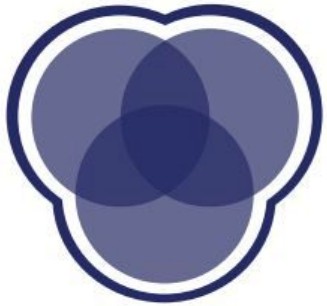


CDS

Cornell Data Science

Data Visualization

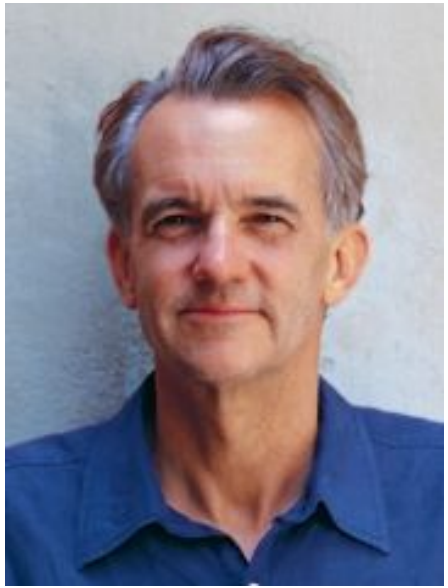
History

Edward Tufte (1942-)

Statistician and Yale professor

Key figure in the field of data visualization

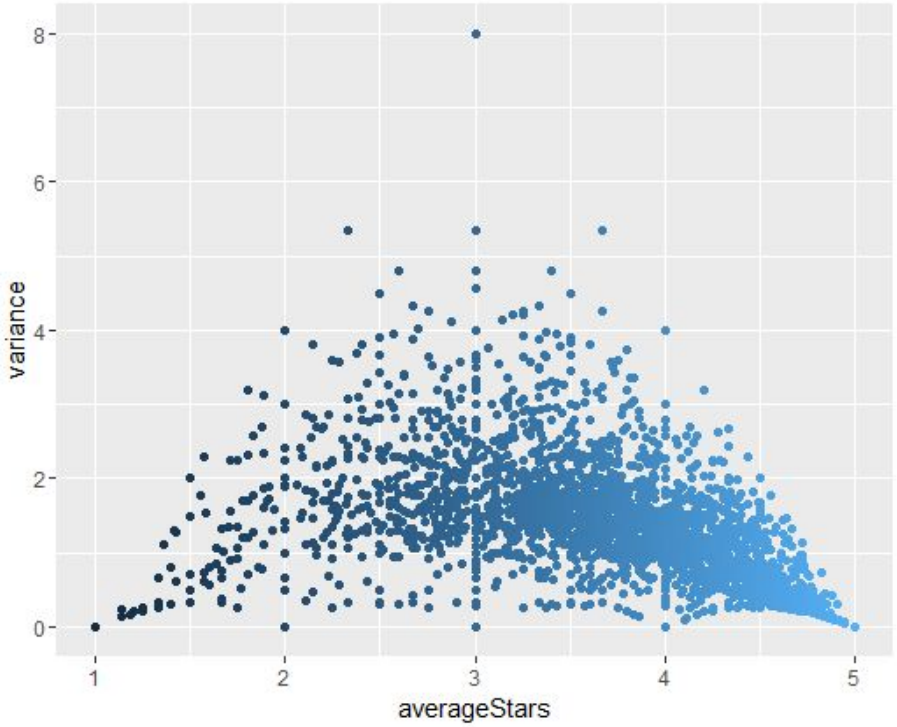
Recommended text:
The Visual Display of Quantitative Information



