Data visualization

Matt Bhagat-Conway

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

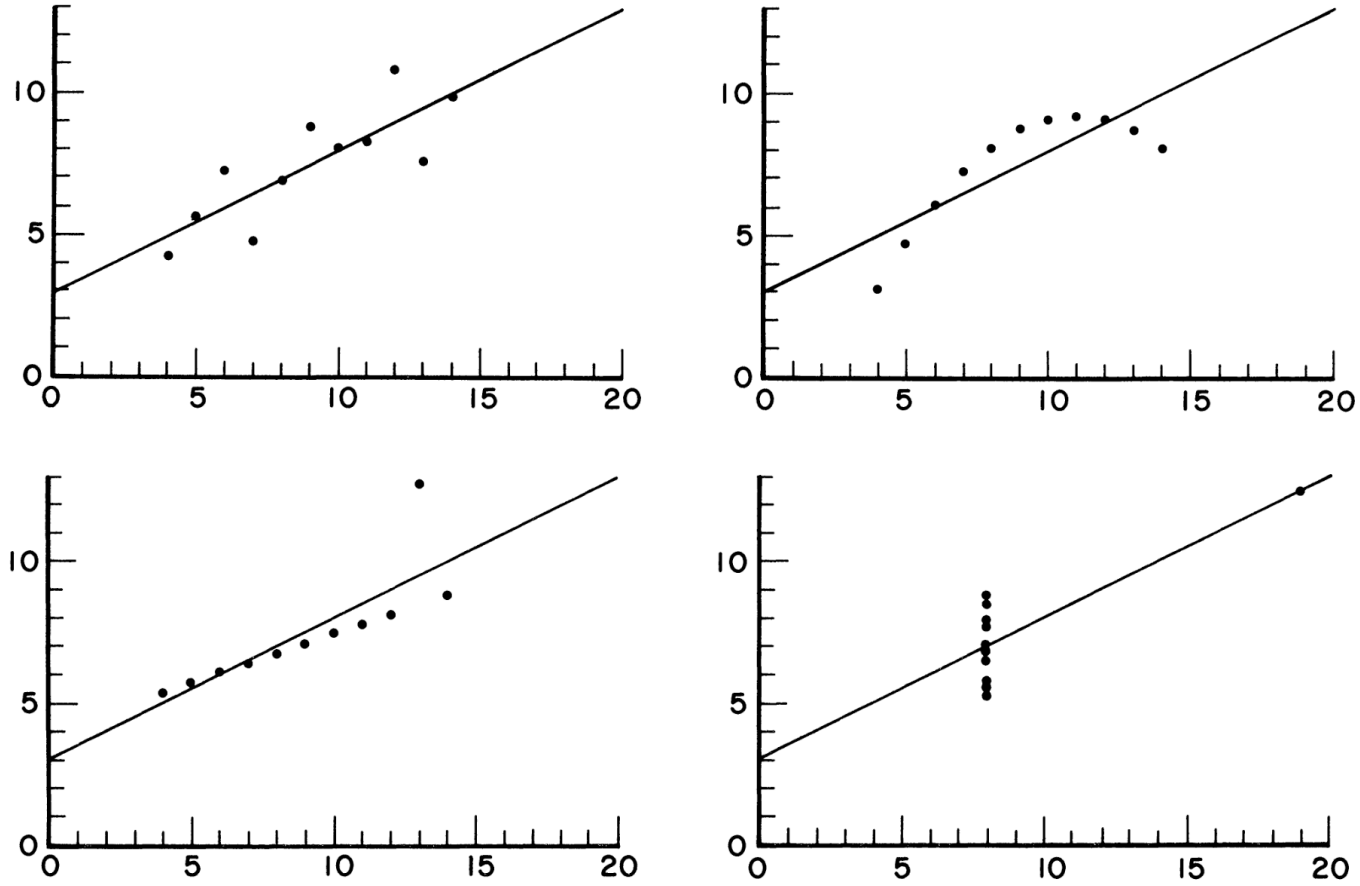
Why visualize data

- Many people are visual thinkers
- Data visualization can remove the need to understand exact numbers and show the big picture

The power of data visualization

- Data visualization helps us make sense of large datasets
- We can make conclusions and hypotheses that would be difficult from looking at the data alone
- But, we can also mislead and misdirect

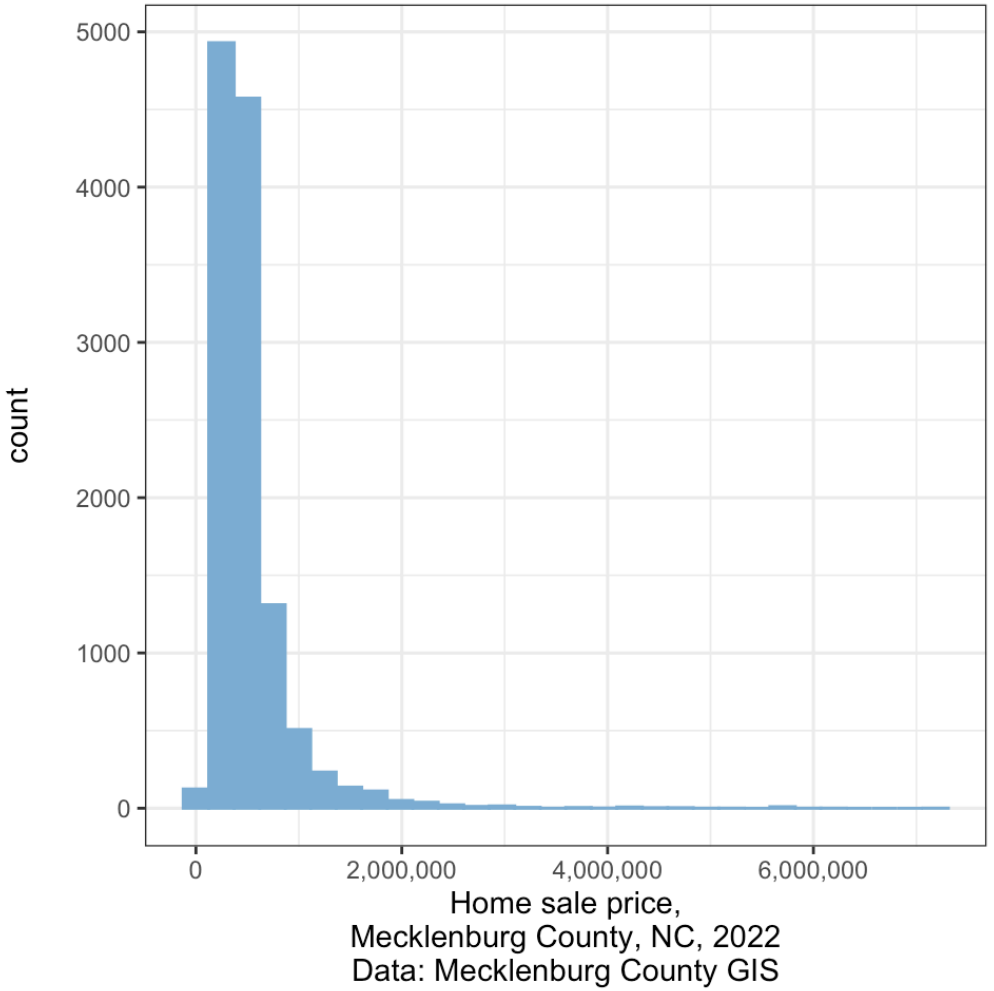
The power of data visualizations: Anscombe's quartet

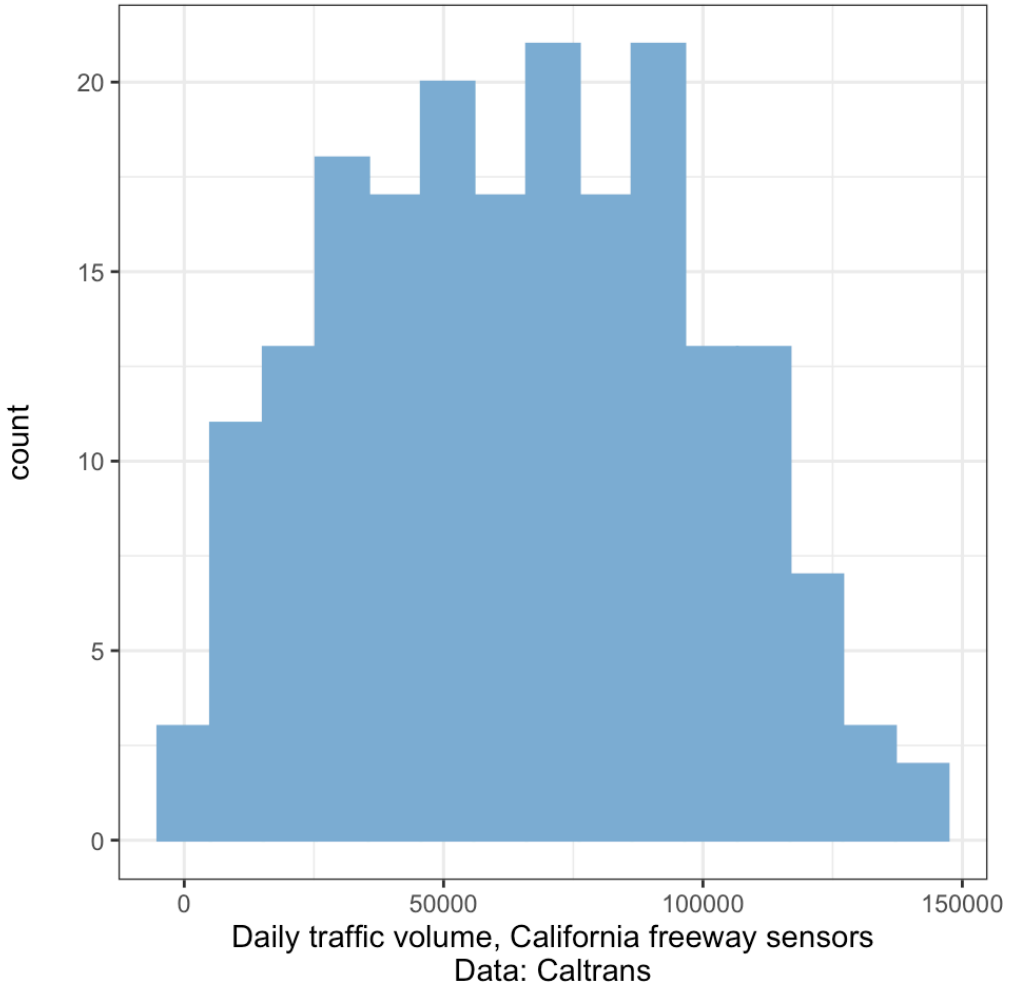


Anscombe (1973)

Histograms

- A histogram visualizes a univariate (one variable) distribution





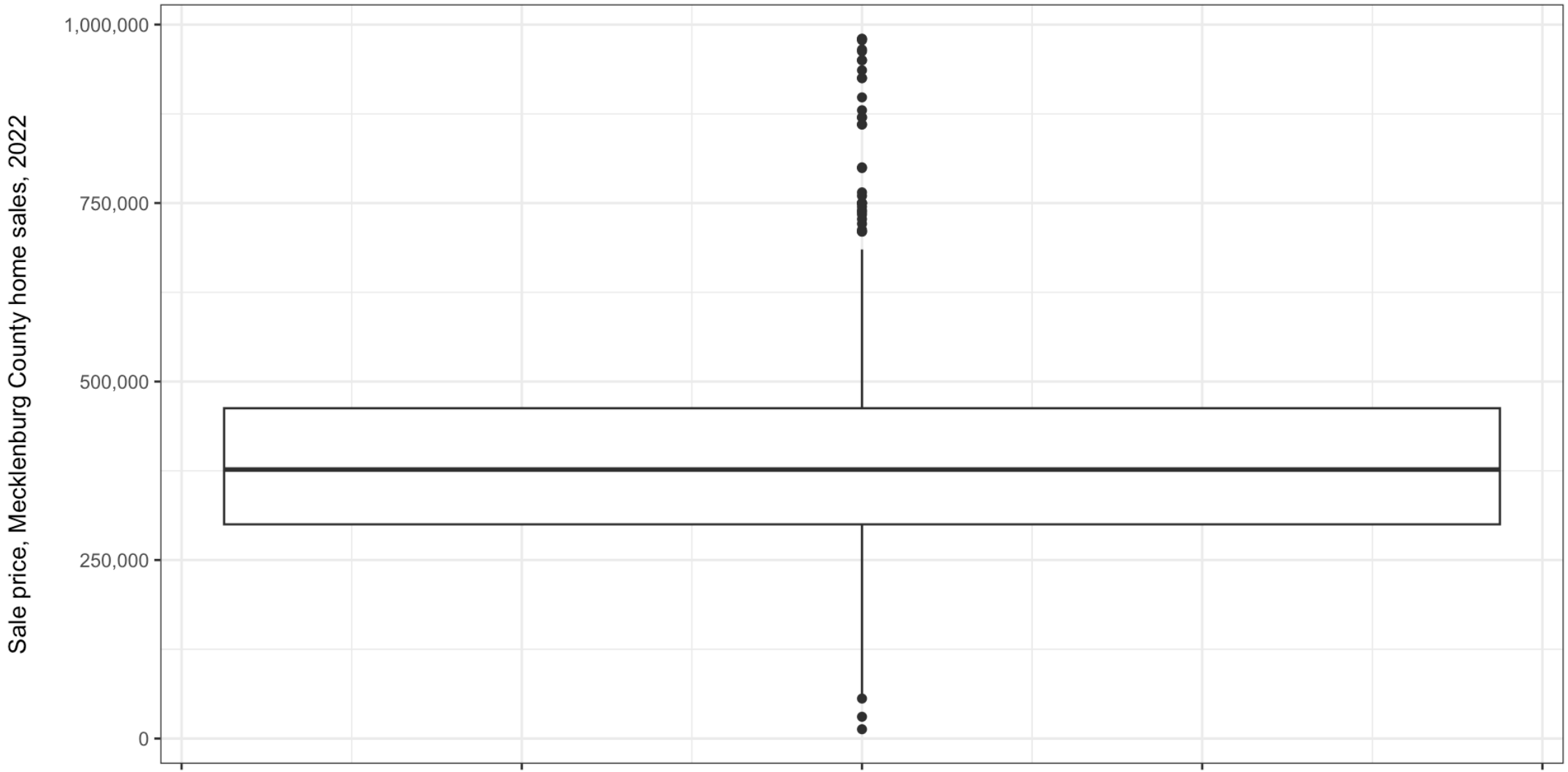
Creating a histogram in Excel

- Select the column you want to create a histogram of
- Choose Insert -> Charts -> Statistical -> Histogram
- By double-clicking on the bars you can edit the number of bins, bin size, how outliers are handled
- Histograms should not have gaps between the bars even though they do by default in Excel
  - After double-clicking on the bars, choose the Fill and Line tab (paint can icon) and set the border to “solid line” to fill the gaps

