

Data Visualization Part 2

Grinnell College

September 8, 2025

We looked at lots of ways to display variables

- one categorical variable \rightarrow bar graph
- one quantitative variable \rightarrow histogram
- categorical + categorical \rightarrow bar graphs
- quantitative + quantitative \rightarrow scatterplot

There is an **association** between two variables when information about one gives information about the other

Eg 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

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

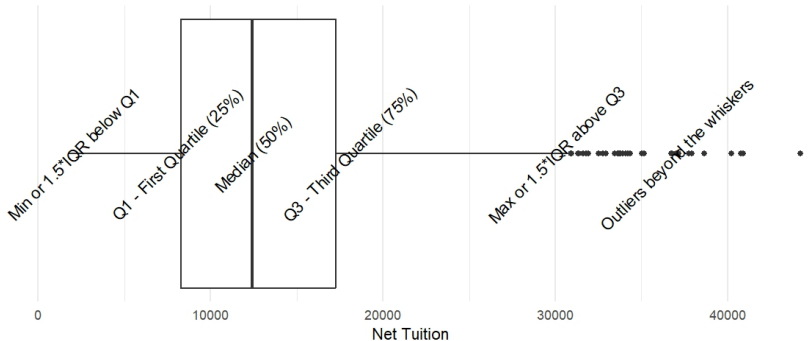
Data: $\{3, 4, 4, 5, 7, 8, 9, 12, 13, 13, 15, 19\}$

- Q_1 is the median of the lower 50% of the data
 $\{3, 4, 4, 5, 7, 8, | 9, 12, 13, 13, 15, 19\} \rightarrow Q_1 = 4.5$
- Q_3 is the median of the upper 50% of the data
 $\{3, 4, 4, 5, 7, 8, | 9, 12, 13, 13, 15, 19\} \rightarrow Q_3 = 13$

