Data Visualization Part 2

Grinnell College

February 3, 2025

Review

We looked at lots of ways to display variables

Some of the graphs we saw:

- one categorical variable $\rightarrow$ bar graph
- one quantitative variable $\rightarrow$ histogram
- categorical + categorical $\rightarrow$ stacked, dodge, conditional bar graph
- quantitative + quantitative $\rightarrow$ scatterplot

Review

There is an **association** between variable when knowing about one variable affect what we know about the other

ex) We saw that public colleges tend to have higher admission rates

The **distribution** of a variable is a description of how frequently certain values of that variable show up in the data

Goals for Today

We are going to continue to look at ways to visualize data.

At the end of today you will be able to:

- describe what a 'percentile' is
- explain the parts of a boxplot
- recognize what types of graphs to use when we have Categorical + Quantitative variables

Percentiles

A **percentile** α is a number such that $\alpha\%$ of our (quantitative) observations fall at or below this number when ranked from smallest to largest

Some percentiles have special names. The *median*, for example, is the 50th percentile.

Other notable percentiles include:

1. Minimum
2. 25th percentile or **first quartile** (Q_1)
3. 75th percentile or **third quartile** (Q_3)
4. Maximum

IQR

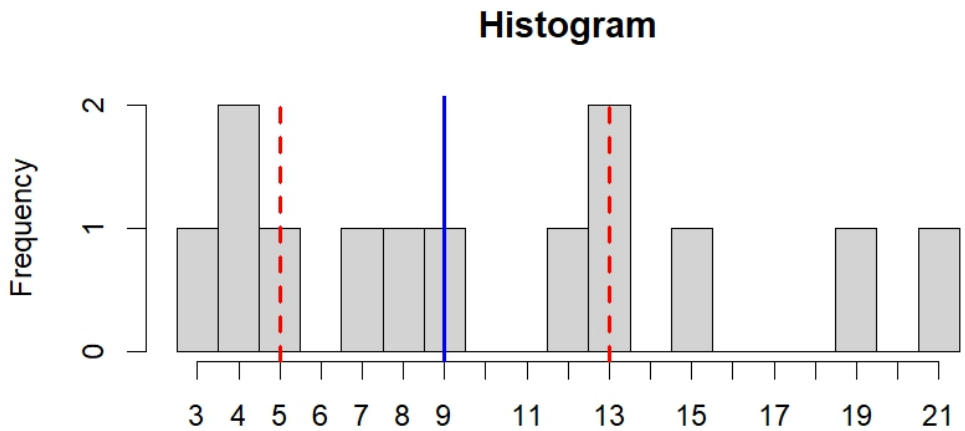
The **interquartile range** or **IQR** is the value of $Q_3 - Q_1$, and gives us the range of the middle 50% of our data

IQR

The **interquartile range** or **IQR** is the value of $Q_3 - Q_1$, and gives us the range of the middle 50% of our data

Data: $\{3, 4, 4, 5, 7, 8, 9, 12, 13, 13, 15, 19, 21\}$

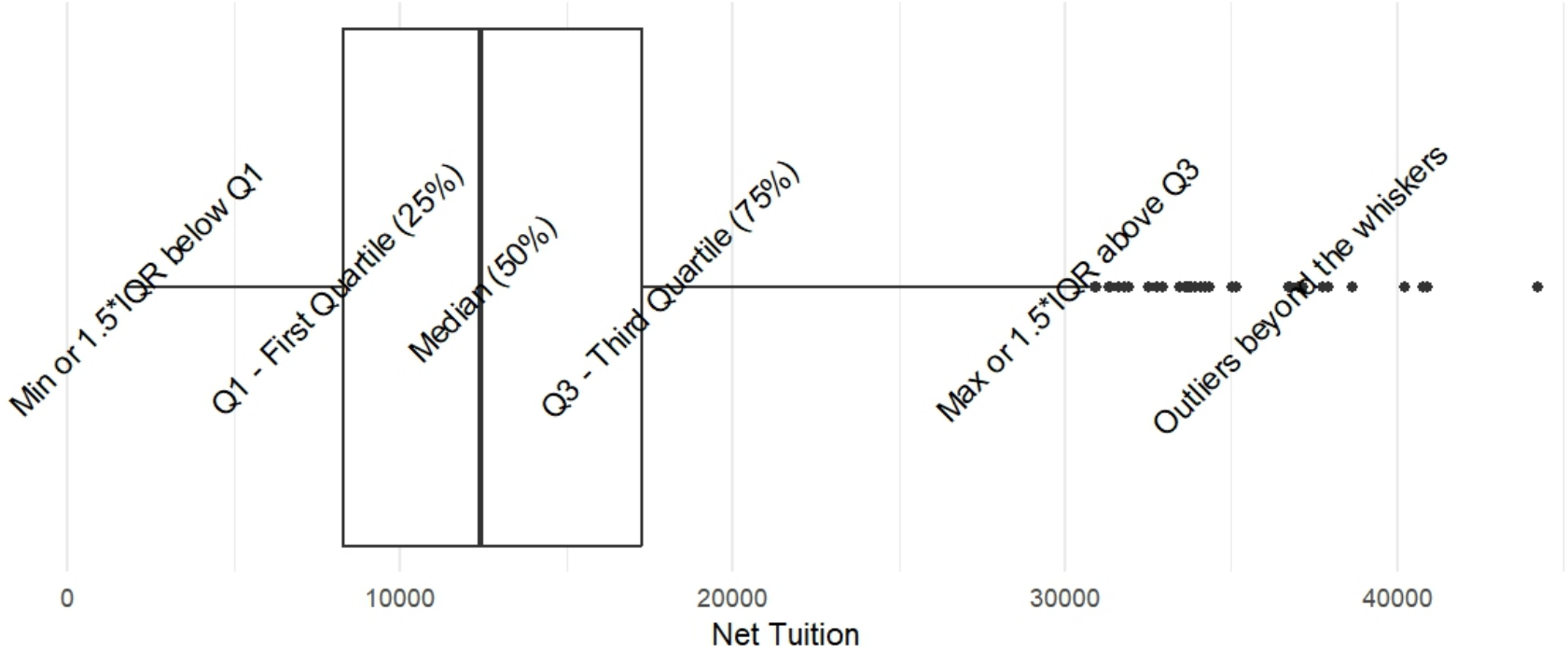
$\{3, 4, 4, 5, 7, 8, 9, 12, 13, 13, 15, 19, 21\} \rightarrow \text{Q1} = 5$
$\{3, 4, 4, 5, 7, 8, 9, 12, 13, 13, 15, 19, 21\} \rightarrow \text{Q3} = 13$
IQR = Q3 - Q3 = 13 - 5 = 8