Boxplots/box and whisker plots

- Box plots visualize data by showing the median, 25th and 25th percentiles, tails, and outliers
- The center line is the median, the top of the box is the 75th percentile, and the bottom of the box is the 25th percentile
- Whiskers *generally* extend to largest/smallest data value less than 1.5 $\times$ interquartile range from the ends of the box
  - Sometimes 5th/95th percentiles
- Points beyond these are plotted individually as outliers

Boxplots/box and whisker plots

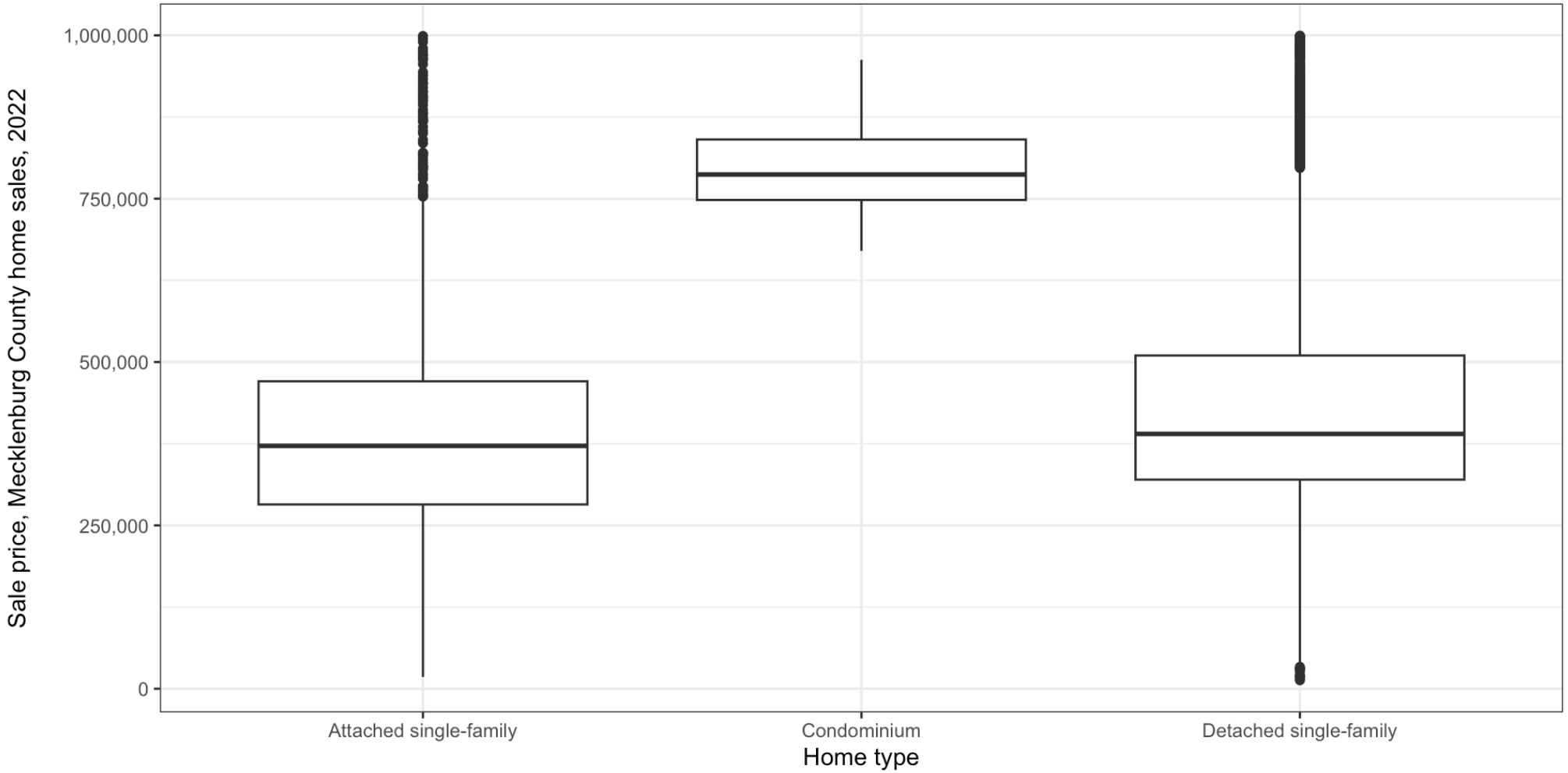


Making a boxplot in Excel

- Select the data you want to create a boxplot for
- Insert -> Charts -> Statistical -> Box and Whisker

Multiple boxplots

- Often, you will see multiple boxplots presented next to one another

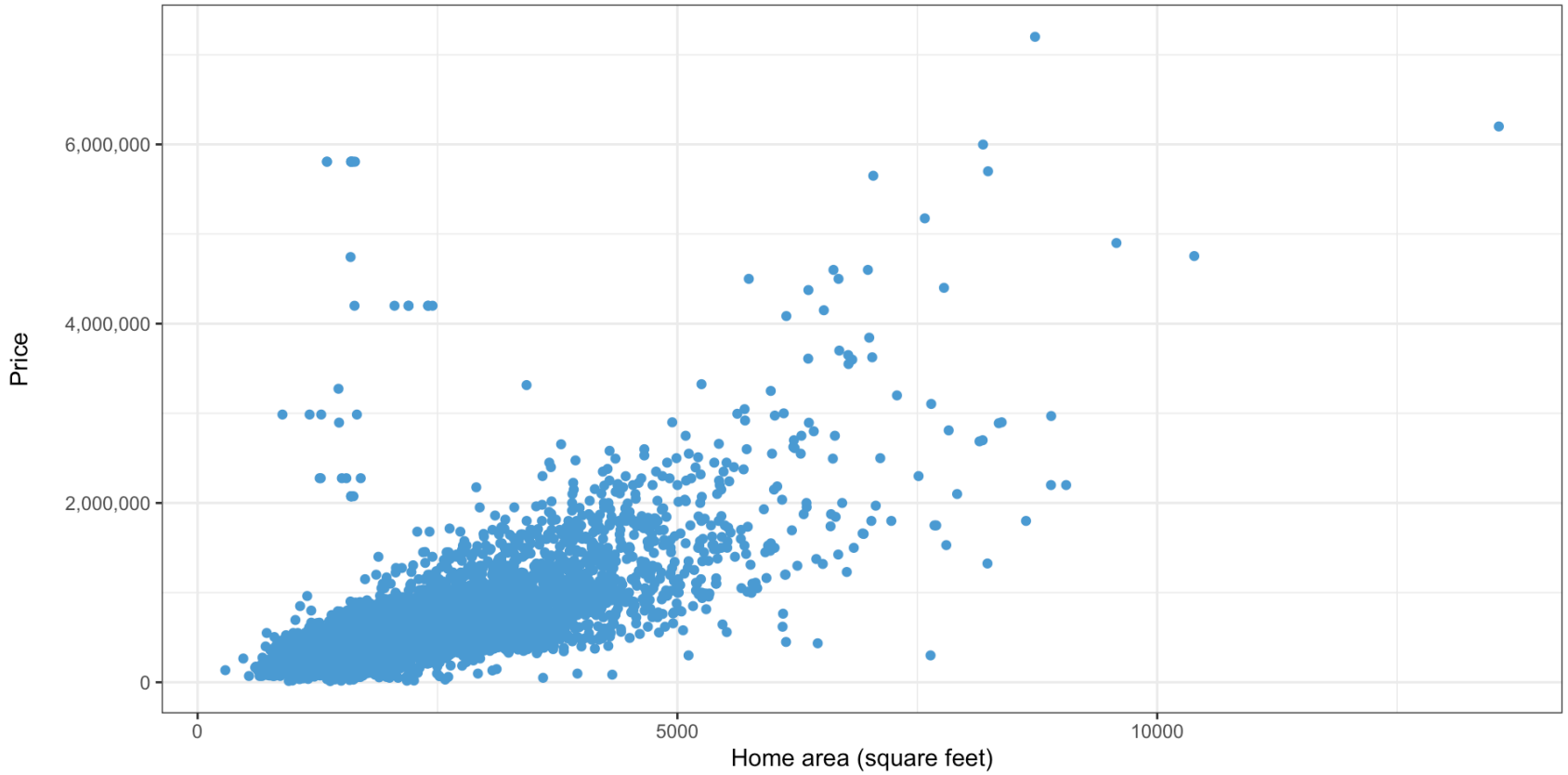


Multiple box plots in Excel

- If you select multiple columns, Excel will make multiple box plots
- Unfortunately, often you want to do this by separating out different observations in the same column
- This is inexplicably difficult in Excel
- Best way I've found is to make a table in Excel, filter by the attributes you want, and copy variables to a new sheet

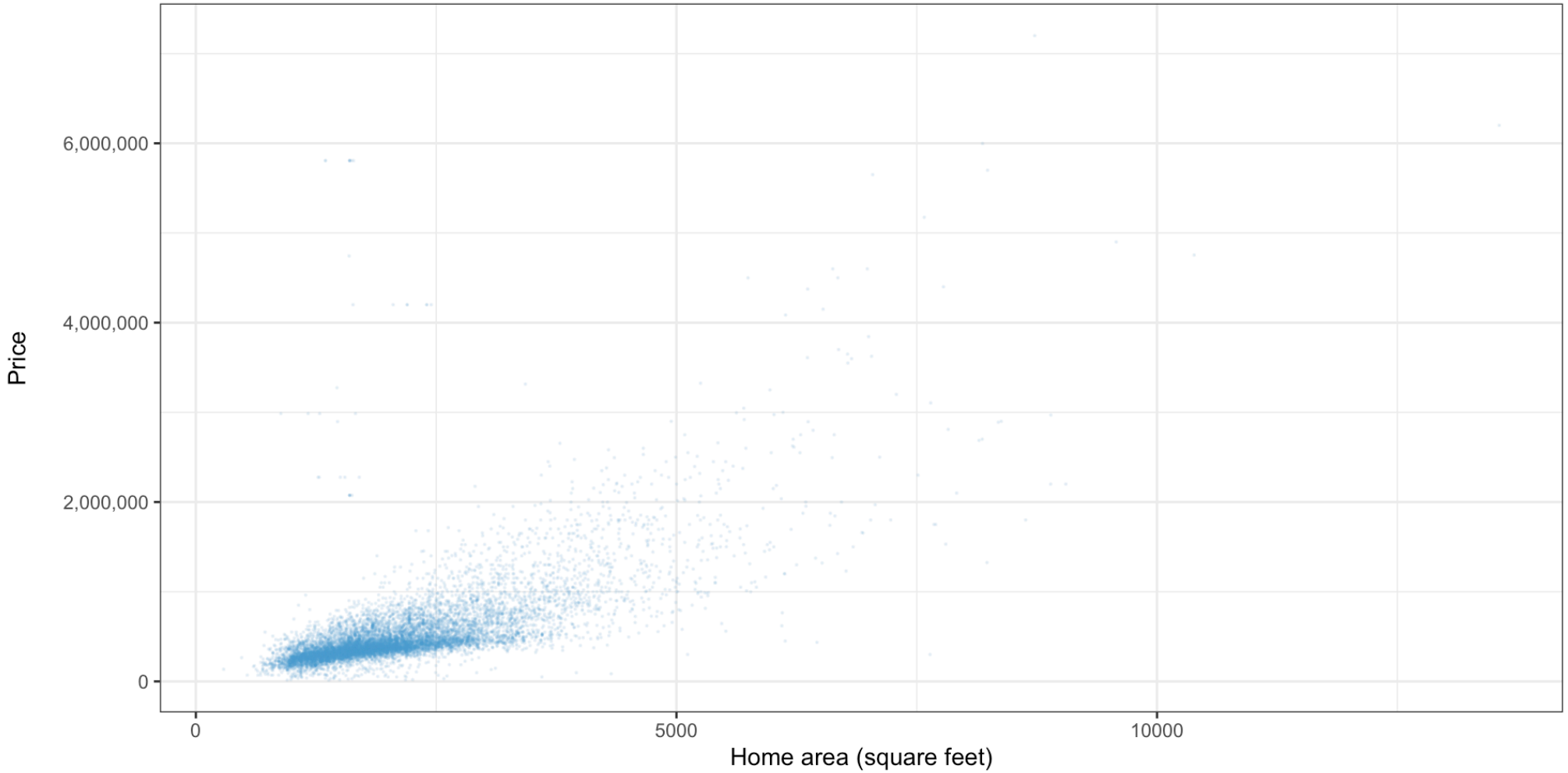
Scatter plots

- Used when you have two continuous variables
- Simply plot the location of each data point
- In a regression context, the independent variable is customarily on the x axis



