

# Statistical Graph Literacy

**Mikhail Popov**  
Analysis



# Data Visualization as Storytelling

— — —

Graphical displays should:

- Show the data
- Induce the viewer to think about the substance rather than graphic design or format
- Avoid distorting the data
- Present many numbers in a small space
- Make large data sets coherent
- Encourage the eye to compare different pieces of data

– Edward R. Tufte, *The Visual Display of Quantitative Information*



# Types of Variables

— — —

*Continuous* variables have “an infinite”\* range of possible values

**Examples:** time, age, weight, lengths (height, distance, time spent online), drug dosage

*Categorical / discrete / qualitative*

*Nominal* variables have two or more categories that do not have an intrinsic order

**Examples:** gender, ethnicity, controls vs test group, operating system

*Ordinal* variables are like nominal, but the categories have an ordering/ranking

**Examples:** Likert (rating) scale, number of visits to a website

\* conceptually, but usually not practically :)



# Things To Look For

— — —

- Title (most plots should have this)
- Axis labels (almost all plots should have this)
- Type(s) of variable(s) being visualized
  - Including ones used to dictate colors, shapes, patterns, sizes, opacities, etc.
- Independent ("*predictor*") variables (e.g. time) are usually on the x axis
- Dependent ("*outcome*" / "*response*" ) variables are usually on the y axis
- Scales (especially log-transformed ones)

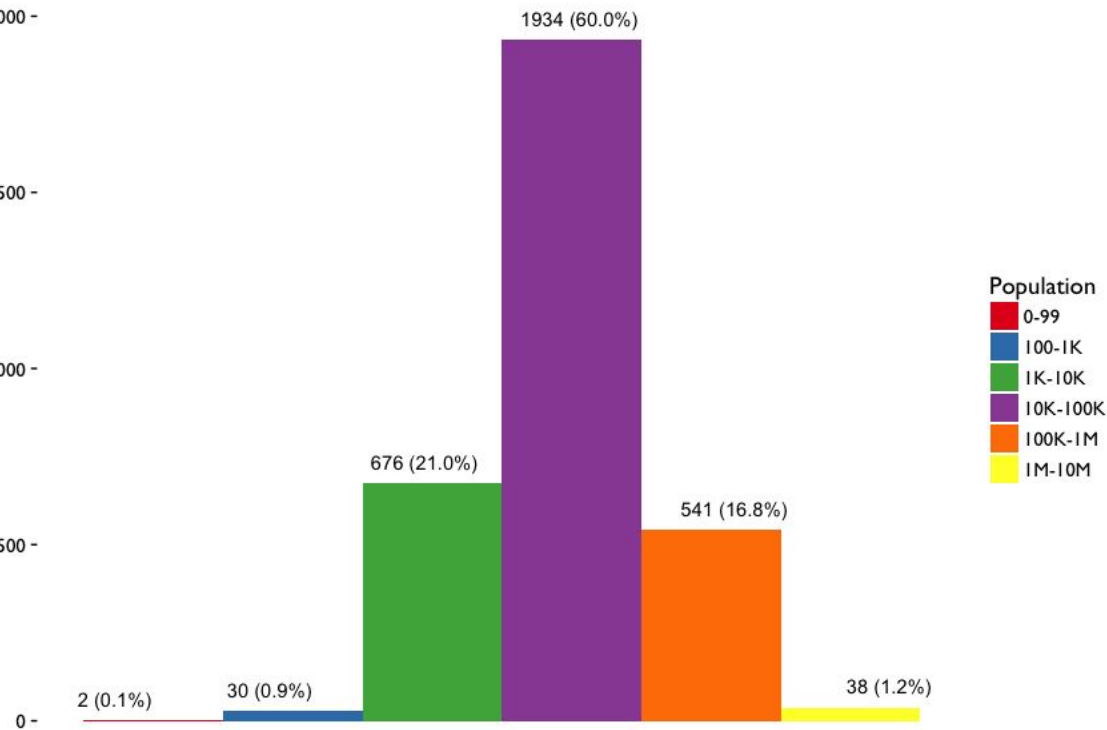


# Some common data visualizations

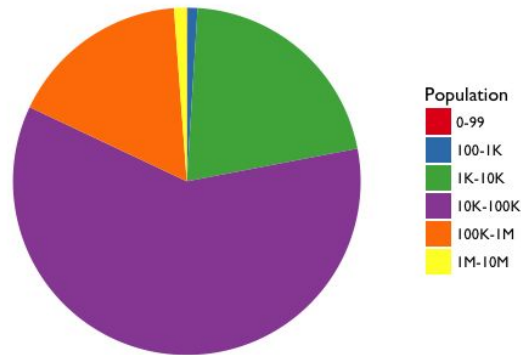


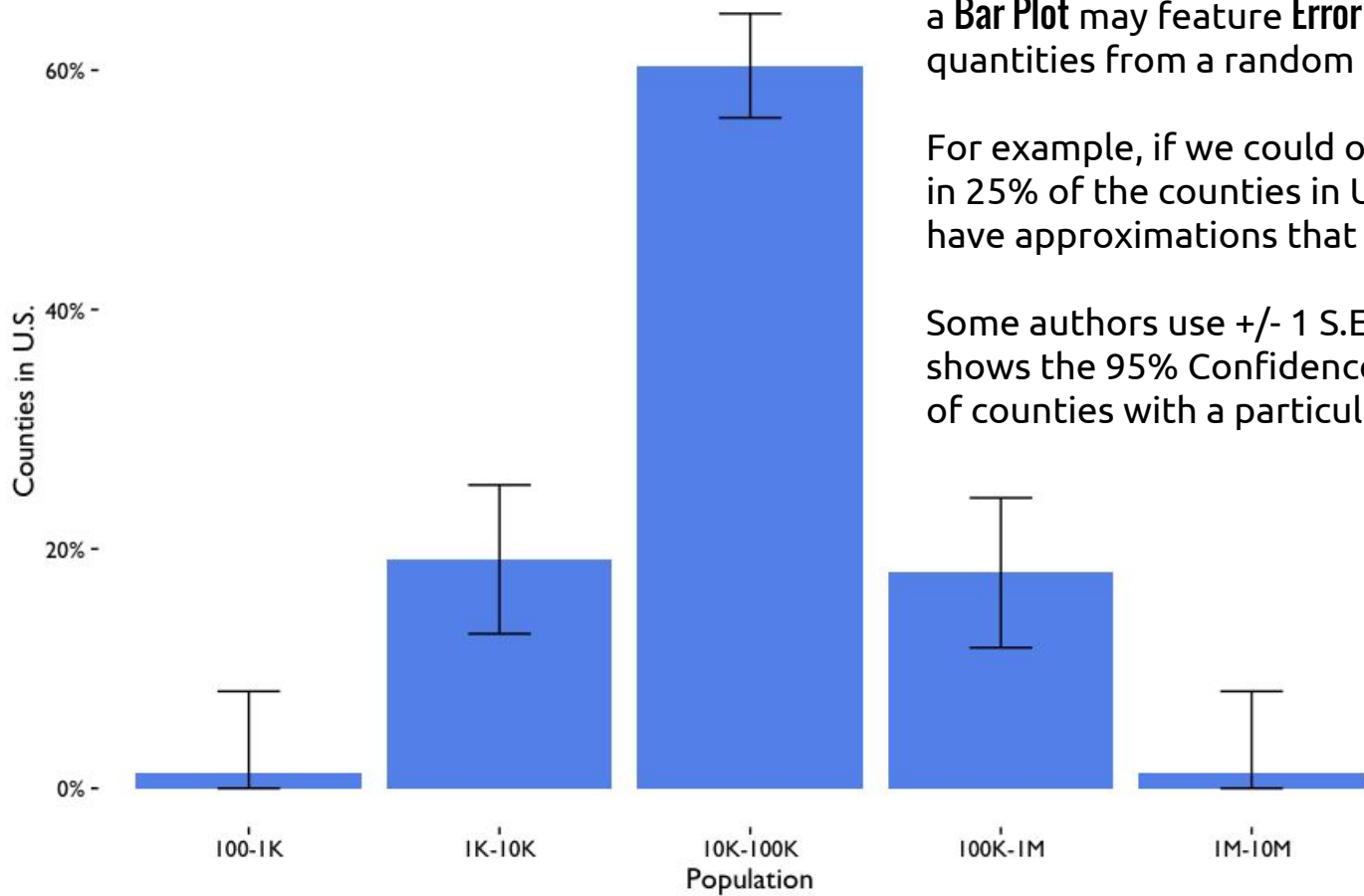
Population	Counties	Proportion
0-99	2	0.001
100-1K	30	0.009
1K-10K	676	0.210
10K-100K	1934	0.600
100K-1M	541	0.168
1M-10M	38	0.012

