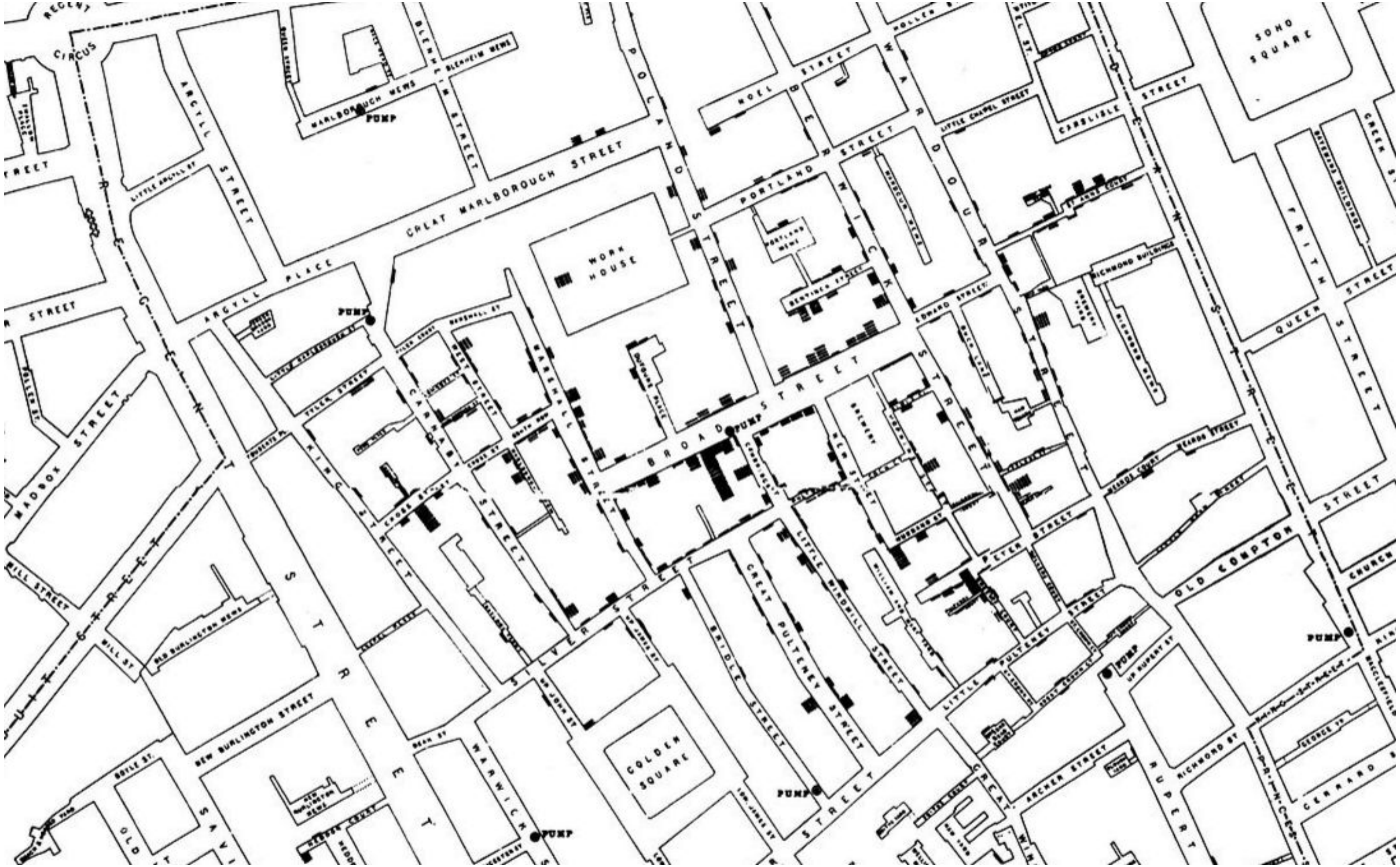
Value of Visualization

Yannet Interian -- USF

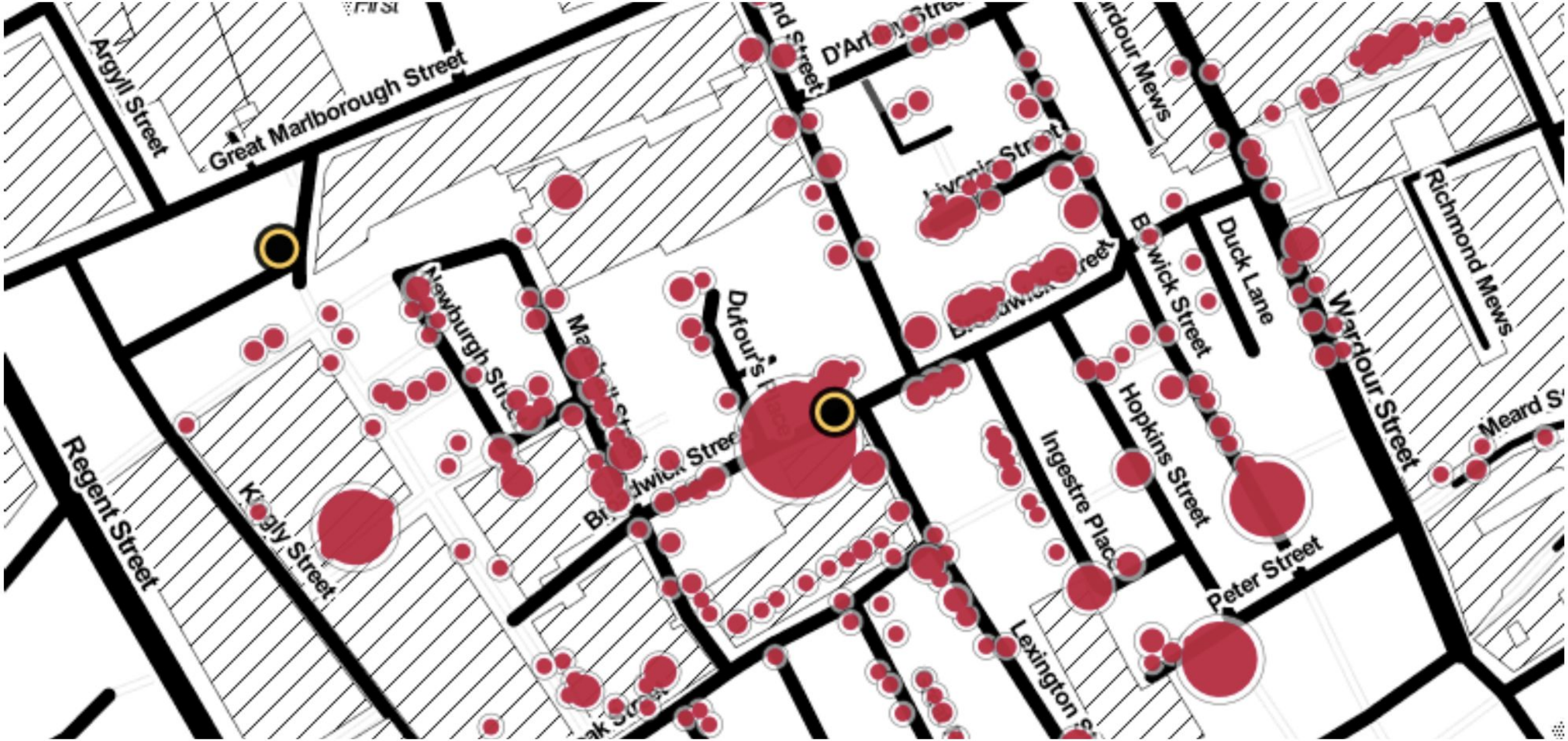
Agenda

- Introduction: Why Visualize
- Data types
- Basic Charts
- Course Admin
- Lab: Advanced ggplot2 (tricks)



http://en.wikipedia.org/wiki/File:Snow-cholera-map-1.jpg

Recreated map with modern tools



http://www.theguardian.com/news/datablog/interactive/2013/mar/15/cholera-map-john-snow-recreated

1854 Cholera Outbreak

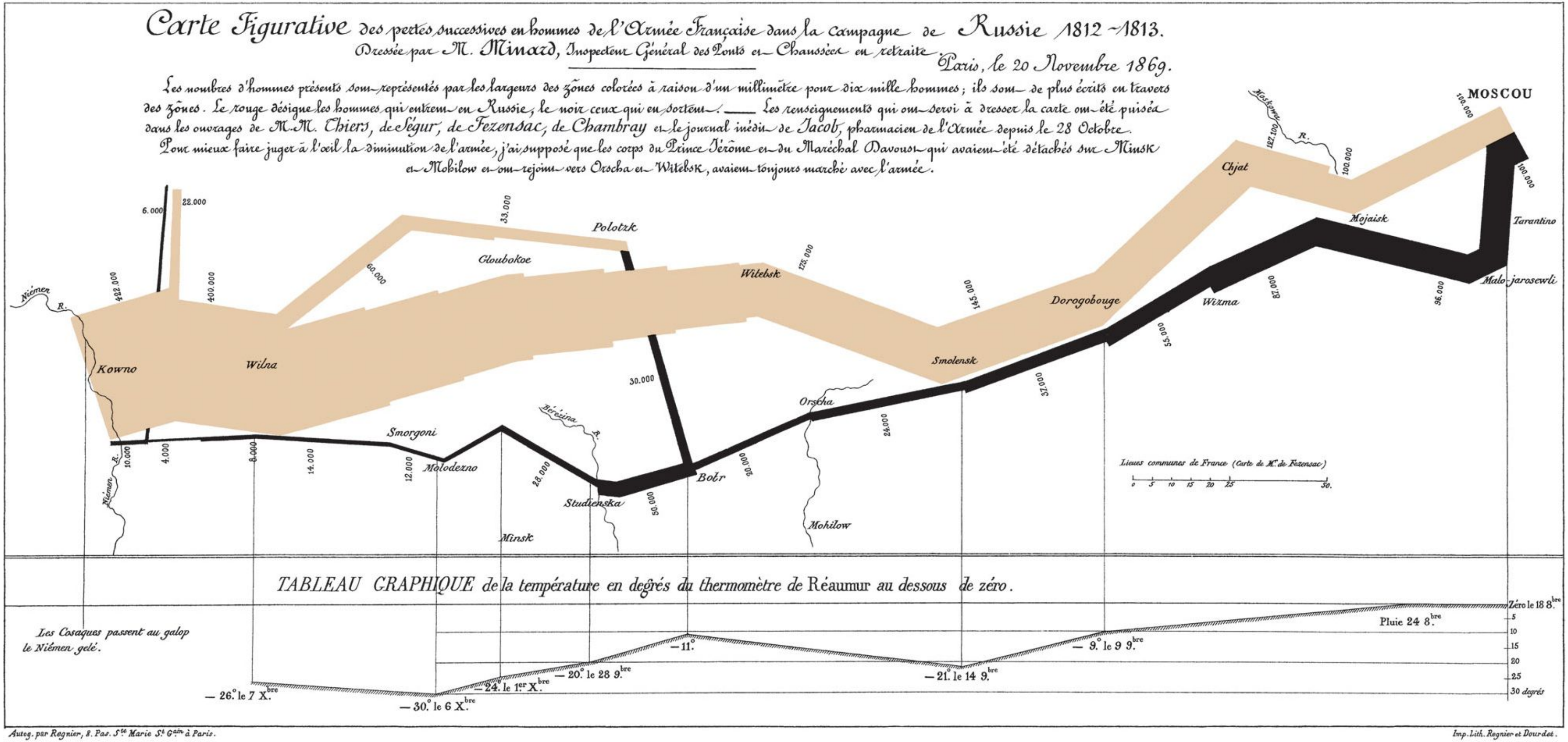
- Tens of thousands people in England were dying of cholera between 1831 and 1854
- Many assumed cholera was caused by breathing vapors or “miasma in the atmosphere”
- People did not have running water or modern toilets
- Terrible cholera outbreak in 1854 in Soho, near where physician John Snow lived
- Tracked down data from hospitals and public records
- Created simple plot of where victims lived and location of water pumps

http://www.ph.ucla.edu/epi/snow/snowcricketarticle.html and
http://www.bbc.co.uk/history/historic_figures/snow_john.shtml

1854 Cholera Outbreak

- John Snow Identified contaminated water pump
- Eventually able to trace many cases to “sherbert” a bubbly drink with a fizzy powder mixed in, served from water coming from the Broad Street area pump
- Pioneered the field of epidemiology

http://www.ph.ucla.edu/epi/snow/snowcricketarticle.html and
http://www.bbc.co.uk/history/historic_figures/snow_john.shtml