Scatter plots: other properties

- When there are many points, reducing the size or opacity can help show trends



Scatter plots in Excel

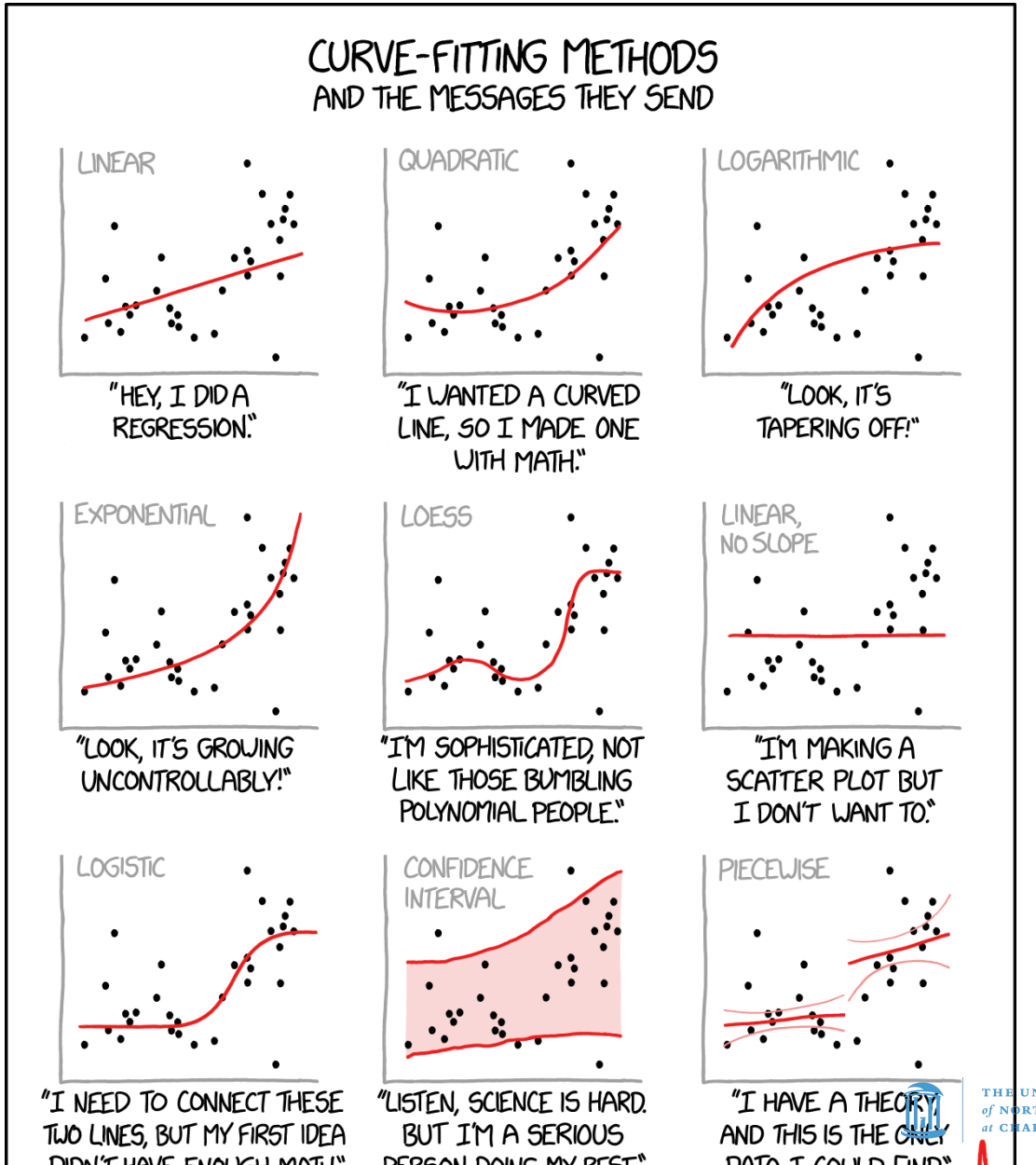
- Select the data you want, then Insert -> Chart -> Scatter plot
- To change the size, select Format tab, then the series name on the left, and open the format pane
- Click the Fill and Line tab (paint bucket), then select Marker, Marker options, Built-in, and edit the size
- Transparency

Trendlines

- It may *occasionally* be useful to add a trendline
- Format -> Add Chart Element -> Trendline

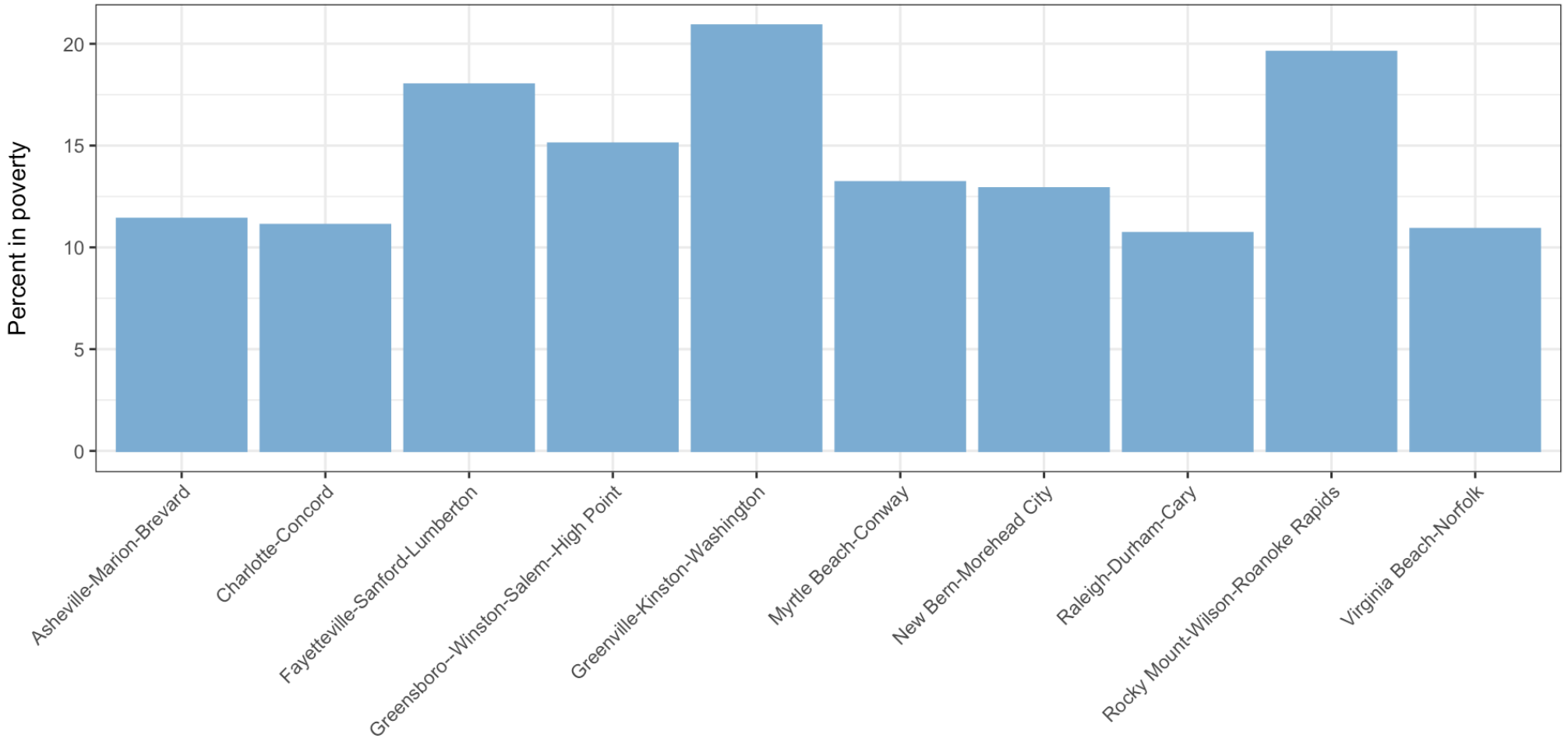
The curse of trendlines

- Excel provides many options for trendlines; don't get carried away



Bar charts

- Used with one continuous and one categorical variable

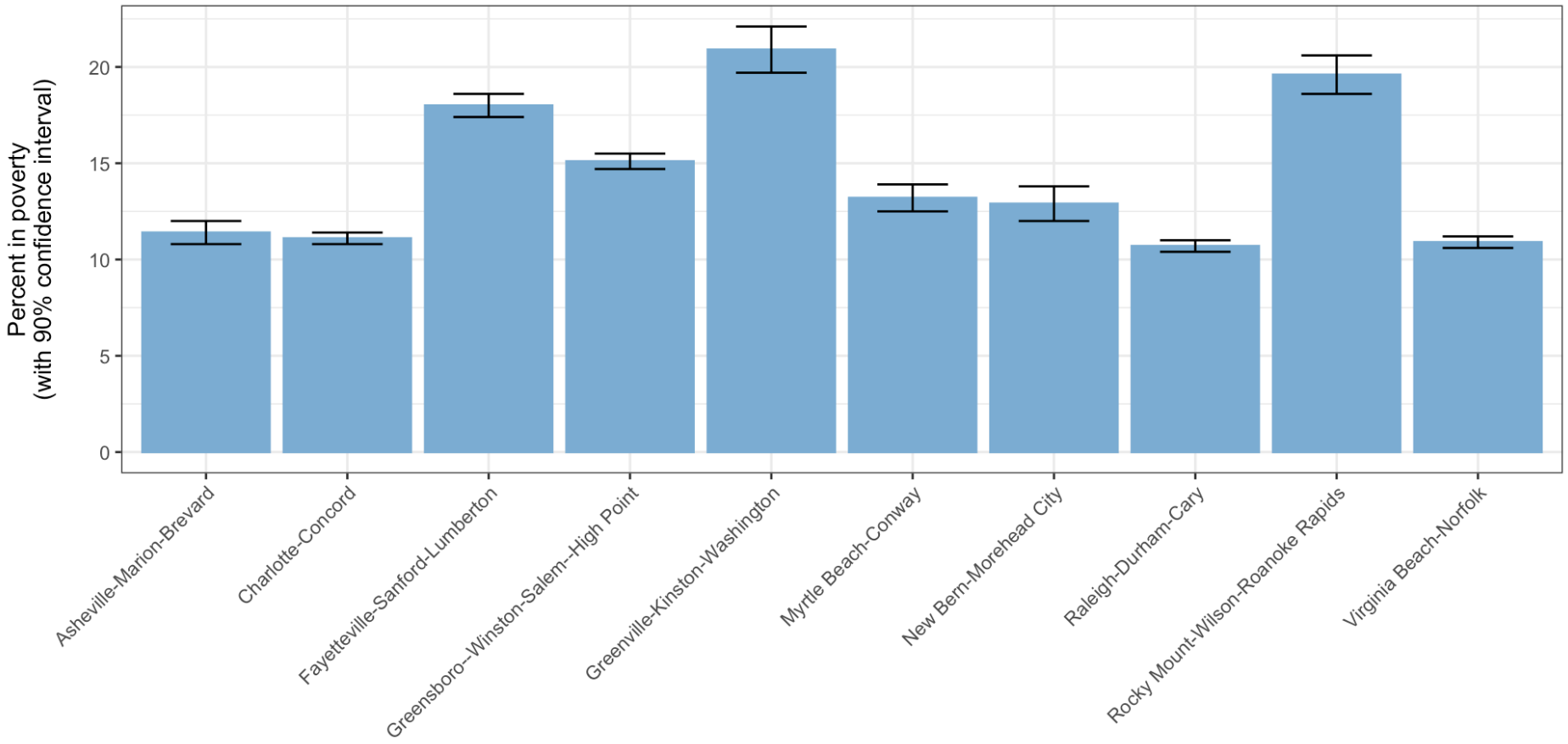


Bar charts in Excel

- Select the column with labels and the column with values
- Insert -> Chart -> Bar chart

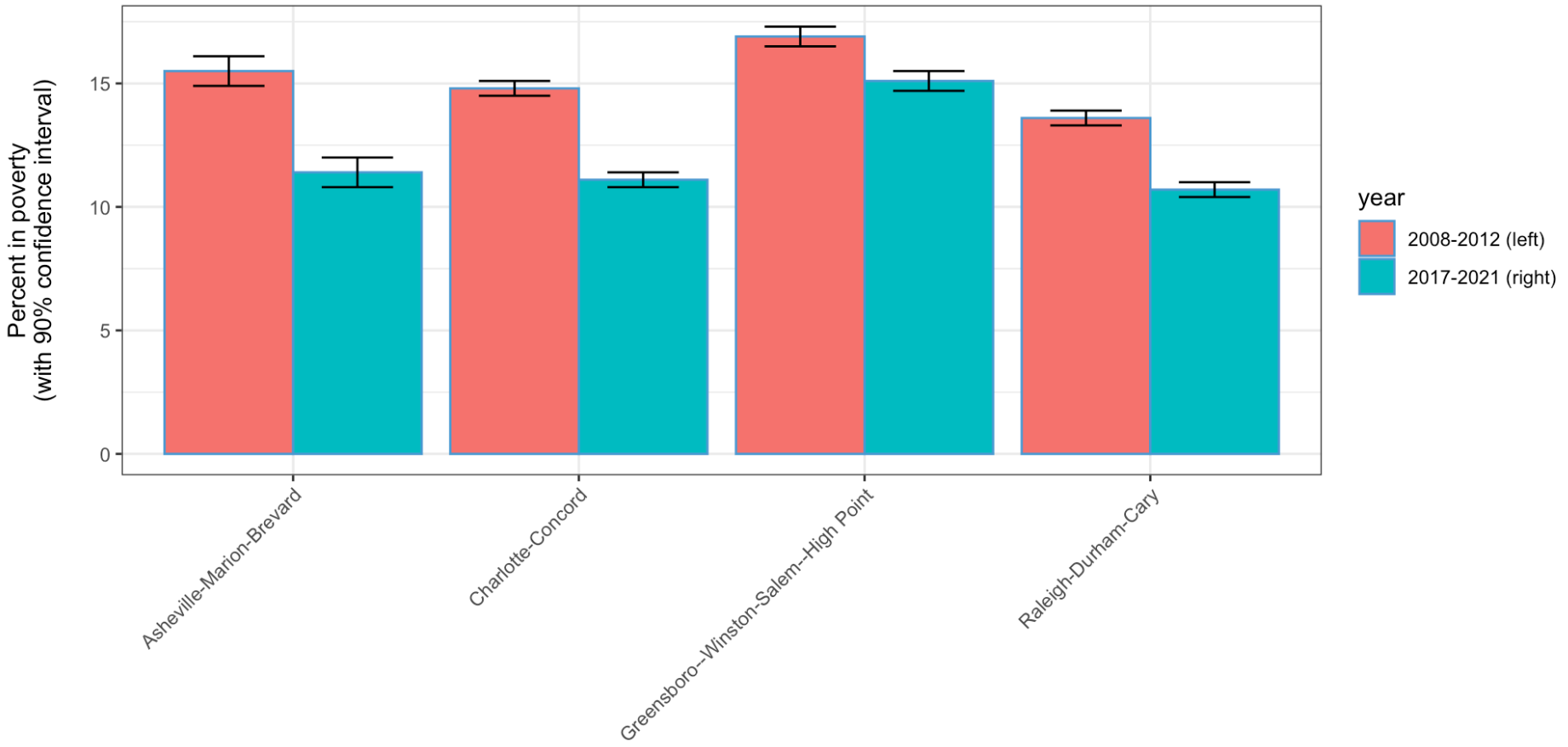
Error bars

- Error bars indicate a margin of error/confidence interval around the tops of the bars
  - Or occasionally just the standard error - be careful here!