A **Pie Chart** (bottom left) and a **Bar Plot** (below) are an easy way to visually compare values in the table on the left. The **Pie Chart** is excellent for 2-4 categories, the table is for 5-8 categories, and the **Bar Plot** works great for comparing more than 8 categories.



Pie Chart of Counties in U.S. by Population





a **Bar Plot** may feature **Error Bars** when using estimated quantities from a random sample

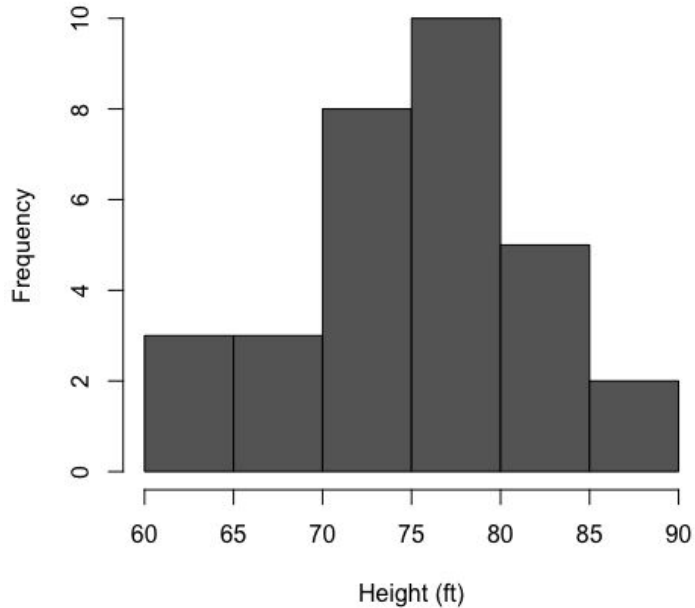
For example, if we could only measure the population in 25% of the counties in U.S. (805), we would only have approximations that have standard errors.

Some authors use  $\pm 1$  S.E. for the bounds. This graph shows the 95% Confidence Interval for the proportion of counties with a particular population size.

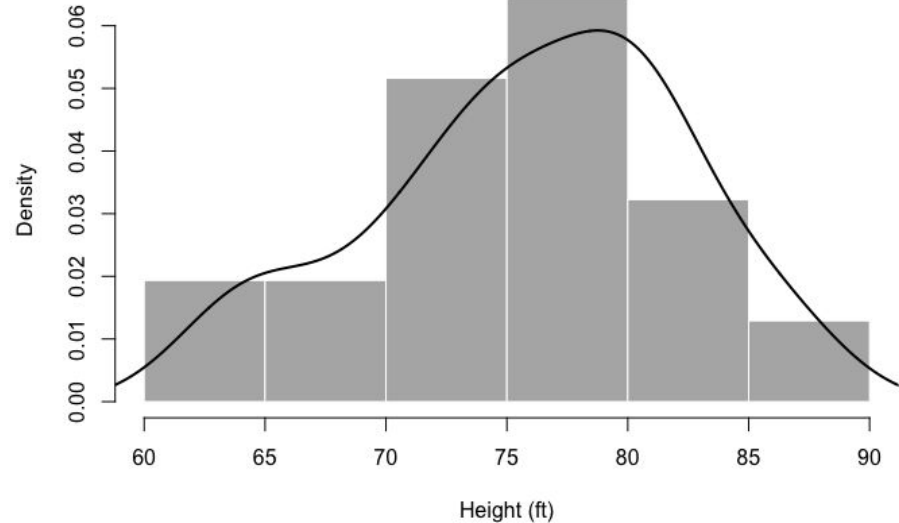


a **Histogram** or a **Density Plot** is a way to visualize a continuous variable's distribution

**Histogram of Height of Black Cheery Trees**



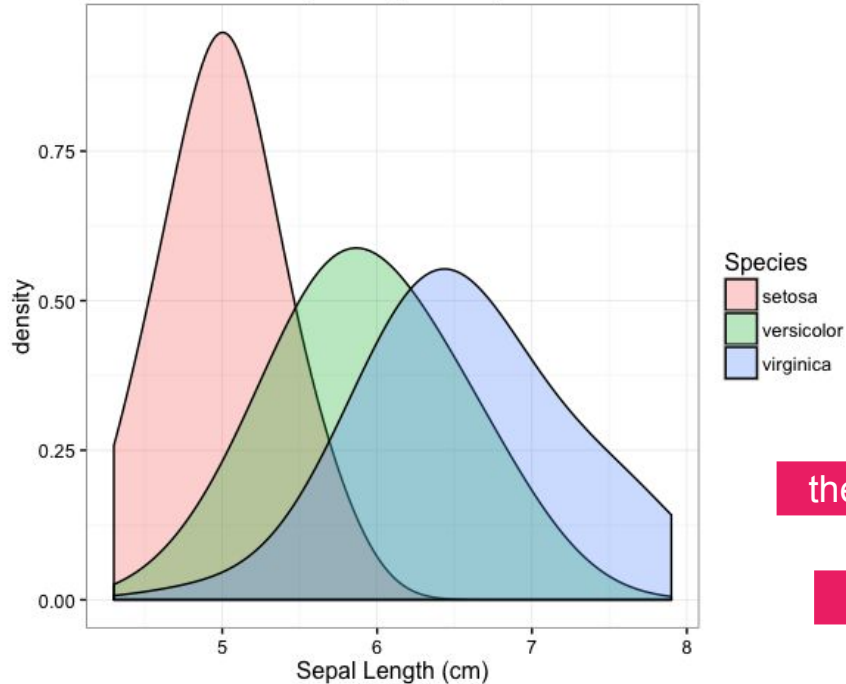
**Histogram & Smooth Density of Height of Black Cheery Trees**