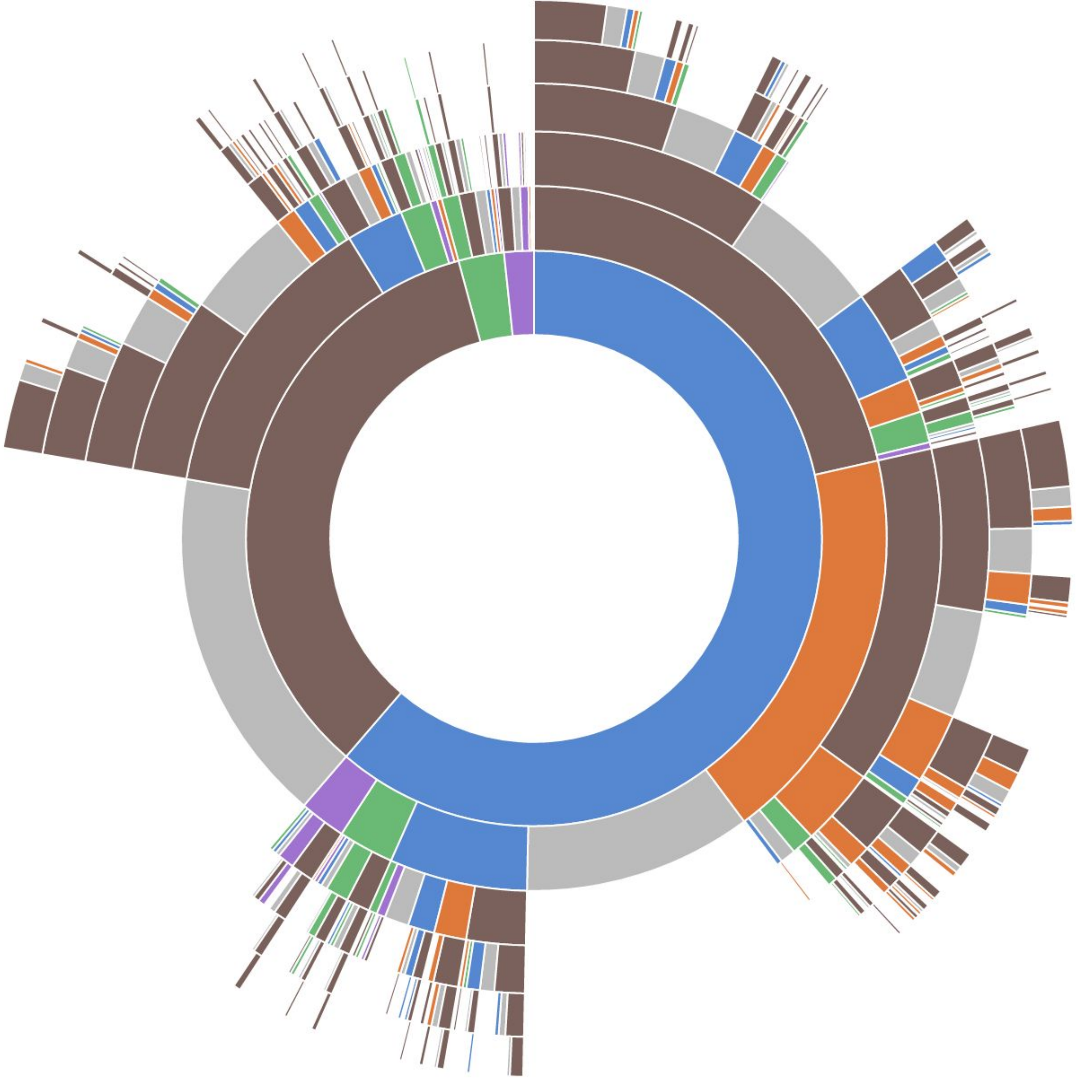


Napoleon’s March on Moscow depicted by M. Minard. Width indicates the number of soldiers. Temperature during the retreat is presented below the map.
https://en.wikipedia.org/wiki/Charles_Joseph_Minard

Napoleon March on Moscow

- Depicts Napoleon's army departing the Polish-Russian border.
- The band illustrates the size of his army at specific geographic points during their advance and retreat.
- It displays six types of data in two dimensions:
  - the number of Napoleon's troops; the distance traveled; temperature; latitude and longitude; direction of travel; and location relative to specific dates.
- This type of band graph for illustration of flows was later called a Sankey diagram.

Sequences sunburst demo

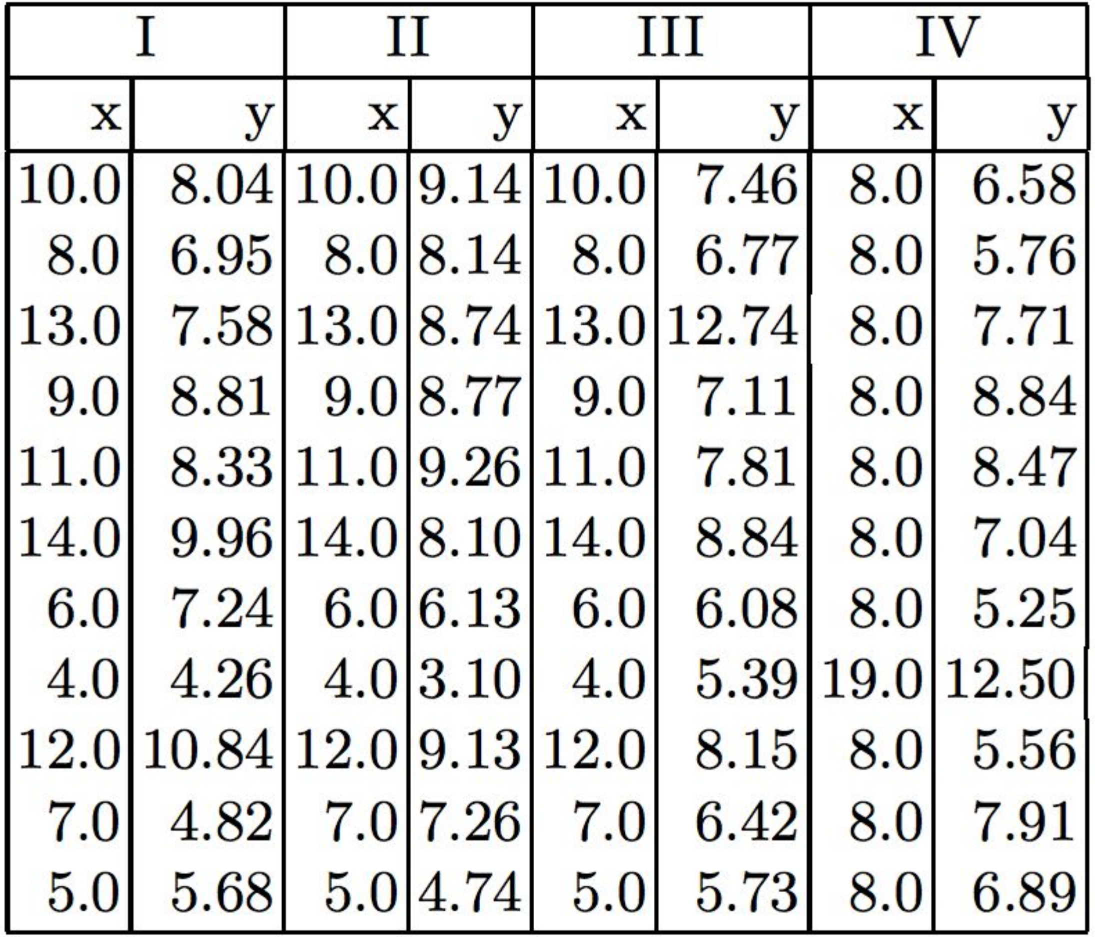


https://bl.ocks.org/kerryrodden/7090426

What is visualization?

What is visualization?

- “Transformation of the symbolic into the geometric” [McCormick et al. 1987]
- “... finding the artificial memory that best supports our natural means of perception.” [Bertin 1967]
- “Information visualization utilizes computer graphics and interaction to assist humans in solving problems.”[Purchase et al., 2008]
- “The use of computer-generated, interactive, visual representations of data to amplify cognition.” [Card, Mackinlay, & Shneiderman 1999]



I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

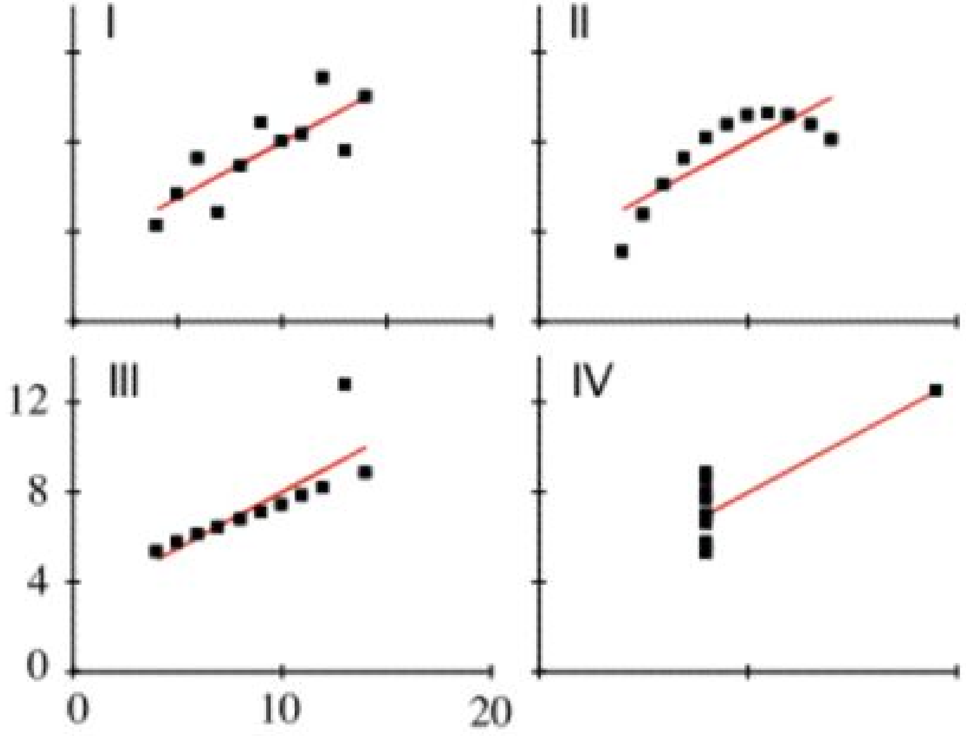
Four datasets with different values and the same statistical profile

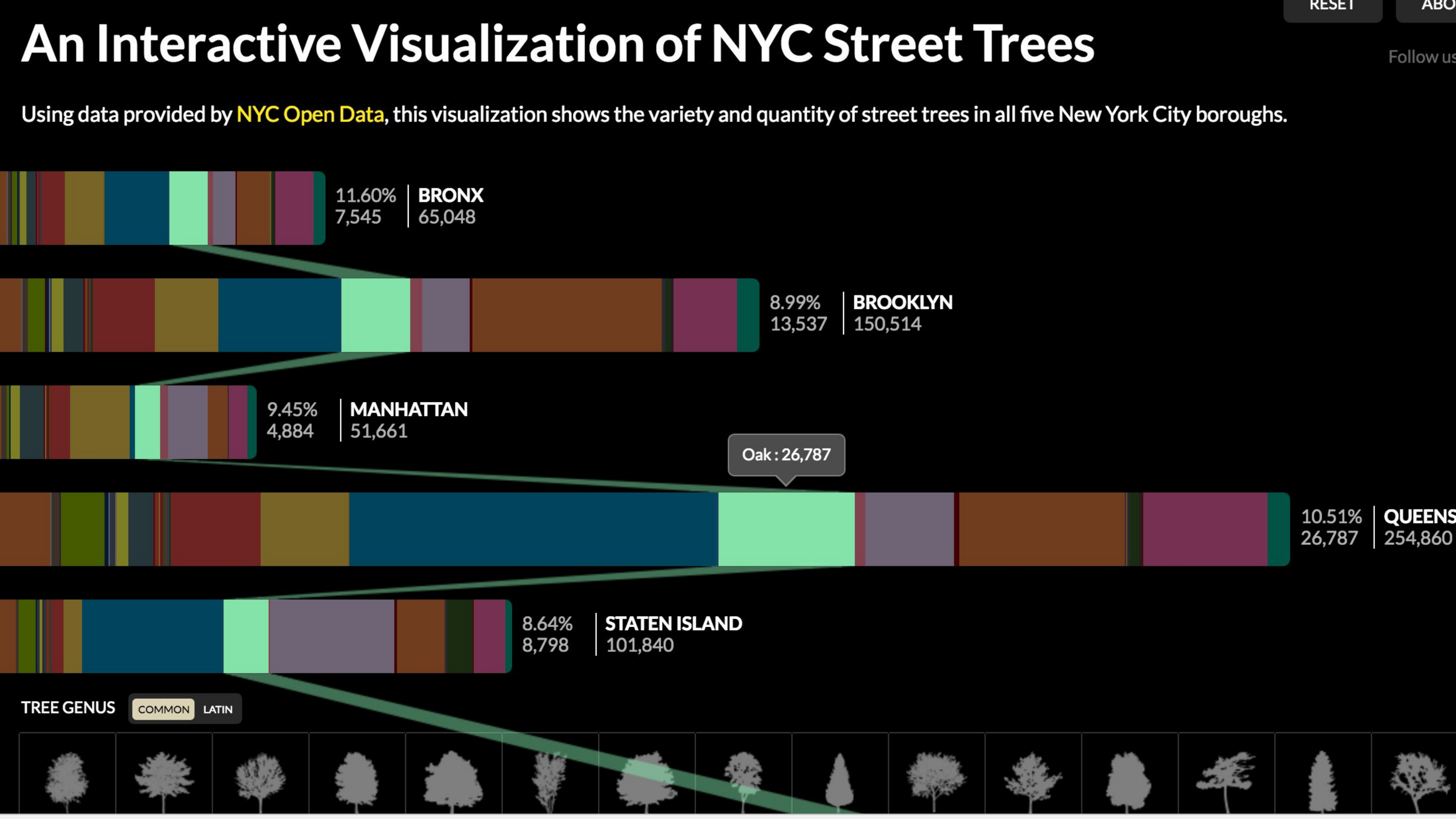
mean(x) = 9.0
mean(y) = 7.5
std(x) = 3.317
std(y) = 2.03

Also same regression line.

