

INFO 1998: Introduction to Machine Learning



CDS Education

We explore, learn, and educate big minds.

Lecture 3: Data Visualization

INFO 1998: Introduction to Machine Learning



CDS Education

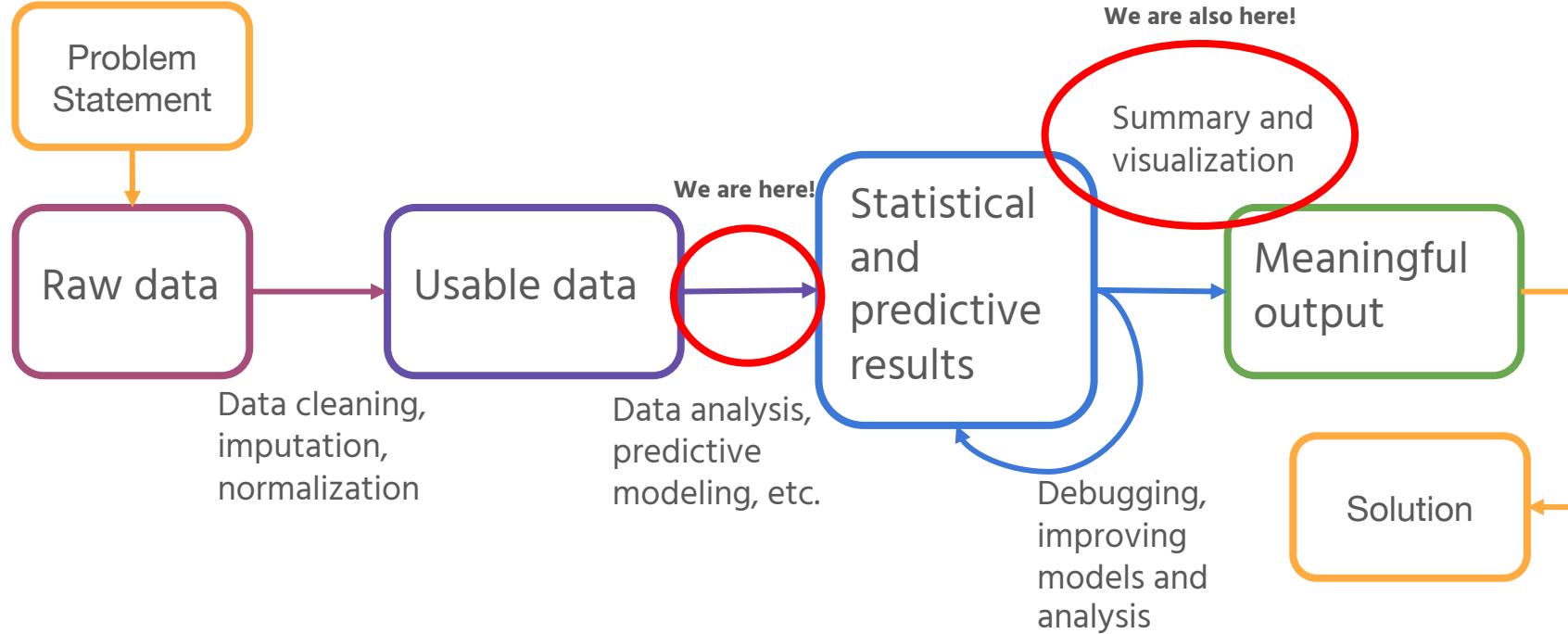
We explore, learn, and educate big minds.

Agenda

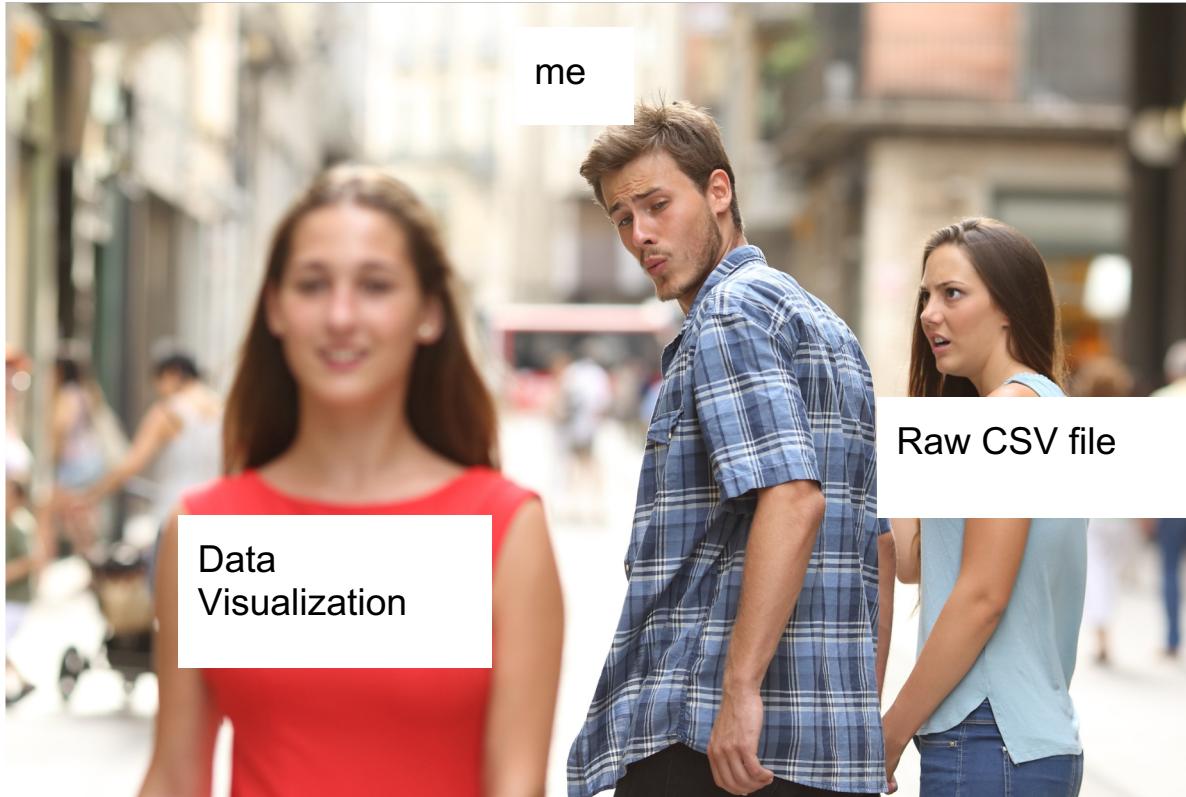
1. Why Data Visualization is Important
2. Data Visualization Libraries
3. Basic Visualizations
4. Advanced Visualizations
5. Challenges of Visualization



The Data Pipeline



Why is Data Visualization Important?



[Source](#)



Why is Data Visualization Important?