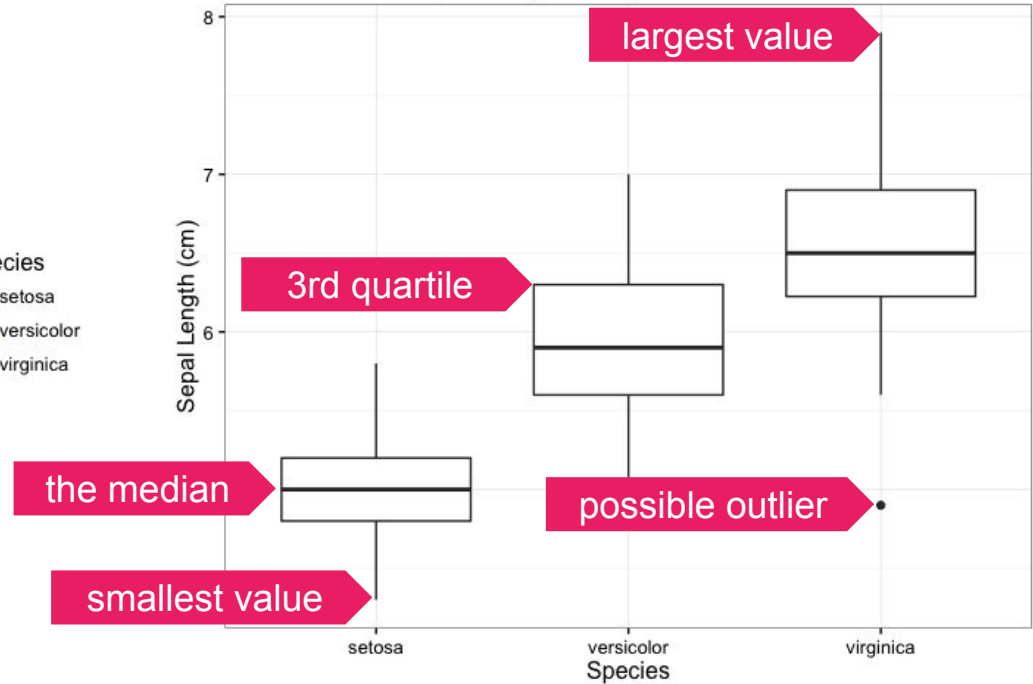
a **Density Plot** makes it easy to compare (smoothed) distributions between groups

Distribution of Sepal Length of Species of Iris flowers



a **Box-and-Whiskers Plot** is a way to vaguely compare distributions between groups

Distribution of Sepal Length of Species of Iris flowers



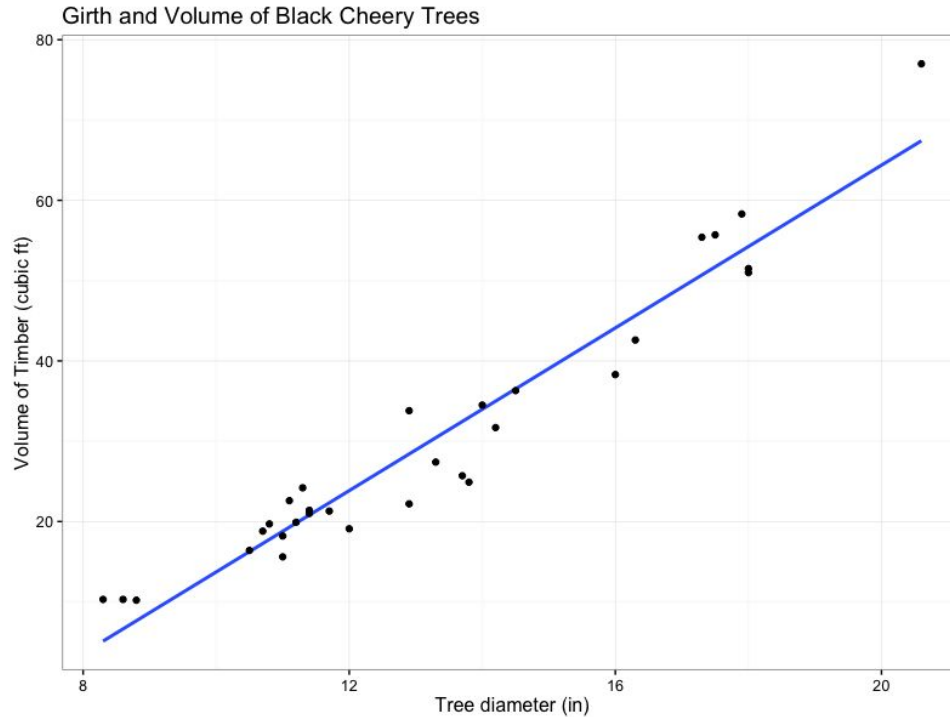
## Violin Plot



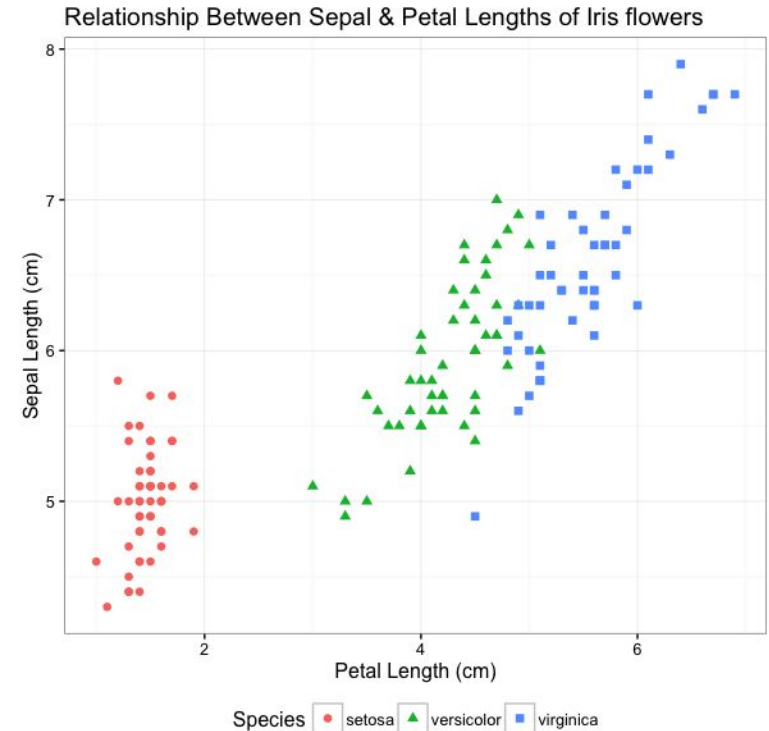
## Violin Plot + Box Plot



## Scatterplot with Trend Line (using Simple Linear Regression)



## Scatterplot with Groups



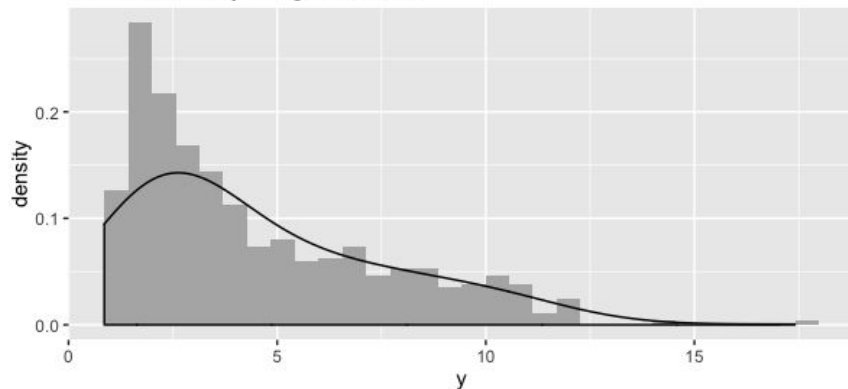
Iris data collected by Anderson, Edgar (1935). The irises of the Gaspé Peninsula, Bulletin of the American Iris Society, 59, 2–5.

Trees dataset: Ryan, T. A., Joiner, B. L. and Ryan, B. F. (1976) The Minitab Student Handbook. Duxbury Press.