Grouped bar charts

- A common variation of the bar chart is the grouped bar chart
- Used when you want to have multiple bars associated with each x-axis value

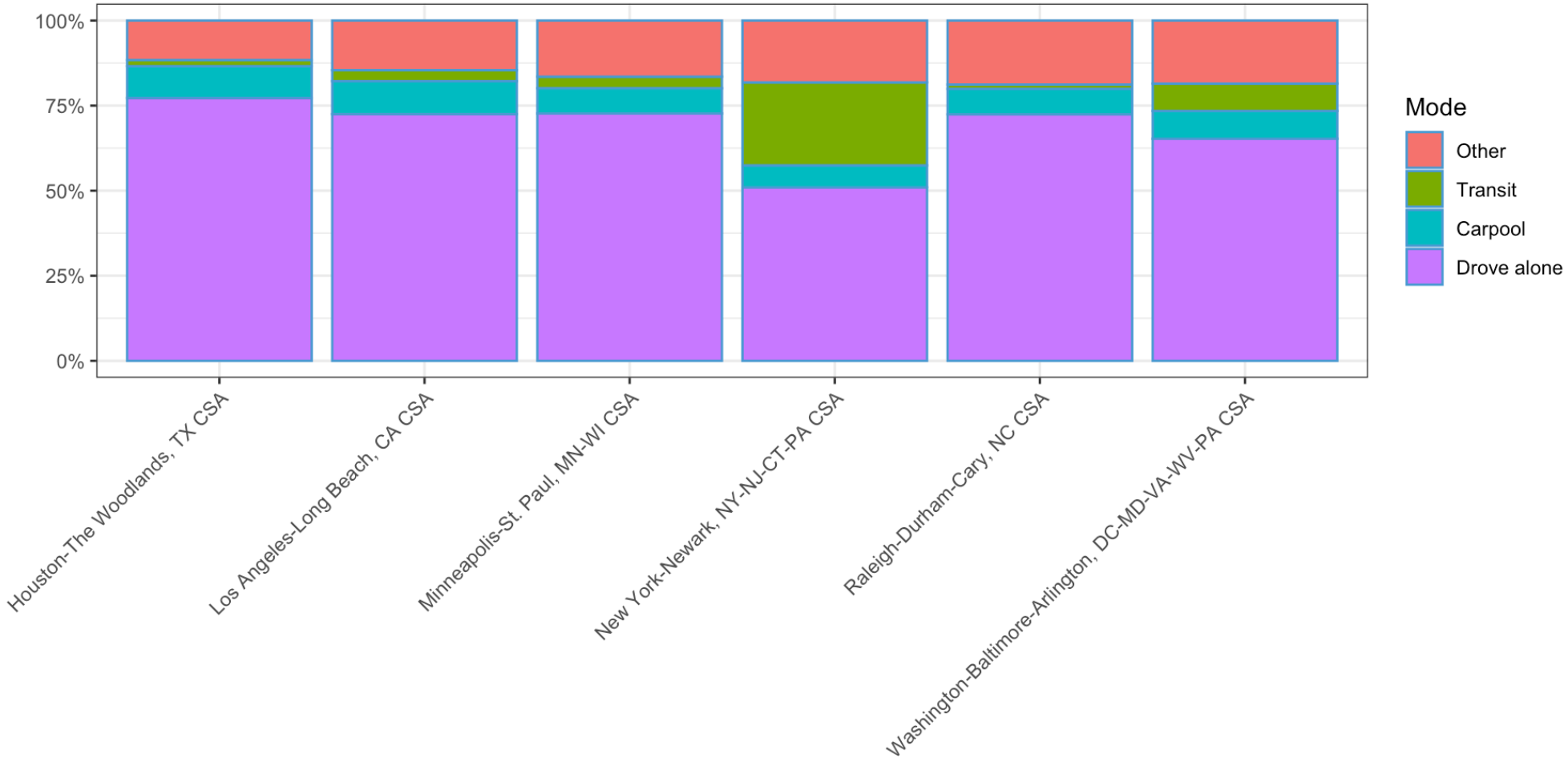


Grouped bar charts in Excel

- If you select multiple columns when creating a bar/column chart in Excel, a grouped bar chart is the default

Stacked bar charts

- Often used with percentages but can be used with other things as well



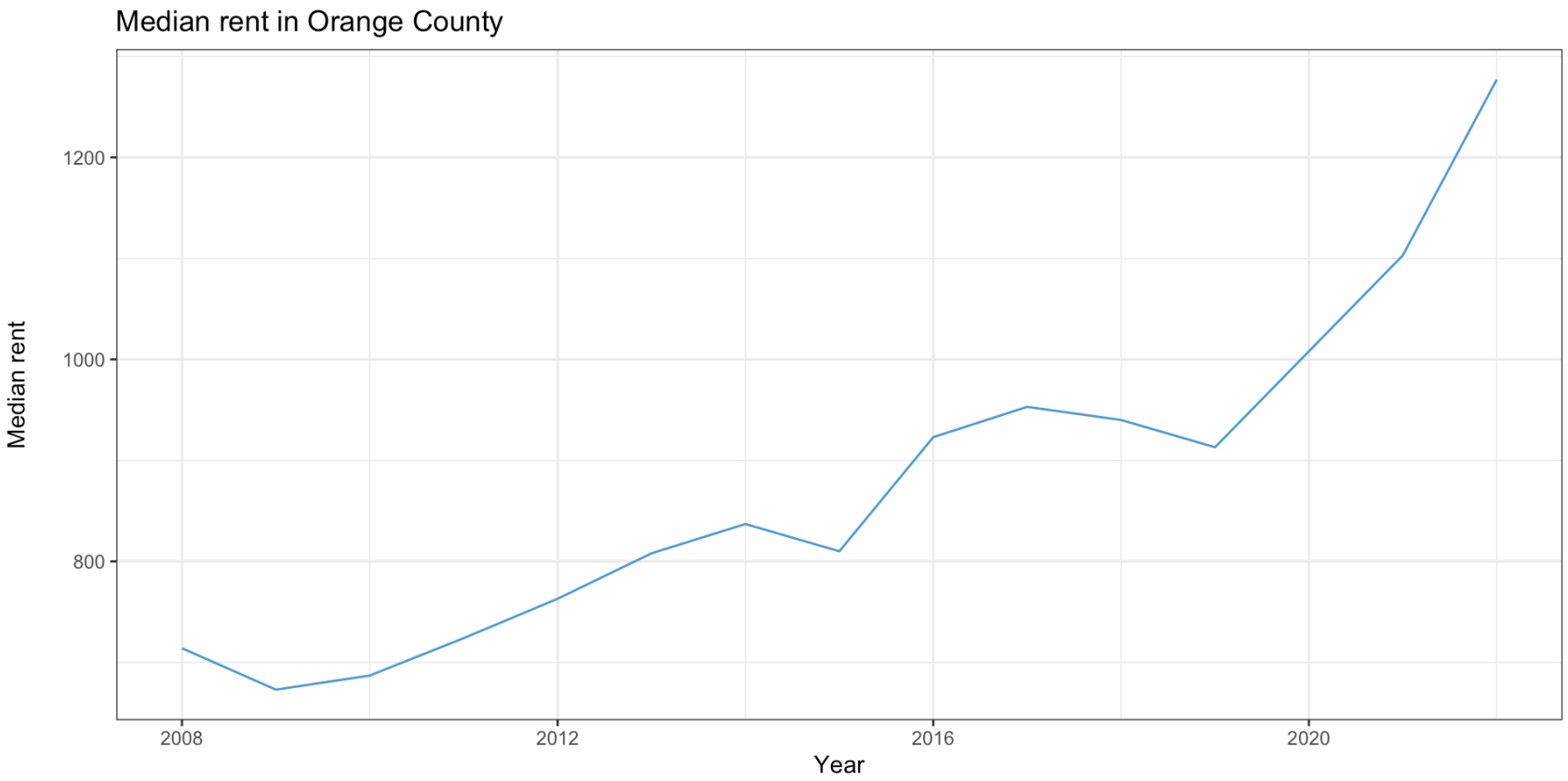
Stacked bar chart in Excel

- Each portion of the stacked bar should be in a different column
- Select all the columns and choose Insert -> Chart -> Stacked Column
  - or 100% stacked column to have Excel normalize to 100% for you

3D bar charts

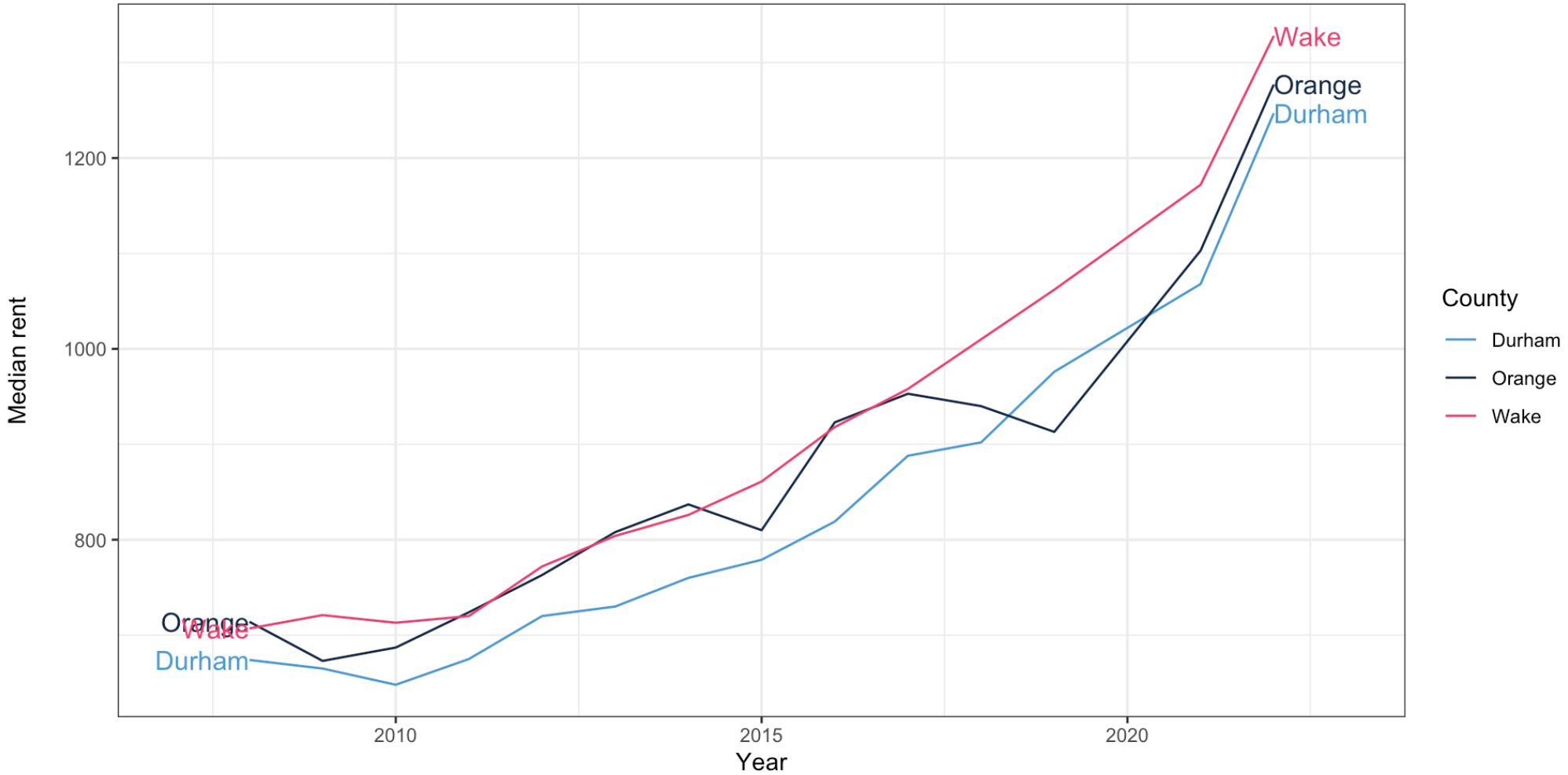
- 3D bar charts look cool but they’re confusing
- Generally, the bar height represents the value being represented, but it is easy to misinterpret the area or volume as the value being represented

