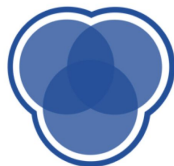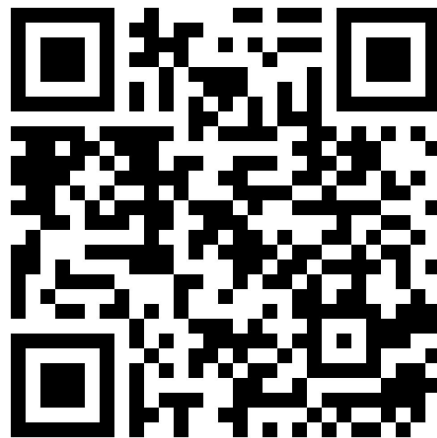# INFO 1998: Introduction to Machine Learning

CDS Education

We explore, learn, and educate big minds.

Lecture 3: Data Visualization

INFO 1998: Introduction to Machine Learning
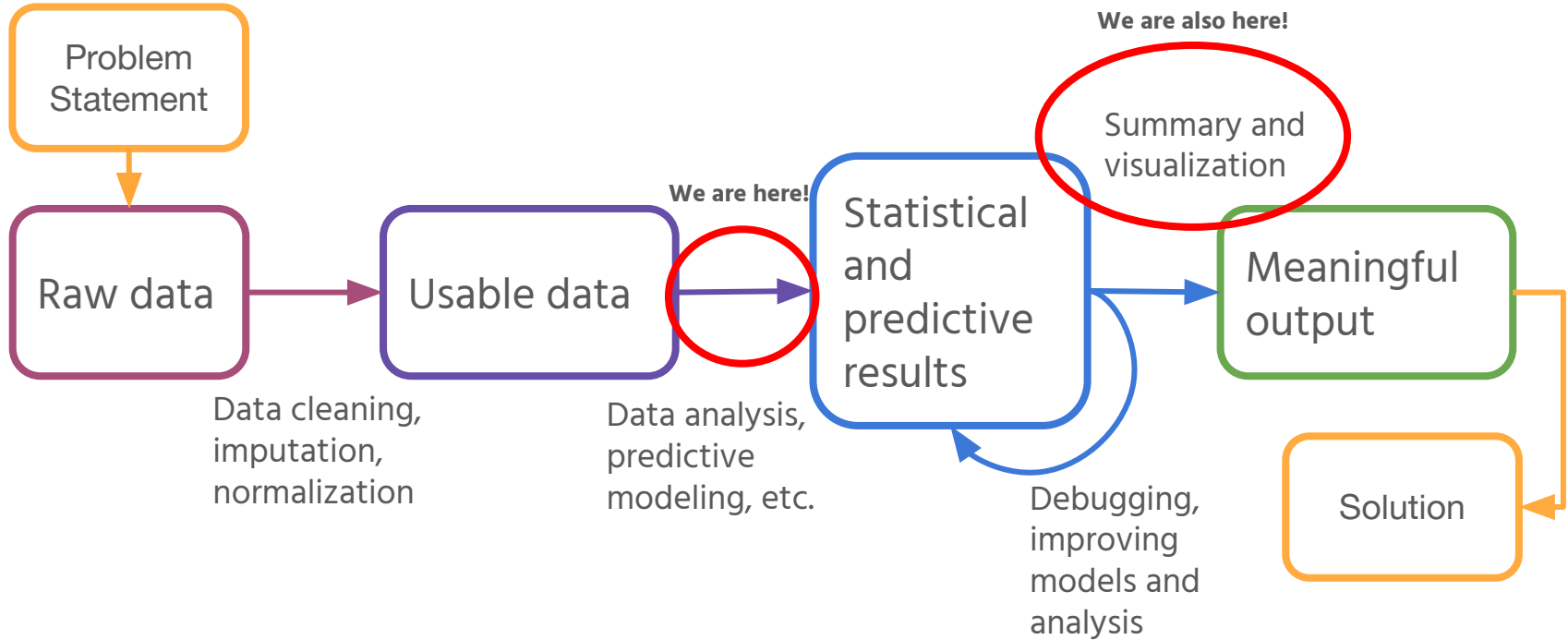
CDS Education