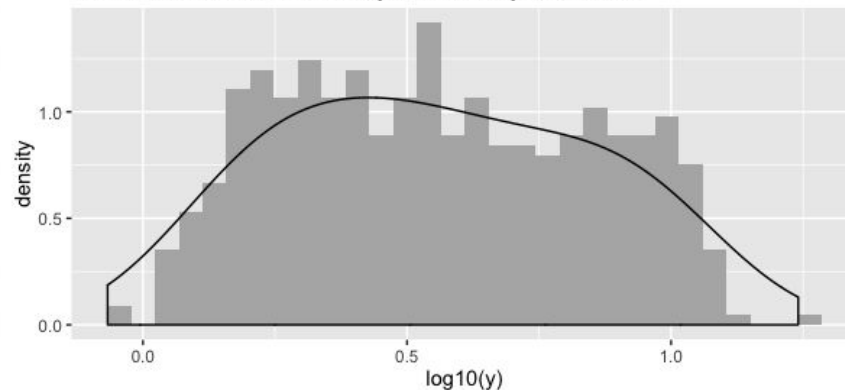


## Histograms and Scatterplots sometimes employ variable transformations for skewed data & nonlinear relationships

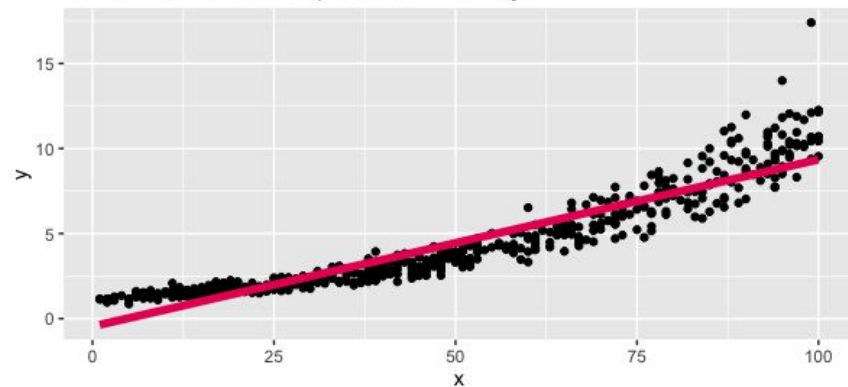
Distribution of  $y$  is right skewed



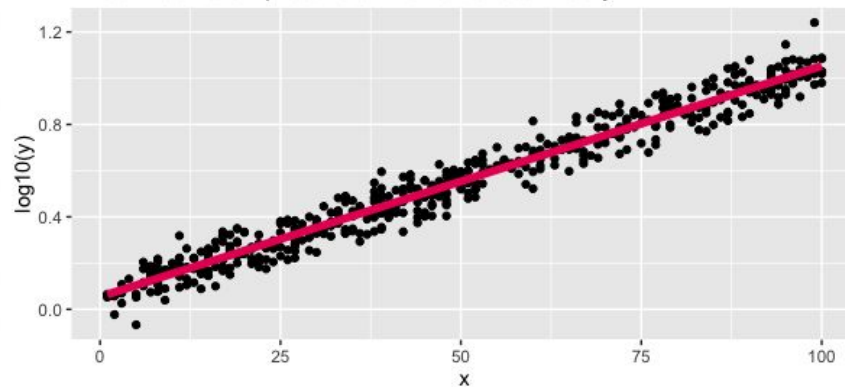
Distribution of transformed  $y$  is Normally distributed



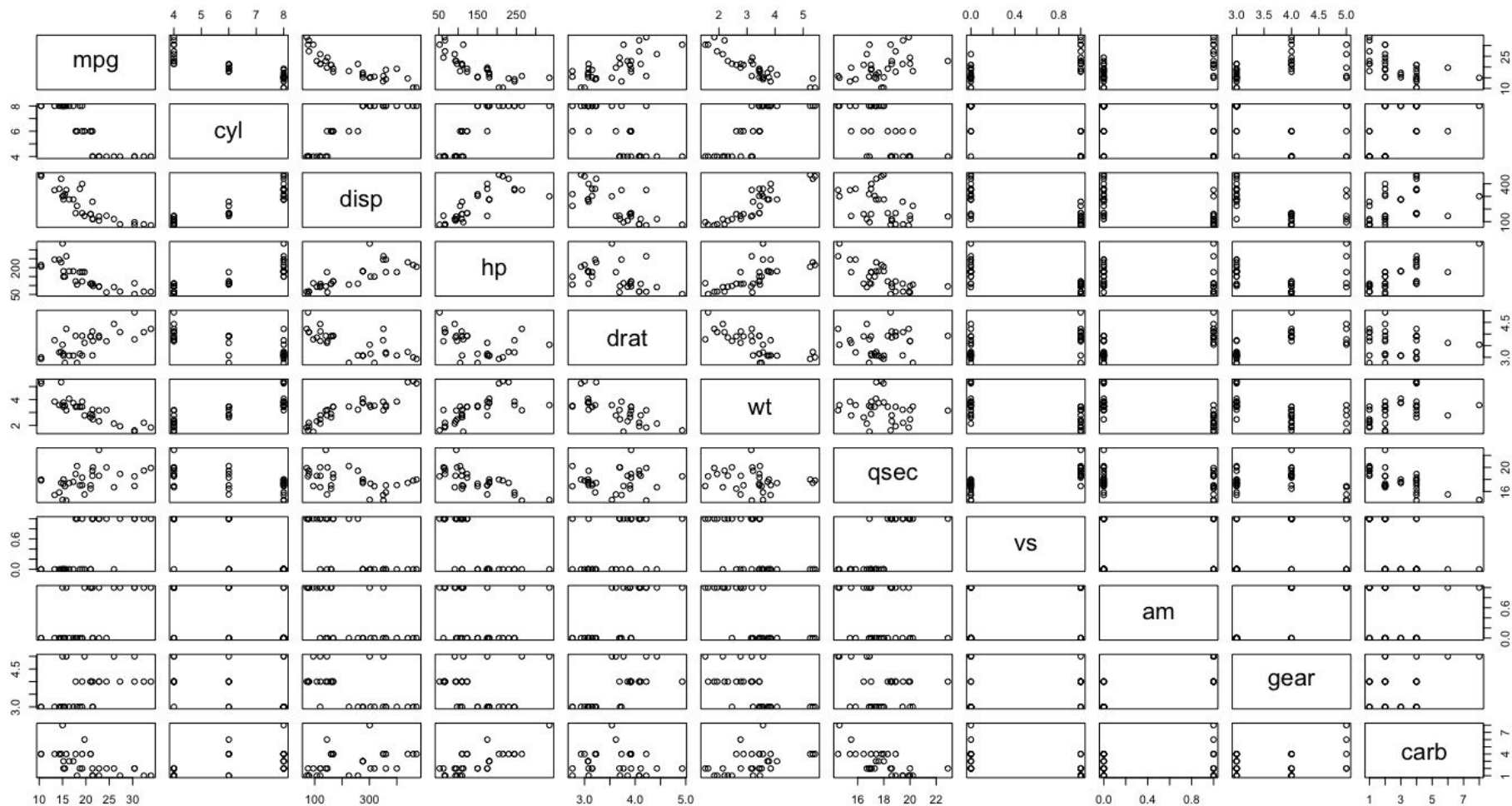
Not a linear relationship between  $x$  and  $y$