Informative

Appealing

Universal

Predictive

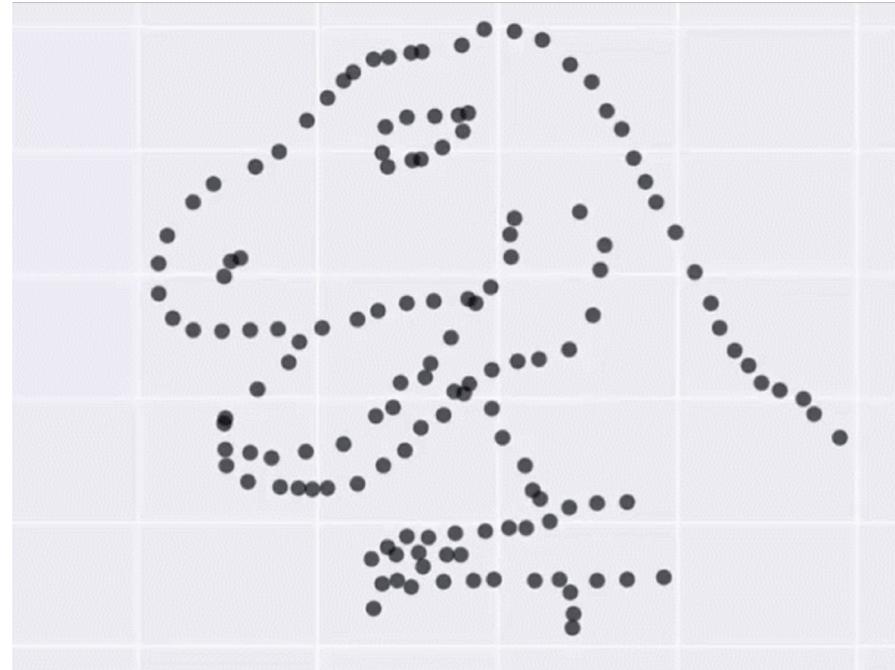


Why is Data Visualization Important?

Same summary stats (mean, median, mode) **but different distributions!**

We need to see how the **actual** data looks!

df.describe() is not enough



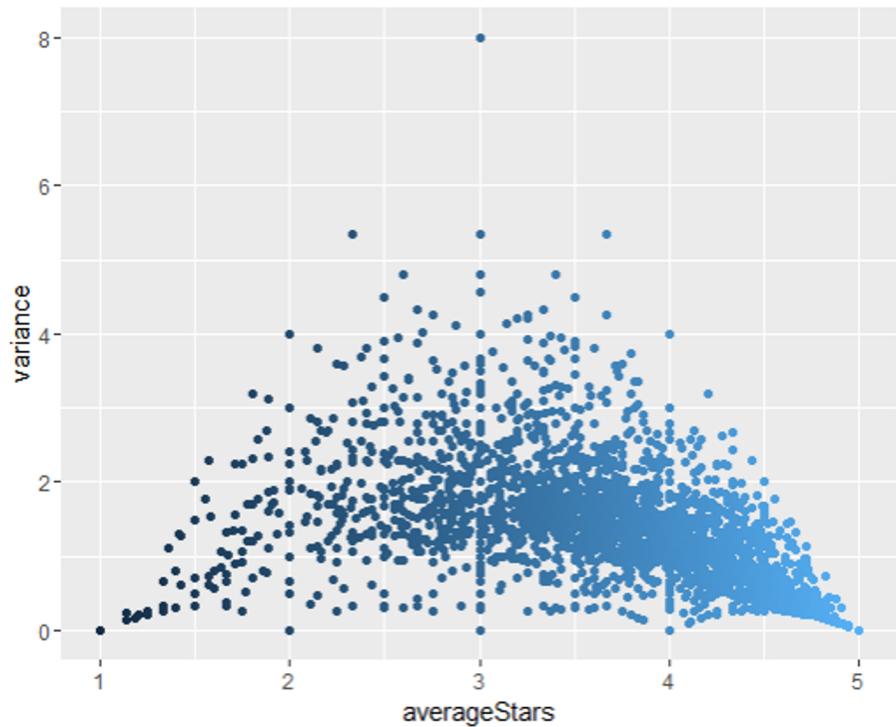
[Source](#)



Data Visualization Simple Example: *Ratings on 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 Libraries

- **matplotlib**
 - Python data visualization package
 - Capable of handling most data visualization needs
 - Simple object-oriented library inspired from MATLAB
 - [Cheatsheet](#)
- **seaborn**
 - Another visualization package built on matplotlib