Line charts


Median rent in Orange County
Median rent
1200
1000
800
2008
2012
2016
2020
Year

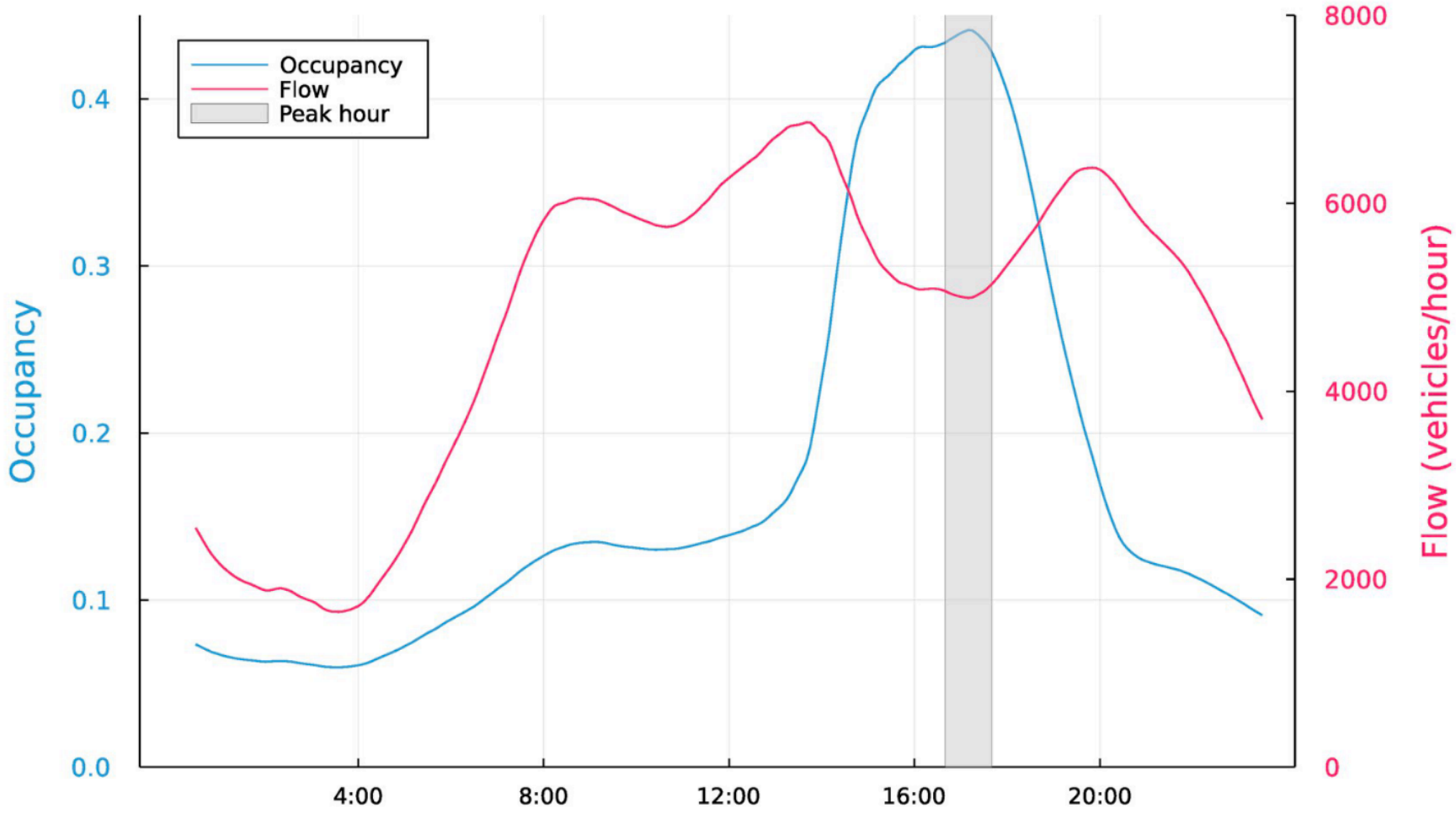
Multiple lines

- It's common to have multiple lines on line charts



Multiple y axes

- Occasionally, you'll even see a graph with multiple y axes



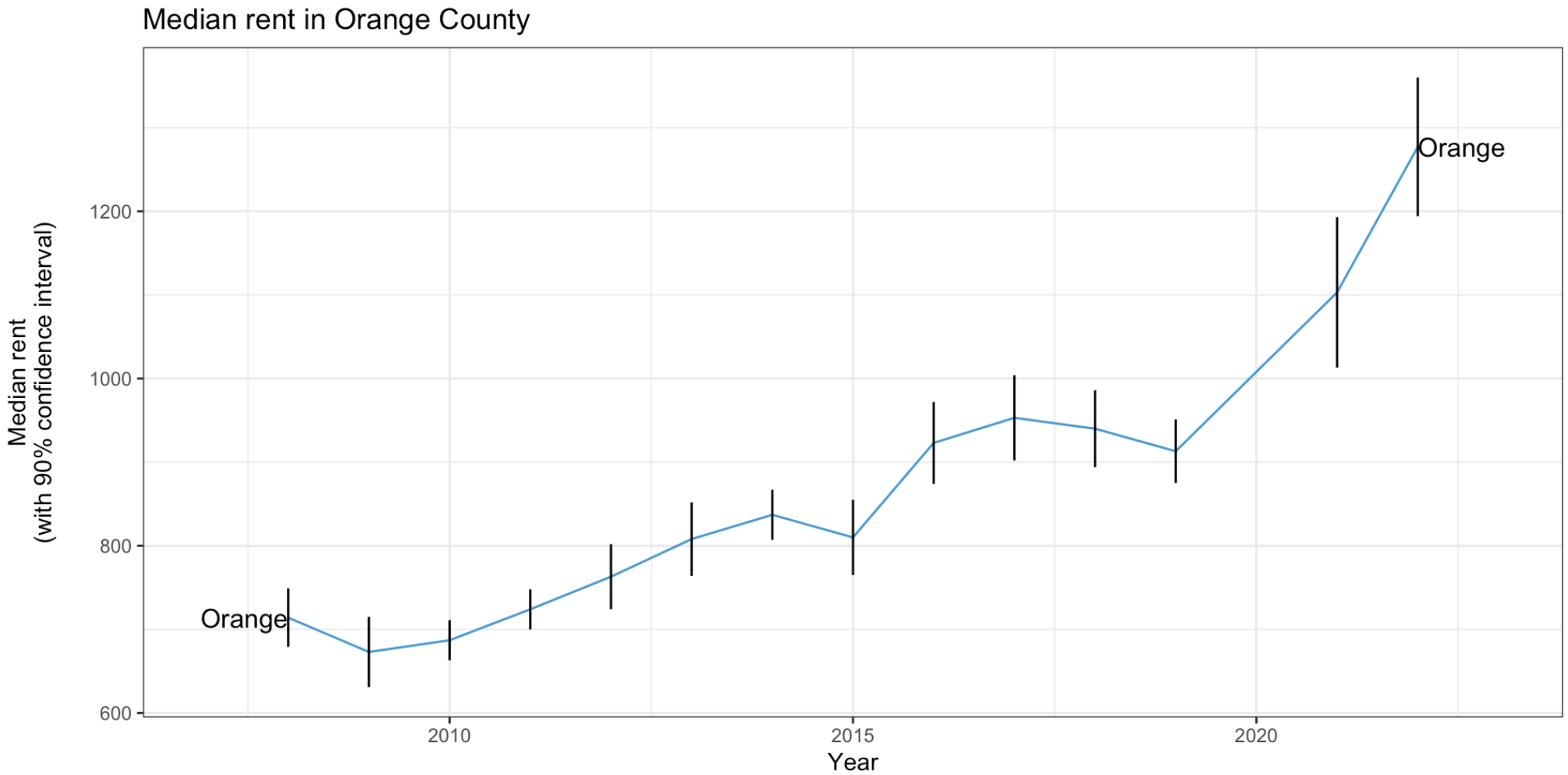
Bhagat-Conway and Zhang (2023)

Line charts in Excel

- In Excel, how you insert a line chart depends on your x axis
- If you want Excel to generate the x axis as 1, 2, 3..., you need to add a line chart
- If you have a column with the x axis values, you want an XY (scatter) plot with lines

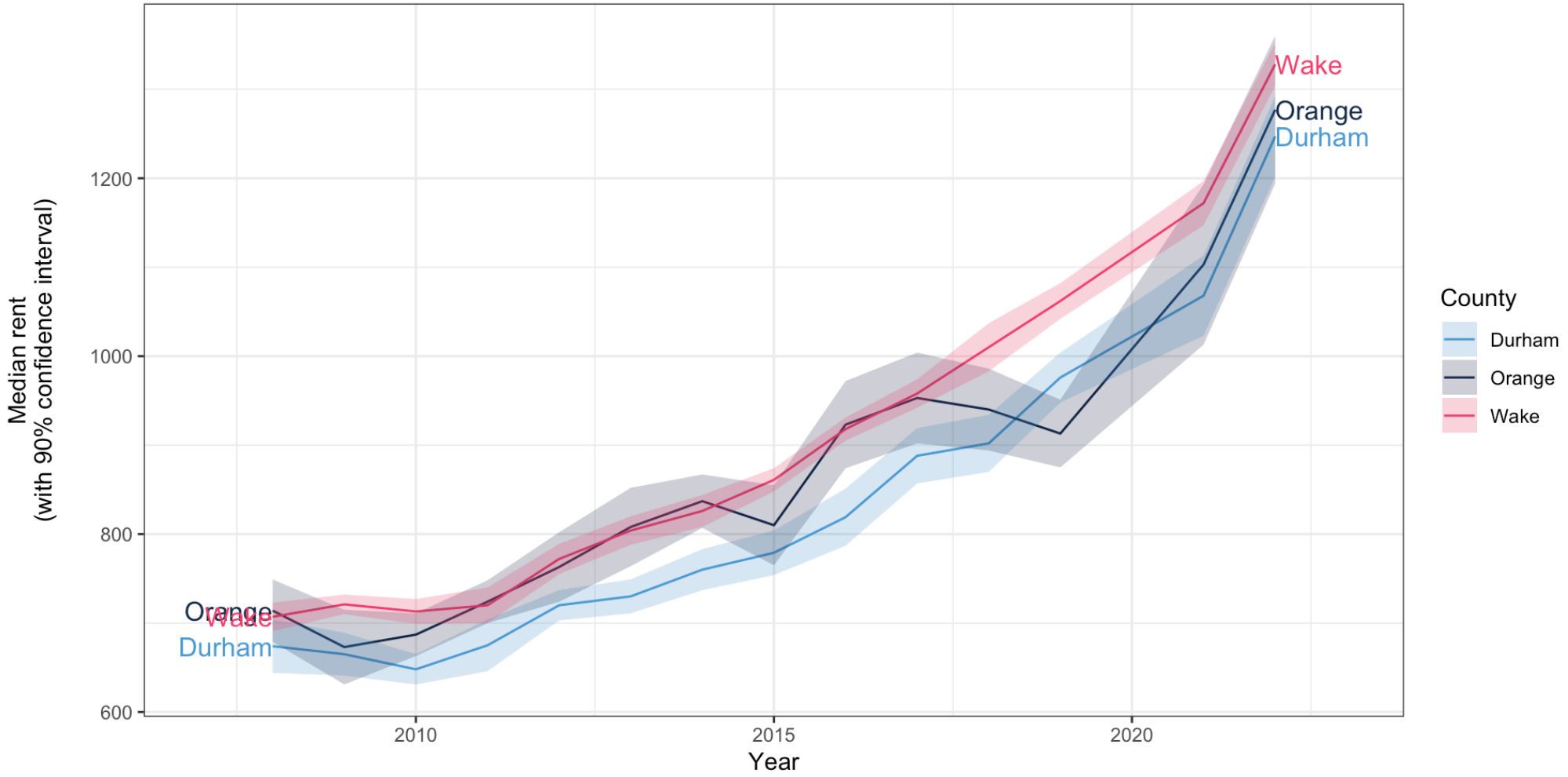
Line charts and error bars

- You'll occasionally see line charts with error bars on them



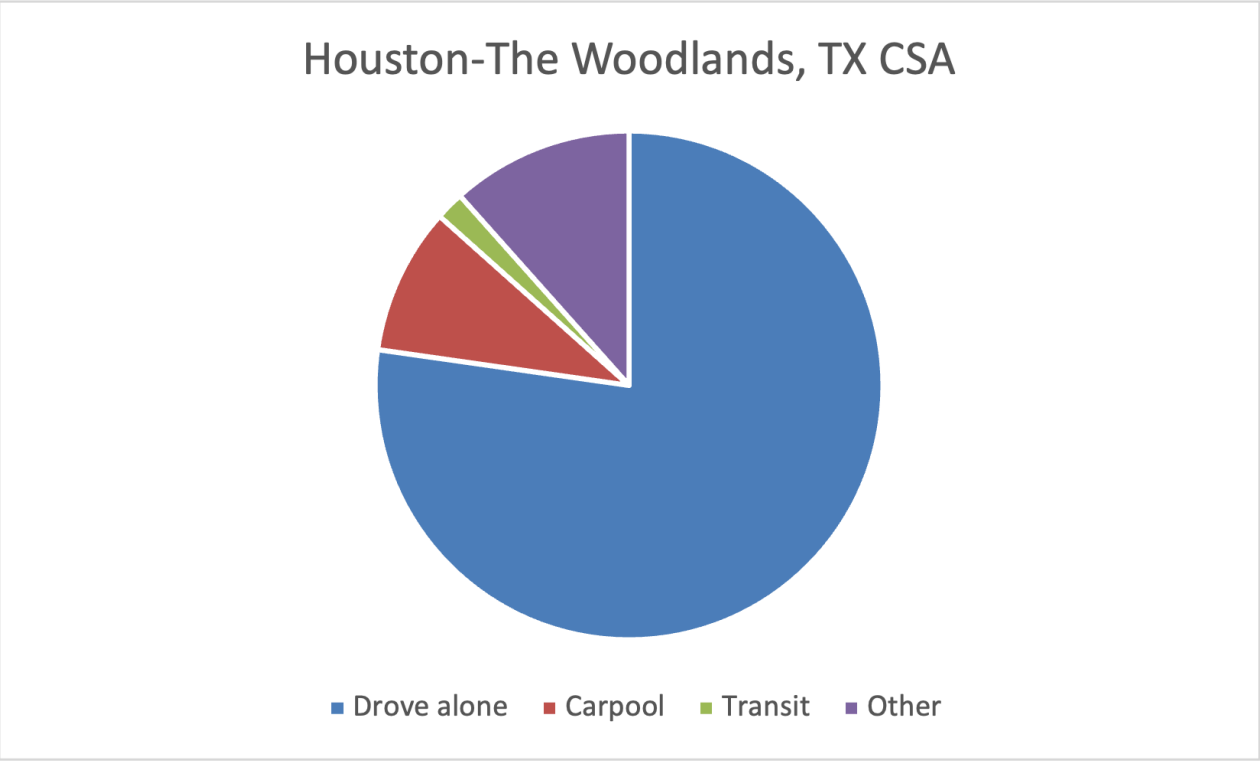
Line charts and error bars

- It's more common to see line charts with error "ribbons"