Linear relationship between  $x$  and transformed  $y$

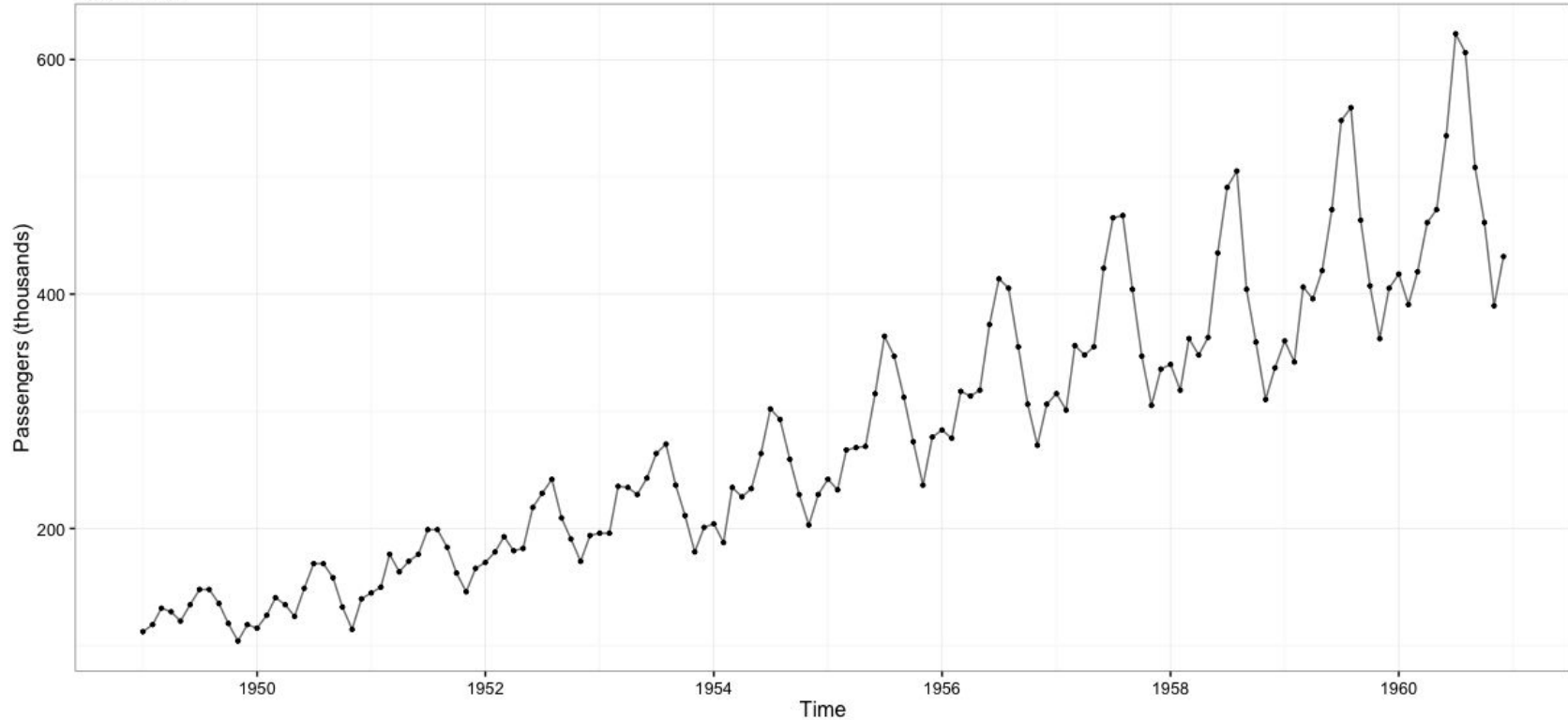


Scatterplot Matrix of 'Motor Trend' car road tests (1974)

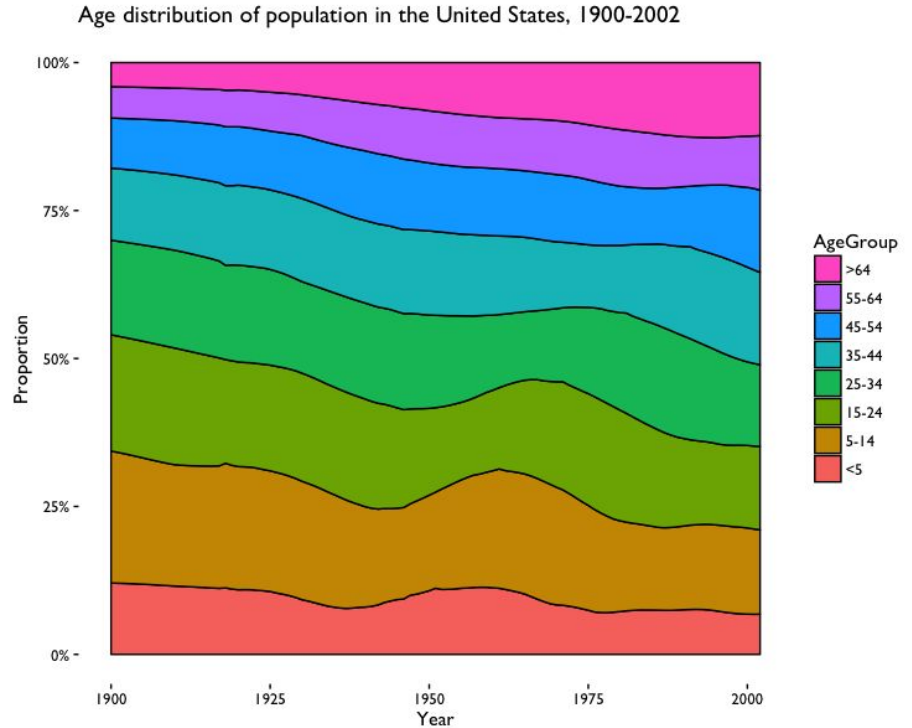
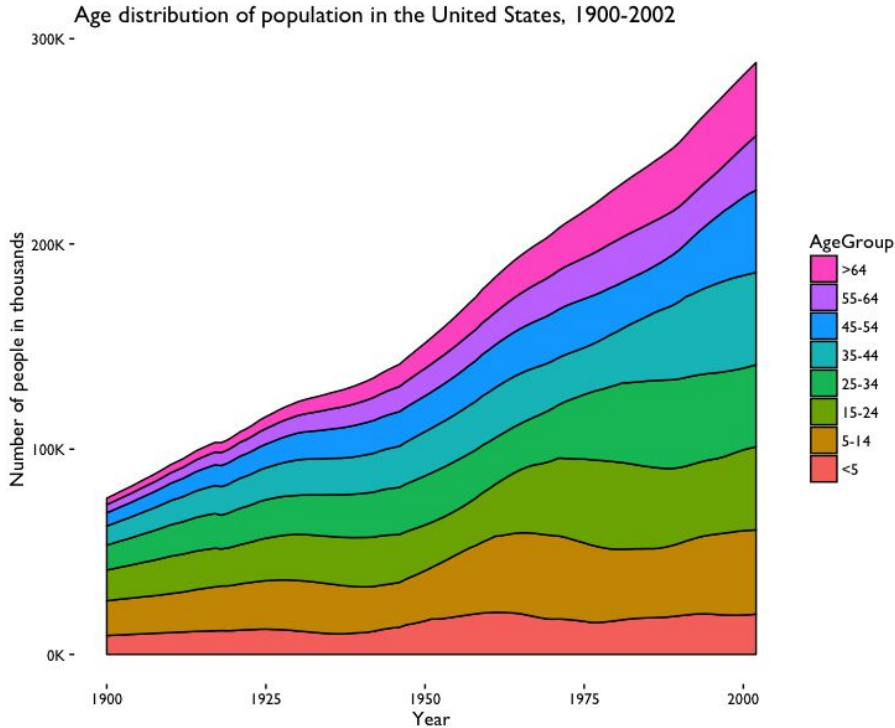


a **Time Series Plot** is a type of scatterplot with time on the x axis, and is a way to visualize changes and patterns over time