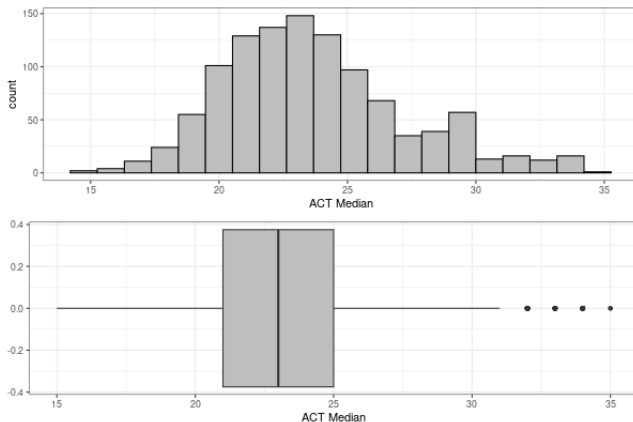
Box plots

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

ex) 2019 College data



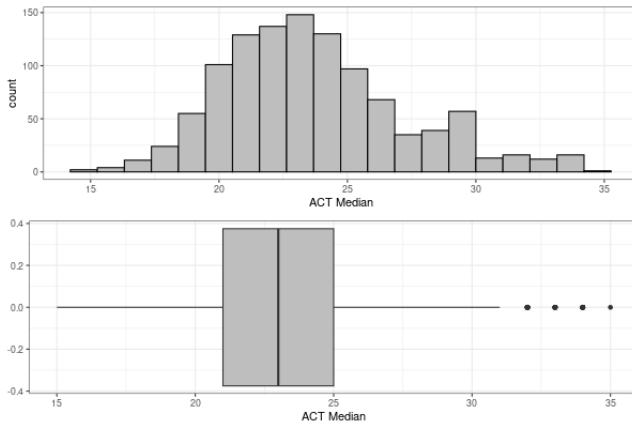
Box plots



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

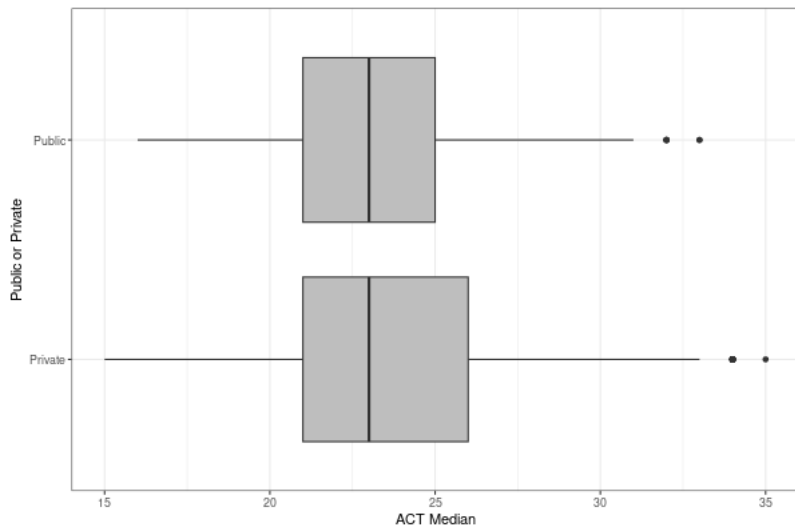
- skew is hard to describe with boxplots
- outliers classification is different (boxplots flag more outliers, usually)
- Boxplots are useless for number of modes

Box plots



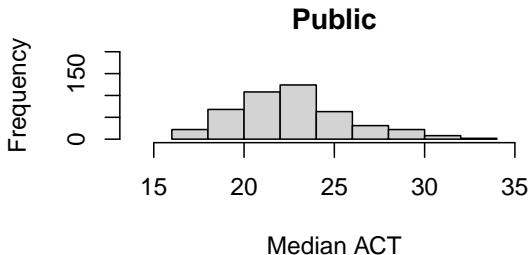
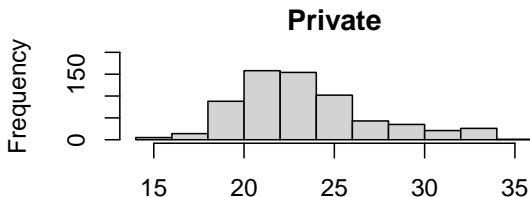
```
> colleges[colleges$Name=="Grinnell College",]$ACT_median  
[1] 32  
> colleges[colleges$Name=="Iowa State University",]$ACT_median  
[1] 25
```

Quantitative + Categorical → Side-by-side Box plots

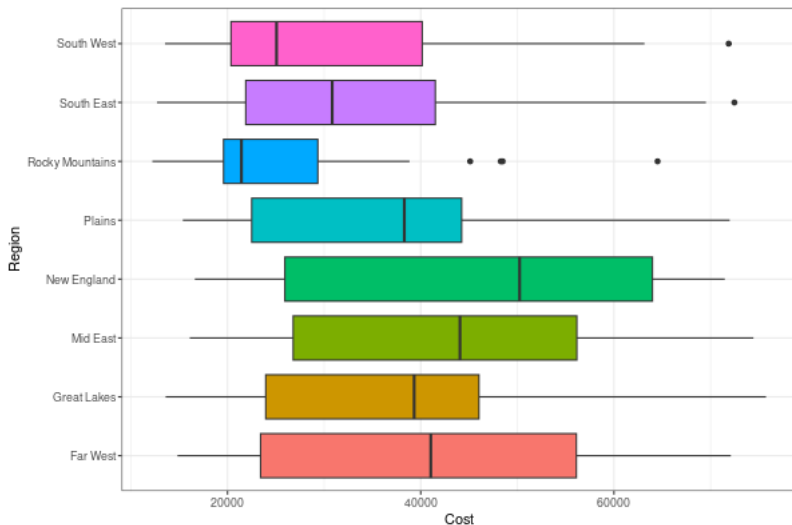


Alternatively....