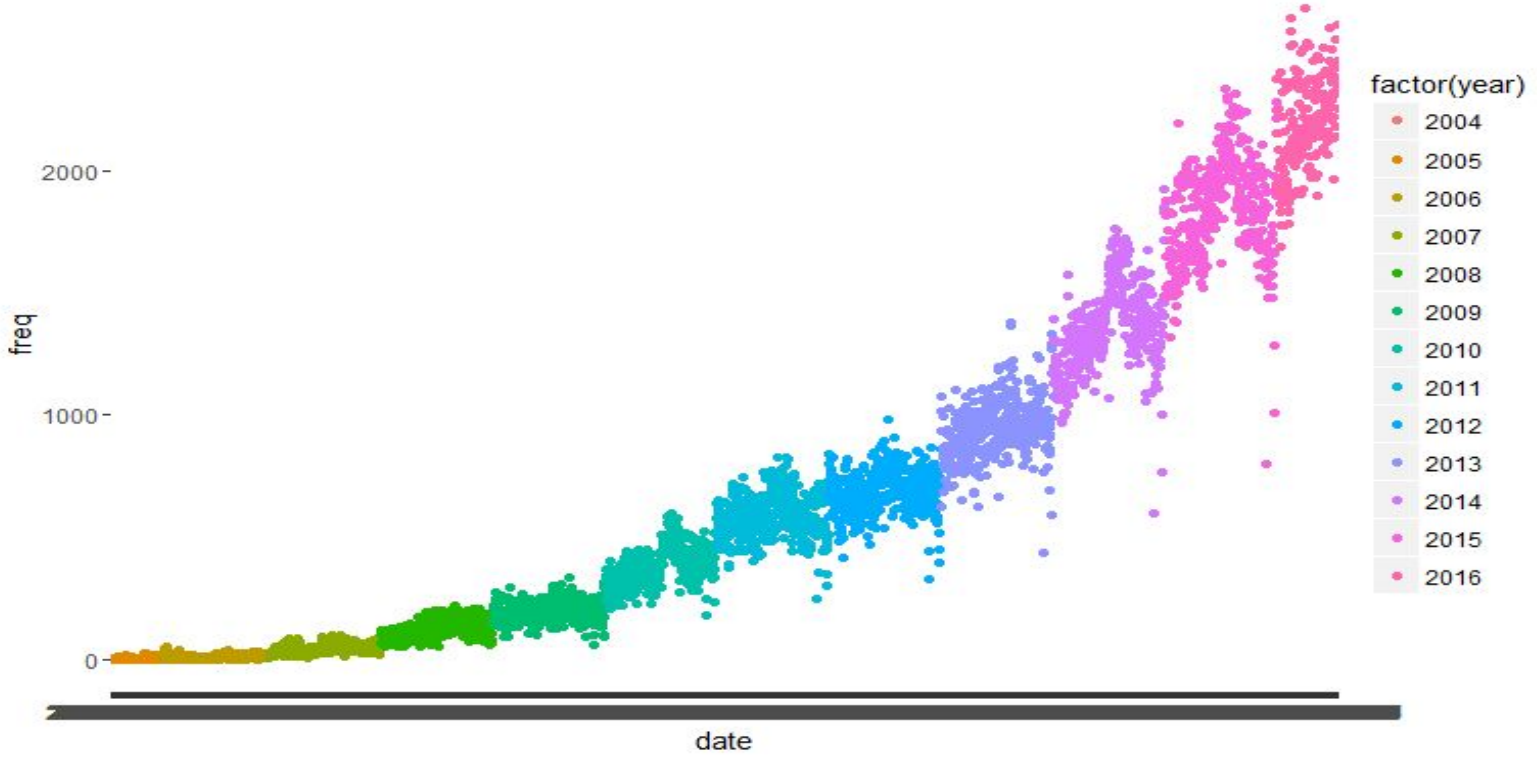
Data Visualization Simple Example: Yelp

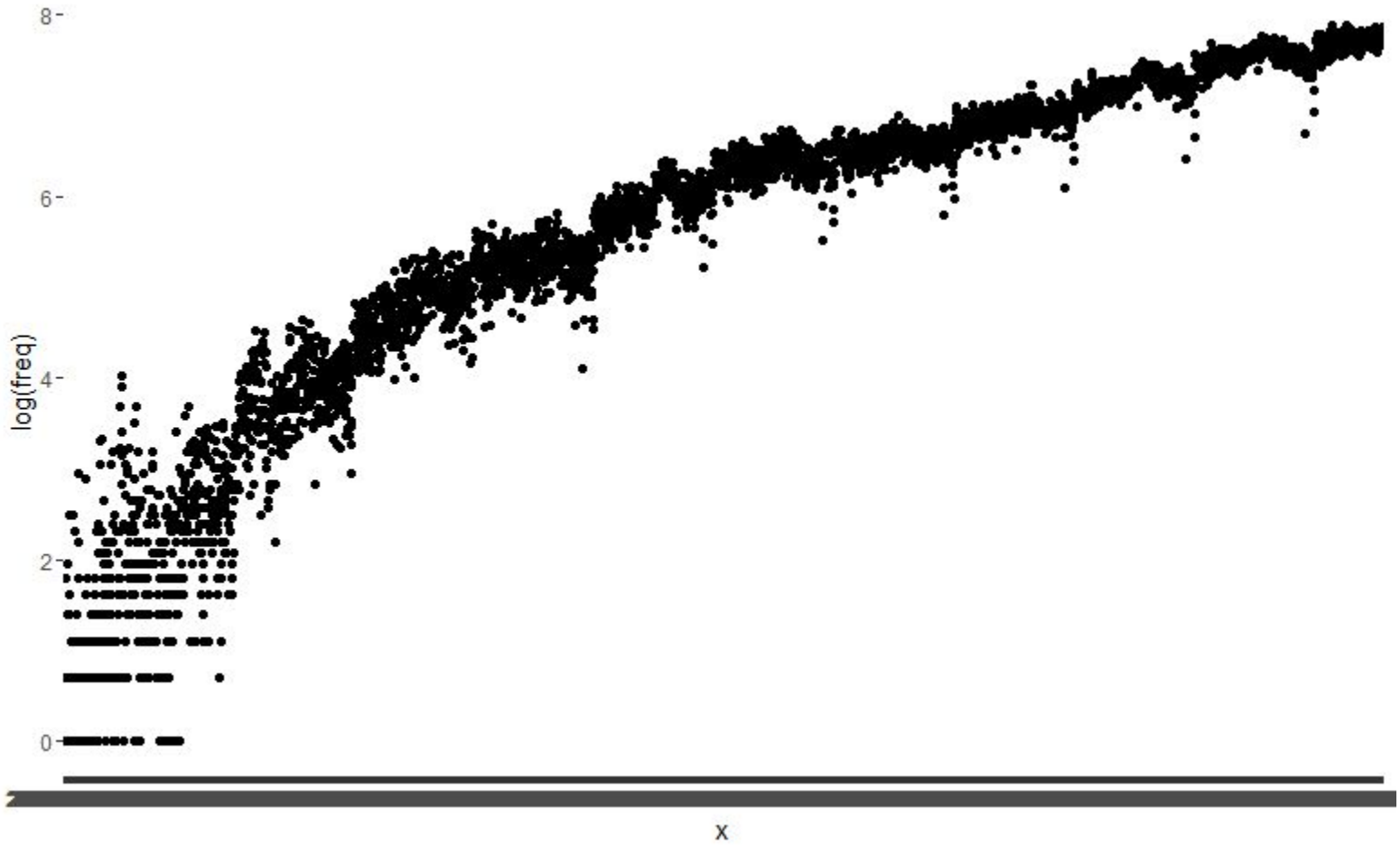
	AVG(stars)	var
AVG(stars)	1.00	-0.43
var	-0.43	1.00