[Anscombe 73]

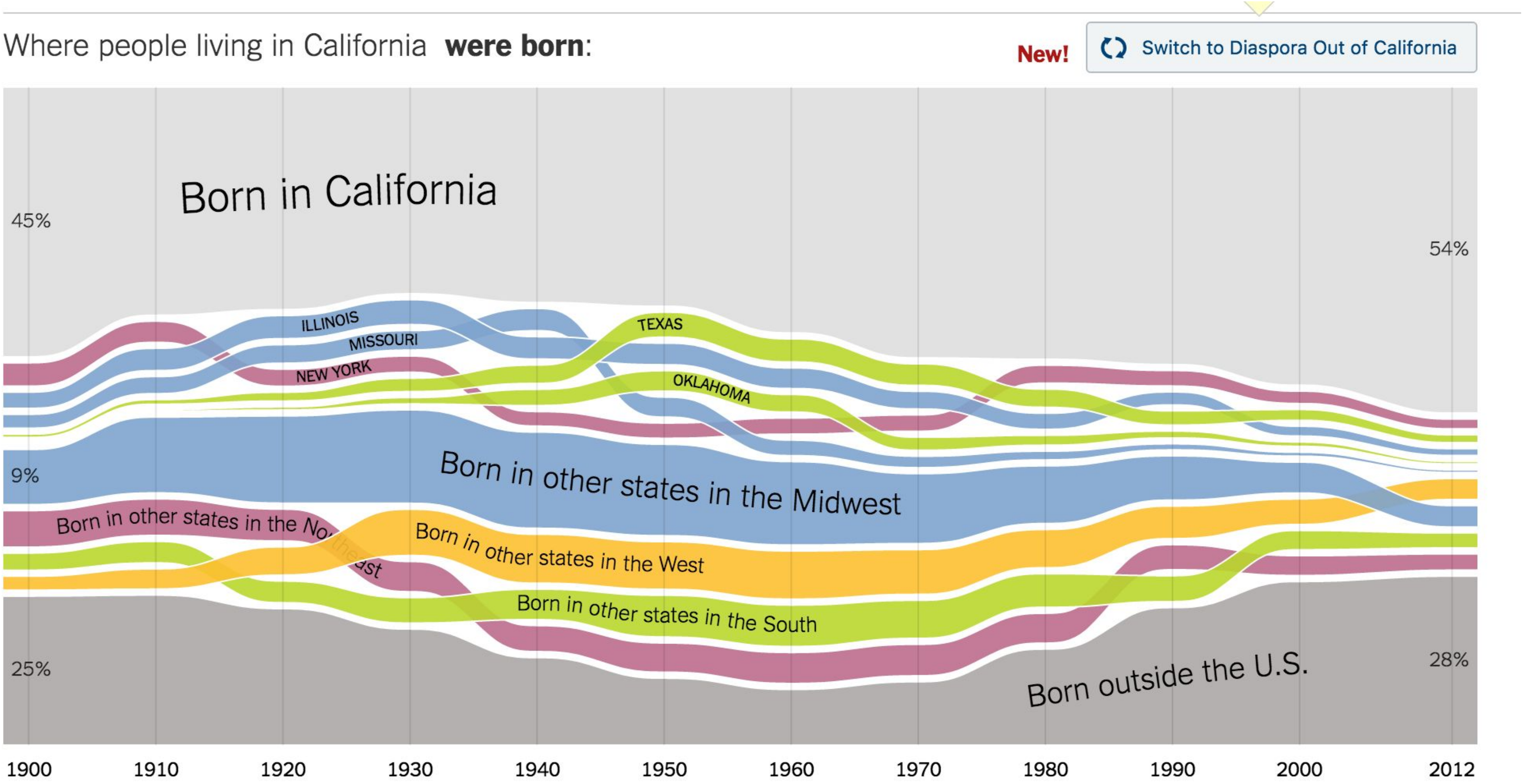
Visualization is much more effective than statistics





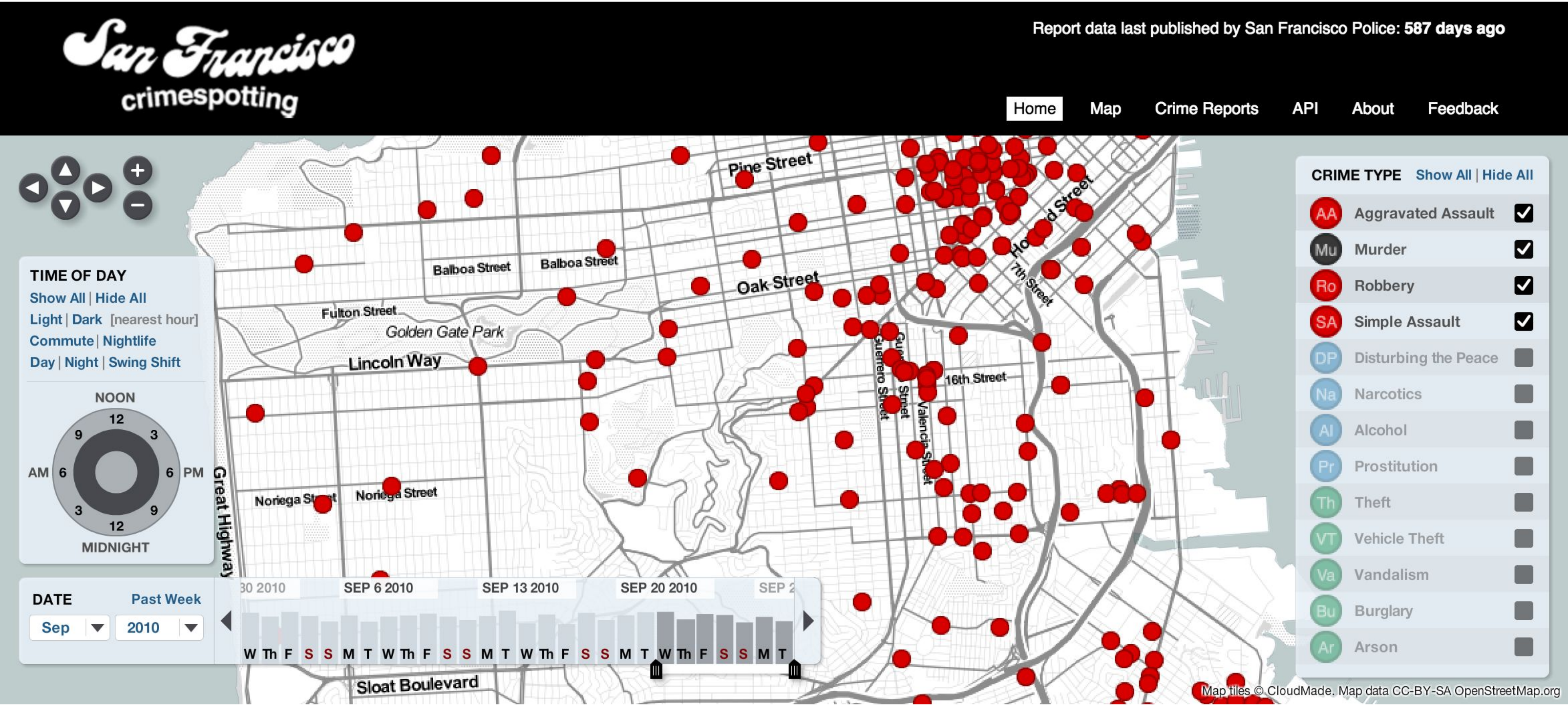
An Interactive Visualization of NYC Street Trees

Using data provided by NYC Open Data, this visualization shows the variety and quantity of street trees in all five New York City boroughs.



Where people living in California **were born**:

http://www.nytimes.com/interactive/2014/08/13/upshot/where-people-in-each-state-were-born.html?abt=0002&abg=0



Report data last published by San Francisco Police: **587 days ago**

DATE Past Week

Sep 2010

CRIME TYPE Show All | Hide All

AA	Aggravated Assault	☑
Mu	Murder	☑
Ro	Robbery	☑
SA	Simple Assault	☑
DP	Disturbing the Peace	□
Na	Narcotics	□
Al	Alcohol	□
Pr	Prostitution	□
Th	Theft	□
VT	Vehicle Theft	□
Va	Vandalism	□
Bu	Burglary	□
Ar	Arson	□

http://sanfrancisco.crimespotting.org/

Why do we create visualizations?

Why do we create visualizations

- Answer questions (or discover them)
- Make decisions
- See data in context
- Expand memory
- Support graphical calculation
- Find patterns
- Present argument or tell a story
- Inspire

Three functions of visualizations

- Record: store information
  - Photographs, blueprints, …
- Analyze: support reasoning about information
  - Process and calculate
  - Reason about data
  - Feedback and interaction
- Communicate: convey information to others
  - Share and persuade
  - Collaborate and revise
  - Emphasize important aspects of data

Data Types

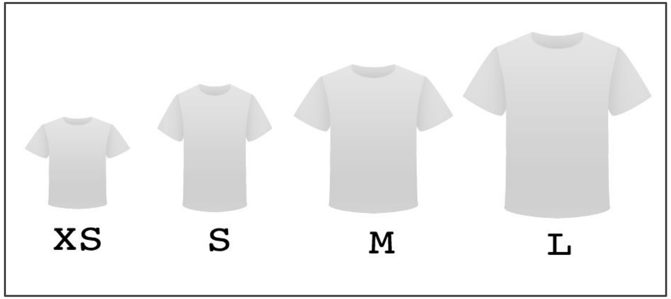
Data Types

In order to visualize data

1. Classify data types
2. Determine which type of plots represent the data types more effectively

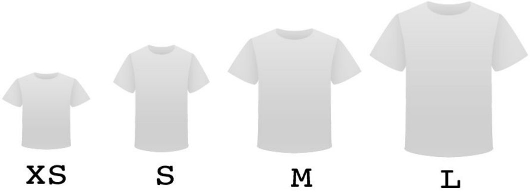
Data Types

- Nominal (categorical)
- Ordinal (categorical)
- Quantitative (numerical)
  - Interval
  - Ratio