Pie charts

- Pie charts are common in government and business but rare in academia
- Can be hard to compare relative sizes
- Should *only* be used with data that totals 100%
- Avoid 3D pie charts for similar reasons as 3D bar charts



Pie charts in Excel

- Let’s use the “Stacked bar” tab to make a pie chart of commute mode in the Triangle
- Select the row with the Triangle
- Insert -> Chart -> 2D pie chart
- The labels are probably wrong; click “Select data” and specify the range for the category labels

Other ~~gimmicks~~ kinds of charts

- There are lots of other kinds of charts that are occasionally useful
- But probably 95% of charting needs are covered by what we've seen today
- It's easy to get carried away

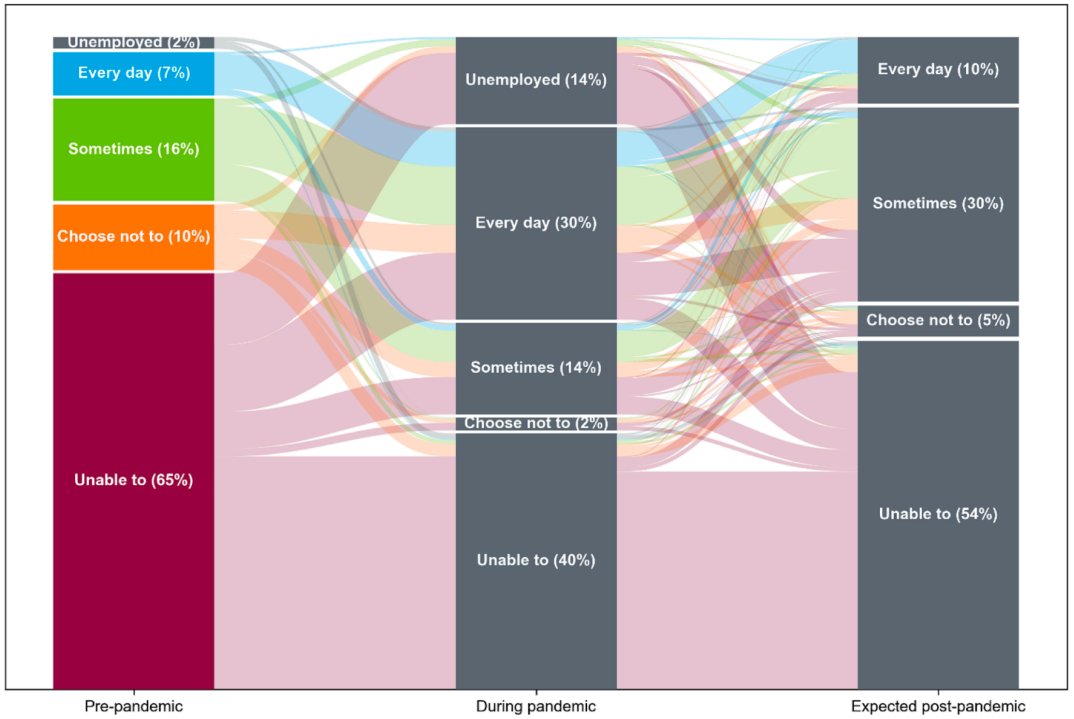


Fig. 1. Sankey Plot of Weighted Telecommuting Outcomes in each Period, Wave 1, COVID Future Panel Survey.

Accessibility concerns: alternative text

- Charts are a graphical element, and are not accessible to all individuals
- Many visually-impaired individuals will use a *screen reader*, software that speaks the text on their screen out loud
- All charts should either have alternative (alt) text, or be fully described in the text of the report
  - Add alt text in Word by right-clicking and choosing “view alt text”
  - Also applies to non-decorative images
- For small charts, you can list the values in the alt text (e.g. the percentages drving alone, etc.)
- For larger charts, you should describe overall trends/findings
- For many organizations, ensuring accessibility is a legal requirement under Section 508 of the Americans with Disabilities Act

Thanks to Dhruti Bhagat-Conway for recommendations on this section

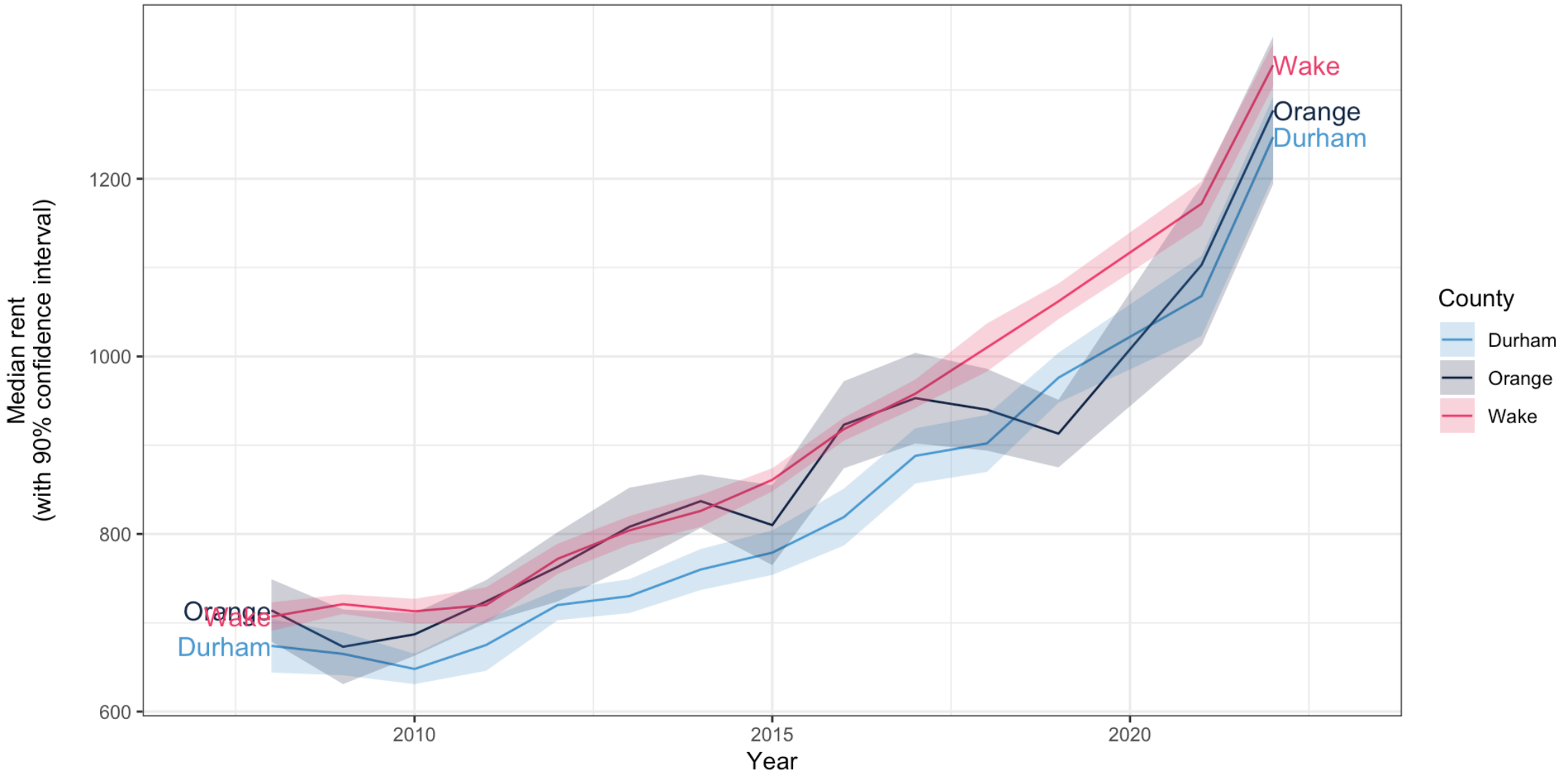
Accessibility concerns: color vision deficiency

- Color vision deficiencies (color-blindness) are very common, especially among White males (~8%)

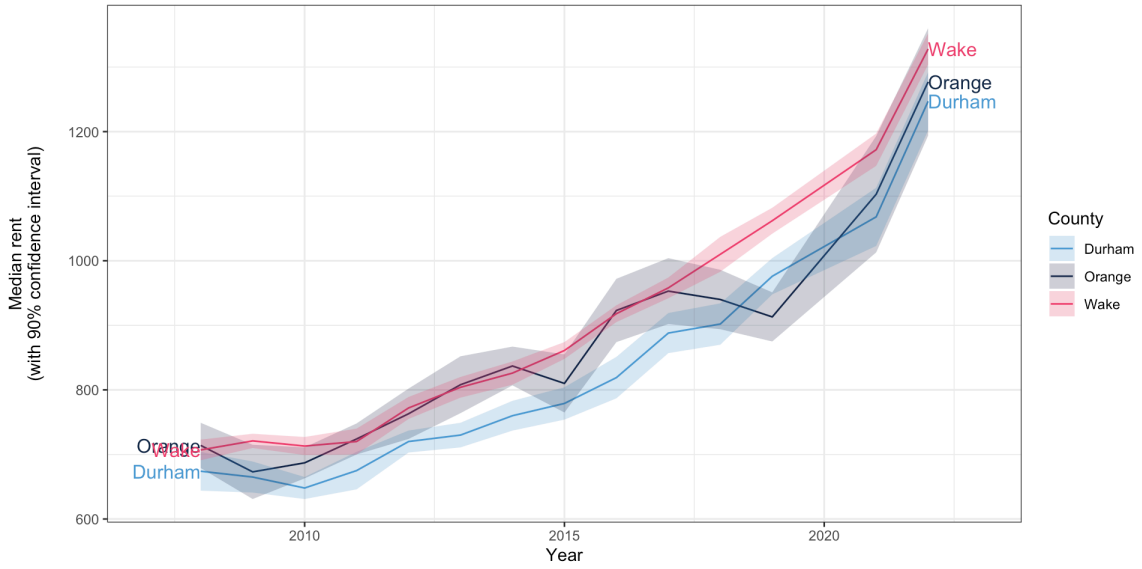
  (*Machado, Oliveira, and Fernandes 2009*)

- This can make charts useless

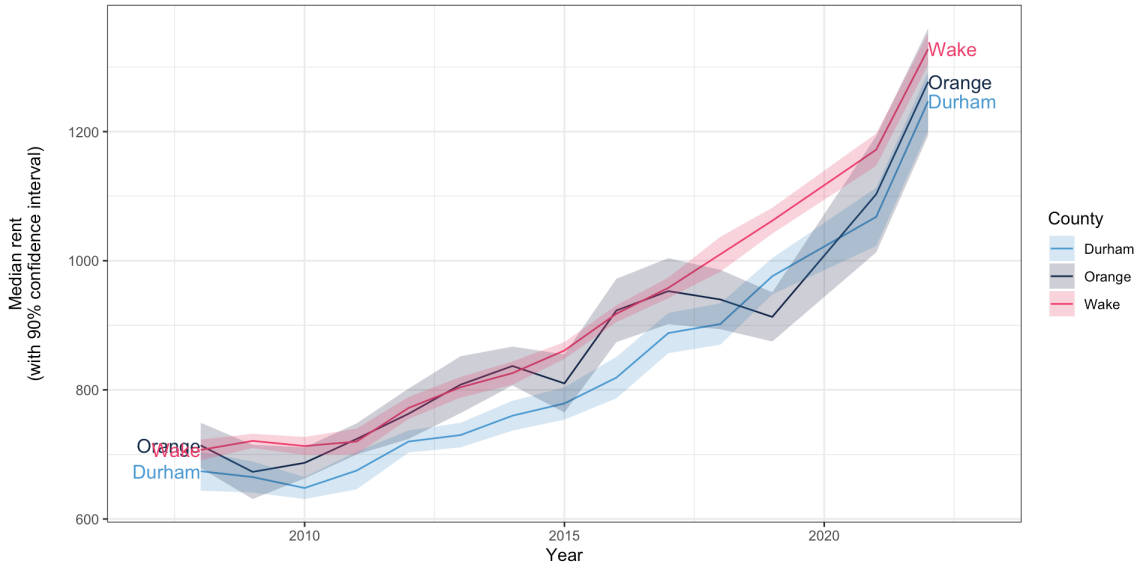
Accessibility concerns: color vision deficiency