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**
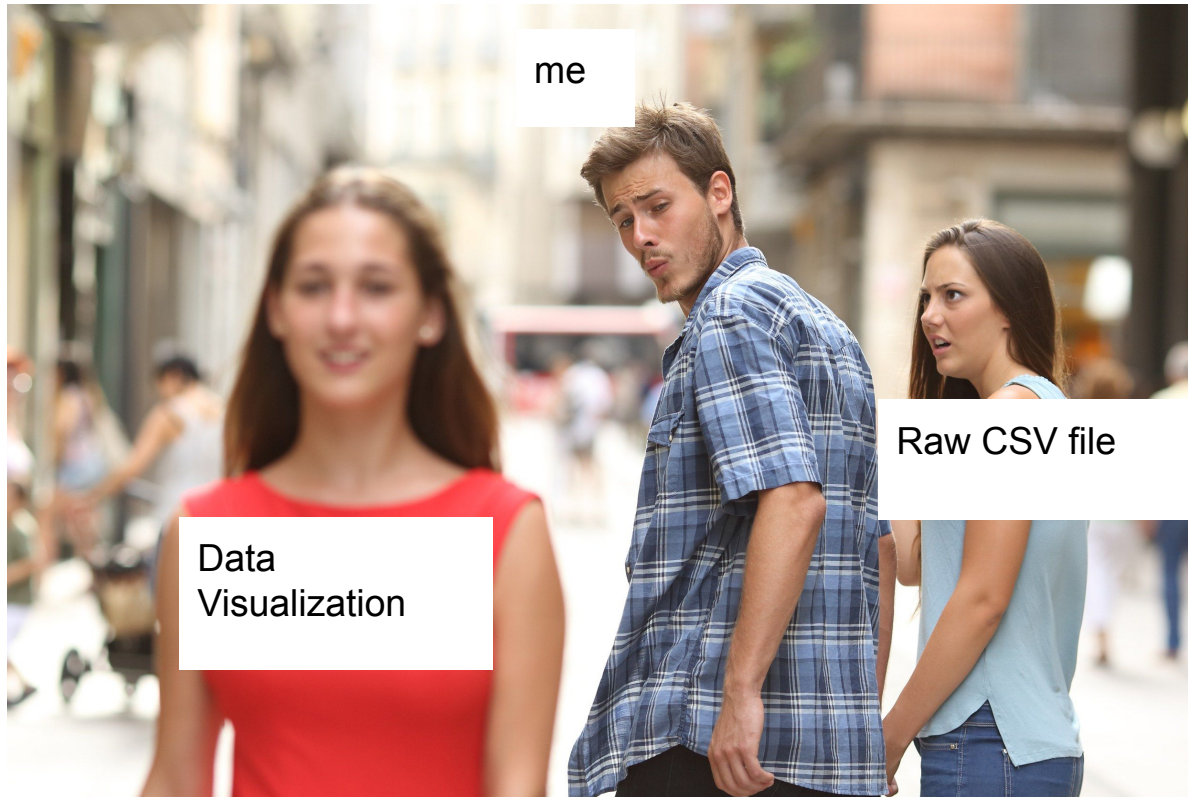
The Data Pipeline

We are also here!

We are here!

| Problem Statement |
| Raw data | → | Usable data | → | Statistical and predictive results | → | Meaningful output | → | Solution |

Summary and visualization

Data cleaning, imputation, normalization

Data analysis, predictive modeling, etc.

Debugging, improving models and analysis

This!!!

Why is Data Visualization Important?