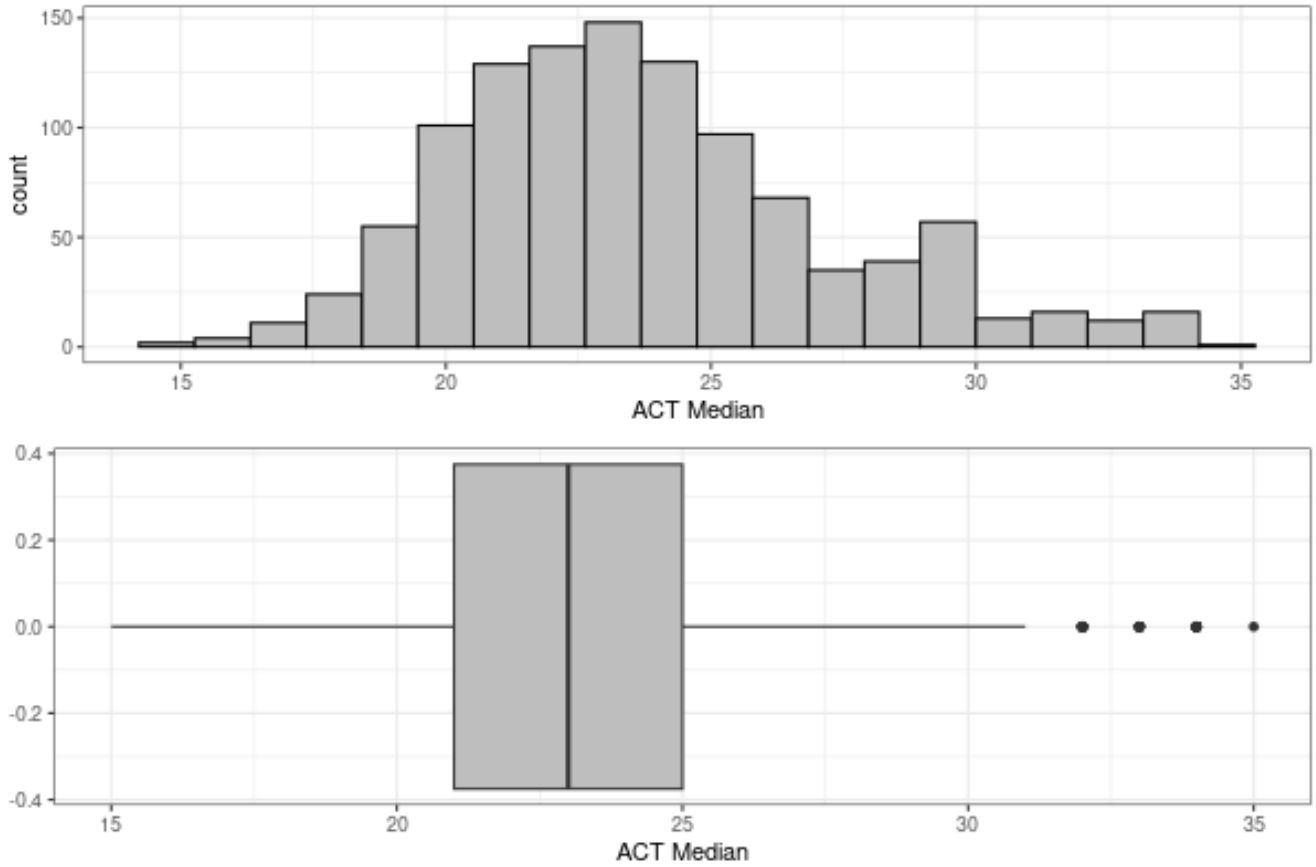
Box plots

A **Box plot** is another way to display a quantitative variable, specifically it displays the 5-number-summary