Question: What do you notice?
What trends do you see?

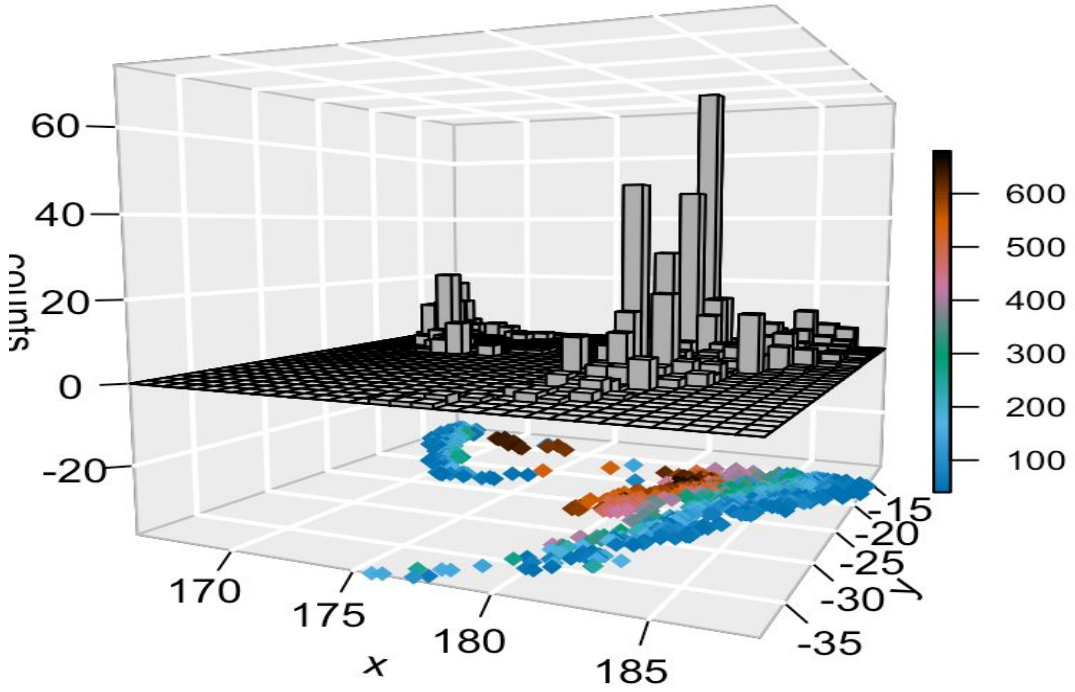


Data Visualization Simple Example: Yelp




8
6
4
2
0
log(freq)
x

3D Plot For Earthquake Data



Why Data Visualization?