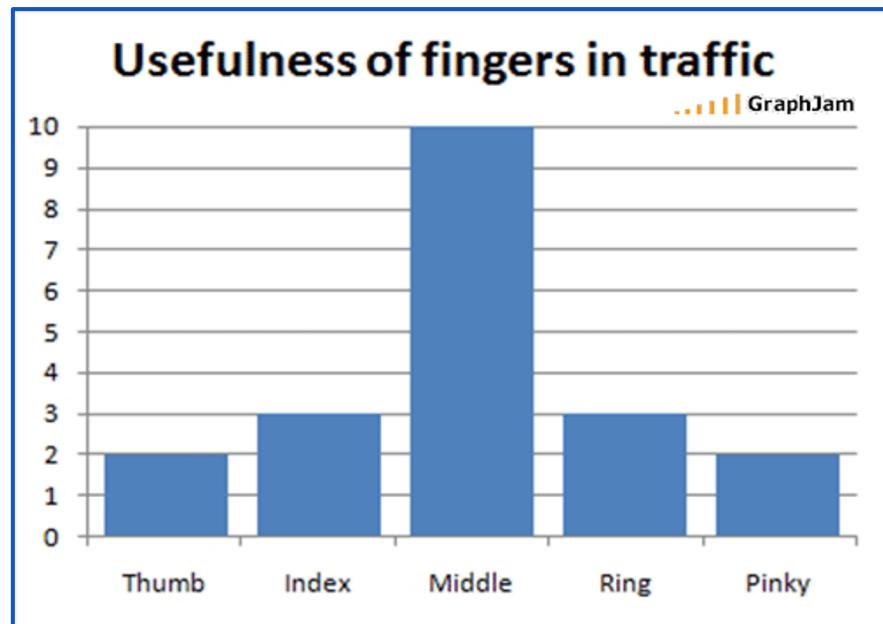


Basic Data Visualizations



Bar Graph

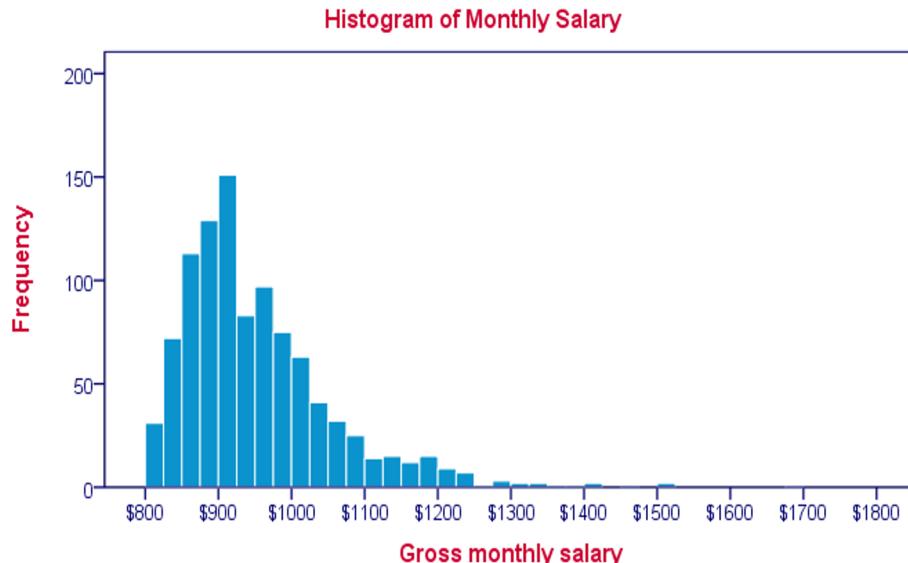
- Represent **magnitude** or **frequency** of discrete variables
- Allows us to compare features



[Source](#)



Histograms



- Used to observe **frequency distribution** of continuous variables
- Data split into **bins**

[Source](#)



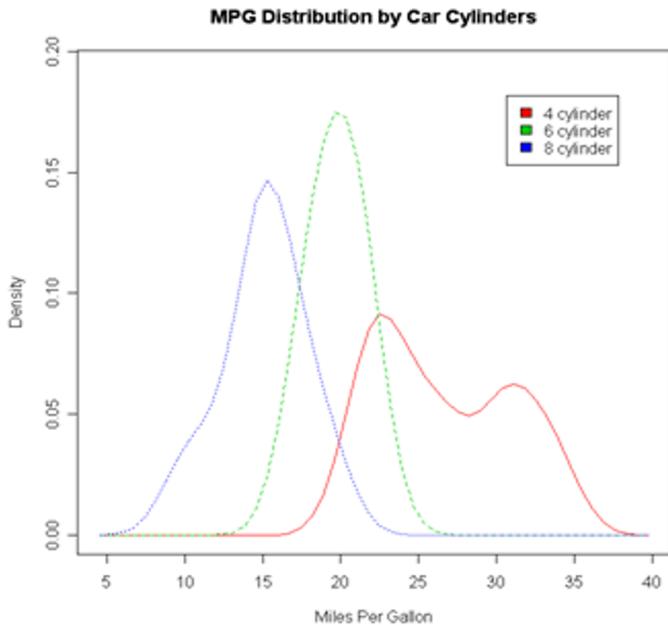
Histograms: Different Bin Sizes



[Source](#)



Density Plot

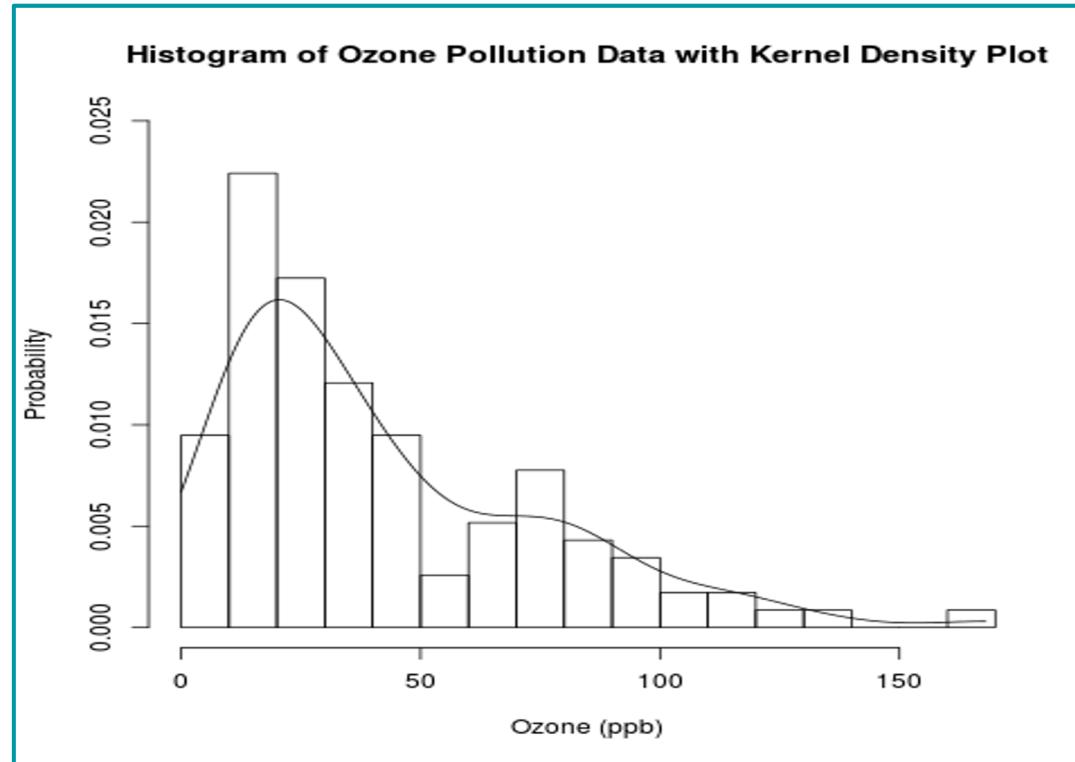


Like a histogram, but **smooths** the shape of the distribution

[Source](#)