ex) 2019 College data



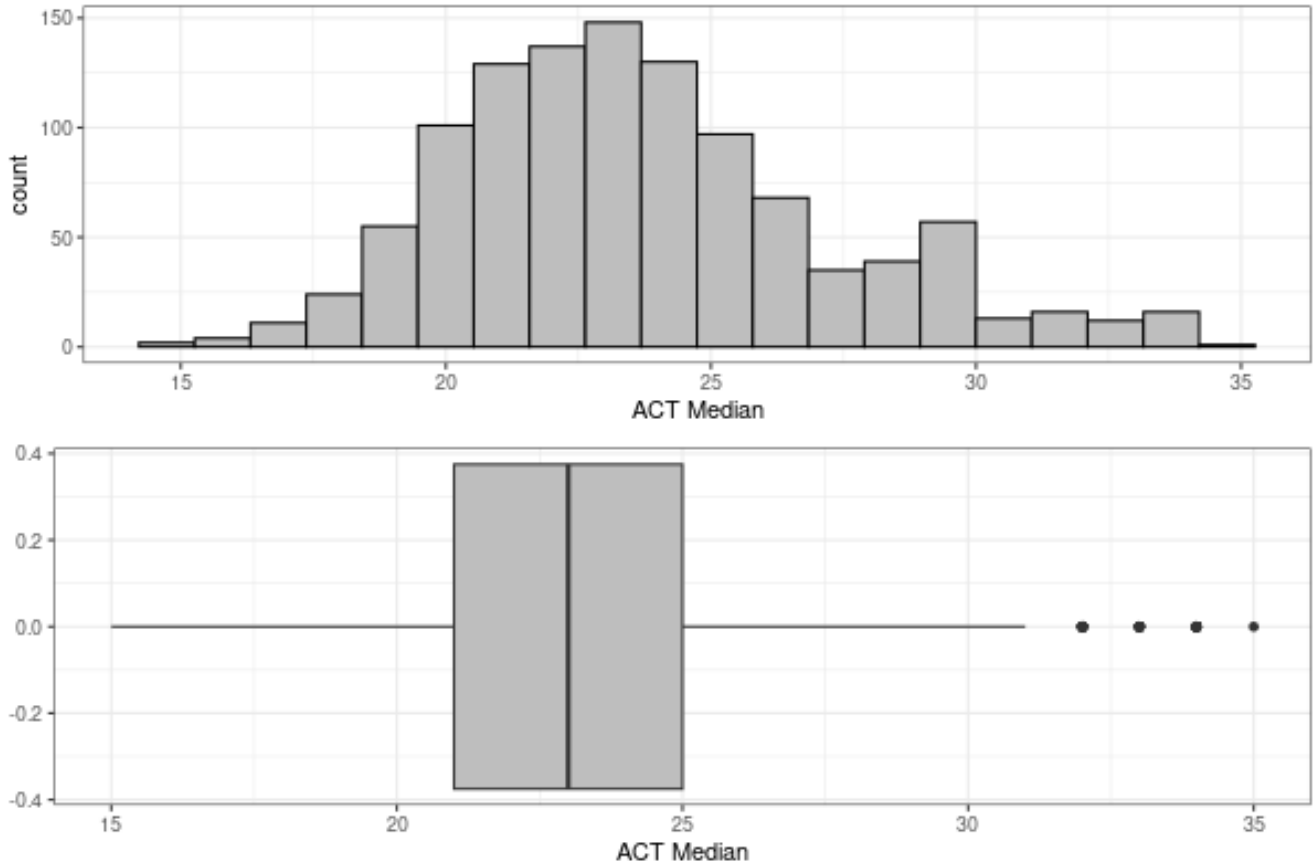
Box plots



Using either will (generally) give us the same distribution description

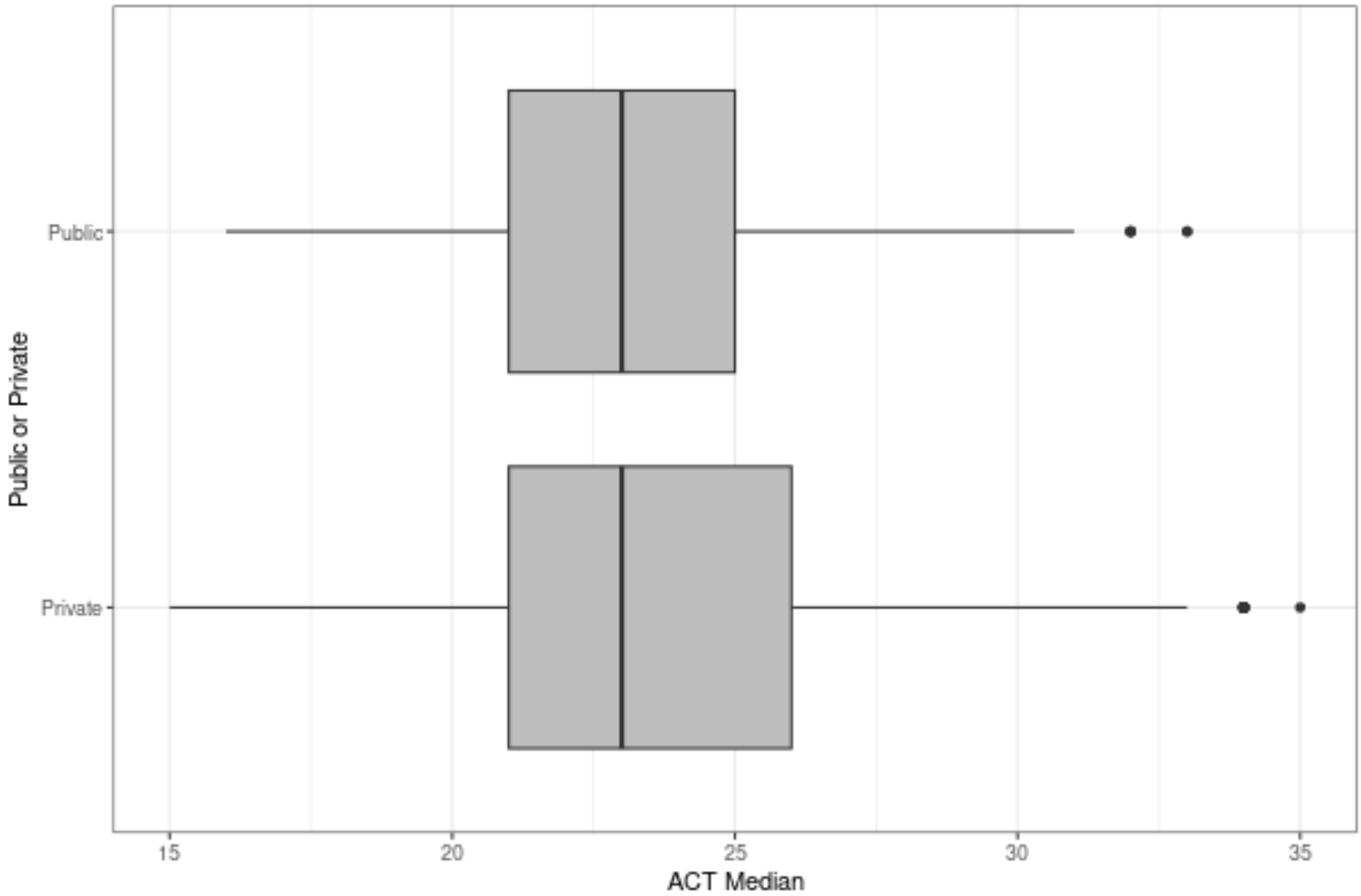
- skew is sometimes harder to describe with boxplots