- **Understanding a dataset**
  - “A picture is worth a thousand words.”
  - A good visualization is worth a thousand charts.
- **Communication of knowledge**
  - Quick and clear transfer of ideas
  - End product must be presented to non-technical people

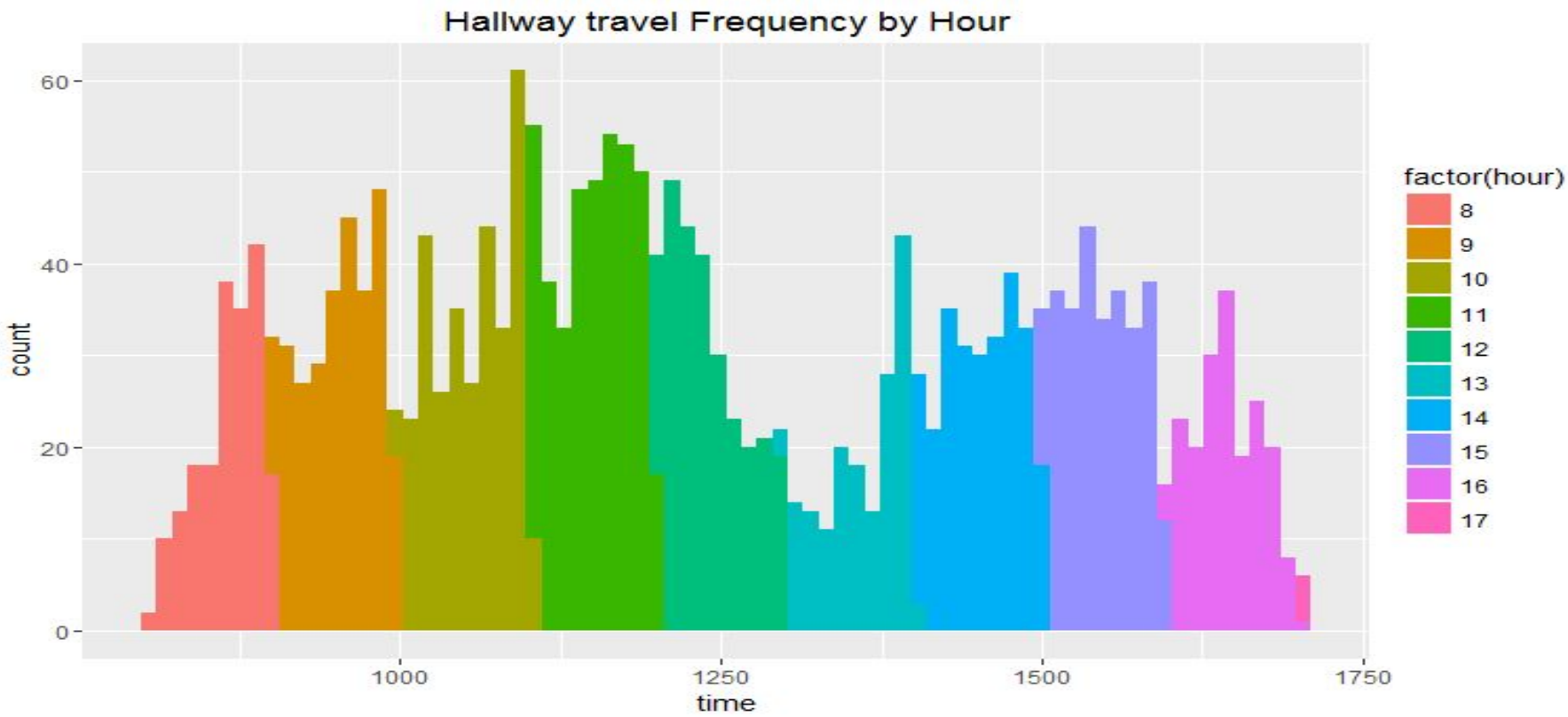
http://www.buildwelliver.com/sites/default/files/styles/project_slider/public/Lecture-Hall_0.jpg?itok=MFuEIFe8

Why is Data Visualization So Powerful?

- **Visual Patterns**
  - We process things visually, yet...
  - Conveying knowledge visually is hard!
    - Trends, discrepancies, and comparative magnitudes
- **Key concepts and insights can be highlighted**
  - Color, size, shape can be used to highlight trends