Histogram vs Density Plot

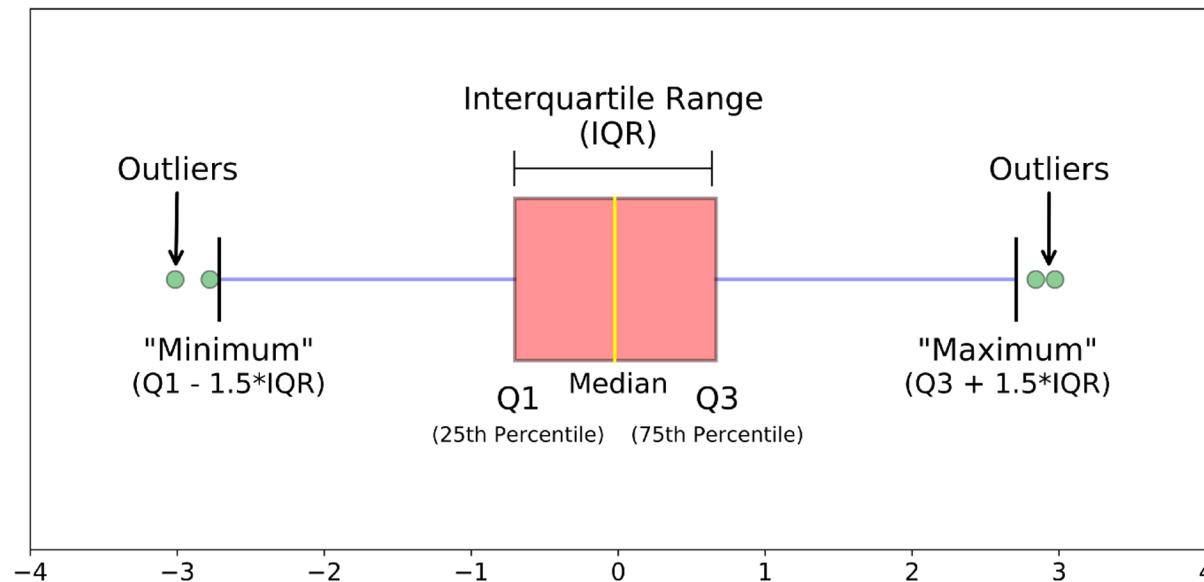


[Source](#)



Boxplot (a.k.a box and whisker plot)

- Summary of data
- Shows **spread** of data
- Gives range, interquartile range, median, and outlier information

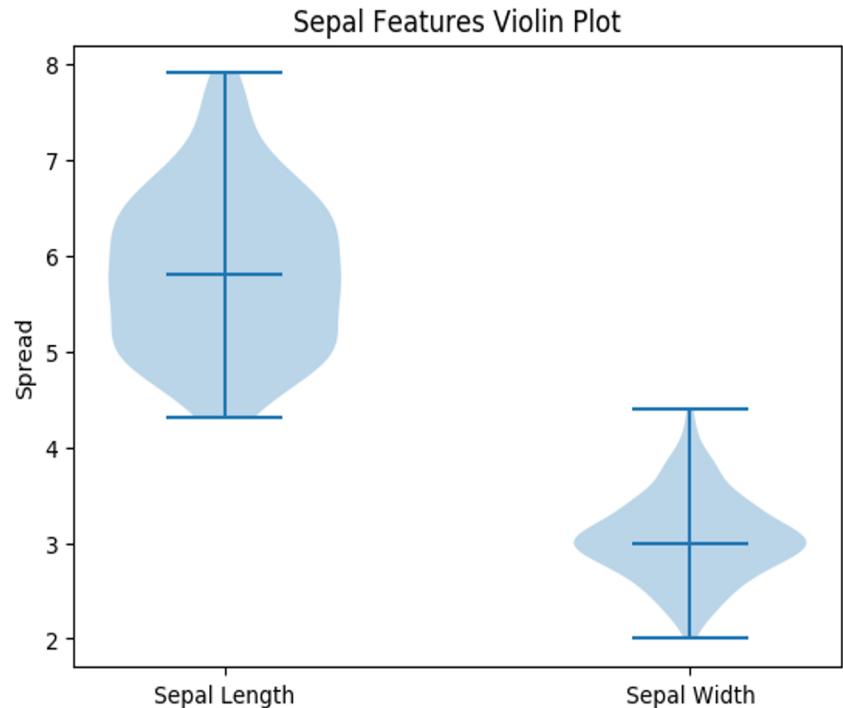


[Source](#)



Violin Plot

- Combination of **boxplot** and **density plot** to show the **spread** and **shape** of the data
- Can show whether the data is **normal** (i.e. is distributed normally)