- outliers classification is different

Box plots

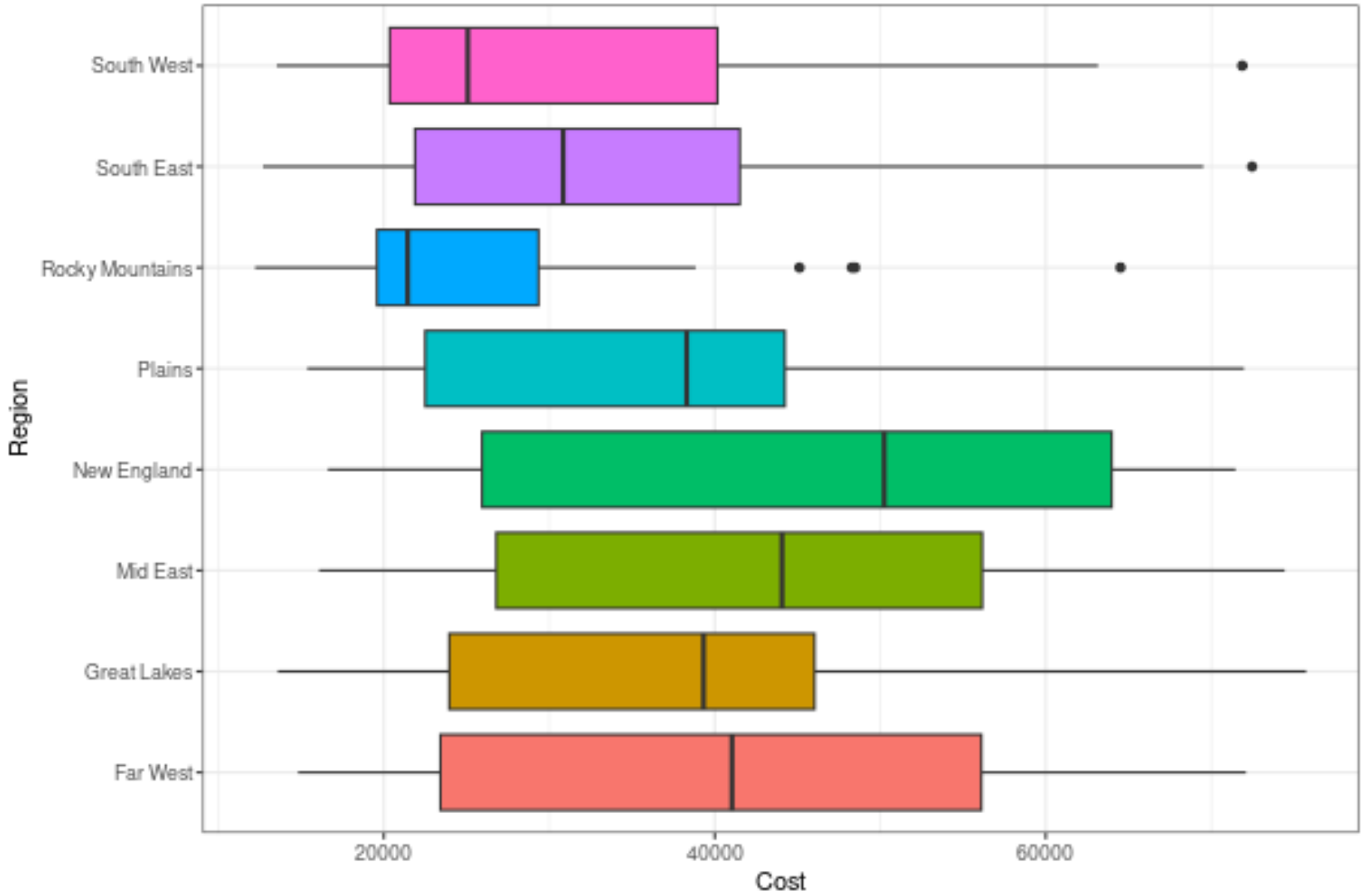



```
> colleges[colleges$Name=="Grinnell College",]$ACT_median
[1] 32
> colleges[colleges$Name=="Iowa State University",]$ACT_median
[1] 25
```