http://www.thrive-team.com/wp-content/uploads/2014/08/Visualization.jpg

Example: Nurse Hallway Travel Frequency

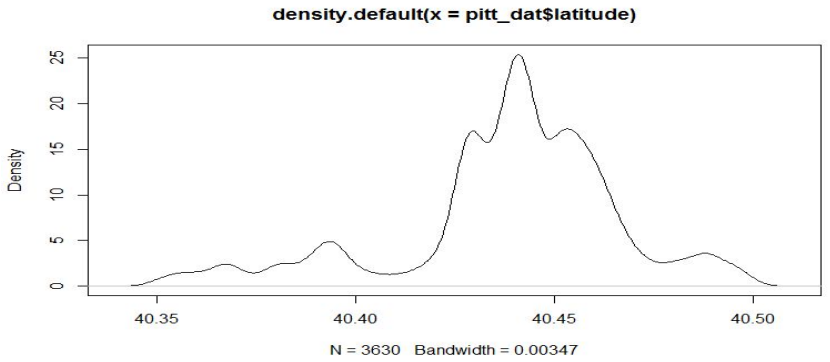


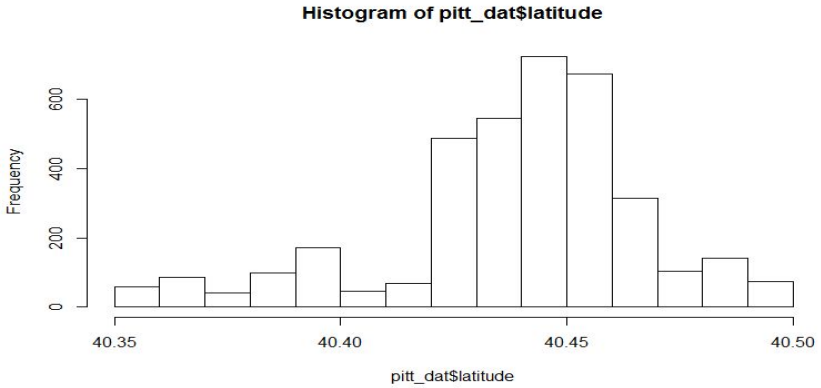
Data Visualization Techniques

- Histogram
- Scatterplot
- Density Plots
- Contour Maps
- 3D plots
- Bar Graphs
- Boxplots
- Heatmaps
- Animation
- Correlation Matrix
- Mosaic Plot

Histogram vs Density Plot

- Histogram shape varies greatly with **bin size**
- Density plot captures overall trend often better
- The “smoothing” of density plot can remove some important details.