Histograms work when there are few categories



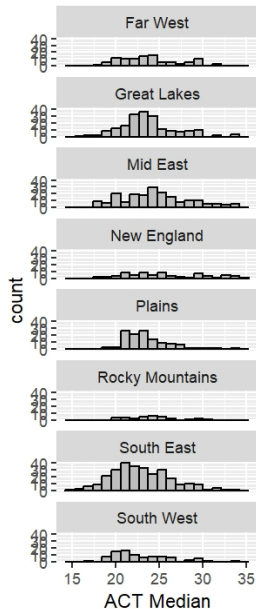
Quantitative + Categorical → Side-by-side Box plots



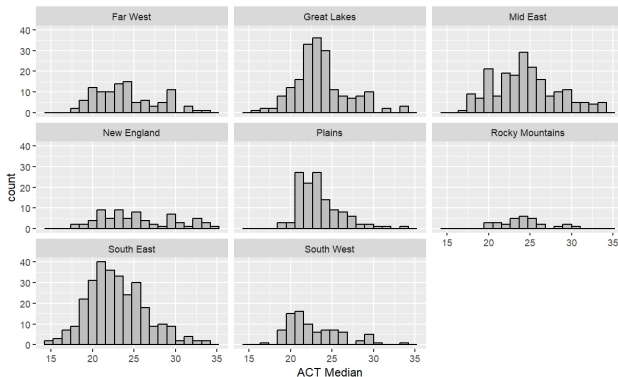
Quantitative + Categorical → Stacked Histograms

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



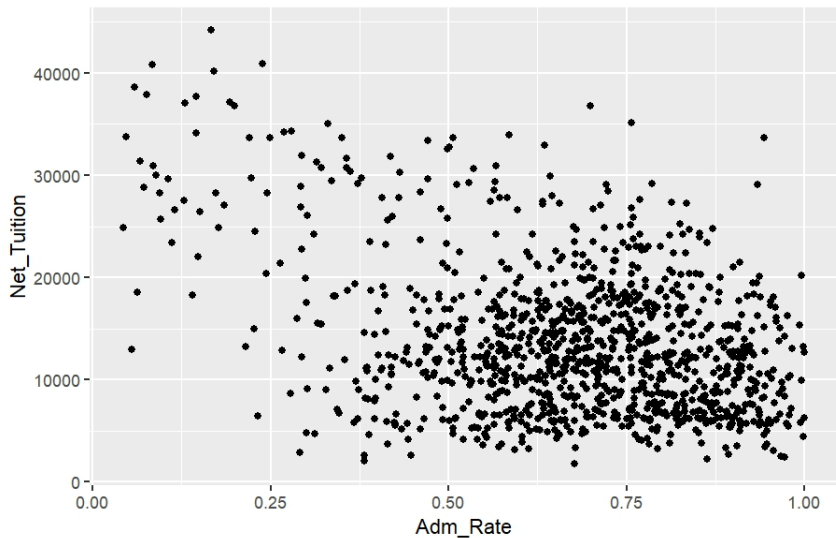
Quantitative + Categorical → Grid of Histograms