Quantitative + Categorical → Side-by-side Box plots



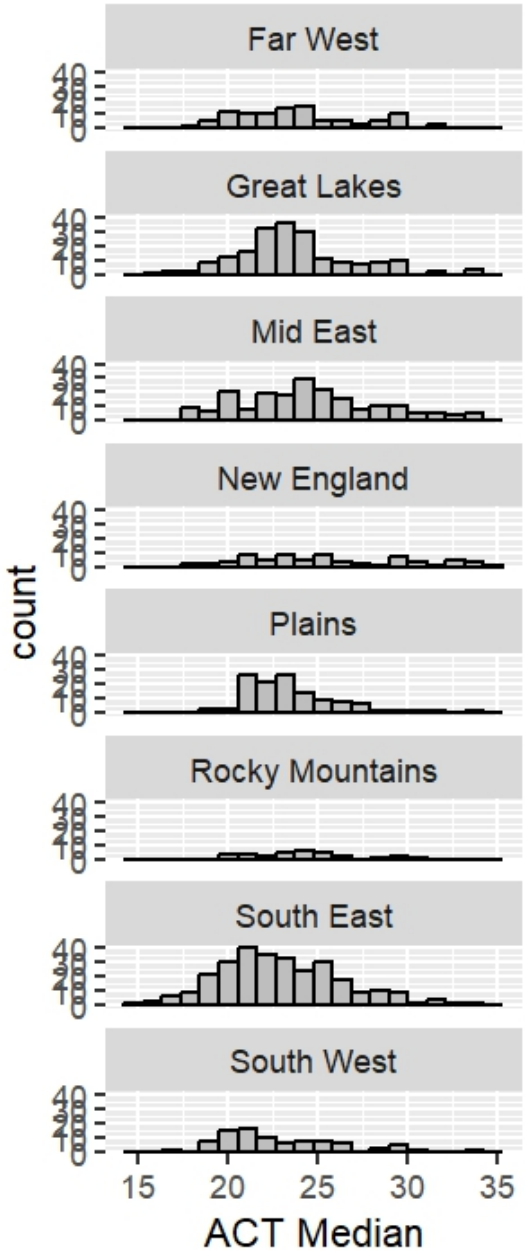
Quantitative + Categorical → Side-by-side Box plots