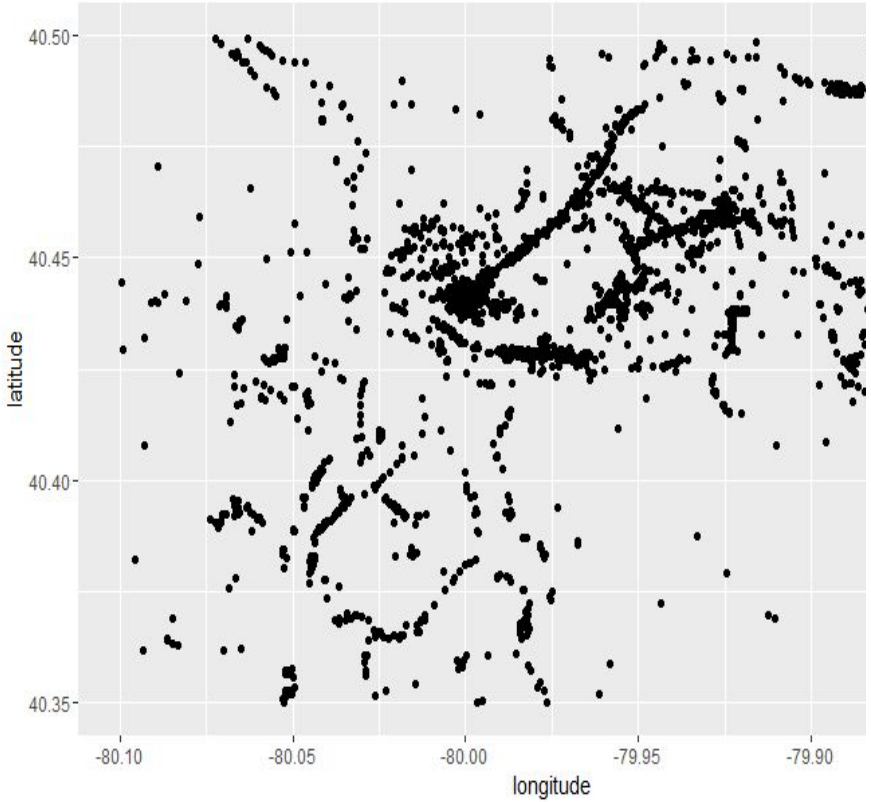


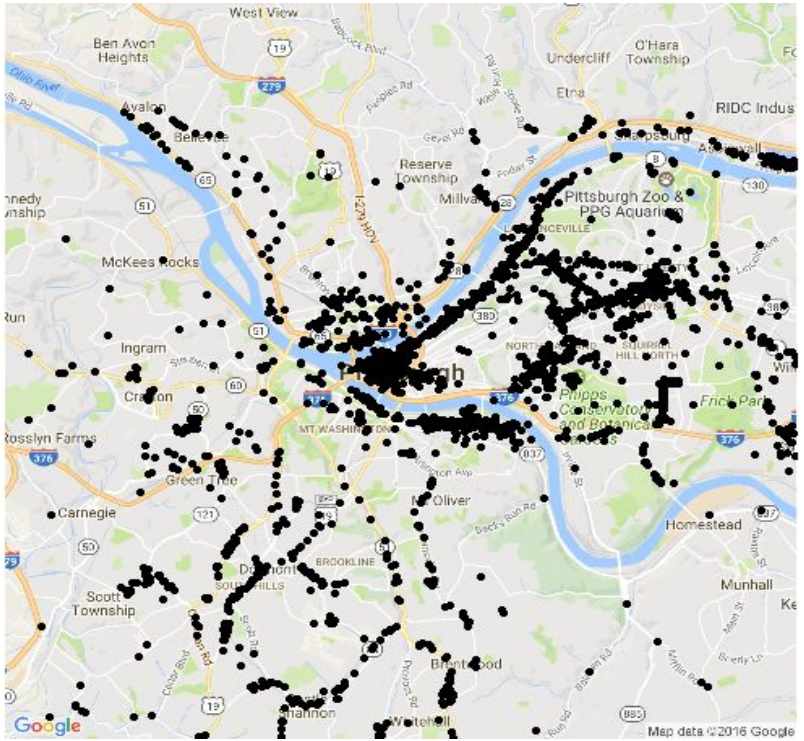
http://vignette2.wikia.nocookie.net/doratheexplorer/images/4/43/Map.png/revision/latest?cb=20131219004126

Using Maps

- **Map visualization assigns contextual information**
  - There are trends not apparent in the data itself
  - If there are longitudes and latitudes in your data, try out geographical visualization
- **Ways of obtaining maps**
  - qmap(), get_map()