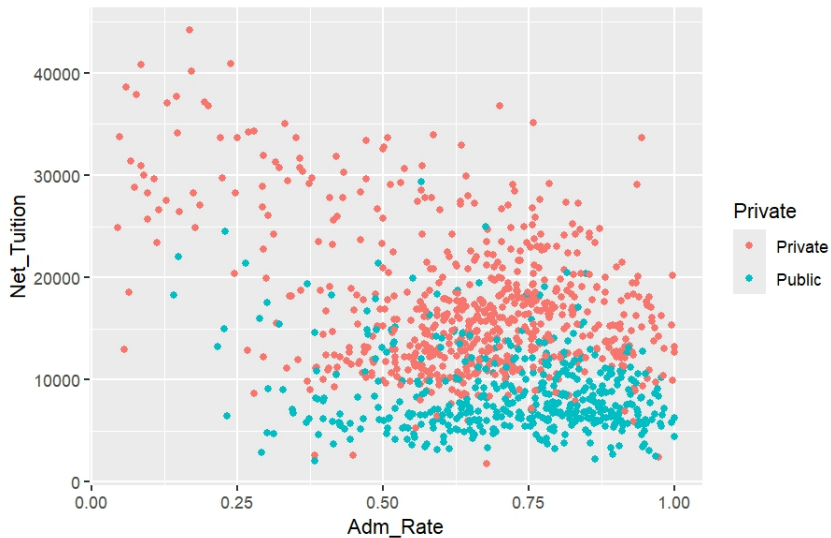
This does not have a special name that I know of... but is another way to display many histograms.