Monthly totals of international airline passengers  
1949 to 1960

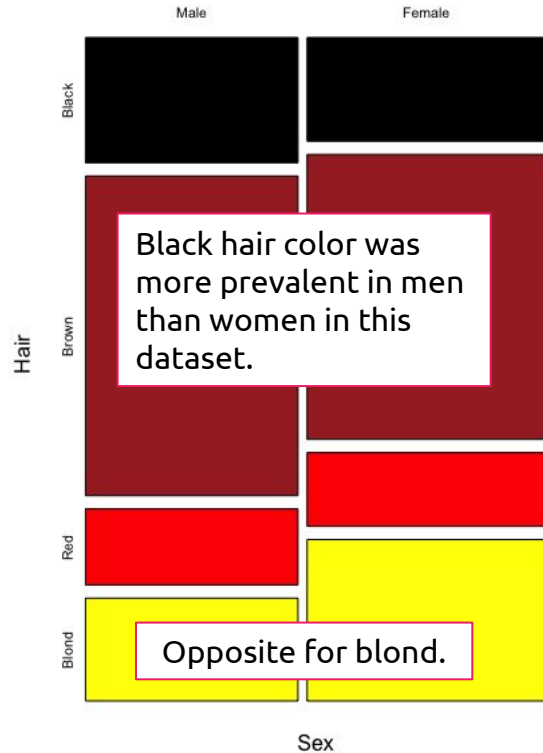


a **Stacked Area Plot** is a way to visualize multiple amounts/proportions changes over time

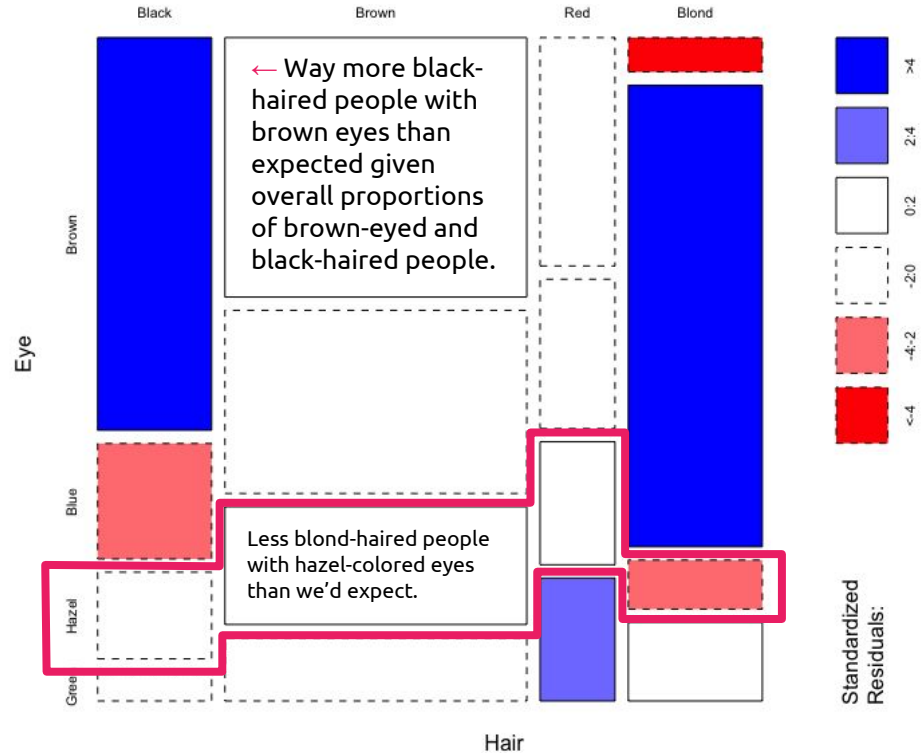


a **Mosaic Plot** is a way to visualize proportions in two or more categorical variables

**Mosaic Plot of Men and Women's Hair Colors**



**Shaded Mosaic Plot of Hair and Eye Colors**





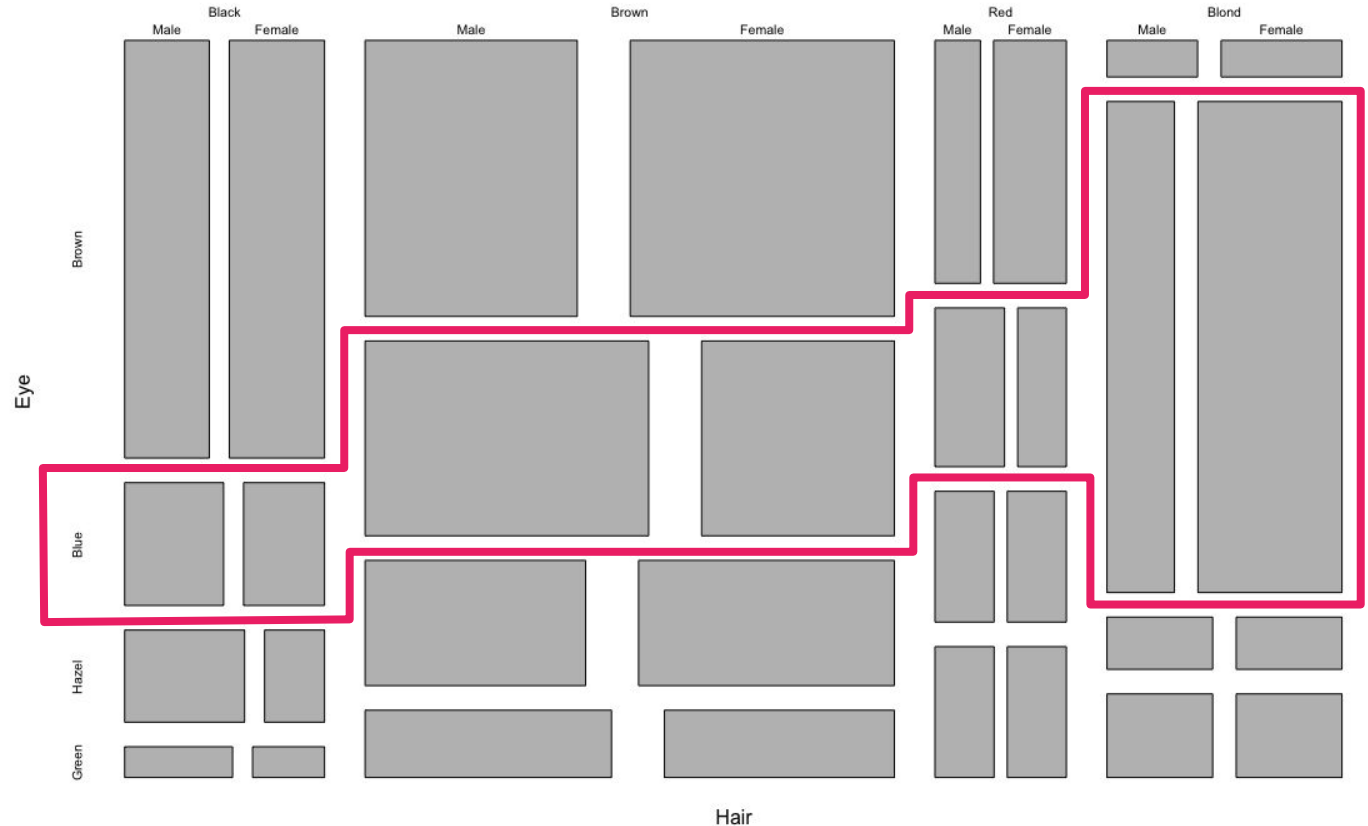
Mosaic Plot of Hair and Eye Colors in Women and Men

## What this tells us...

Blond was the most prevalent hair color among those with blue eyes.

More brown-haired men had blue eyes than brown-haired women.

More blond-haired women had blue eyes than blonde-haired men.



## What this tells us...

On average, only half of the sessions last longer than about 10 seconds.