Data Types

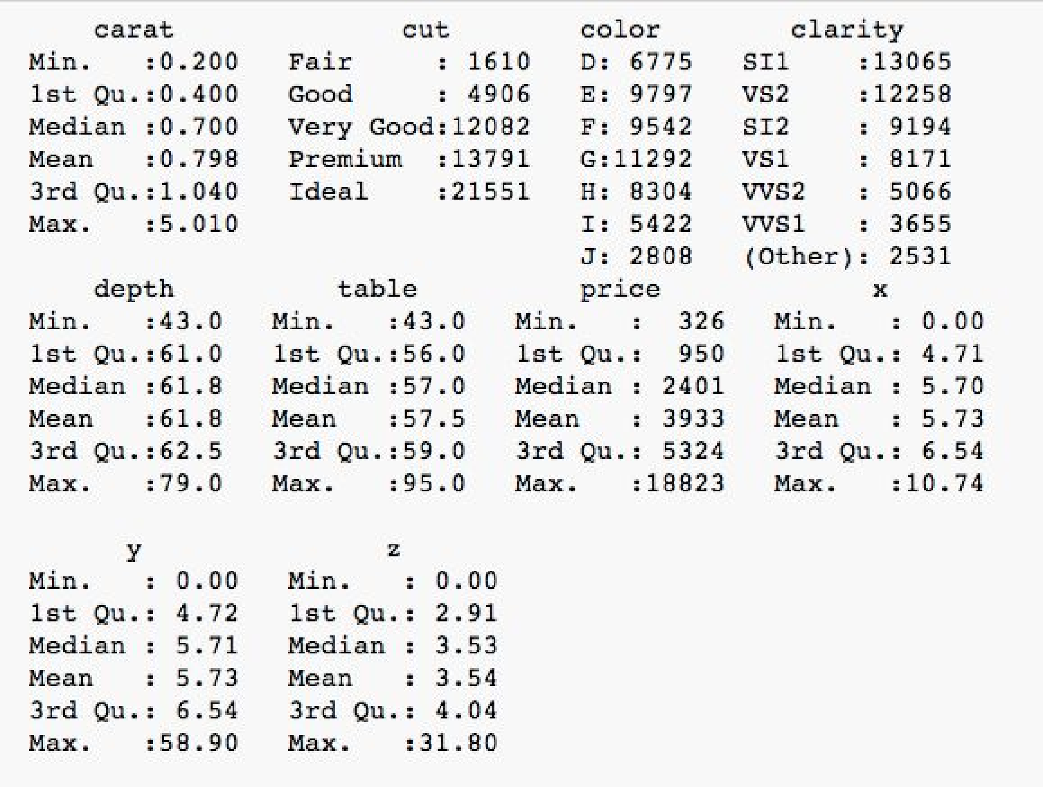
- Nominal (categorical)
  - Fruits, apple, bananas
- Ordinal (categorical)
  - Shirt size (XS, S, M, L)
  - Phone call quality (bad, good, excellent)
- Quantitative (numerical)
  - Interval (Zero is arbitrary)
    - Dates, locations
    - Differences can be compared
  - Ratio (Zero is fixed)
    - Zero indicates that there is none of that variable
    - Measurements
    - Counts and amounts



Nominal, Ordinal, Quantitative

- Nominal
  - Operations: ==, !=
- Ordinal
  - Operations: ==, !=, <, >
- Quantitative
  - Interval (Zero is arbitrary)
    - Operations: ==, !=, <, >, -
  - Ratio (Zero is fixed)
    - Operations: ==, !=, <, >, -, +, /

Dataset: diamonds (ggplot2)



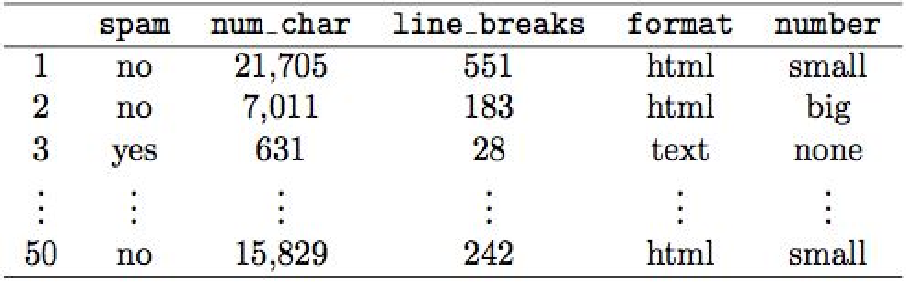
```
    carat                  cut          color        clarity    
 Min.   :0.200   Fair     : 1610   D: 6775   SI1    :13065  
 1st Qu.:0.400   Good     : 4906   E: 9797   VS2    :12258  
 Median :0.700   Very Good:12082   F: 9542   SI2    : 9194  
 Mean   :0.798   Premium  :13791   G:11292   VS1    : 8171  
 3rd Qu.:1.040   Ideal    :21551   H: 8304   VVS2   : 5066  
 Max.   :5.010                     I: 5422   VVS1   : 3655  
                                   J: 2808   (Other): 2531  
    depth           table           price             x         
 Min.   :43.0   Min.   :43.0   Min.   :  326   Min.   : 0.00  
 1st Qu.:61.0   1st Qu.:56.0   1st Qu.:  950   1st Qu.: 4.71  
 Median :61.8   Median :57.0   Median : 2401   Median : 5.70  
 Mean   :61.8   Mean   :57.5   Mean   : 3933   Mean   : 5.73  
 3rd Qu.:62.5   3rd Qu.:59.0   3rd Qu.: 5324   3rd Qu.: 6.54  
 Max.   :79.0   Max.   :95.0   Max.   :18823   Max.   :10.74  

       y                z         
 Min.   : 0.00   Min.   : 0.00  
 1st Qu.: 4.72   1st Qu.: 2.91  
 Median : 5.71   Median : 3.53  
 Mean   : 5.73   Mean   : 3.54  
 3rd Qu.: 6.54   3rd Qu.: 4.04  
 Max.   :58.90   Max.   :31.80  
```

- Prices and quality information of 54k diamonds.
- Diamond quality: carat, cut, color, clarity
- Physical measurements: depth, table, x,y,z

What type of variables?

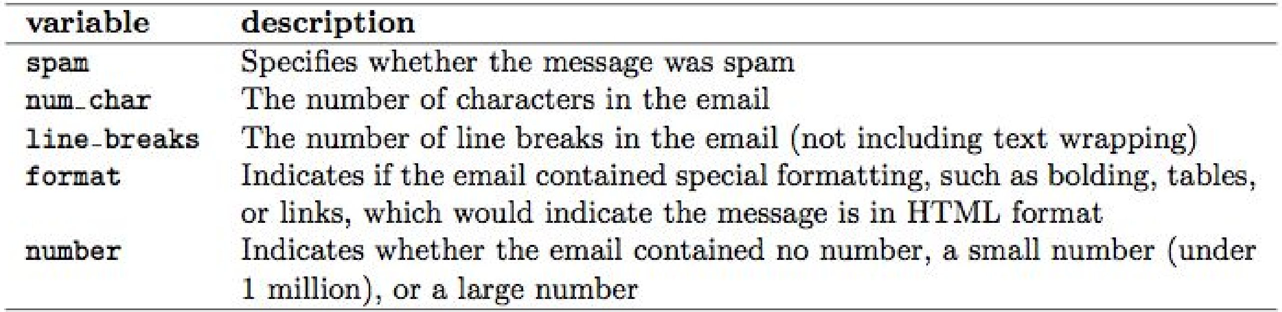
Dataset: Email spam

What type of variables?



	spam	num_char	line_breaks	format	number
1	no	21,705	551	html	small
2	no	7,011	183	html	big
3	yes	631	28	text	none
⋮	⋮	⋮	⋮	⋮	⋮
50	no	15,829	242	html	small

Table 1.3: Four rows from the email50 data matrix.



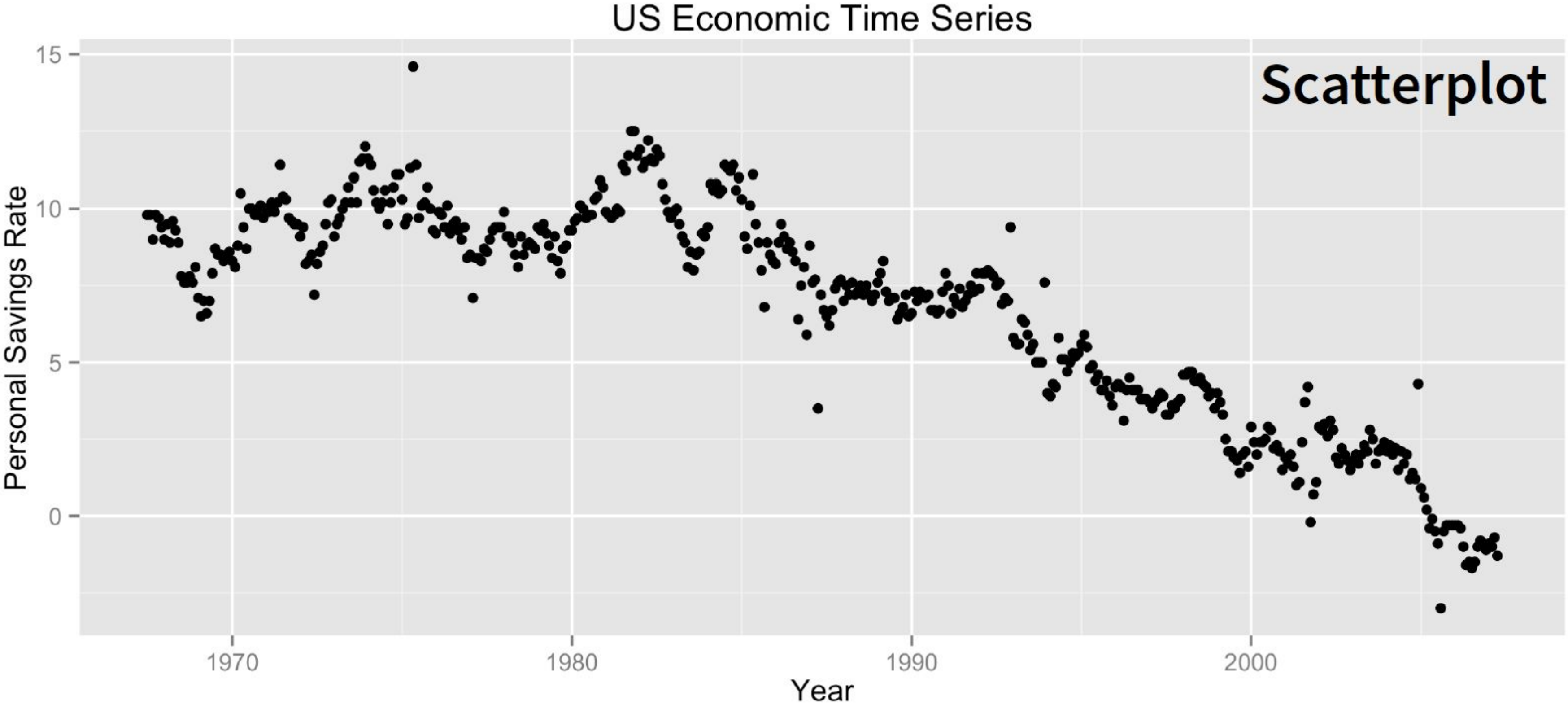
variable	description
spam	Specifies whether the message was spam
num_char	The number of characters in the email
line_breaks	The number of line breaks in the email (not including text wrapping)
format	Indicates if the email contained special formatting, such as bolding, tables, or links, which would indicate the message is in HTML format
number	Indicates whether the email contained no number, a small number (under 1 million), or a large number

Basic Charts (ggplot2)

Basic Charts

- Scatterplot
- Line Chart
- Area Chart
- Bar Chart
- Box Plot
- Multi-Line Chart
- Small Multiples Chart
- Histogram
- Stacked Bar Chart
- Stacked Area Chart

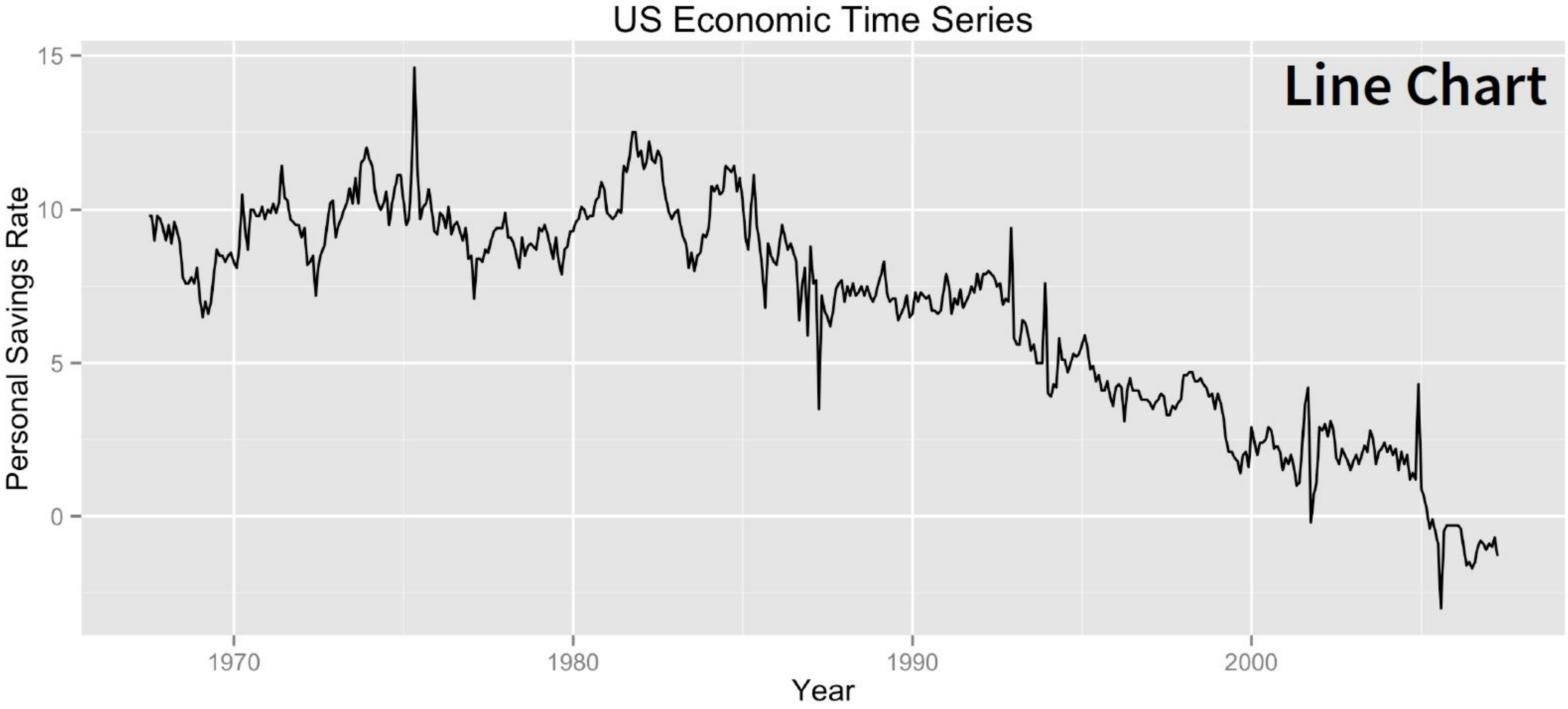
What type of variables can be encoded with each chart?