Why is Data Visualization Important?
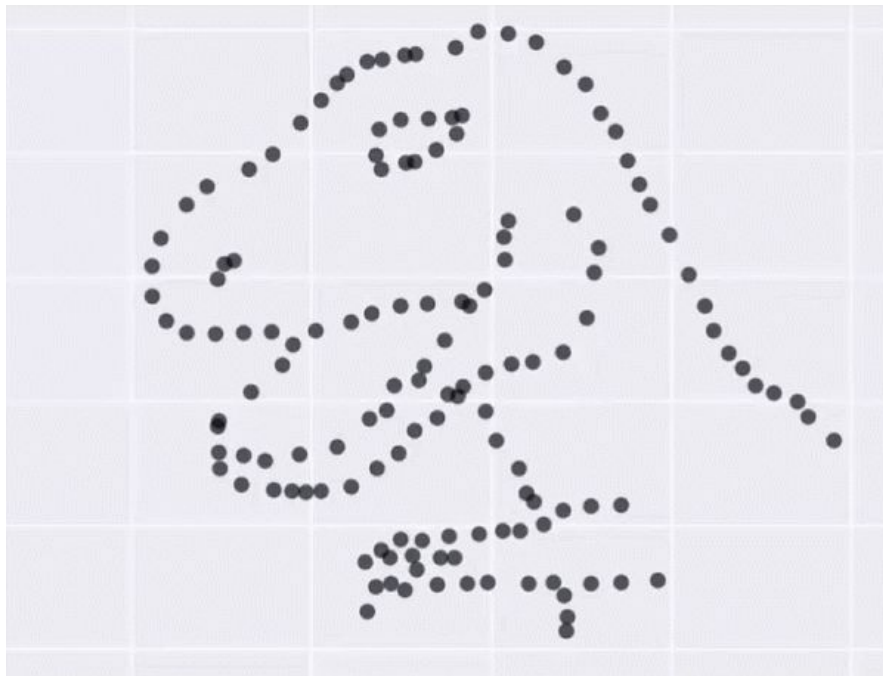
Informative

Appealing

Universal

Predictive

Why is Data Visualization Important?

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

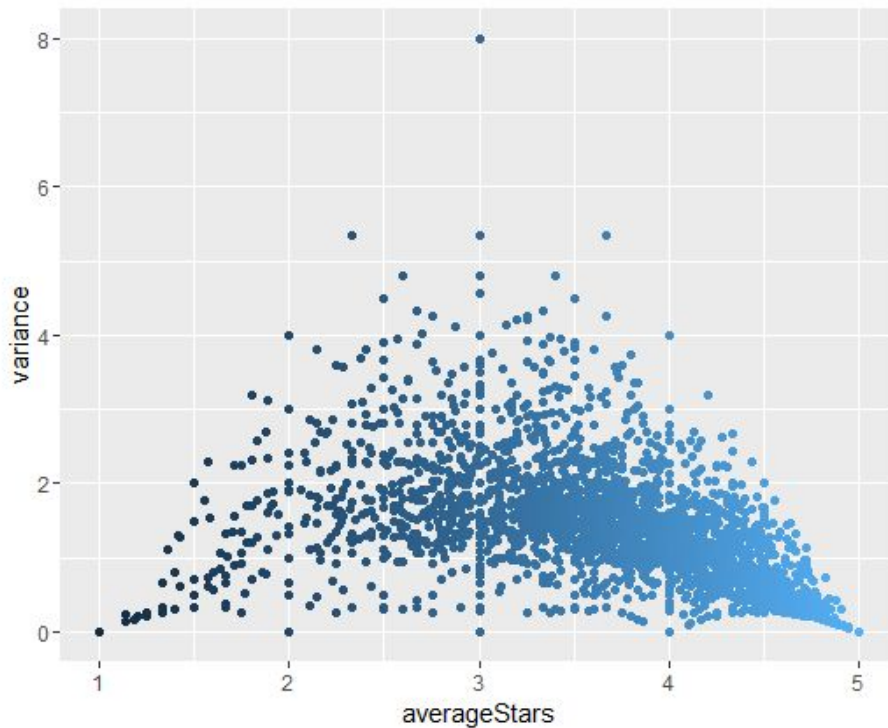We need to see how the **actual** data looks!