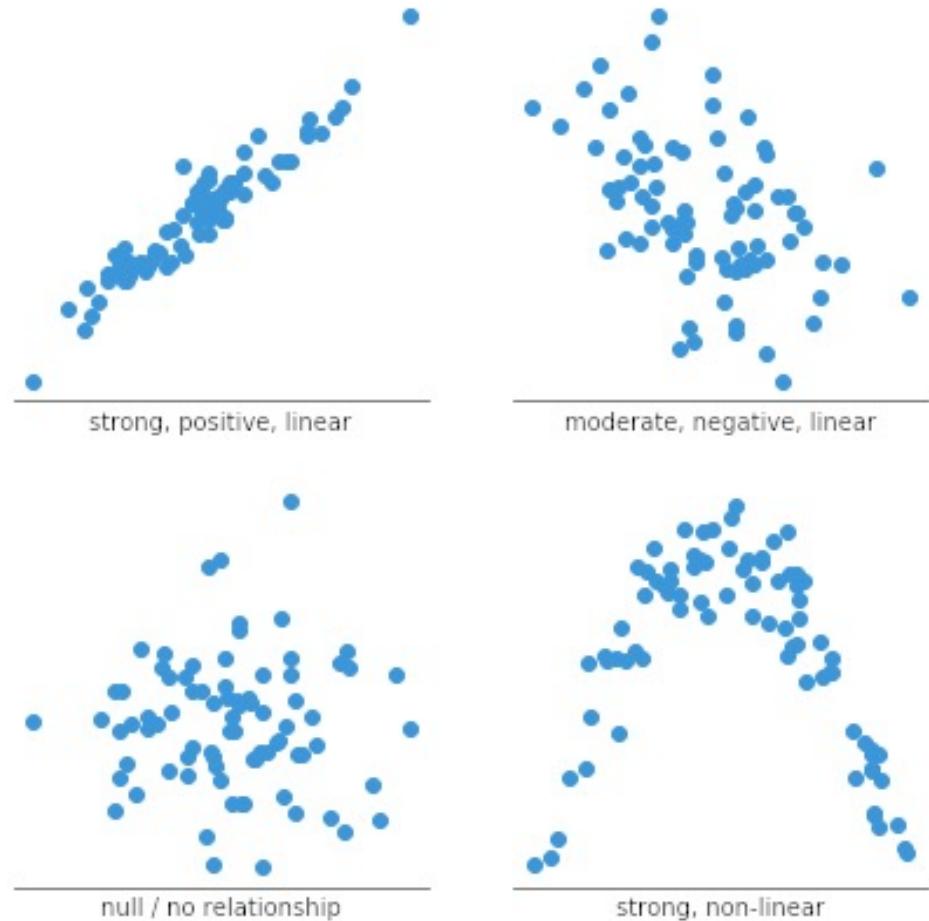


Advanced Data Visualizations



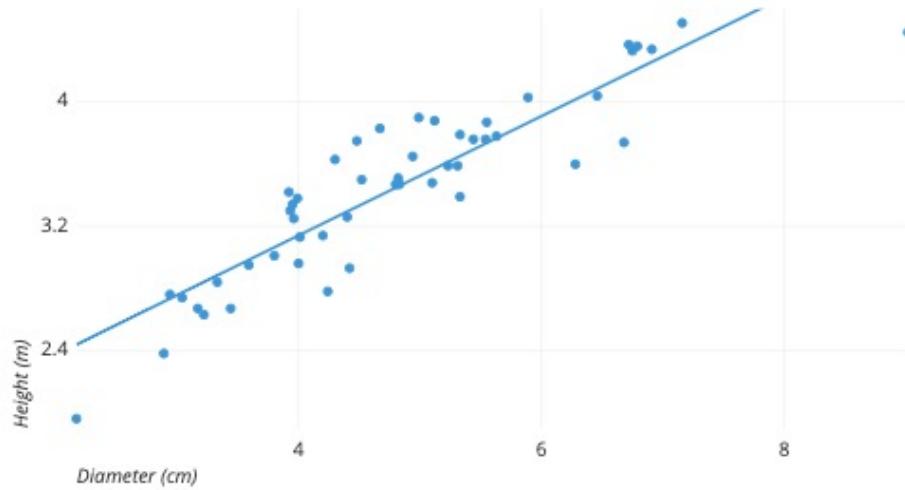
Scatterplot

- See **relationship** between two features
- Can be useful for **extrapolating** information

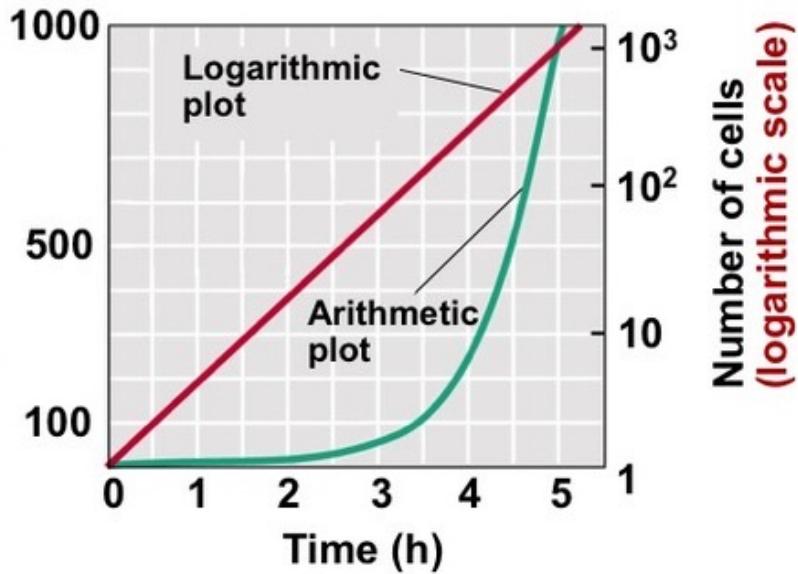


Scatterplot – more ways

- Line of best fit

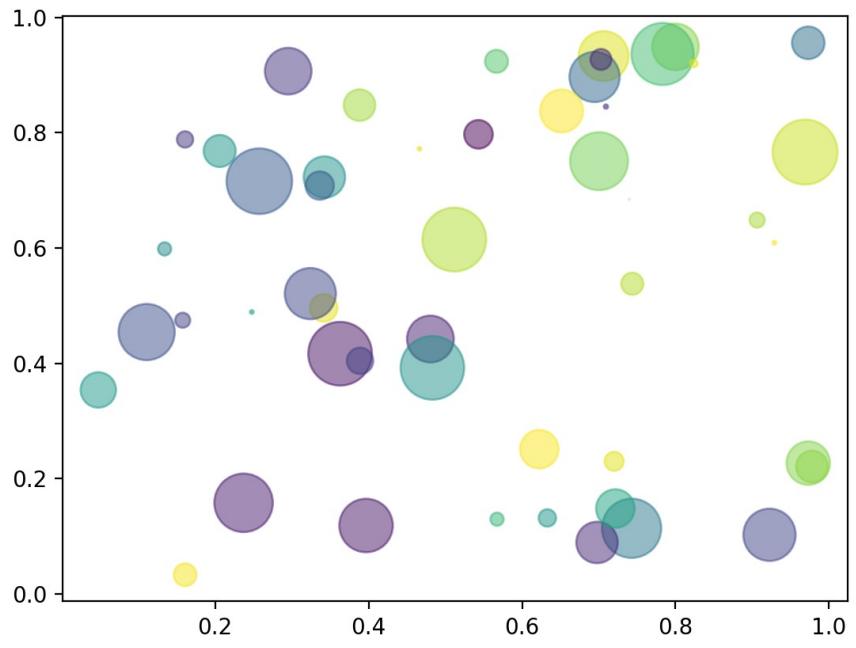
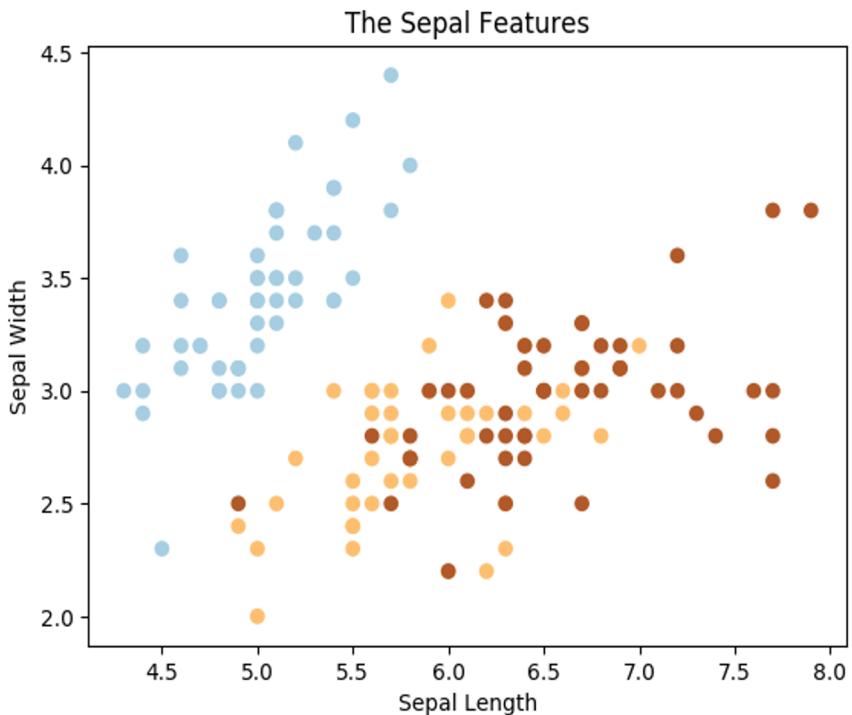


- Number of cells
(arithmetic scale)