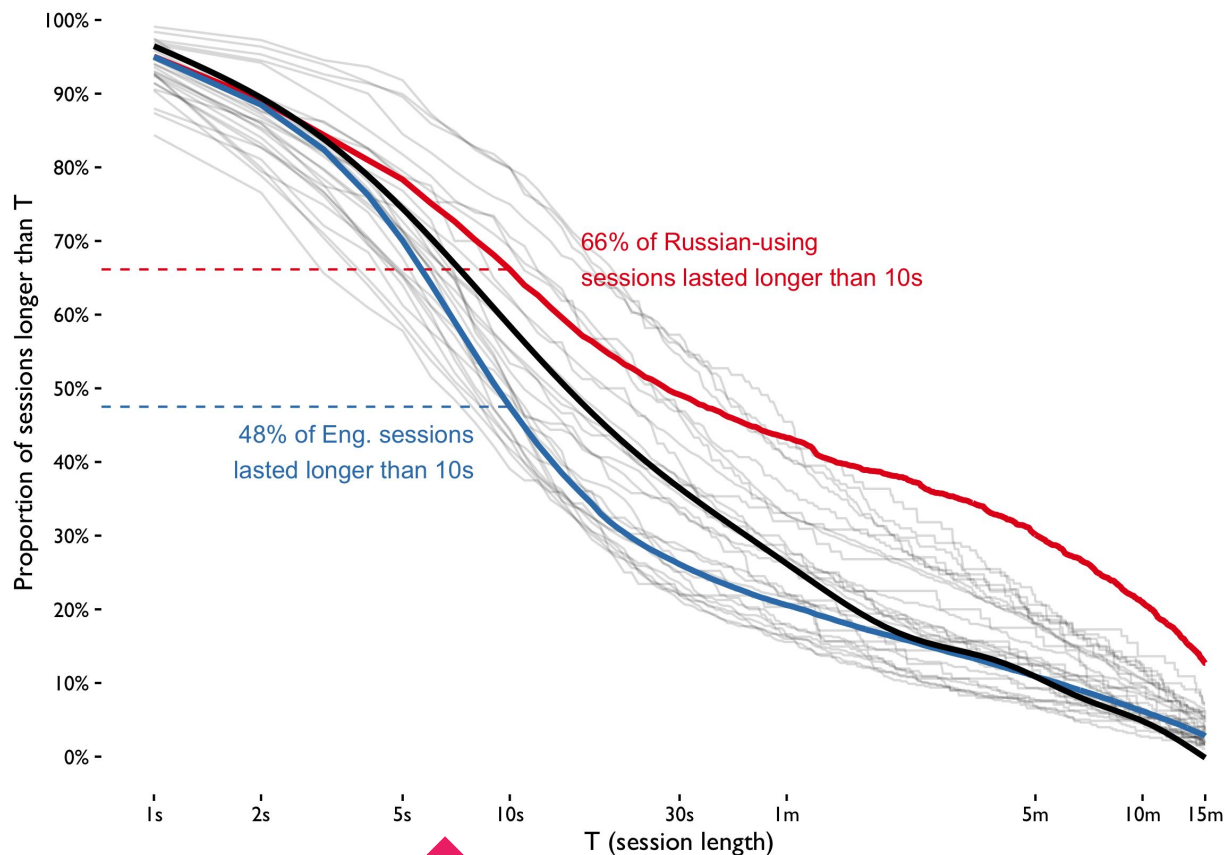
Users of different languages have different session lengths.

Russian-preferring users' sessions that are longer than English-preferring users' sessions.

Steeper means faster drop-out -- such as shorter time-to-clickthrough.

### Session "survival" curve by most preferred language

108.6K sessions; top 30 languages by volume of sessions; black curve is smoothed median



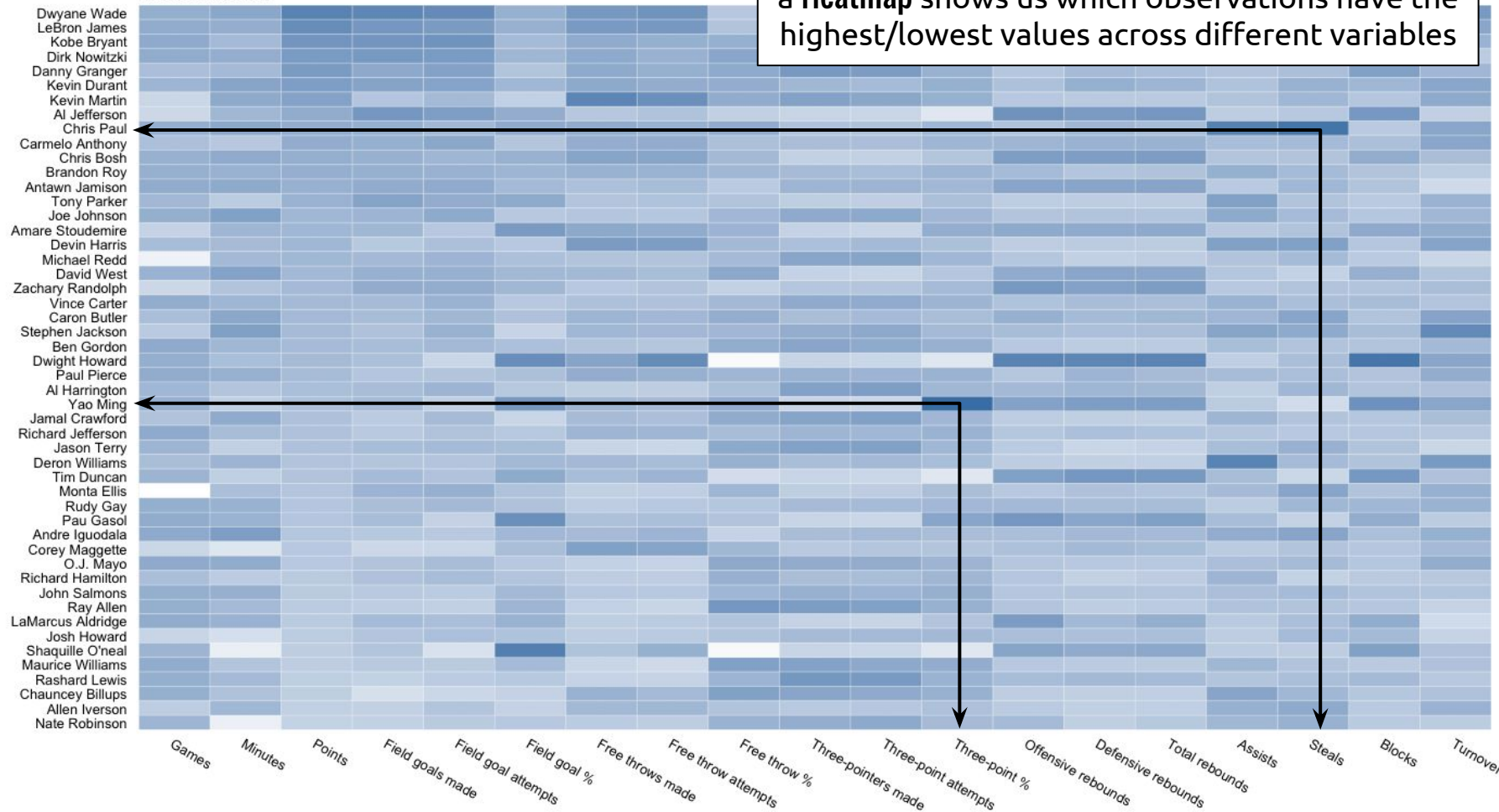
Caution! This is a log-transformed time axis.