df.describe() is not enough

Data Visualization Simple Example: *Ratings on Yelp*

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

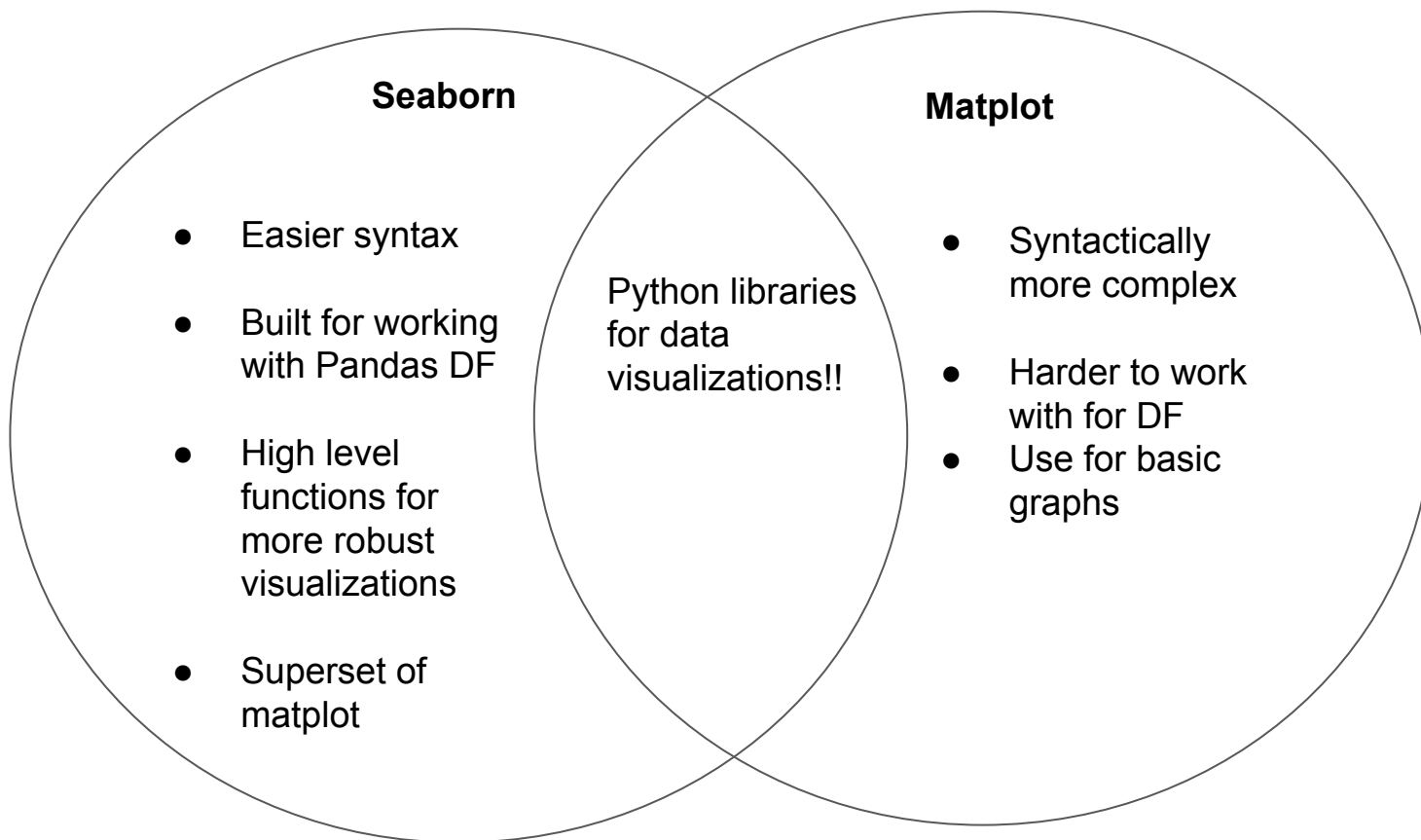**Question:** What do you notice? What trends do you see?

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

Seaborn vs Matplot

Seaborn

Matplot

- Easier syntax

- Built for working
  with Pandas DF

- High level
  functions for
  more robust
  visualizations

- Superset of
  matplot

Python libraries
for data
visualizations!!