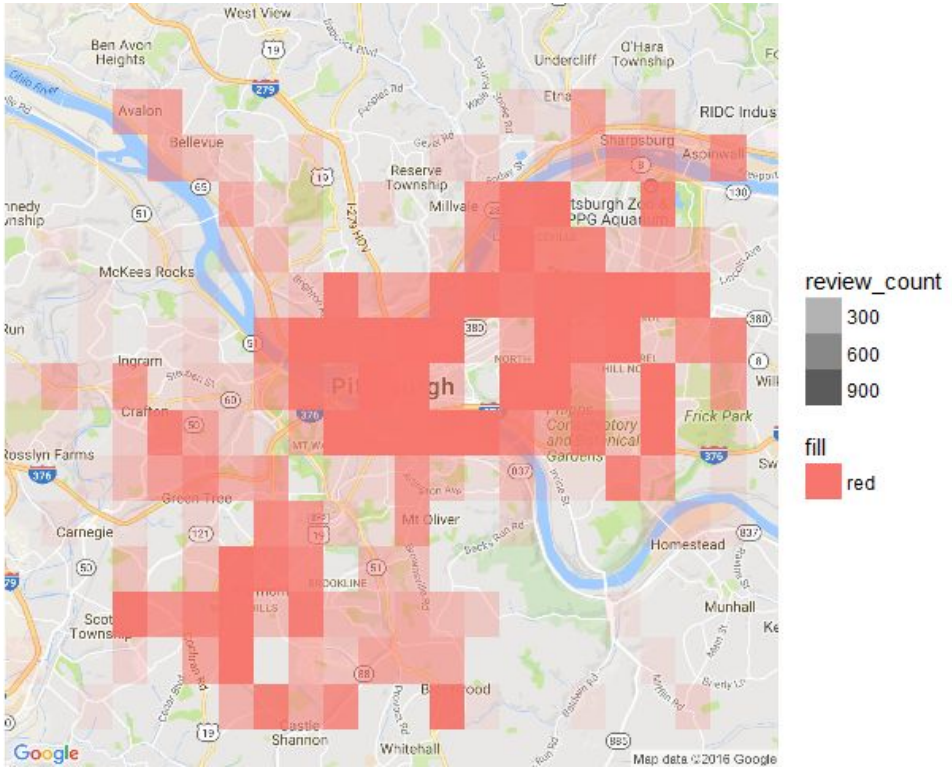
Example: Pittsburgh Data





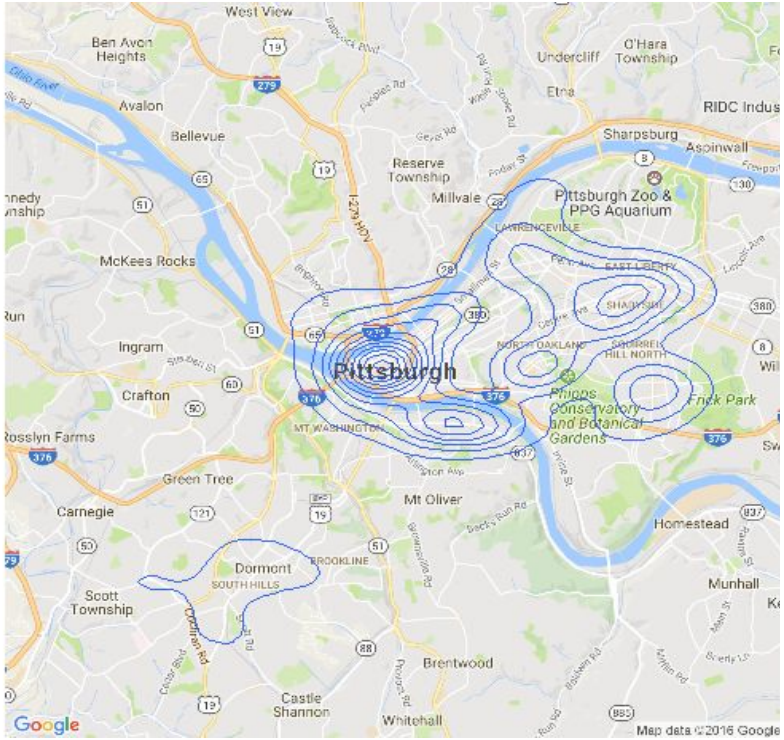
Visualization Technique: Heatmap

- Can yield insights for cluster analysis
- “Hot Spot” Analysis
- Can be very powerful when used on a map



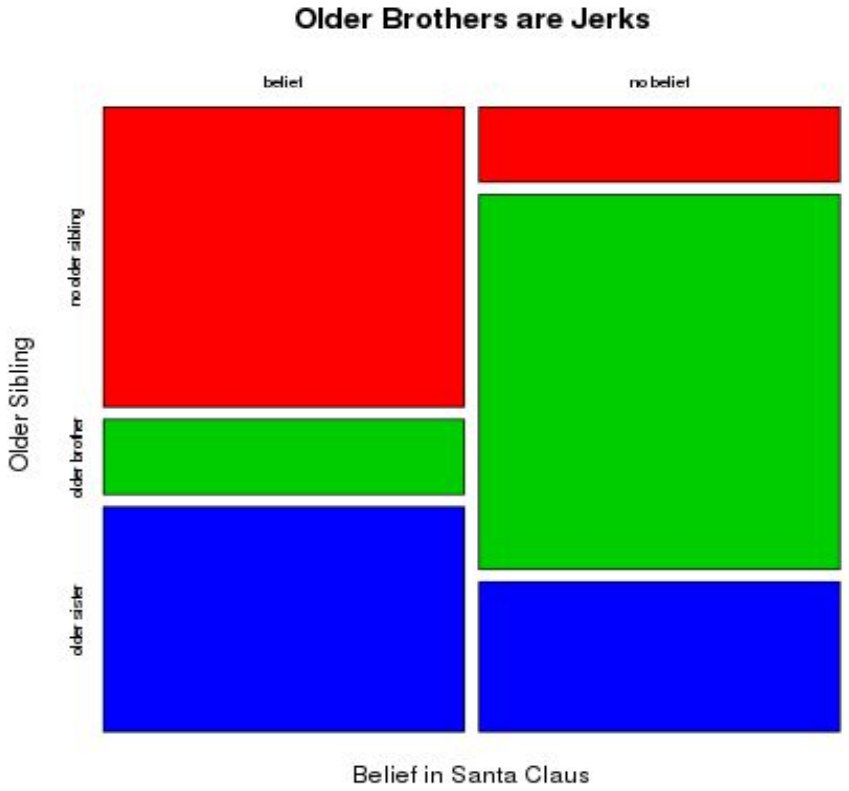
Visualization Technique: Contour map

- Similarly useful for cluster analysis
- **Kernelized Smoothing**
  - Bandwidth adjustable
- Good for exploring:
  - **Gaussian Mixture Models**
  - **Gaussian Naive Bayes**



Visualization Technique: Mosaic Plot

- Categorical data can be frustrating!
- Mosaic plot allows visual for categoricals
- Use function `mosaicplot()`



What (Amazing) Visual Packages does R offer?

- **ggplot2 package**
  - This package is the mother of all R visual tools
  - Cheatsheet: https://www.rstudio.com/wp-content/uploads/2015/12/ggplot2-cheatsheet-2.0.pdf
- **Plotly**
  - Can complement `ggplot2`'s lack of interactive interface
- **Animation**