Scatterplot

Scatterplot using R and `ggplot2` with the `economics` dataset

Line Chart

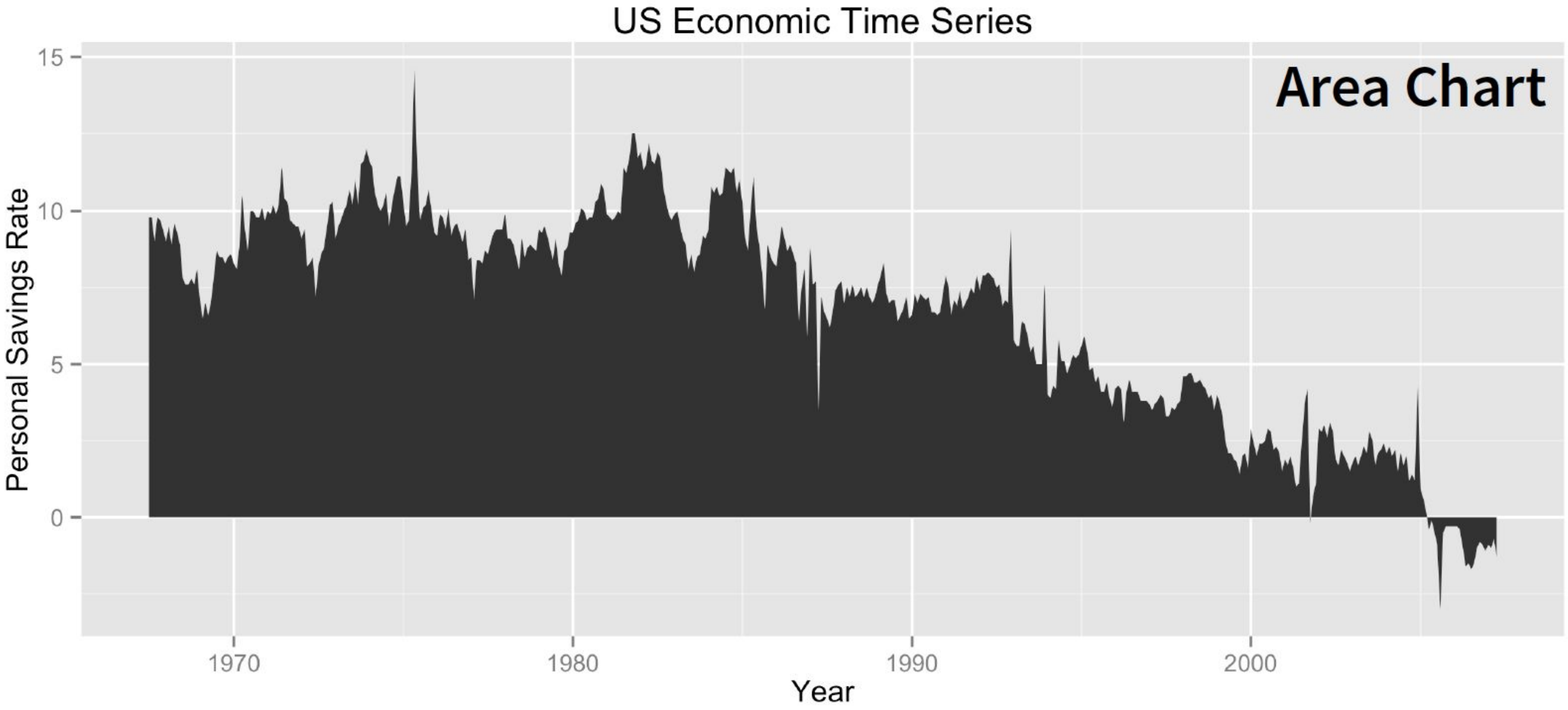


Line chart using R and ggplot2 with the economics dataset

How to use line charts

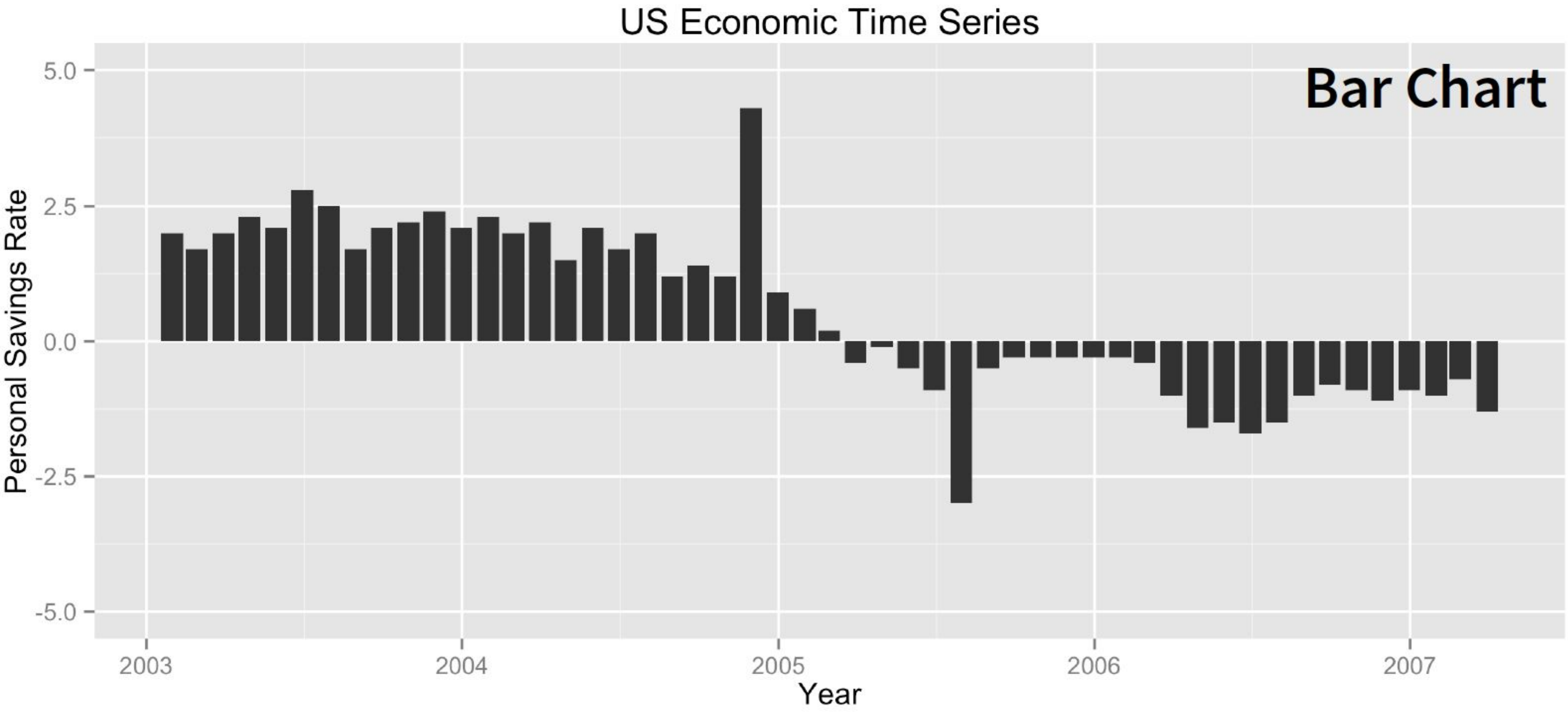
- When to use them
    a. Time series (chronological), dates, months, sequence of stages of a project, sequence of meters along on a gas pipeline,
    b. To detect trends and patterns, not to give people exact quantitative readings.
    c. Don’t use line charts with categorical data on the x-axis
- Scale
    a. As line charts are not really intended to give people exact numbers, forcing zero scaling is not necessary and can make it considerably more difficult to detect said trends and patterns.
- Dimension order:
    a. There should be some logical order to the dimensions on the x-axis

http://www.axisgroup.com/bar-charts-vs-line-charts/



Area Chart

Area chart using R and `ggplot2` with the `economics` dataset



Bar chart using R and `ggplot2` with the `economics` dataset

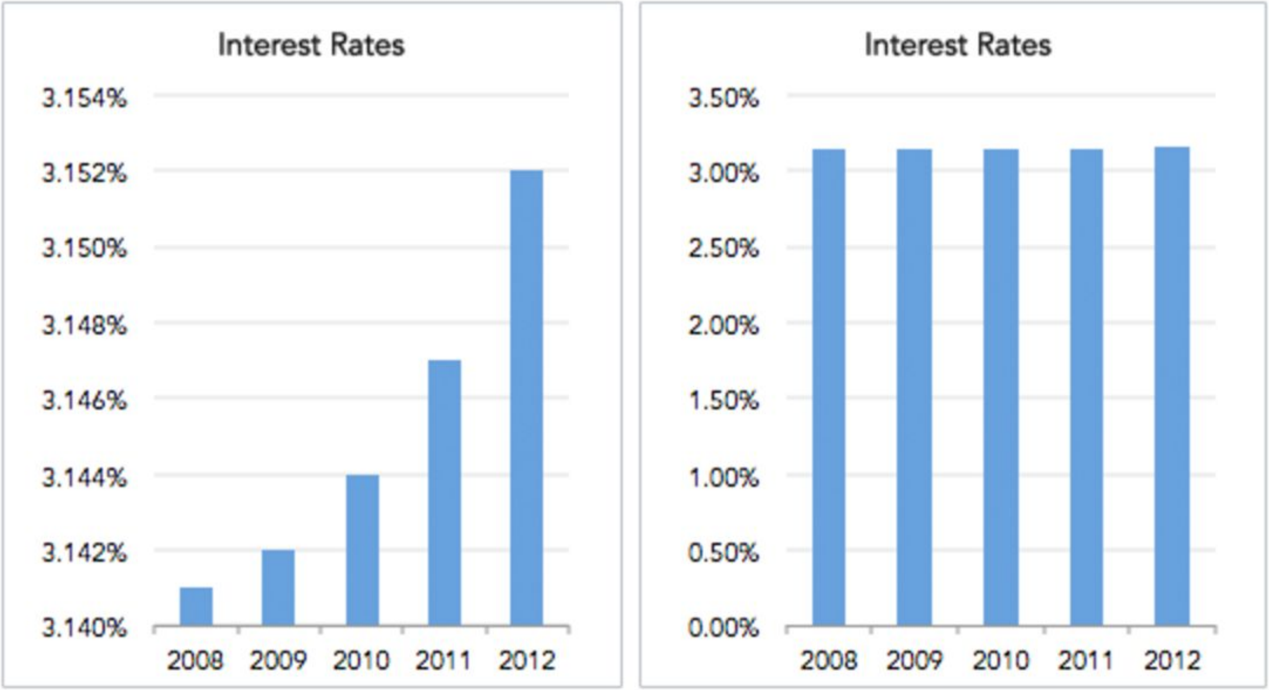
How to use Bar Charts

- When to use them:
    a. Bar charts should be used for comparing specific x-axis values
- Orientation:
    a. Use horizontal labels
    b. If needed use horizontal bar chart, so the text can read left-to-right
- Start the Y-Axis value at 0
- Space between bars should be ½ bar width.
- Order categories alphabetically, sequentially, or by value.