- Syntactically
  more complex

- Harder to work
  with for DF
- Use for basic
  graphs

Basic Data Visualizations

Bar Graph

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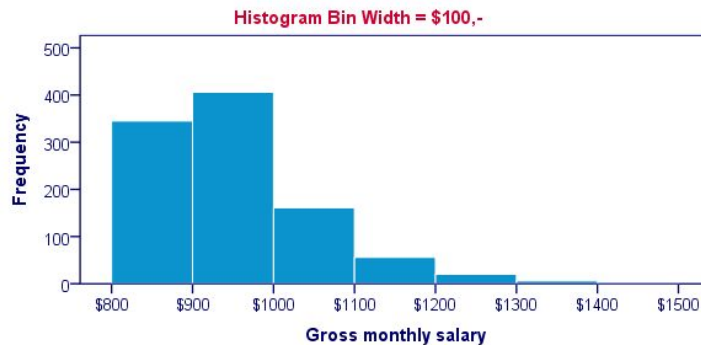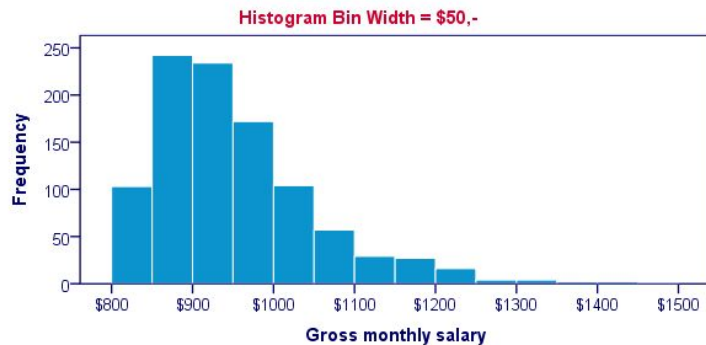
- Allows us to compare features
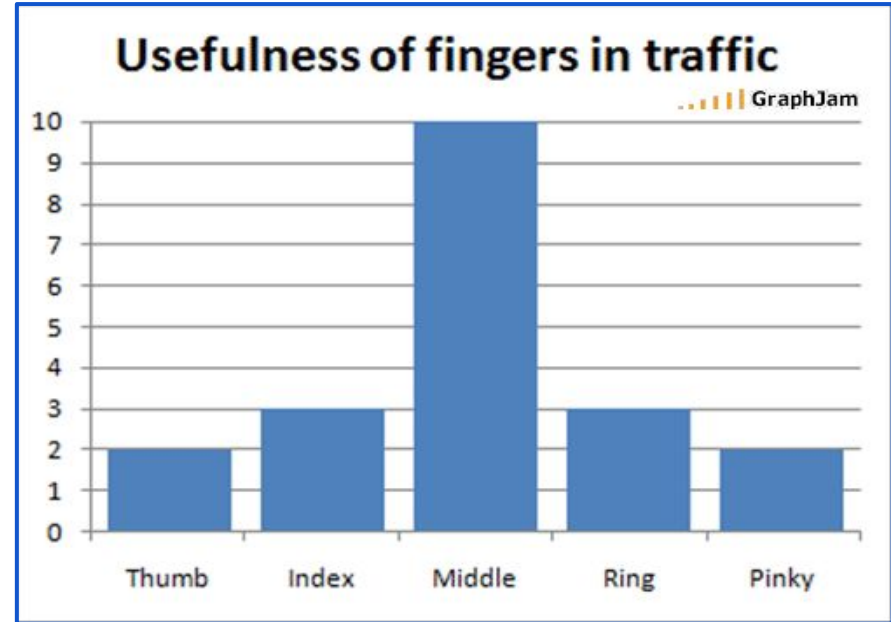
Source