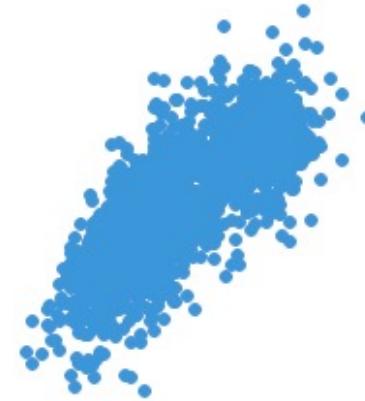


Scatterplot – more ways

- Line of best fit
- Demonstrate clusters
- Bubble chart



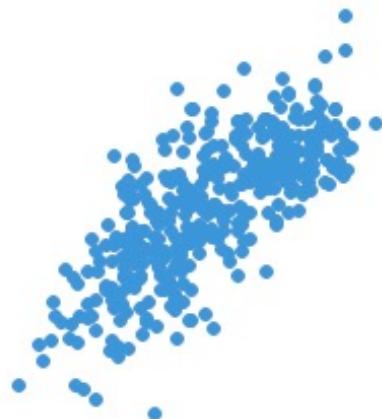
Original data, 1500 points



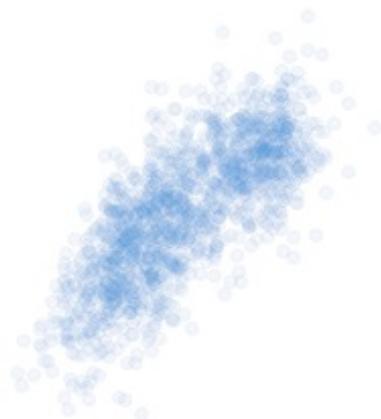
Scatterplot - Overplotting

- Only sample a random selection
- Change dot form (eg. add transparency)
- Use heatmap

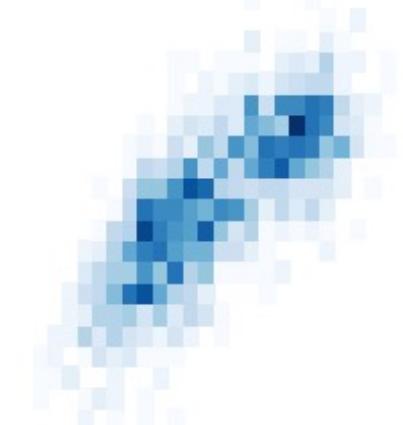
Sampled data, 400 points



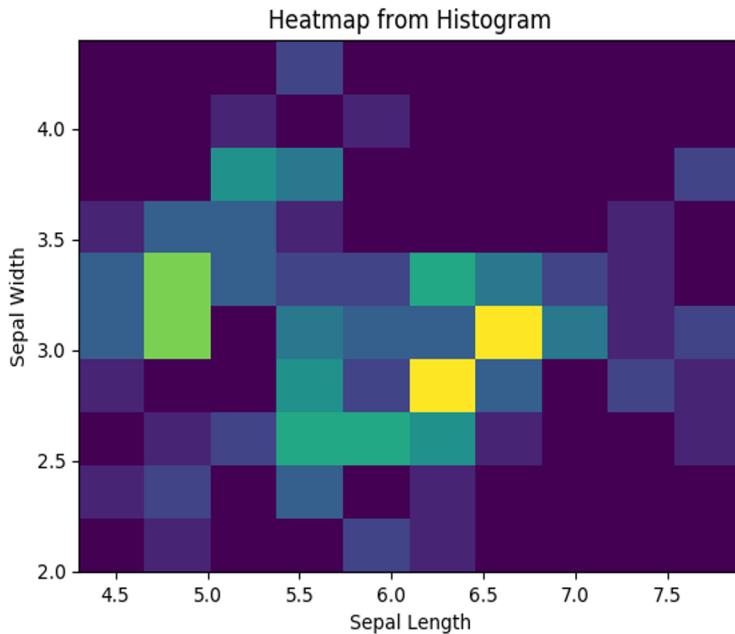
Plot w/ Transparency



Plot as 2-d histogram



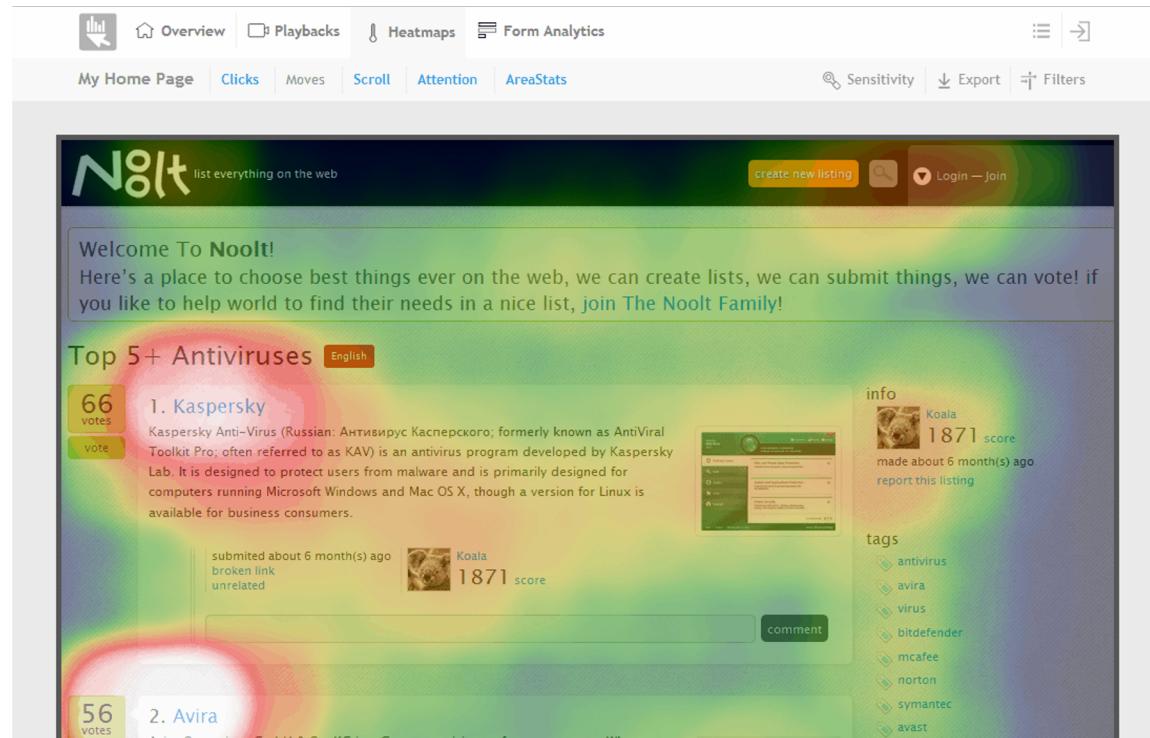
Heatmap



- Varying degrees of one metric are represented using **color**
- Especially useful in the context of **maps** to show geographical variation