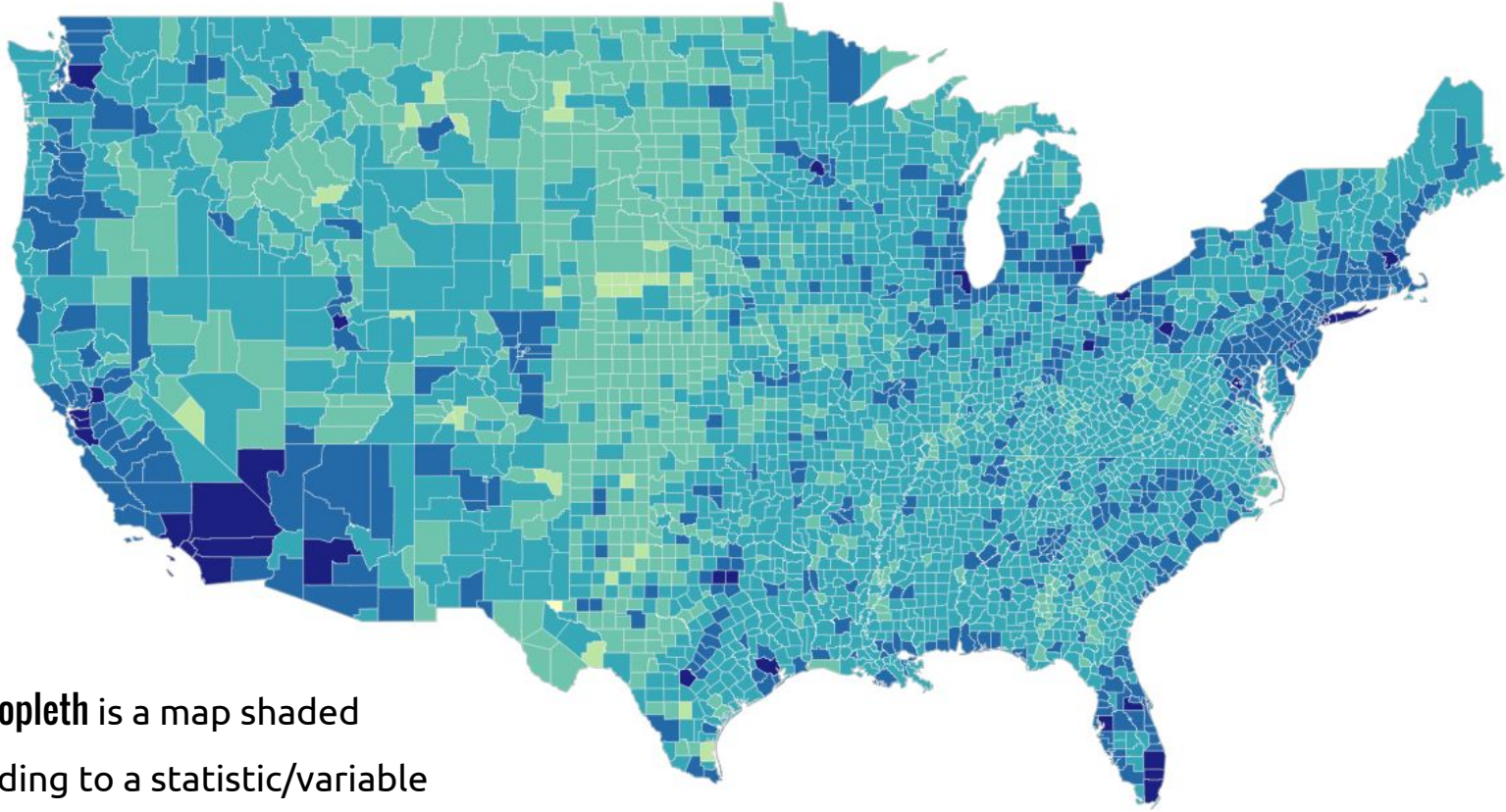
Source: Wikipedia Portal Event Logging Data



NBA per game performance of top 50 scorers  
2008-2009 season

a **Heatmap** shows us which observations have the highest/lowest values across different variables



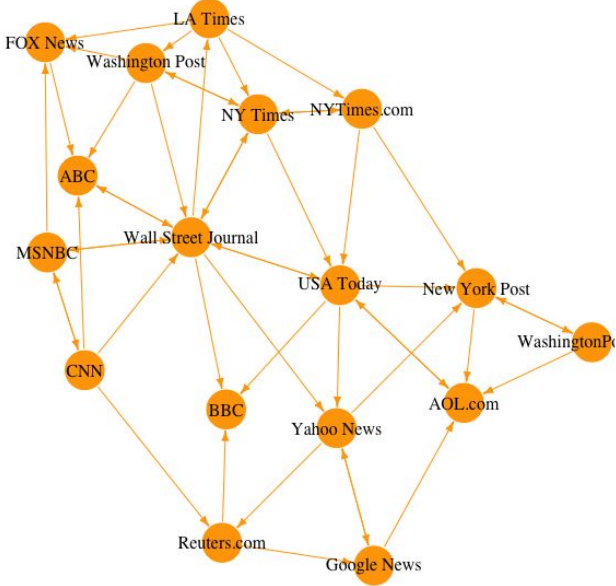


a **choropleth** is a map shaded  
according to a statistic/variable

Population 0-99 100-1K 1K-10K 10K-100K 100K-1M 1M-10M

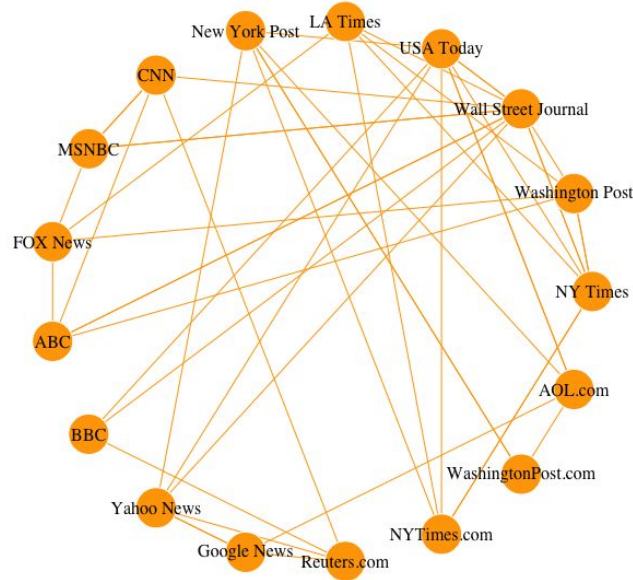


Directed Network Graph



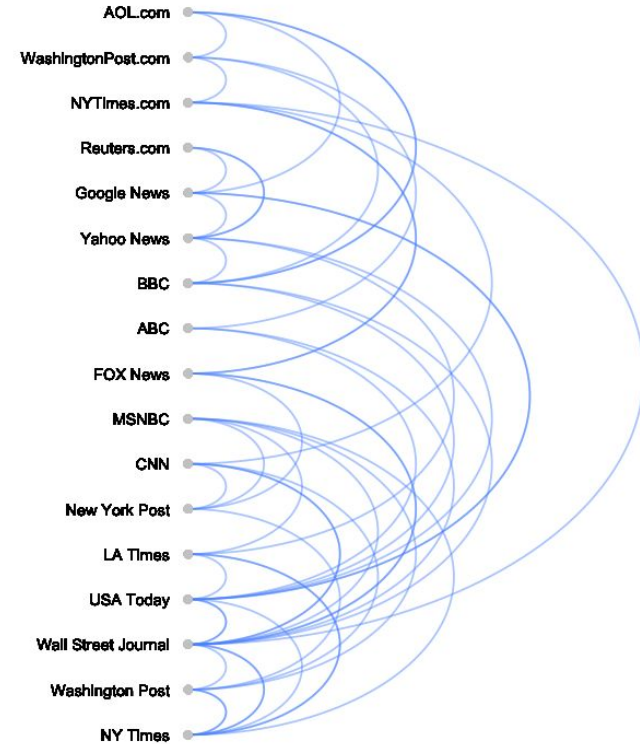
Links and mentions between news sources

Undirected Network Graph (Circular Layout)



Links and mentions between news sources

Arc Diagram Plot



Links and mentions between news sources

...and many more ways to visualize networks!



# Questions?