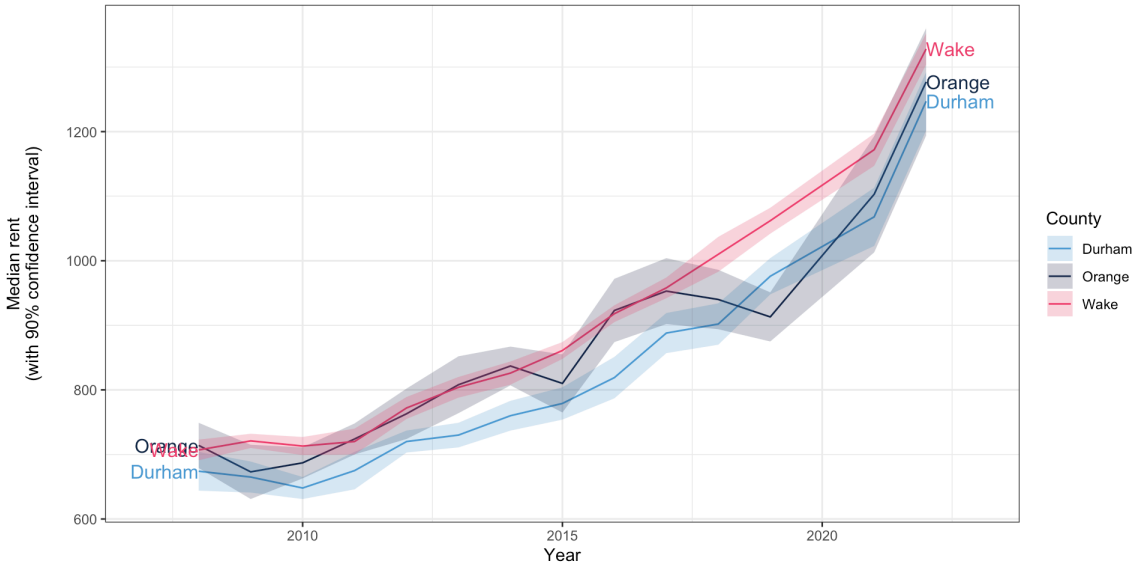
Deuteranomaly (red-green): very common



Protanomaly (red-green): less common



Tritanomaly (blue-yellow): rare



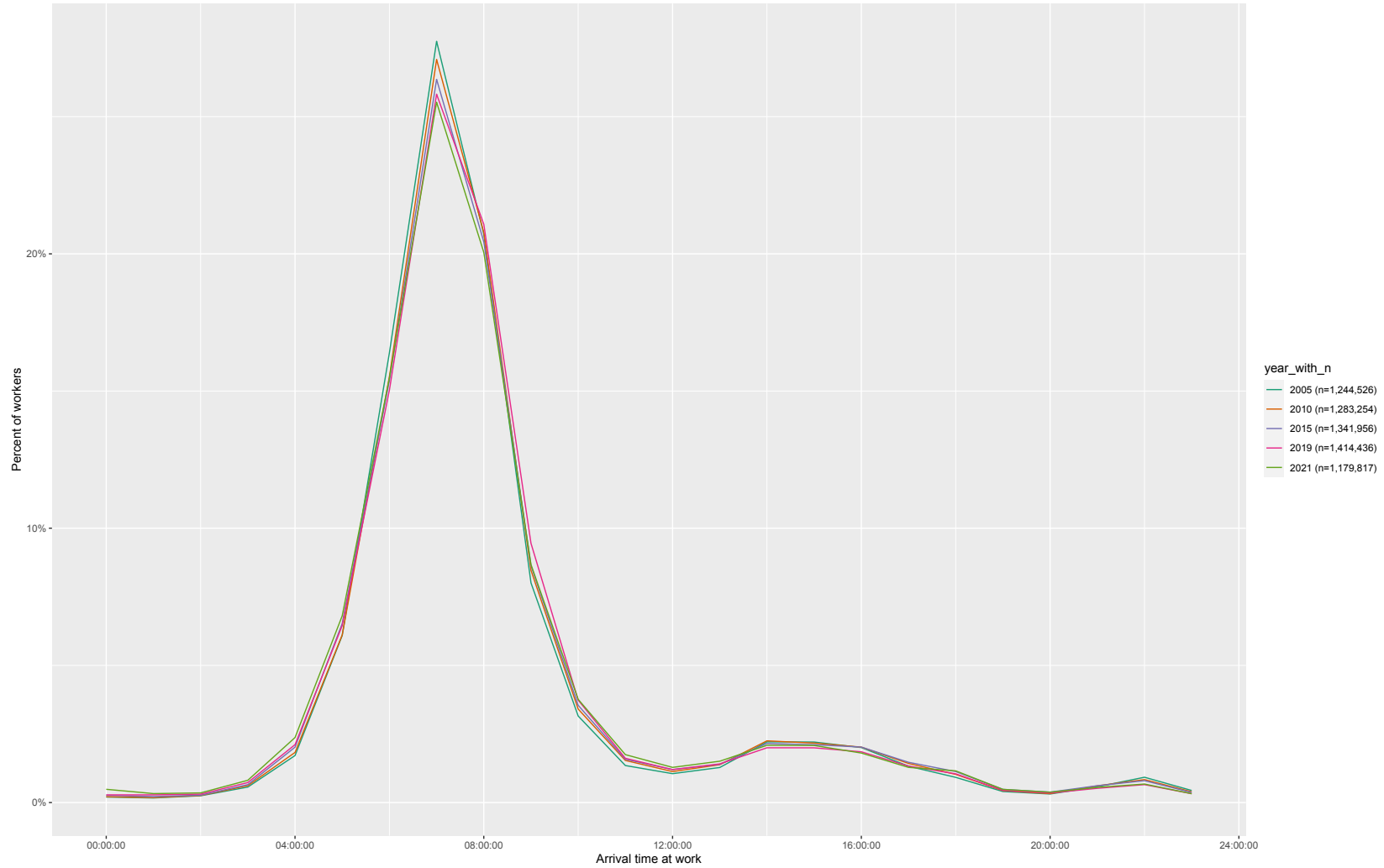
What to do about it

- Avoid presenting information only using color
- Avoid using red and green together
- Vary saturation and brightness in addition to color

Contrast

- Can you read this?
- Text and graphical elements should provide enough contrast that they are easily distinguishable
- You can check the contrast of two colors using the WebAIM contrast checker
- Or generate color schemes using Coolers