Package: ggplot2

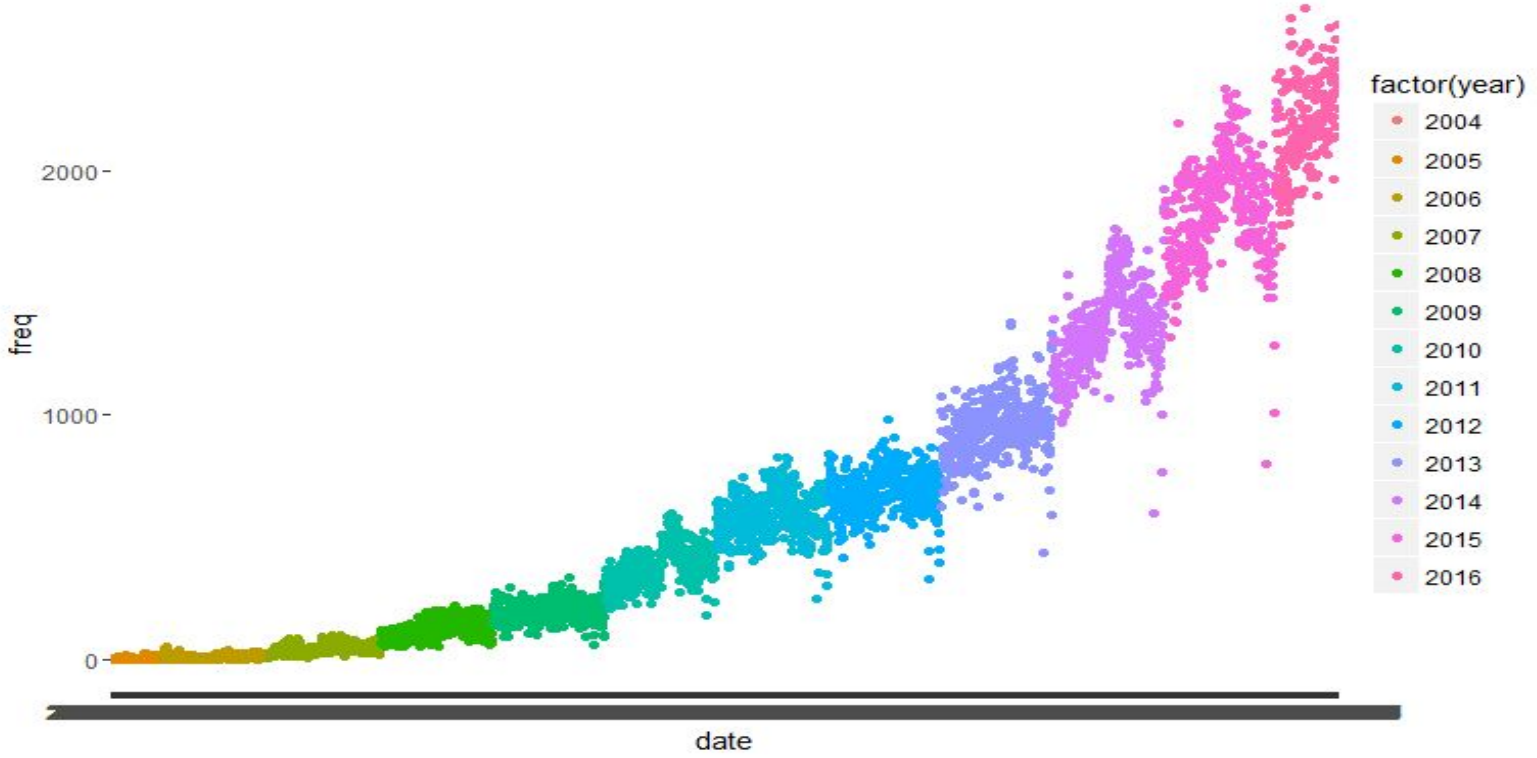
- Polished package that uses an intuitive language.
- Structure:
  - Specify data, and mapping
  - Specify type of visualization
  - Specify any modification
  - Example: ggplot(data = dat, aes(x = x, y = y)) + geom_point(color = factor(group), data = dat) + coord_flip()

https://upload.wikimedia.org/wikipedia/commons/thumb/8/8a/Plotly_logo_for_digital_final_(6).png/220px-Plotly_logo_for_digital_final_(6).png

Package: Plotly

- Originally a python tool
- Plotly is interactive. (Interactivity matters!)
- Can just use ggplotly() function to make a ggplot into an interactive plotly display!

Example Revisited with Plotly: Yelp



Package: Animation

- The animation package is very easy to use
  - `saveHTML`, `saveGIF` function
- Allows easy comparison of several similar plots
- Can take up a lot of storage for an animation of considerable length
- Code demonstration

Additional packages:

- Shiny: A very powerful alternative to animation
- Allows interactive visualization tools that allows quick comparison
- https://shiny.rstudio.com/
- ggvis - allows interaction with google charts

Challenges of Visualization

- Data with high dimensions
- Finding the right visualization for a given dataset
- Often time-consuming and impedes production process

Coming Up

Your problem set: Unleash your creativity by visualizing a data set

Next week: Making predictions using linear and logistic regression

See you then!