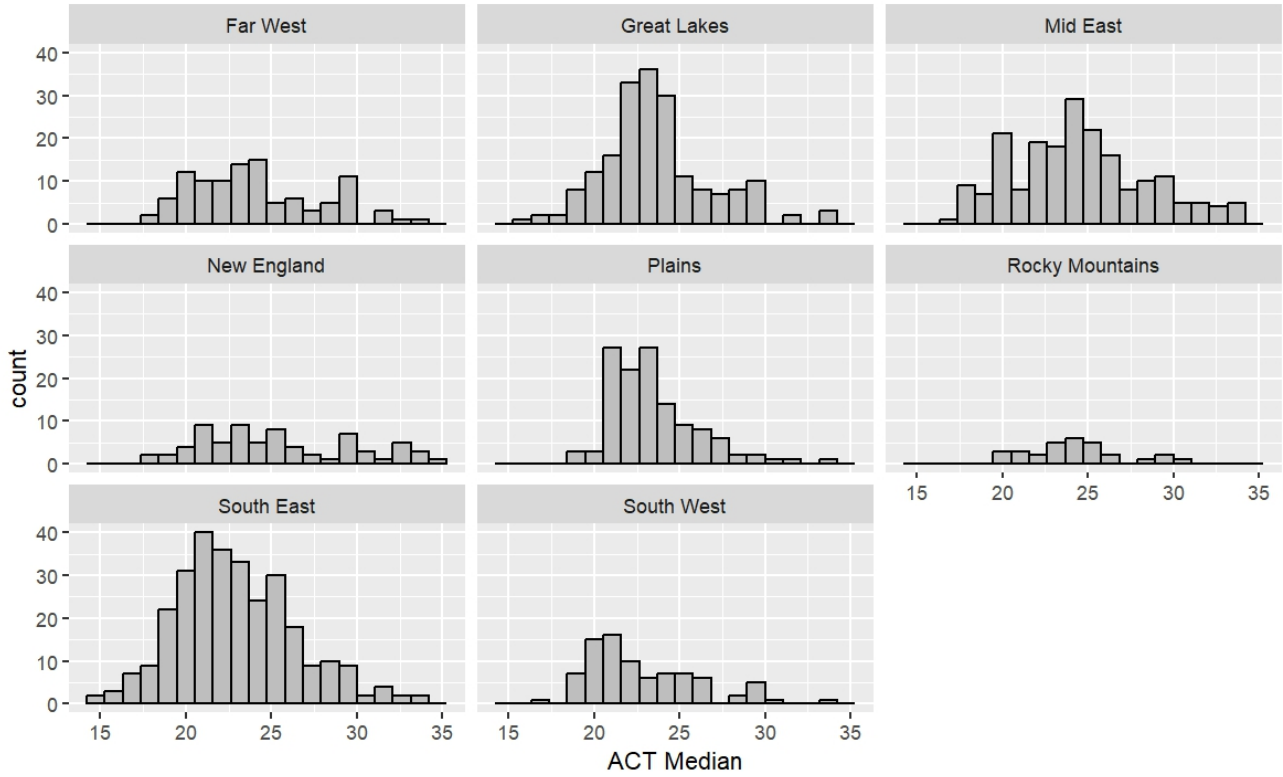
Quantitative + Categorical → Stacked Histograms

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 really hard to read and compare



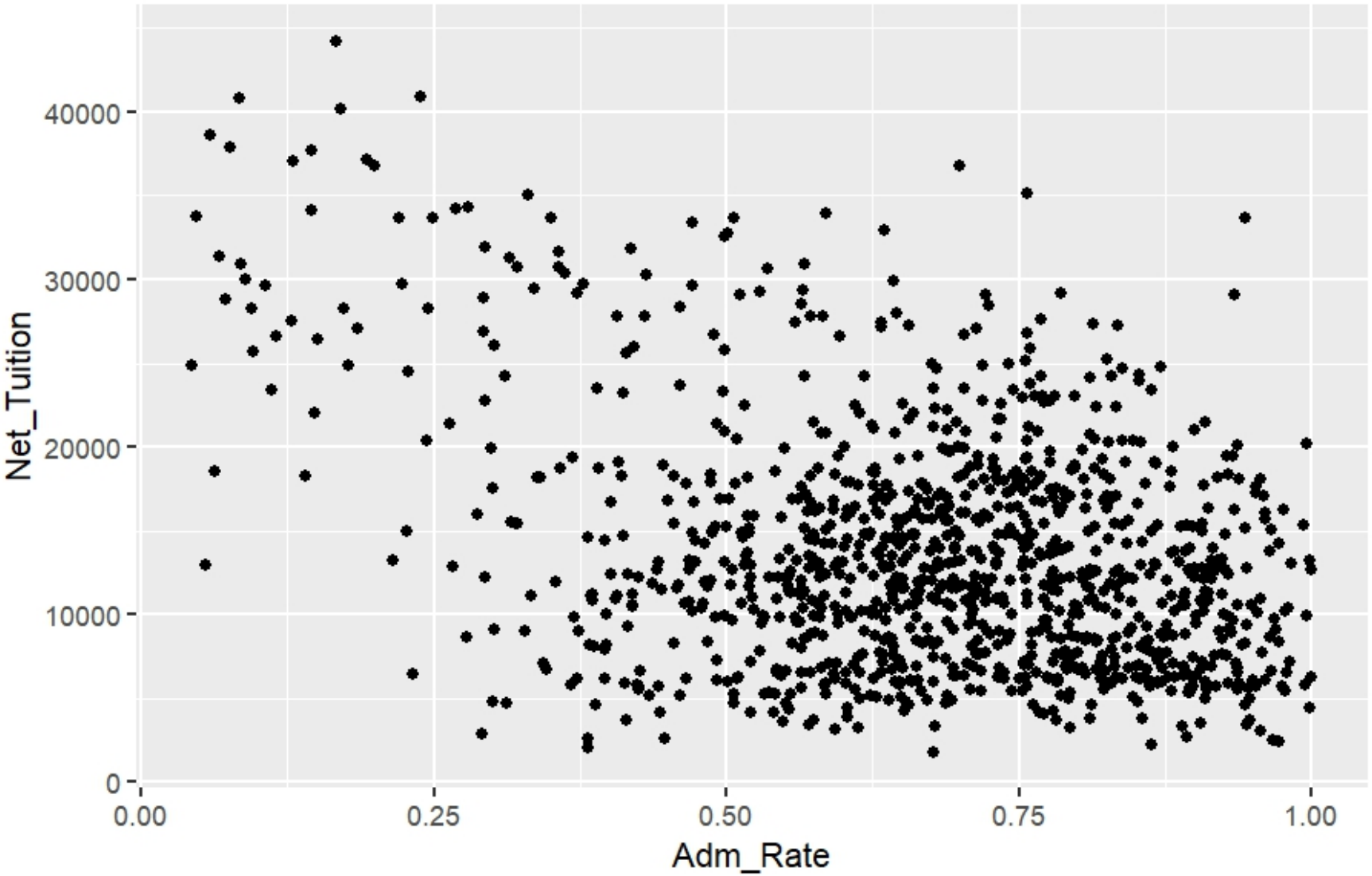
Quantitative + Categorical → Grid of Histograms