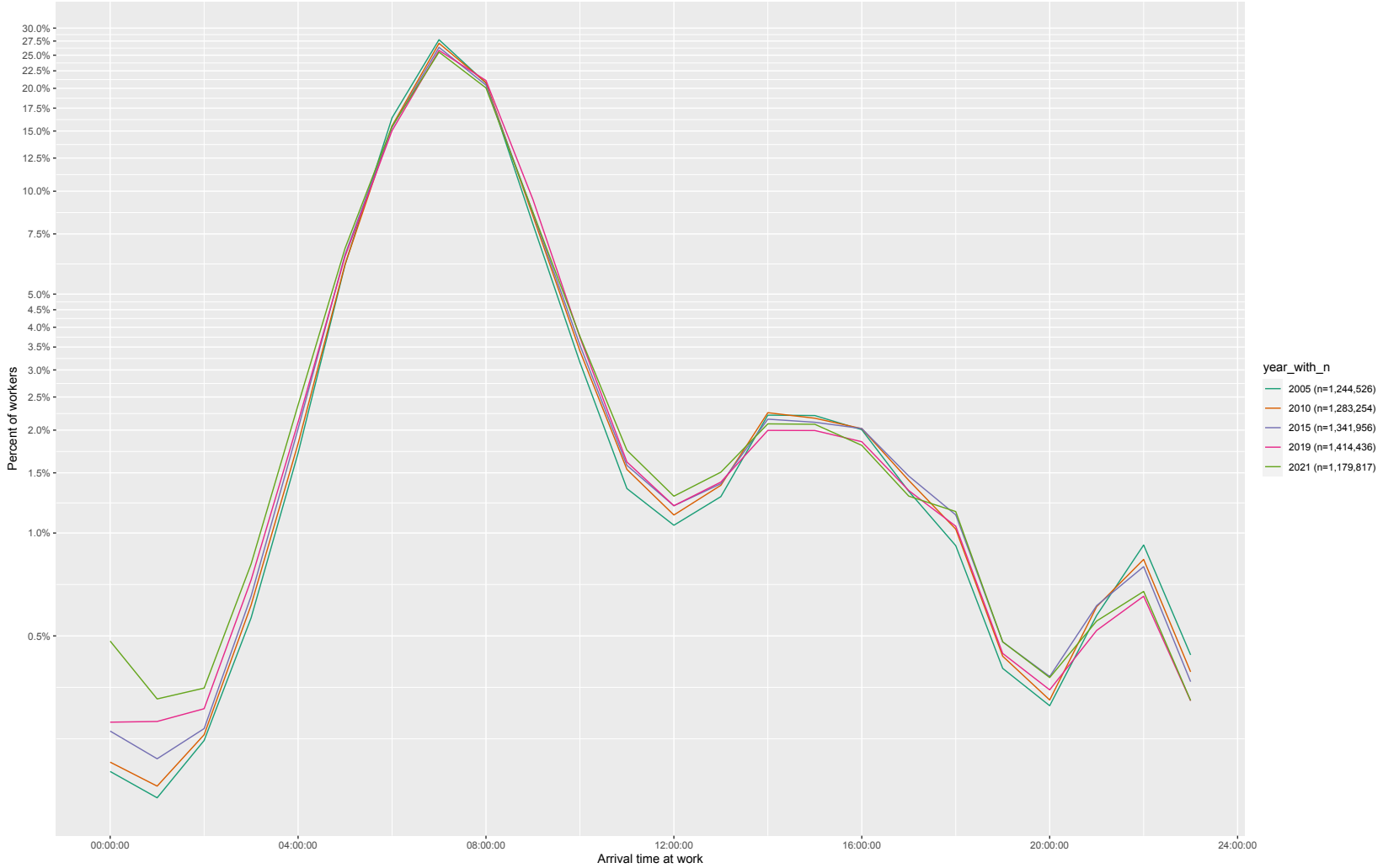
Logarithmic (log) scales



Linear scale

Palm and Bhagat-Conway, under review; data: IPUMS

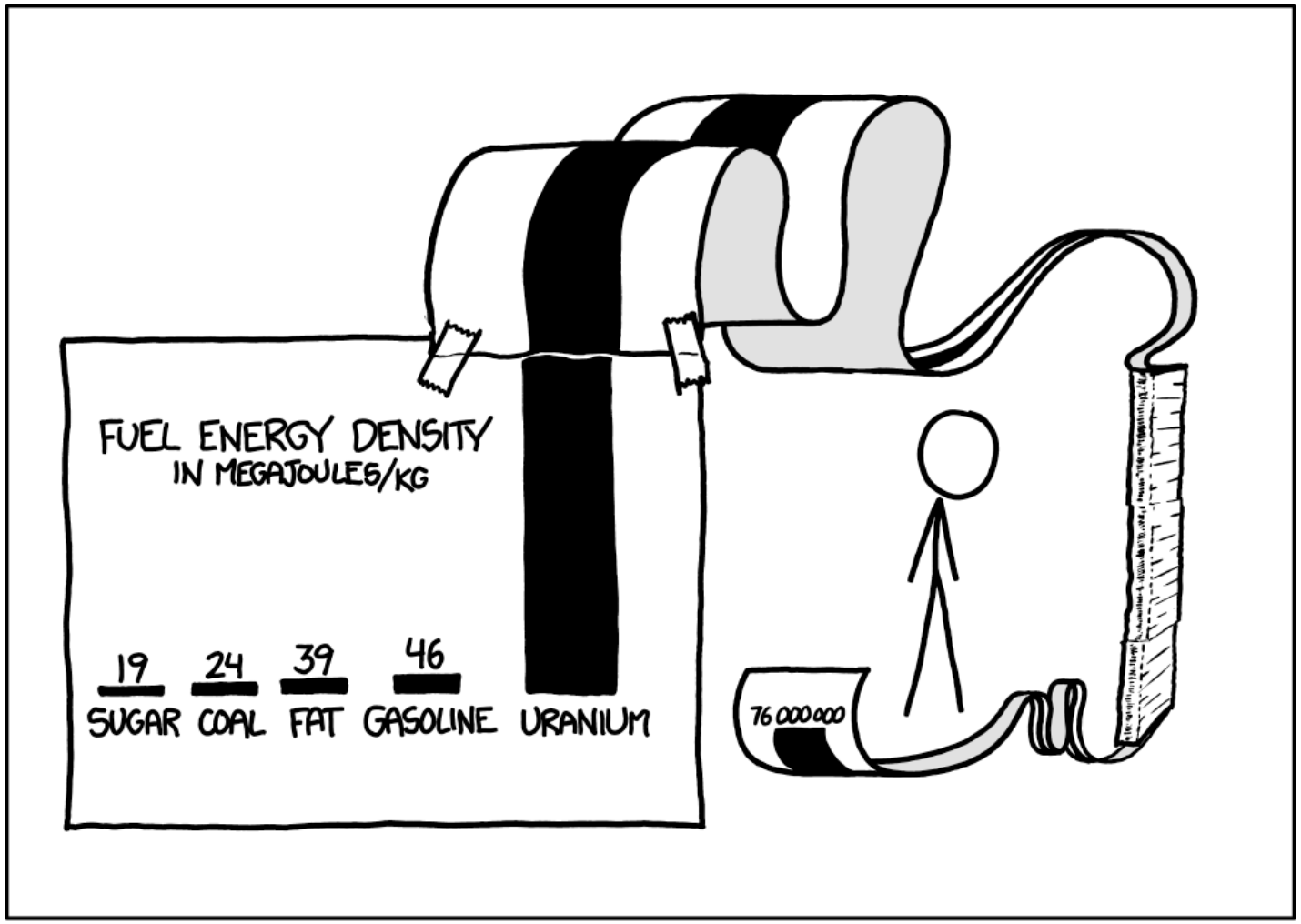
Logarithmic (log) scales



Logarithmic scale

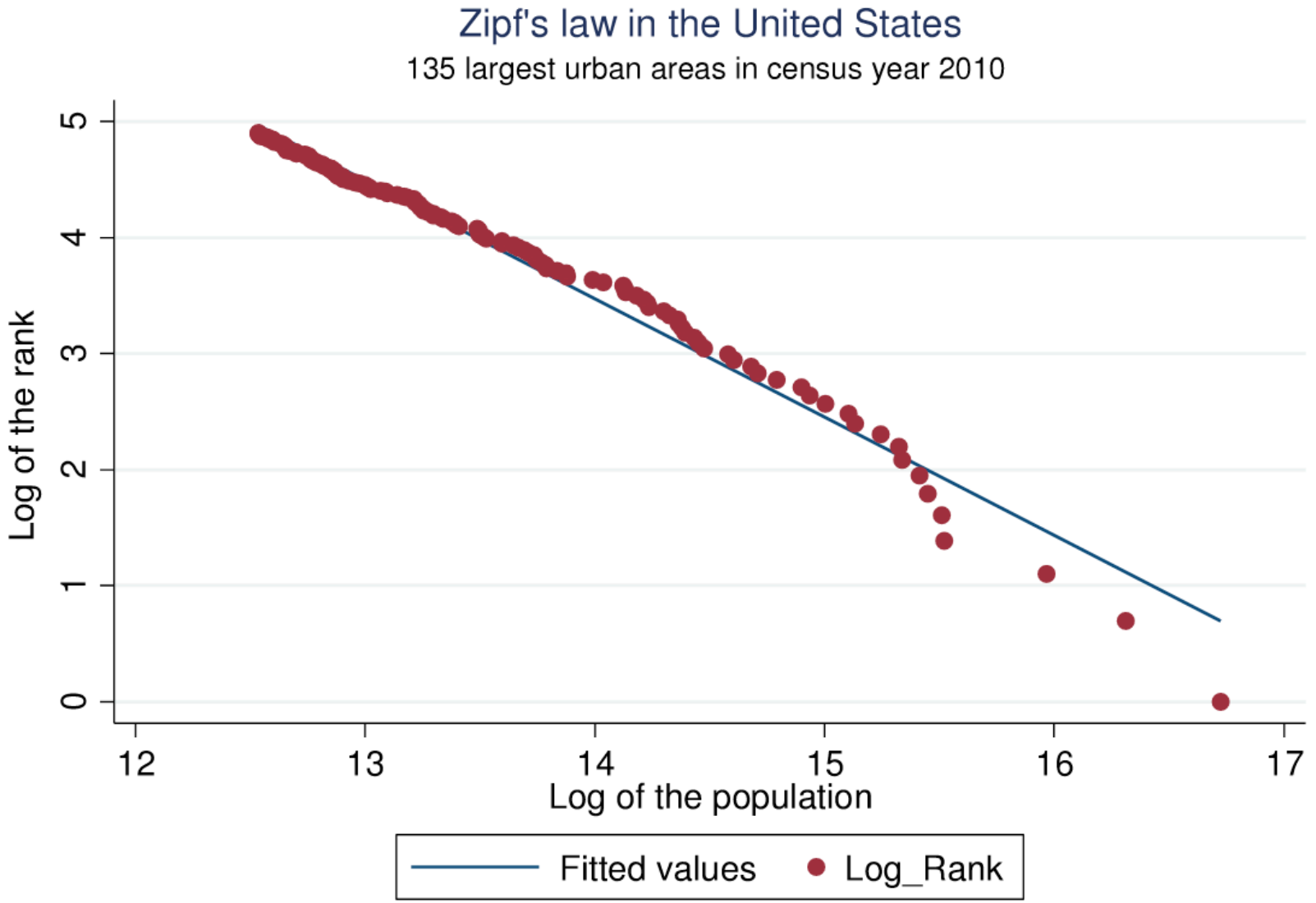
Palm and Bhagat-Conway, under review; data: IPUMS

Logarithmic (log) scales



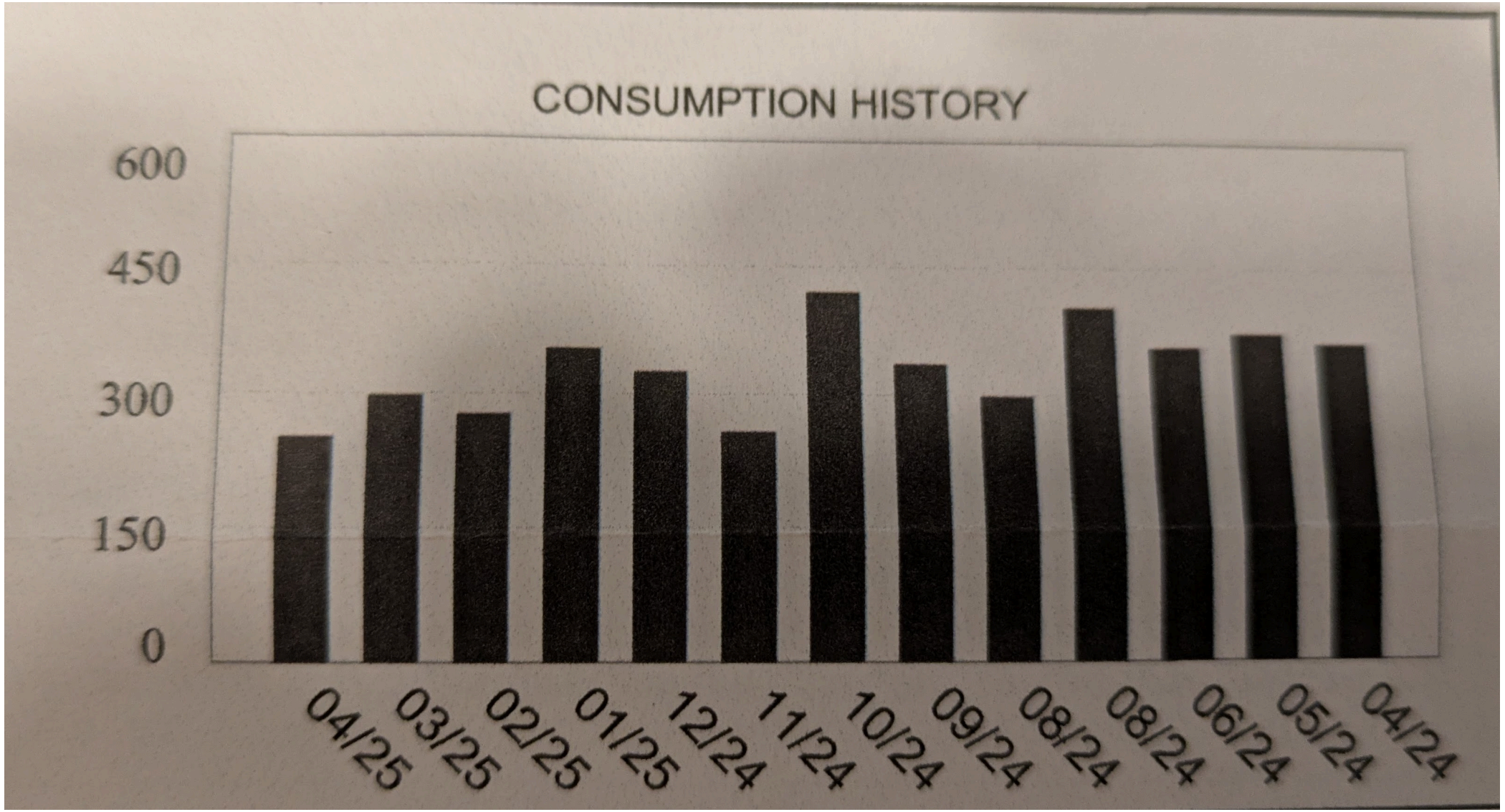
SCIENCE TIP: LOG SCALES ARE FOR QUITTERS WHO CAN'T FIND ENOUGH PAPER TO MAKE THEIR POINT *PROPERLY.*

Log-log scales

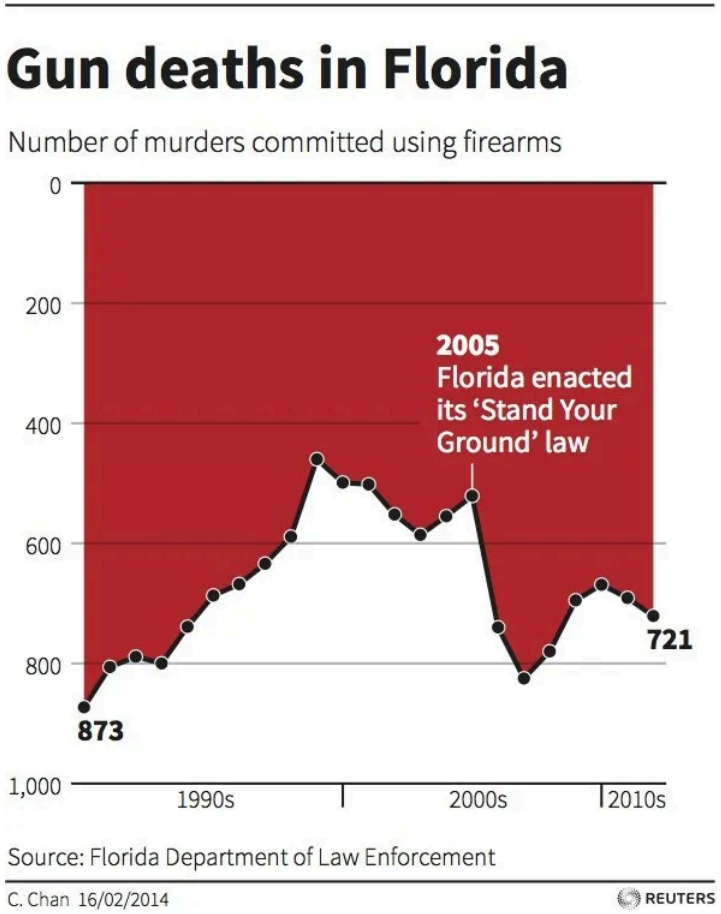


Arshad, Hu, and Ashraf (2018)

Other ways axes can mislead

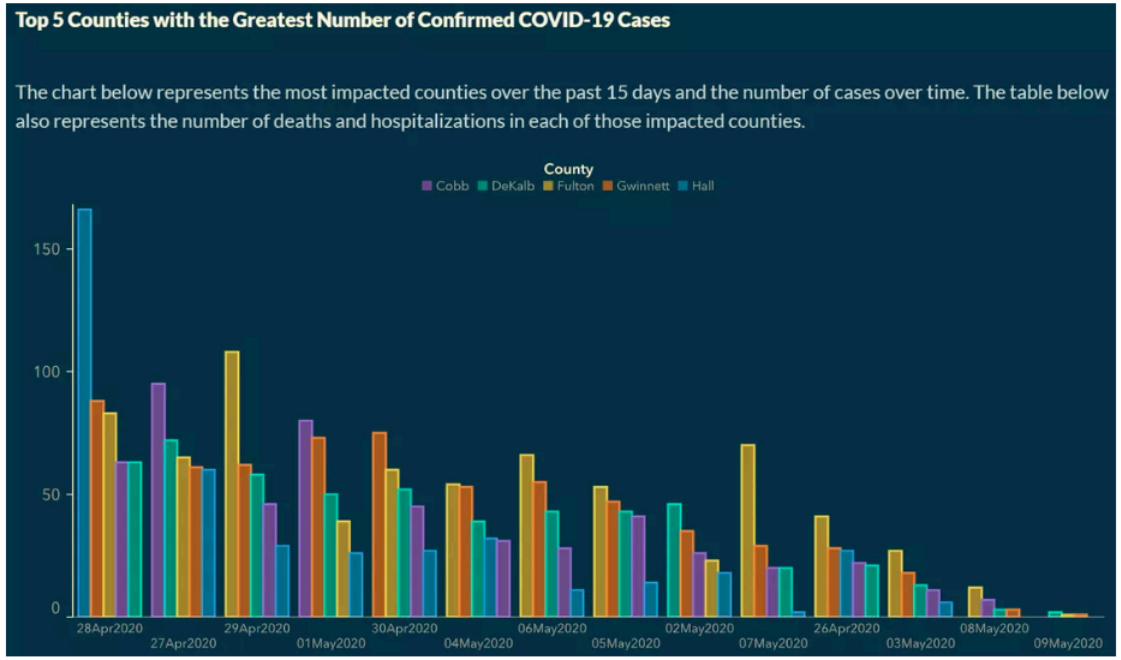


More misleading axes



Reuters via LiveScience

And more...



Georgia Dept of Public Health via Columbia Climate Law/Atlanta Journal-Constitution

Chart formatting in Excel

- Most tools to customize your chart are on the “Chart Design” and “Format” toolbars
- Double-clicking on elements of a chart lets you format and edit them

Advanced charting tools

- In R, the `ggplot` package is widely known and used, and very powerful
- `matplotlib` is the go-to plotting library for Python

References

Anscombe, F. J. 1973. “Graphs in Statistical Analysis.” *The American Statistician* 27 (1): 17–21. https://doi.org/10.2307/2682899.

Arshad, Sidra, Shougeng Hu, and Badar Nadeem Ashraf. 2018. “Zipf’s Law and City Size Distribution: A Survey of the Literature and Future Research Agenda.” *Physica A: Statistical Mechanics and Its Applications* 492 (February): 75–92. https://doi.org/10.1016/j.physa.2017.10.005.

Bhagat-Conway, Matthew Wigginton, and Sam Zhang. 2023. “Rush Hour-and-a-Half: Traffic Is Spreading Out Post-Lockdown.” *PLoS One*.

Machado, G. M., M. M. Oliveira, and L. Fernandes. 2009. “A Physiologically-based Model for Simulation of Color Vision Deficiency.” *IEEE Transactions on Visualization and Computer Graphics* 15 (6): 1291–98. https://doi.org/10.1109/TVCG.2009.113.