http://www.axisgroup.com/bar-charts-vs-line-charts/

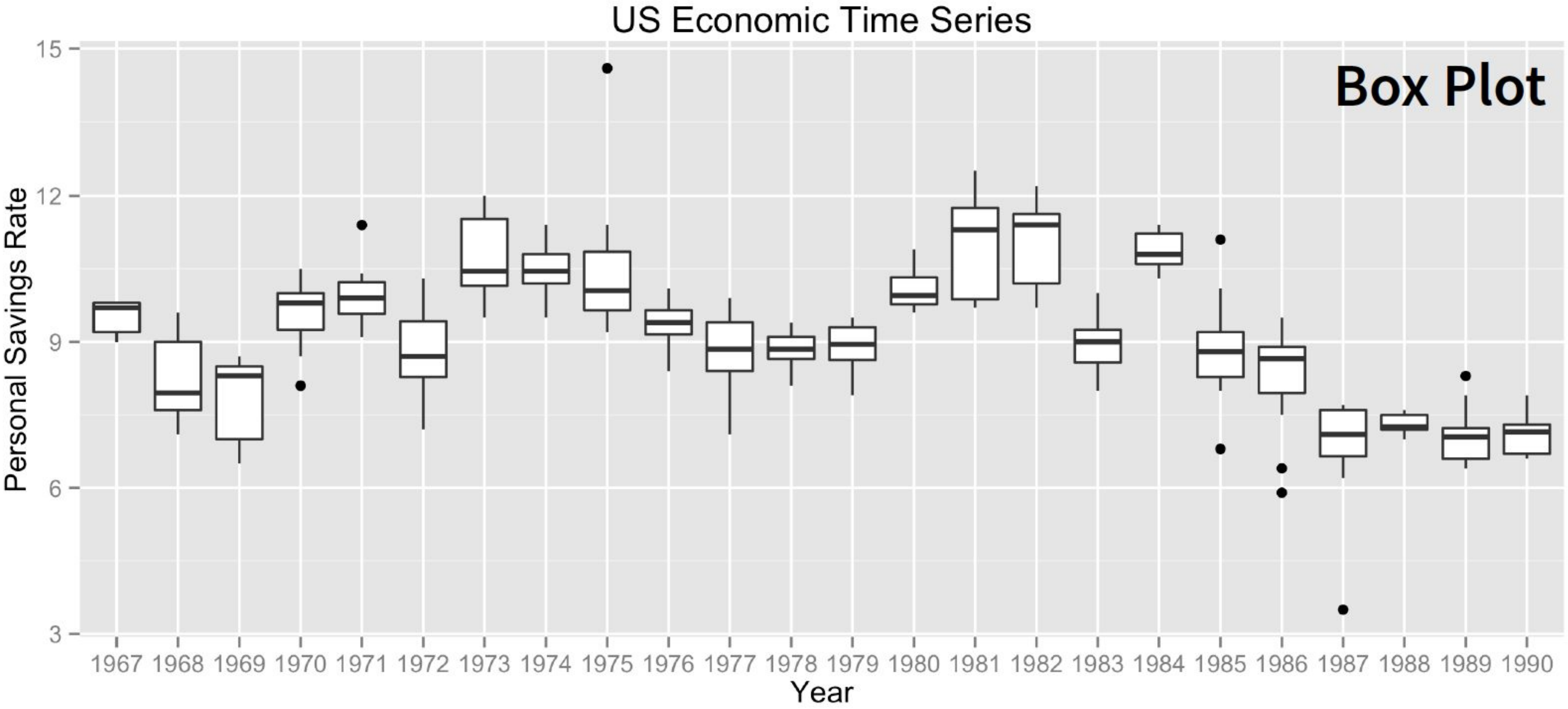
How to lie with bar charts: Truncated Y-Axis

Same Data, Different Y-Axis



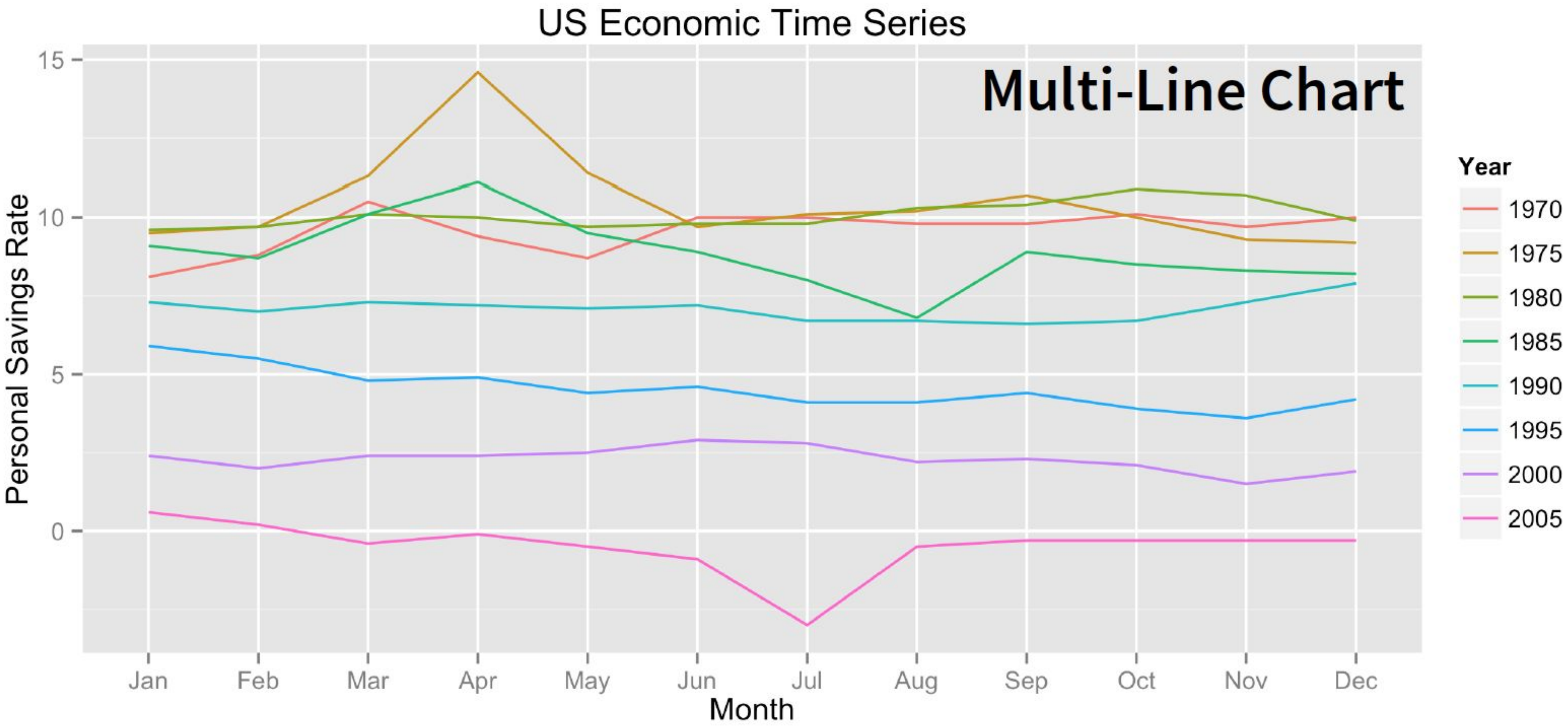
https://blog.heapanalytics.com/how-to-lie-with-data-visualization/