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

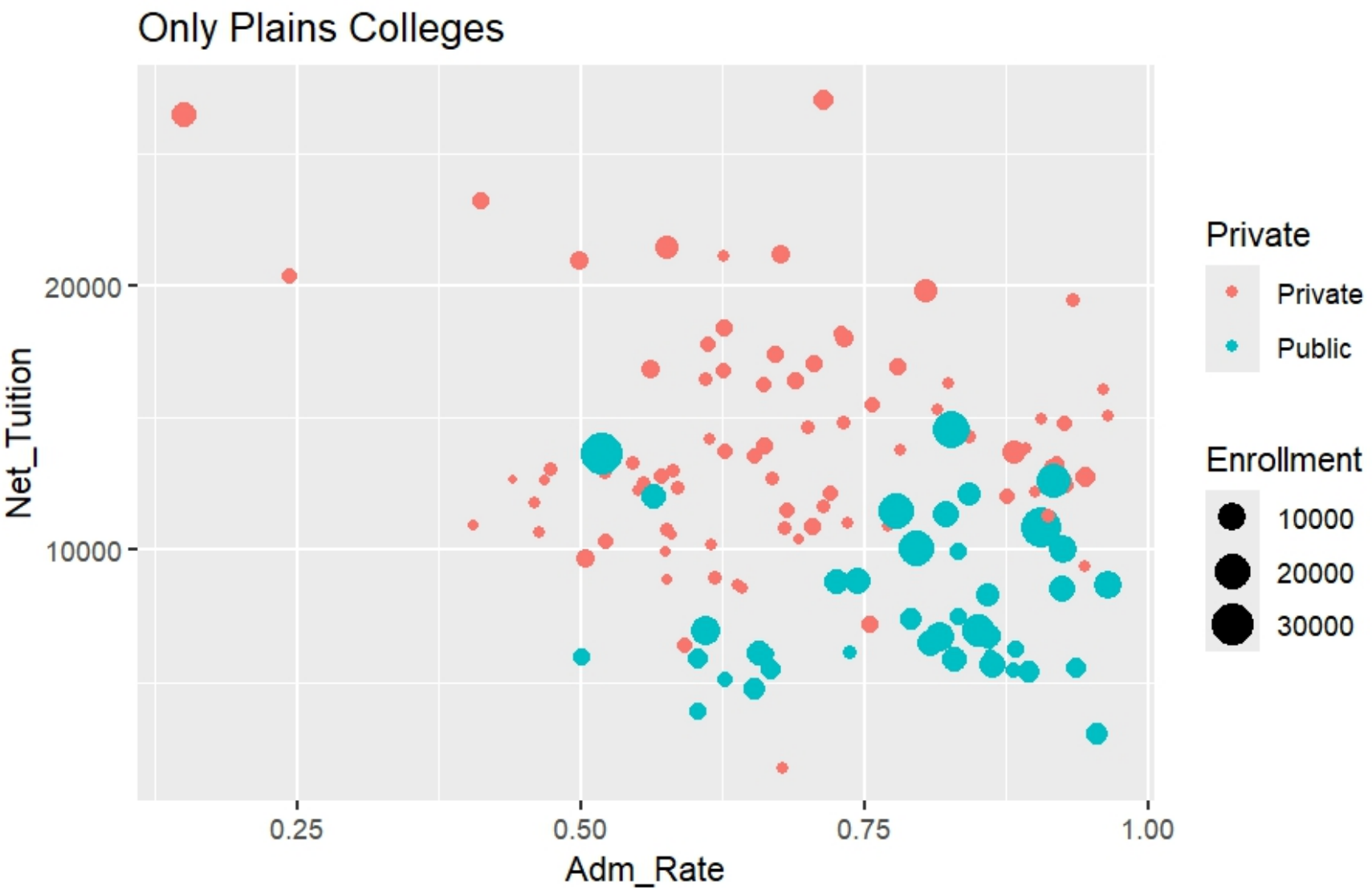
- easier to read the individual histograms
- still harder to compare each group than if we had just used box plots

