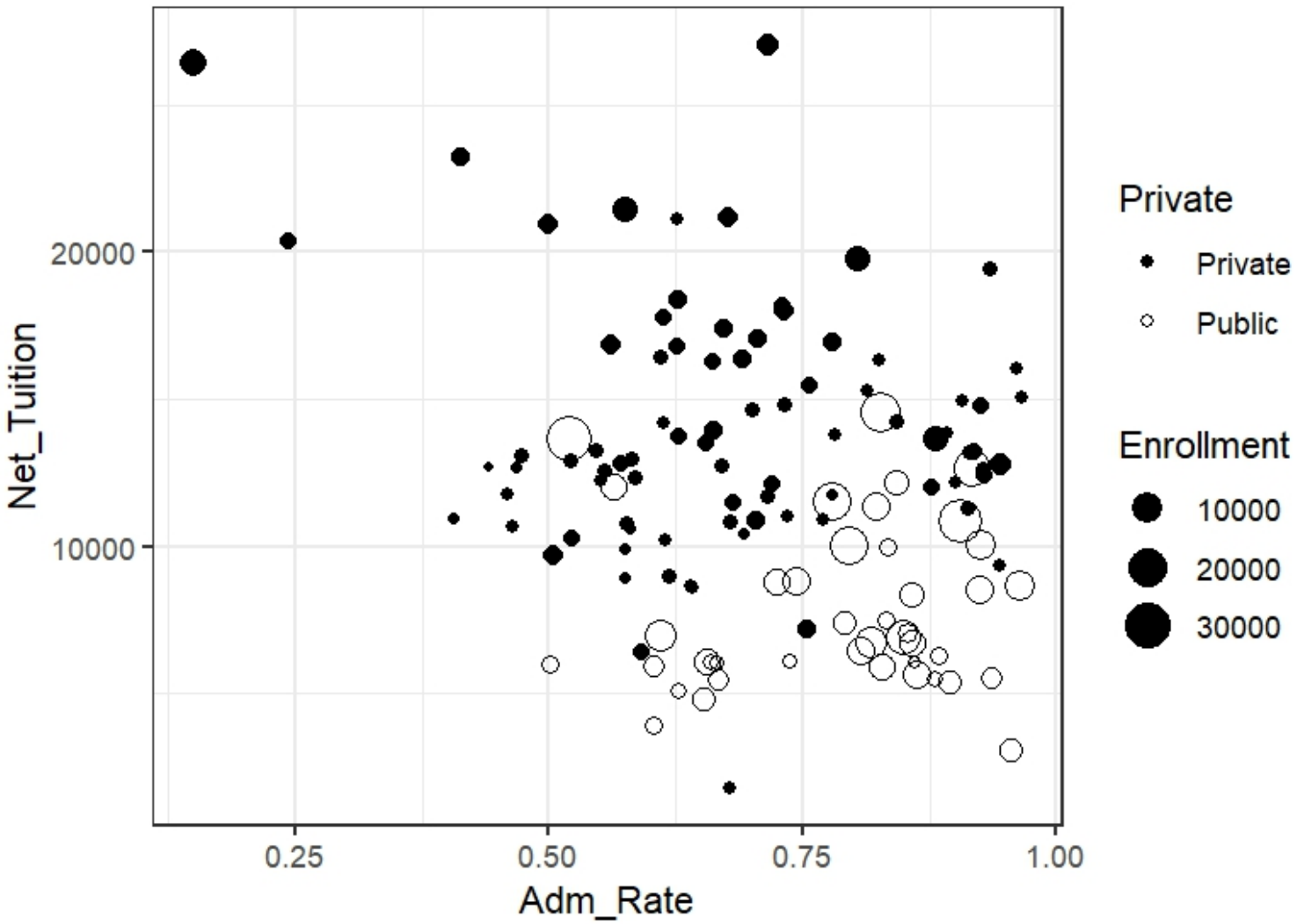
Association?

Even More Variables – Scatterplot



Association?

Even More Variables – Scatterplot



Association?

Even More Variables – Scatterplot

BAD EXAMPLE!!

Even More Variables – Scatterplot

BAD EXAMPLE!!

Even More Variables – Scatterplot

Better?

Reflection

- What determines whether we use the following:
  - mean & standard deviation
  - median & IQR
- What is an association?
- What are some ways in which information for multiple variables were included in our graphs?