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

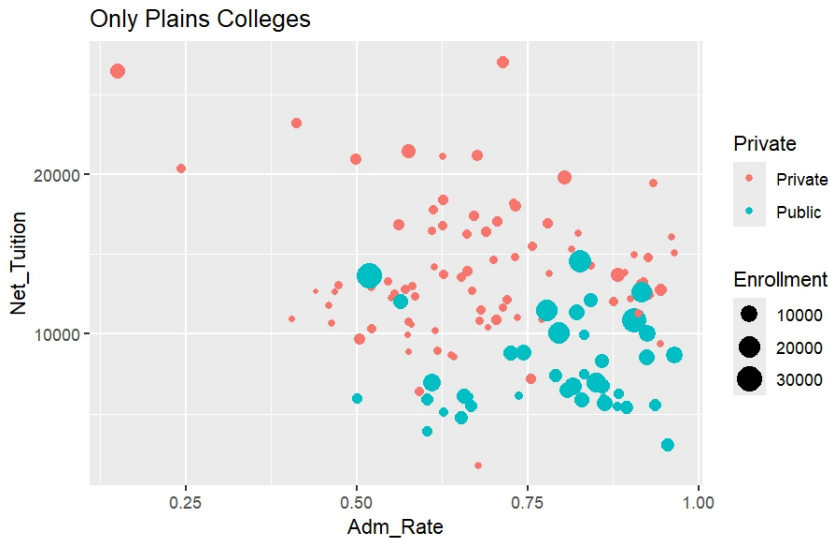
Even More Variables?!? – Scatterplot



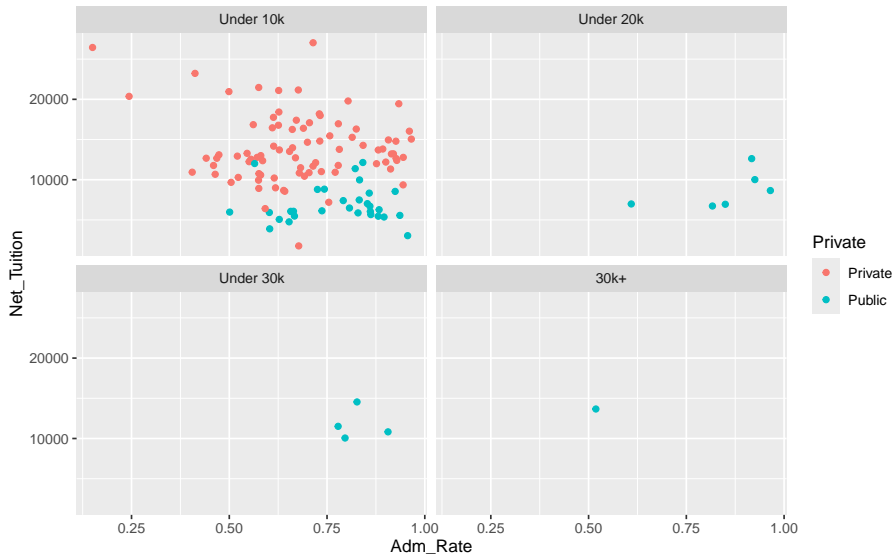
Even More Variables?!? – Scatterplot



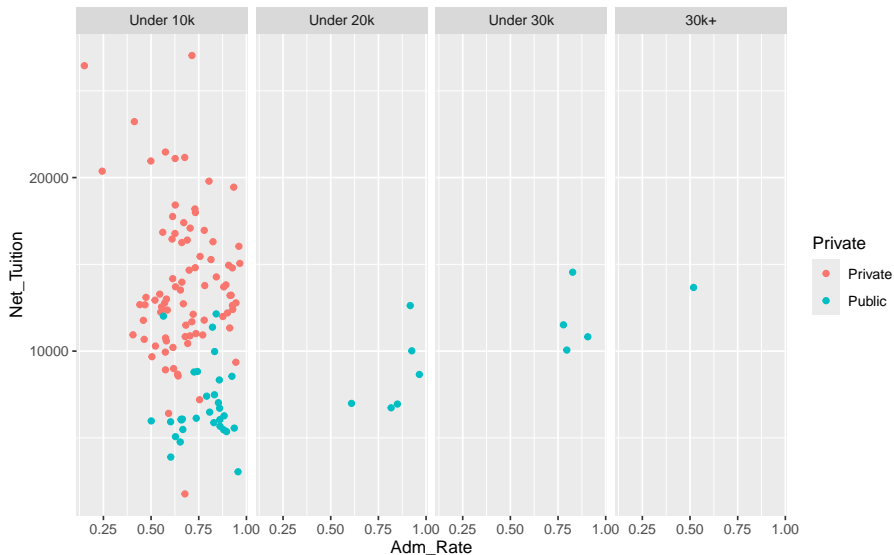
Even More Variables?!? – Scatterplot



Even More Variables?!? – Scatterplot

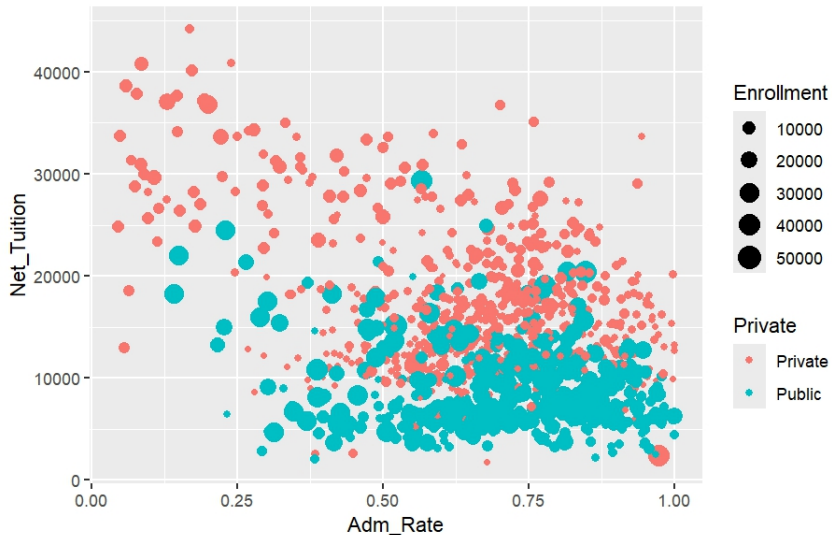


Even More Variables?!? – Scatterplot



Even More Variables?!? – Scatterplot

BAD EXAMPLE!!



Even More Variables?!? – Scatterplot

BAD EXAMPLE!!



Even More Variables?!? – Scatterplot

Better?



- 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?

Work on a lab (time permitting)

Wednesday will be notes on how to make these and possibly data visualization part 3 (accessibility)

First quiz is one week from today