Even More Variables?!? – Scatterplot



Even More Variables?!? – Scatterplot

Even More Variables?!? – Scatterplot



Even More Variables?!? – Scatterplot

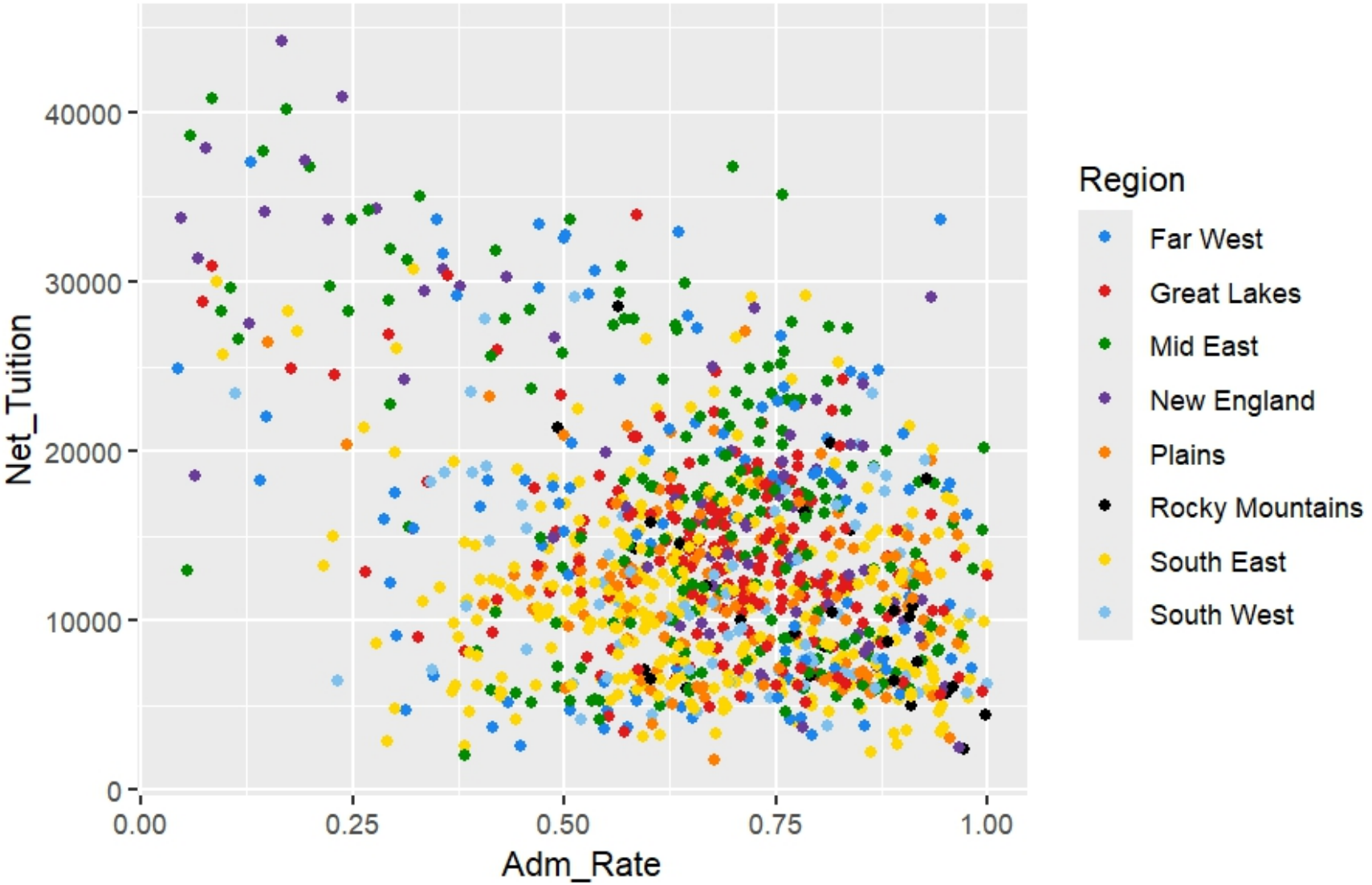
BAD EXAMPLE!!

Even More Variables?!? – Scatterplot

BAD EXAMPLE!!

Even More Variables?!? – Scatterplot

Better?



Review

- What is a percentile?
- What is the IQR, how do we calculate it in terms of Q1 and Q3?
- Is it easier to compare many groups using side-by-side box plots or histograms?

What's Next?

Today we will work on a lab that puts the data visualization information into practice.

Wednesday we will start looking at how to make pretty graphs using an R package called "ggplot2"

The first homework has been assigned $\rightarrow$ see course page

- Similar questions to lab
- population, parameter, sample, statistic, observation
- data visualization basics