Heatmap - Click Density / Website Heatmaps

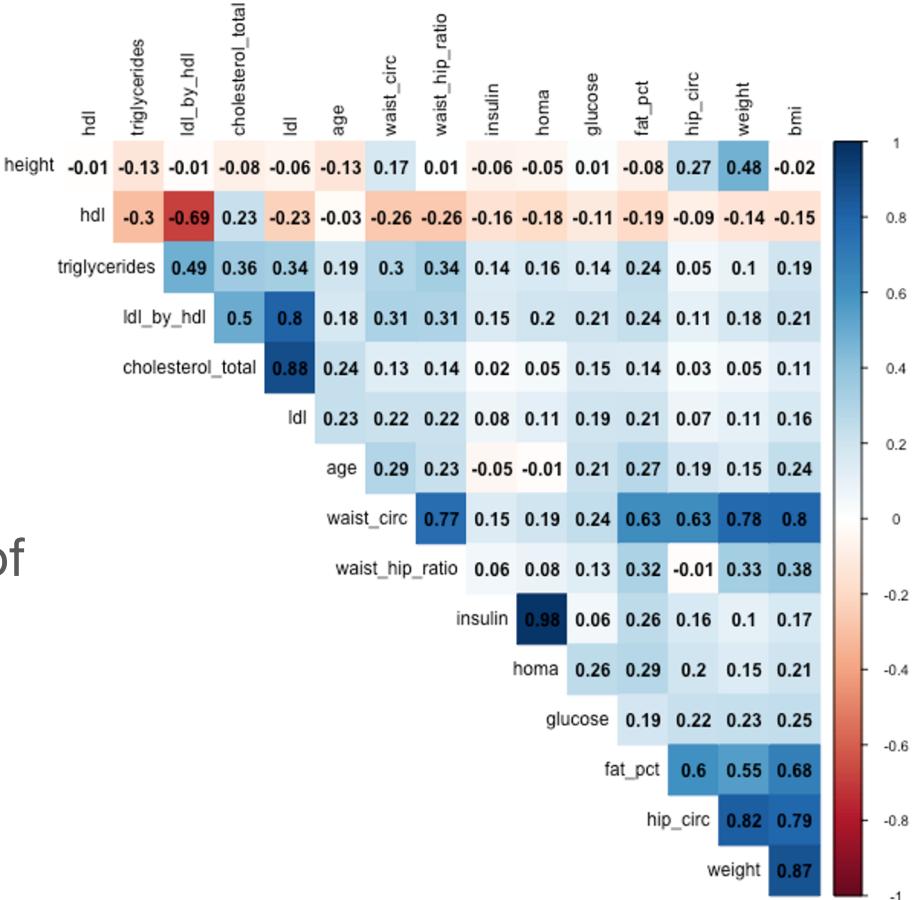


Correlation Plots

- 2D matrix with all variables on each axis
- Entries represent the **correlation coefficients** between each pair of variables

```
[[ 1.          -0.10936925  0.87175416  0.81795363]
 [-0.10936925  1.          -0.4205161   -0.35654409]
 [ 0.87175416 -0.4205161   1.          0.9627571 ]
 [ 0.81795363 -0.35654409  0.9627571   1.        ]]
```

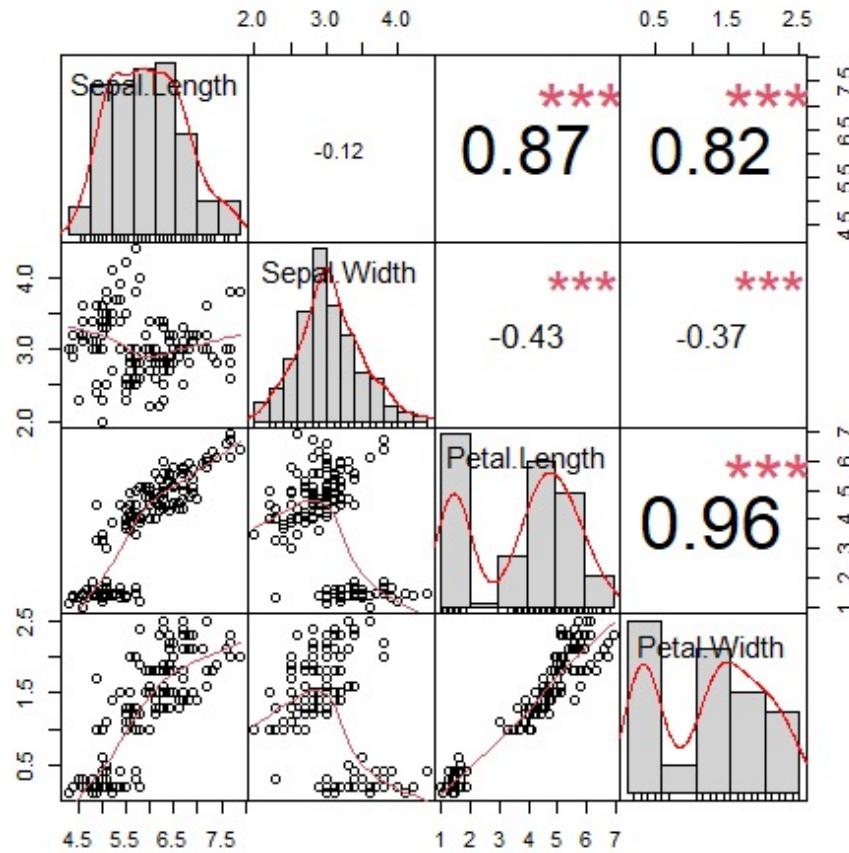
Why are all entries on the diagonal '1'?



Source



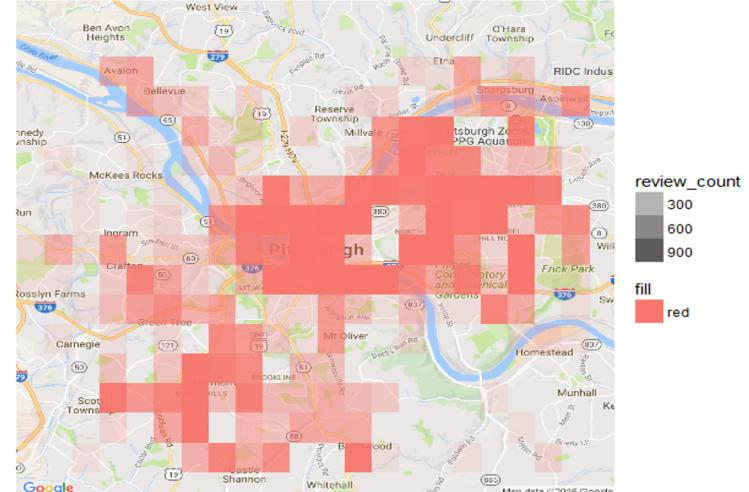
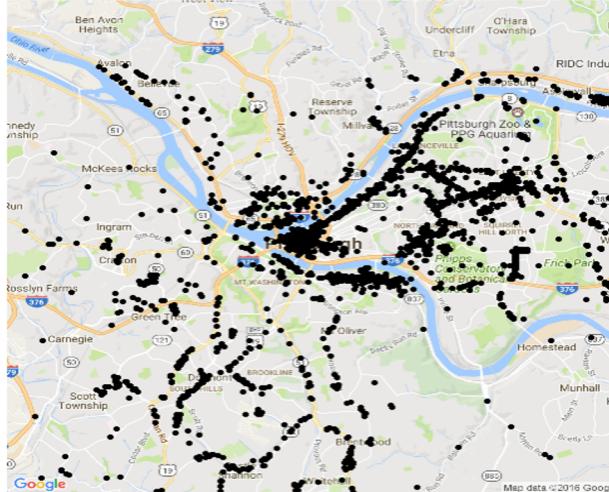
Correlation Plots