Box Plot

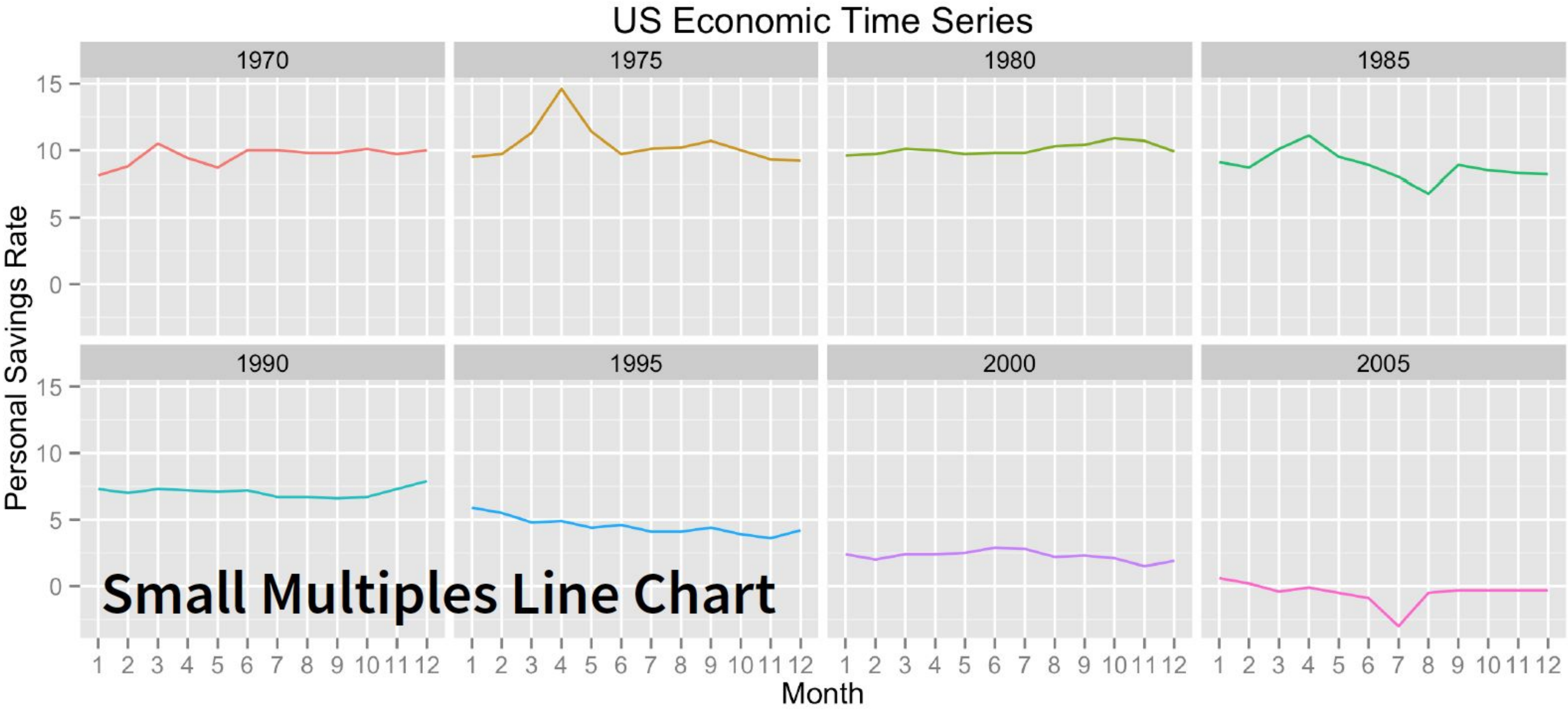


Box plot using R and `ggplot2` with the `economics` dataset

Multi-Line Chart



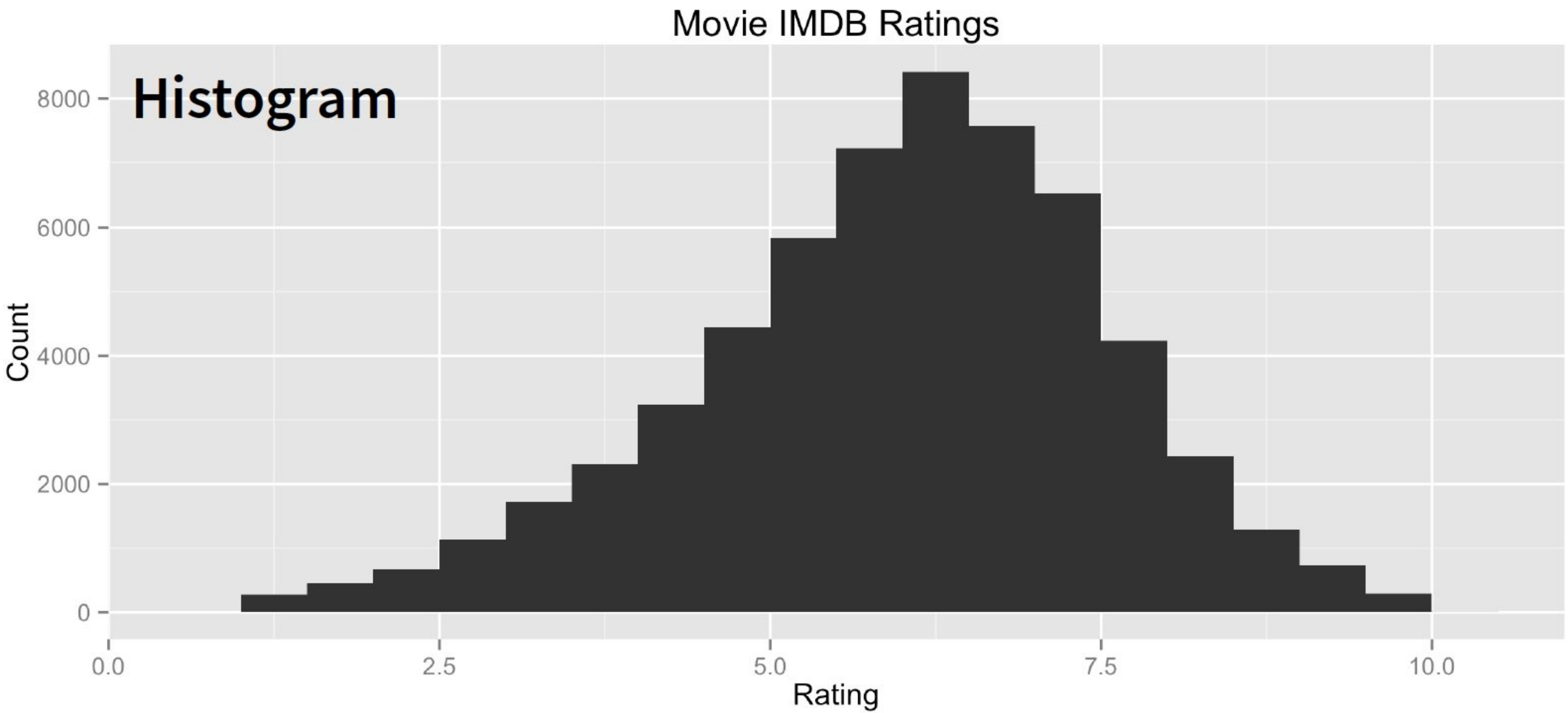
Multi-line chart using R and `ggplot2` with the `economics` dataset



Small Multiples Line Chart

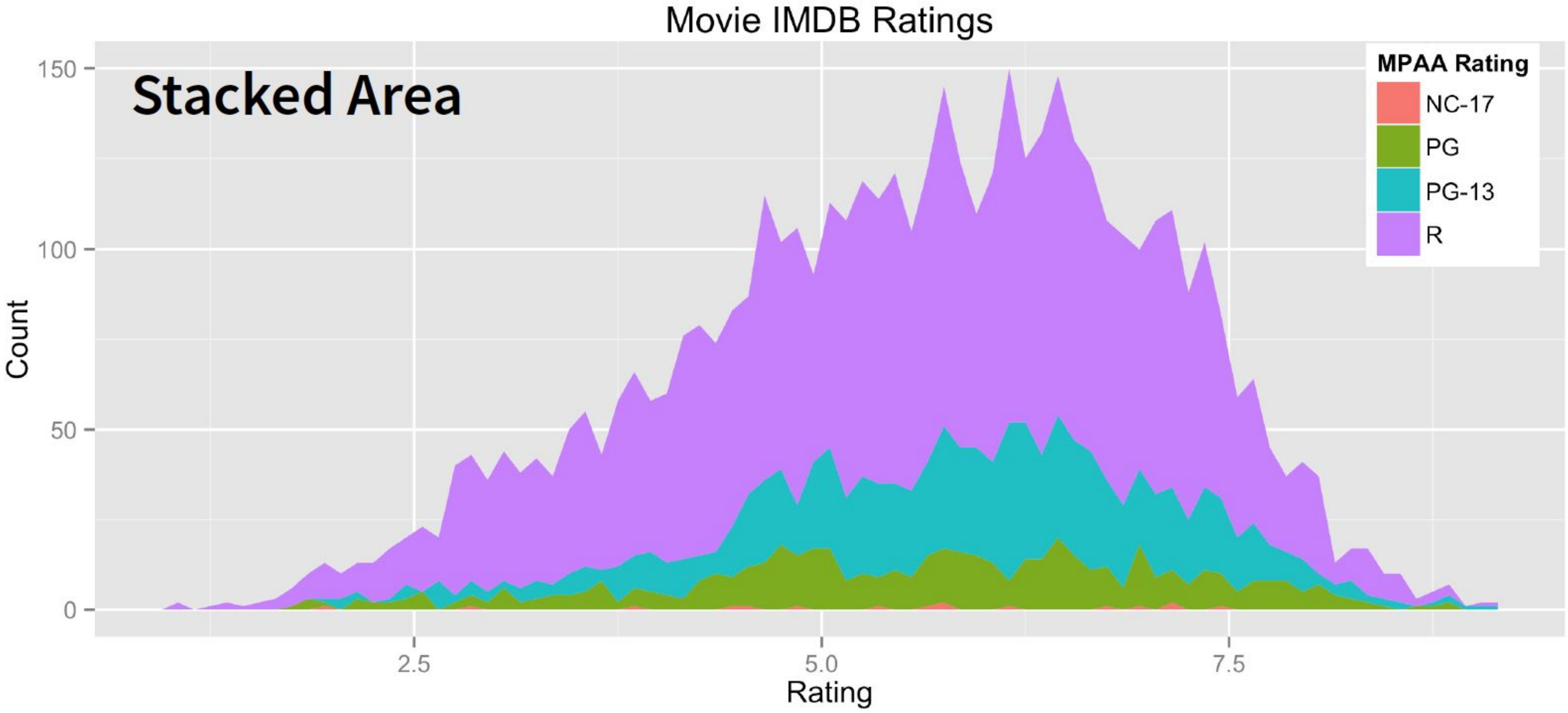
Small multiples chart using R and `ggplot2` with the `economics` dataset

Histogram

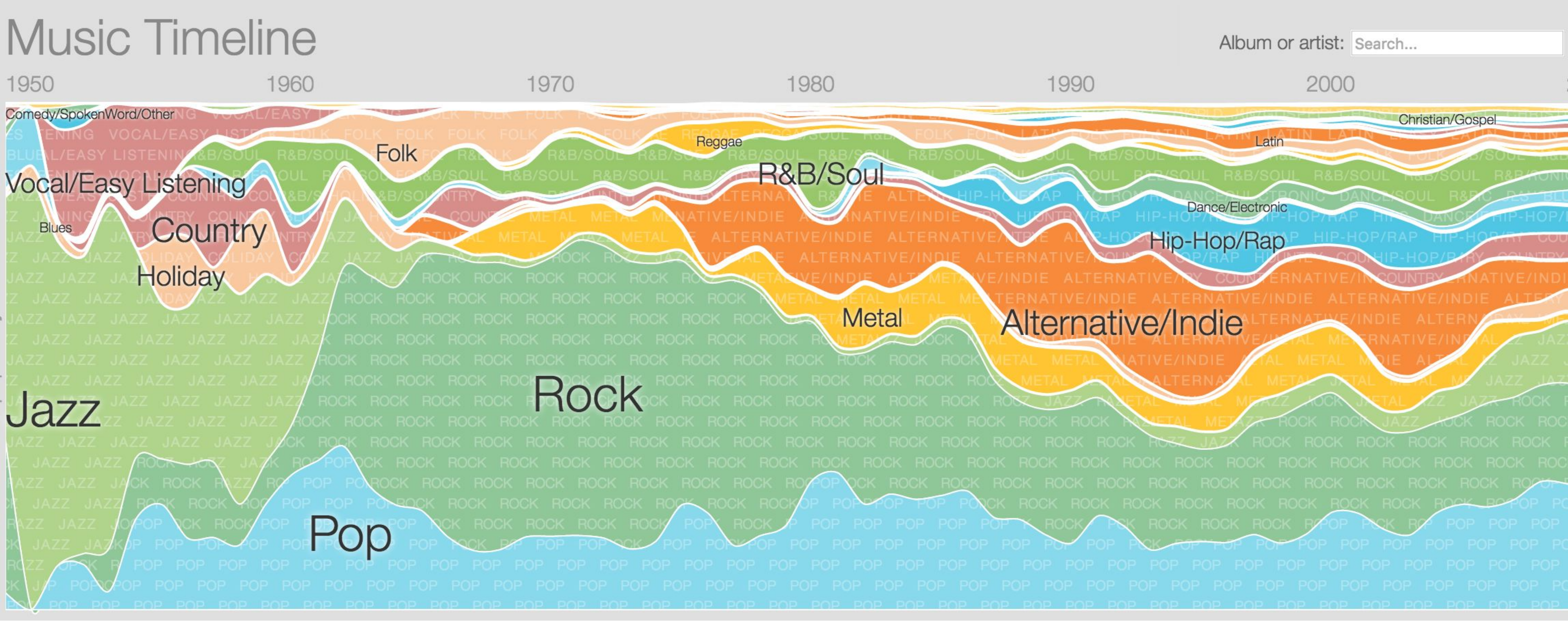


Histogram using R and ggplot2 with the movies dataset

Stacked Area



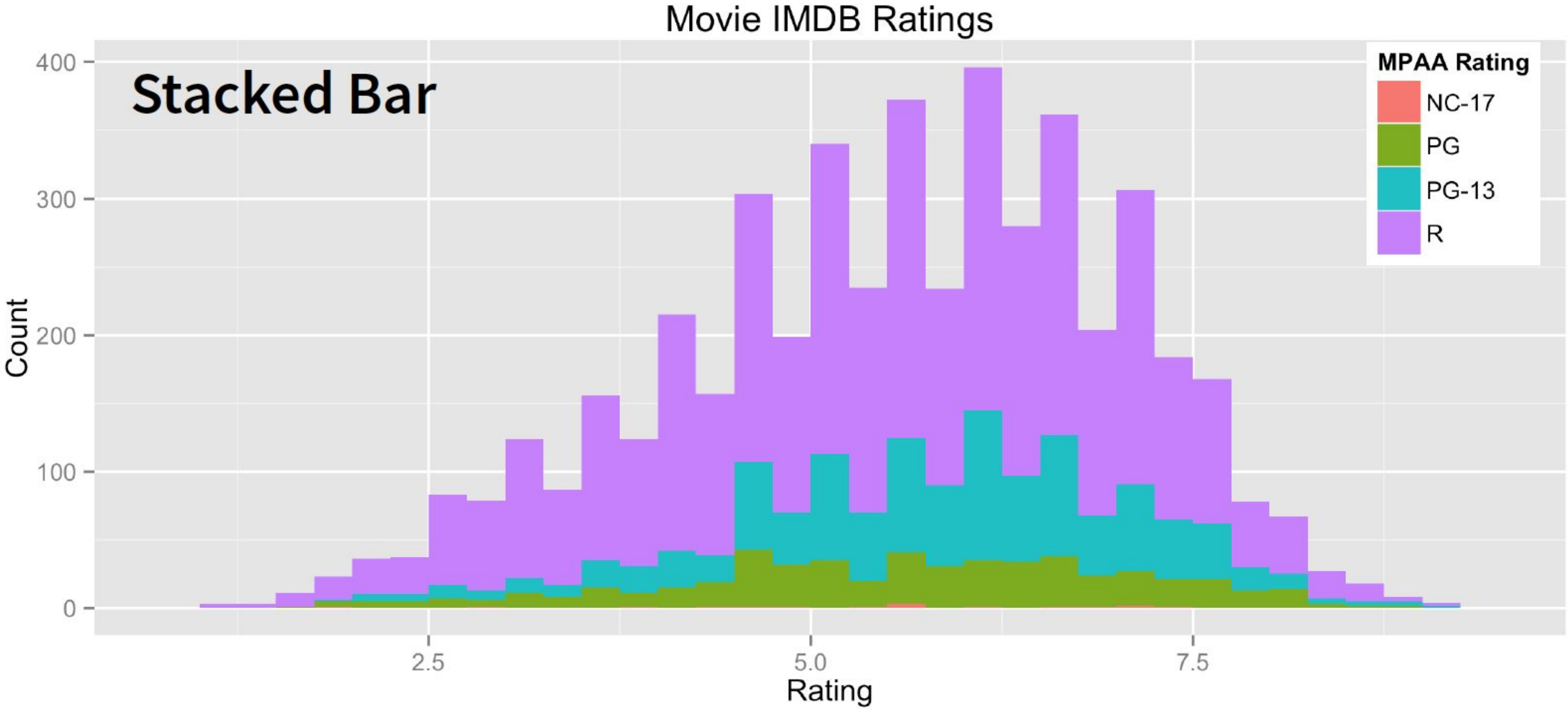
Stacked area chart using R and `ggplot2` with the `movies` dataset