Using Maps

➤ Map visualization → contextual information

- Trends are not always apparent in the data itself
- Eg. Longitudes + Latitudes → Geographical Map



Demo



Challenges of Visualization

Higher Dimension

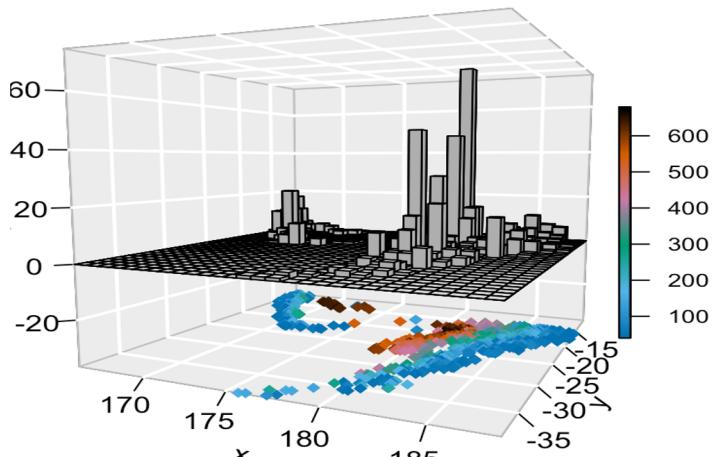
Non-Trivial

Time Consuming

**Hard to Show
Uncertainty**

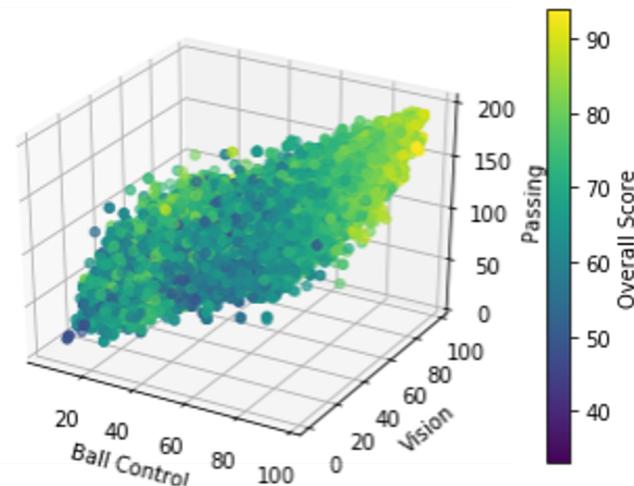


High Dimensional Data



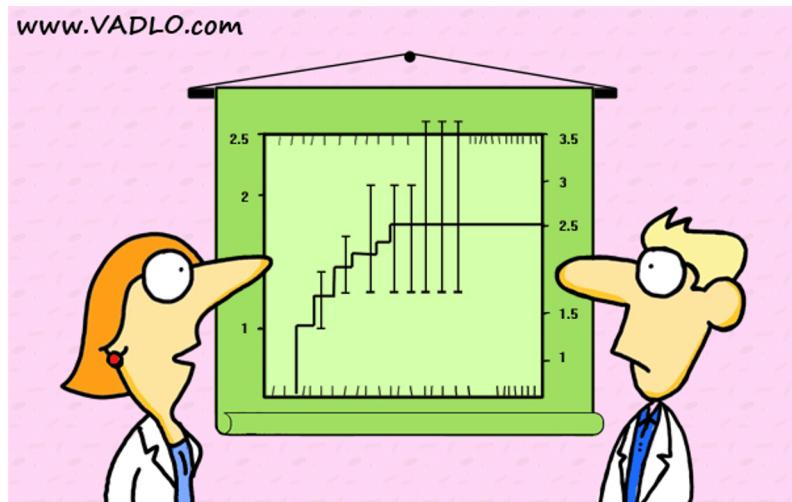
4D Plot For Earthquake Data

- Color, time animations, or point shape can be used for higher dimensions
- There is a limit to the number of features that can be displayed

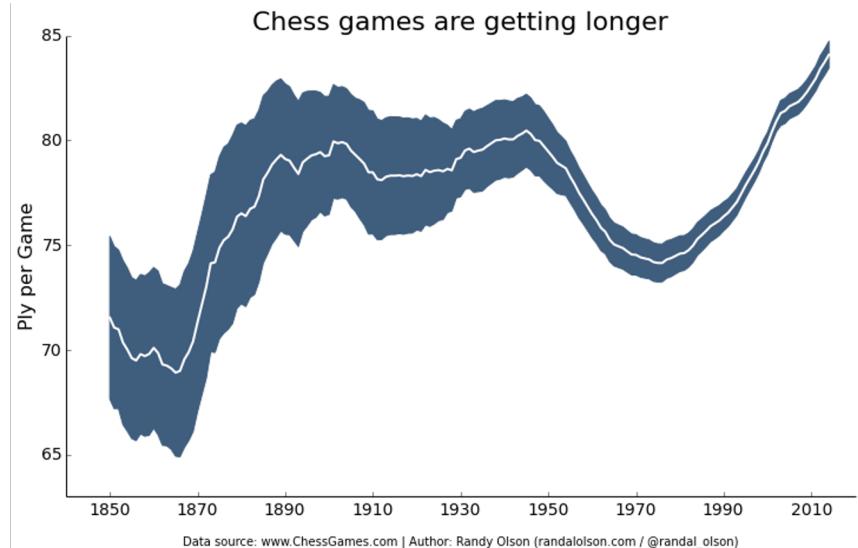


Error Bars

- Used to show uncertainty
- Usually display 95 percent confidence interval



"Did you really have to show the error bars?"



Coming Up

- **Assignment 3:** Due at 4:30pm (ET) on March 9th, 2022
- **Next Lecture:** Fundamentals of Machine Learning
- **Welcome back from Feb. Break!**
- **Data Scraping Workshop October 16th (timing + room TBD)**
- **Enroll in ED! Check ED before writing emails! Post Questions on ED!**



CDS Education

We explore, learn, and educate big minds.