Music Timeline

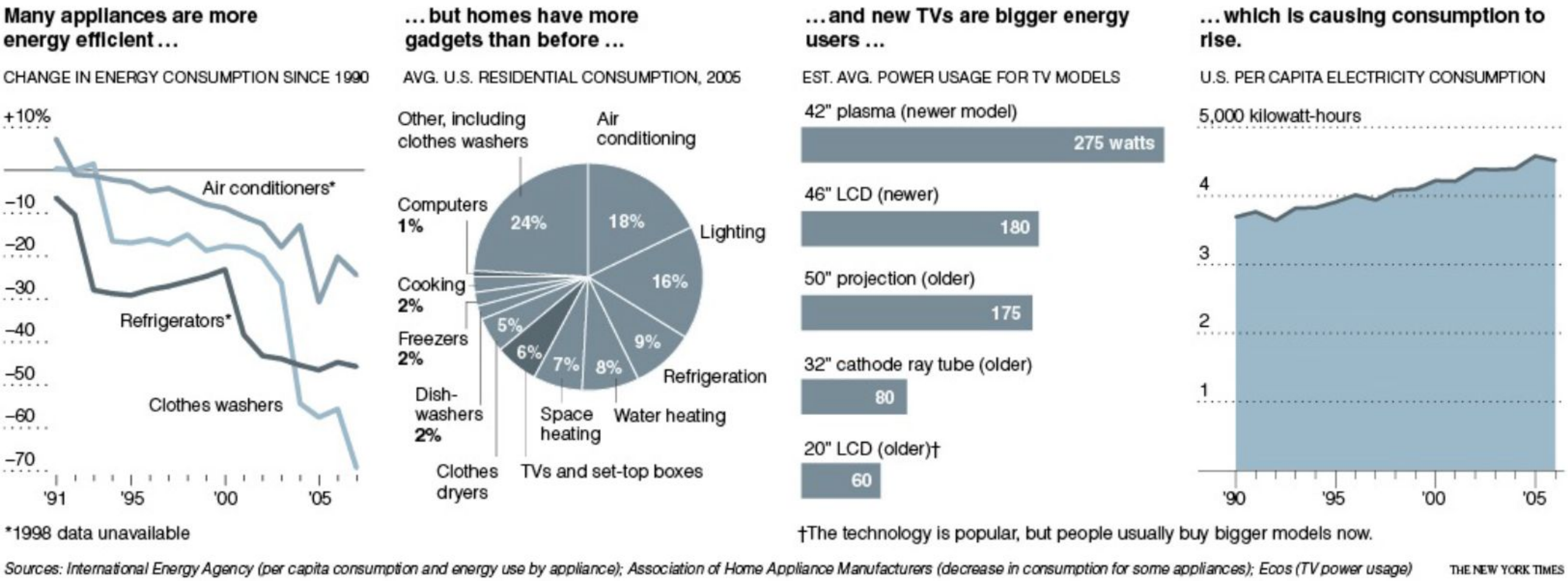
Album or artist: Search...

Demo: https://research.google.com/bigpicture/music/#



Stacked bar chart using R and ggplot2 with the movies dataset

Can you identify each type of chart?



Many appliances are more energy efficient …

CHANGE IN ENERGY CONSUMPTION SINCE 1990

*1998 data unavailable

… but homes have more gadgets than before …

AVG. U.S. RESIDENTIAL CONSUMPTION, 2005

… and new TVs are bigger energy users …

EST. AVG. POWER USAGE FOR TV MODELS

†The technology is popular, but people usually buy bigger models now.

… which is causing consumption to rise.

U.S. PER CAPITA ELECTRICITY CONSUMPTION

Sources: International Energy Agency (per capita consumption and energy use by appliance); Association of Home Appliance Manufacturers (decrease in consumption for some appliances); Ecos (TV power usage)

THE NEW YORK TIMES

http://www.nytimes.com/imagepages/2009/09/19/business/20090920EFFICIENCY-graphic-ready.html

Course Admin

Course Topics (approx)

1. Value of Visualization. Advanced ggplot2. Visualization Design.
2. Intro to Shiny. Visual Perception and Principles of Design.
3. Tableau. Multivariate Data Visualization.
4. Interaction. Text Data Visualization.
5. Design and Evaluation. Temporal Data.
6. Geospatial Data. Hierarchical Data.
7. Redesign Contest, Prototype Demonstrations.
8. Final Project Presentations.

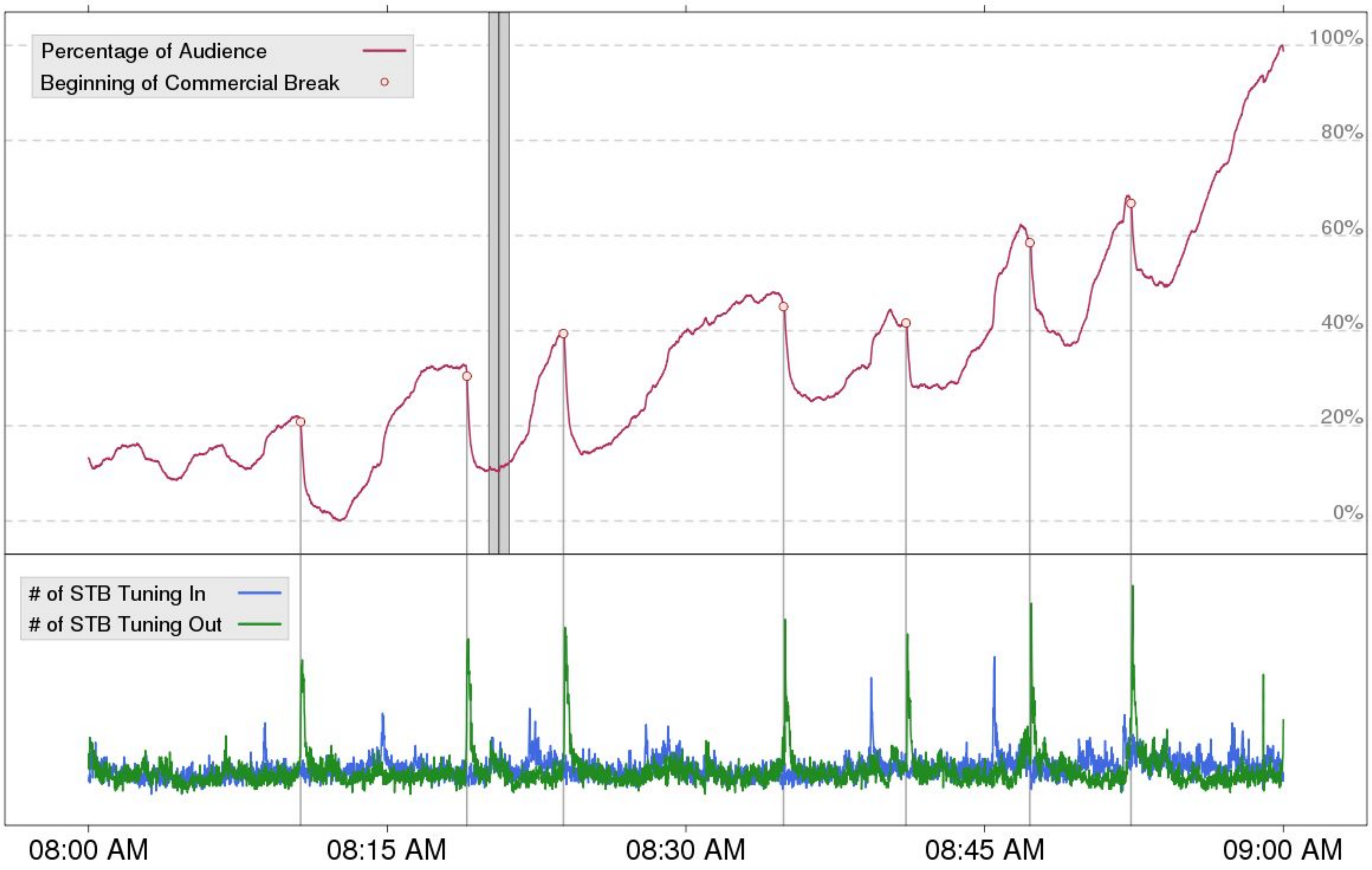
Work

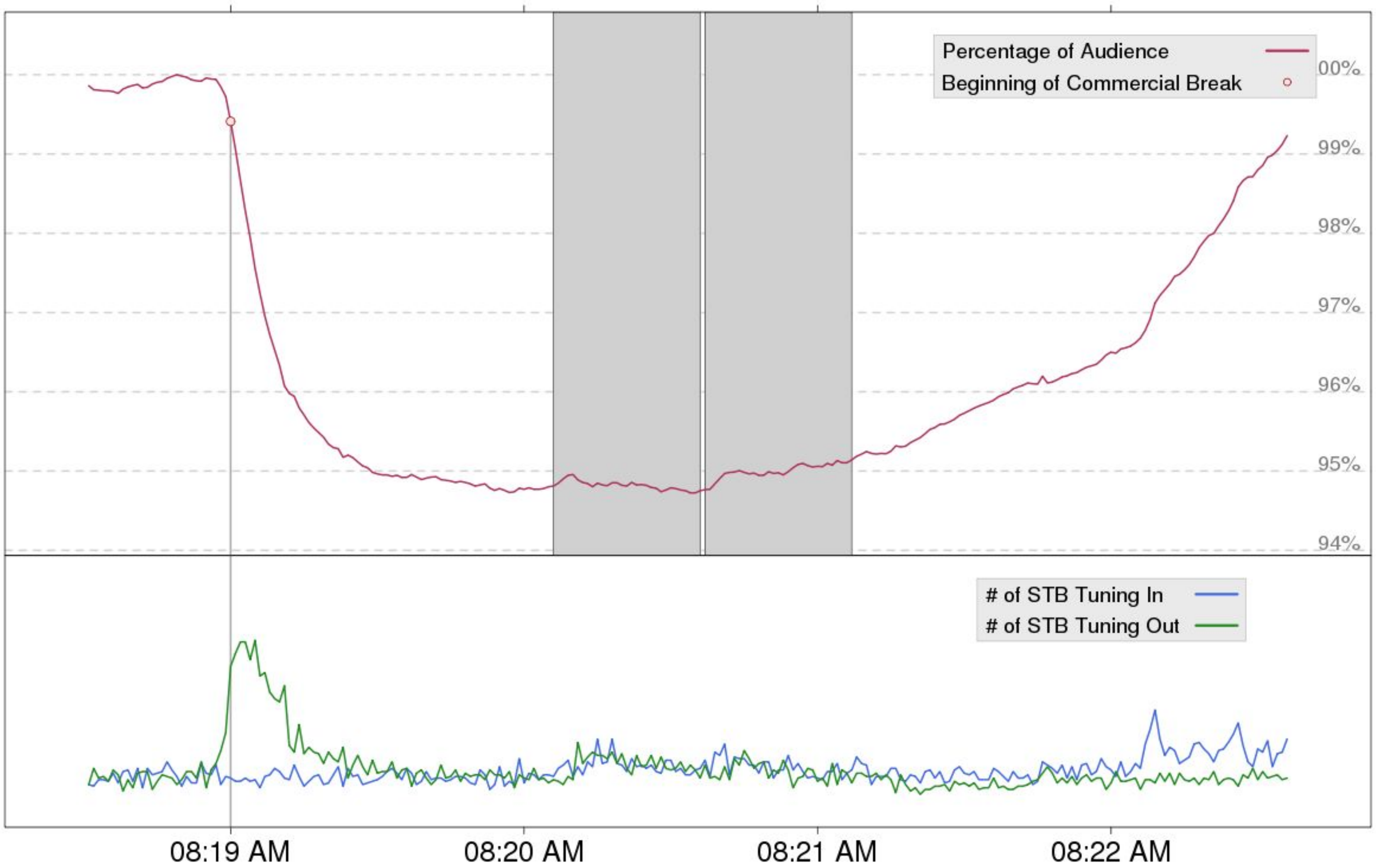
- **Participation (20%)**
  - Weekly participation assignments
  - In-class discussions, exercises, commenting on prototypes from other students
  - Graded on a pass / fail basis
- **Homework (30%)**
  - 4-5 homework assignments
- **Project (50%)**
  - Groups of 2
  - You can select your team

Project

- Select a data set and multiple visualization techniques, develop prototypes
- Rework the prototypes based on peer evaluations
- Students will demonstrate their final projects during a presentation or poster session
- Projects in Shiny
  - You can use D3 but you have to learn it on your own

ggplot2 Lab


Percentage of Audience
Beginning of Commercial Break
100%
80%
60%
40%
20%
0%
# of STB Tuning In
# of STB Tuning Out
08:00 AM
08:15 AM
08:30 AM
08:45 AM
09:00 AM


Percentage of Audience
Beginning of Commercial Break
00%
99%
98%
97%
96%
95%
94%
# of STB Tuning In
# of STB Tuning Out
08:19 AM
08:20 AM
08:21 AM
08:22 AM