Histograms

Histogram of Monthly Salary

- Used to observe **frequency distribution** of continuous variables

- Data split into **bins**

Histograms: Different Bin Sizes

Density Plot
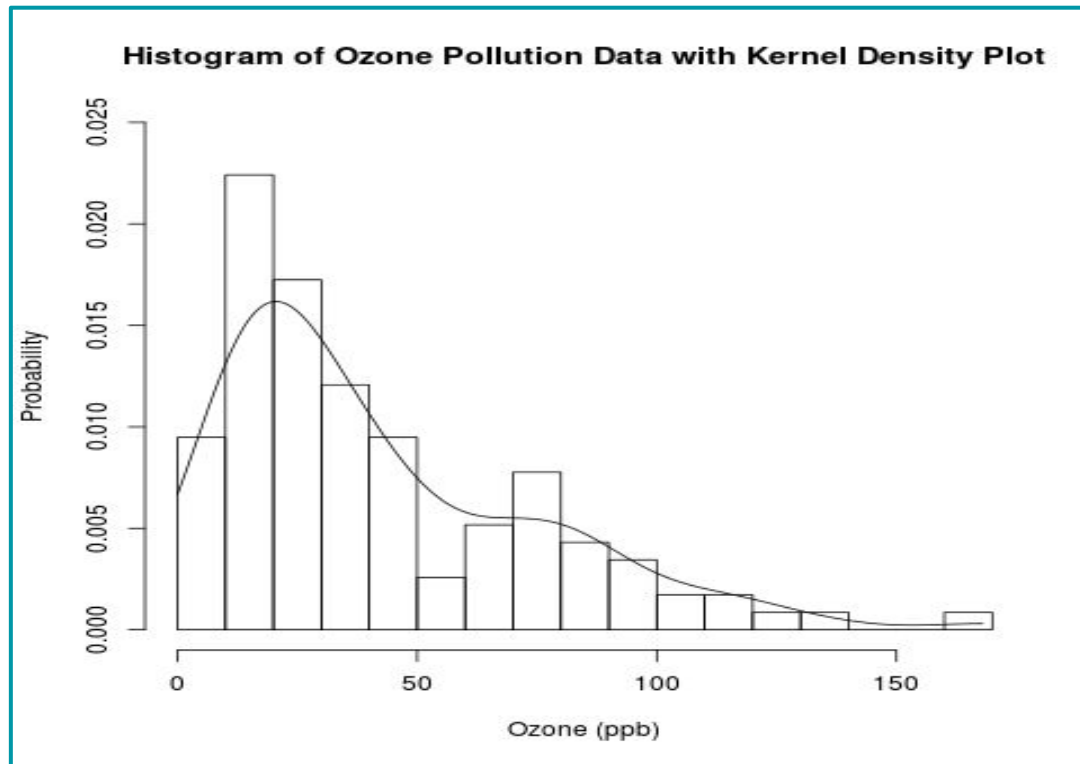

MPG Distribution by Car Cylinders
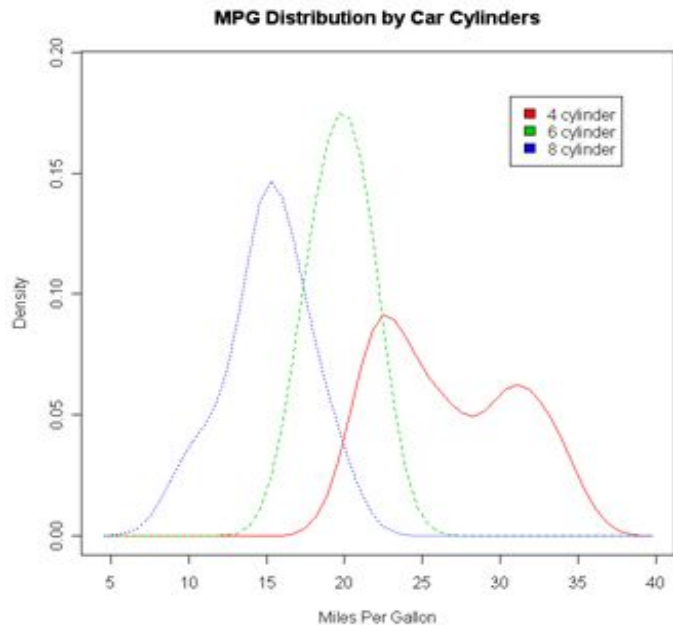
Like a histogram, but **smooths** the shape of the distribution

Histogram vs Density Plot

Histogram of Ozone Pollution Data with Kernel Density Plot
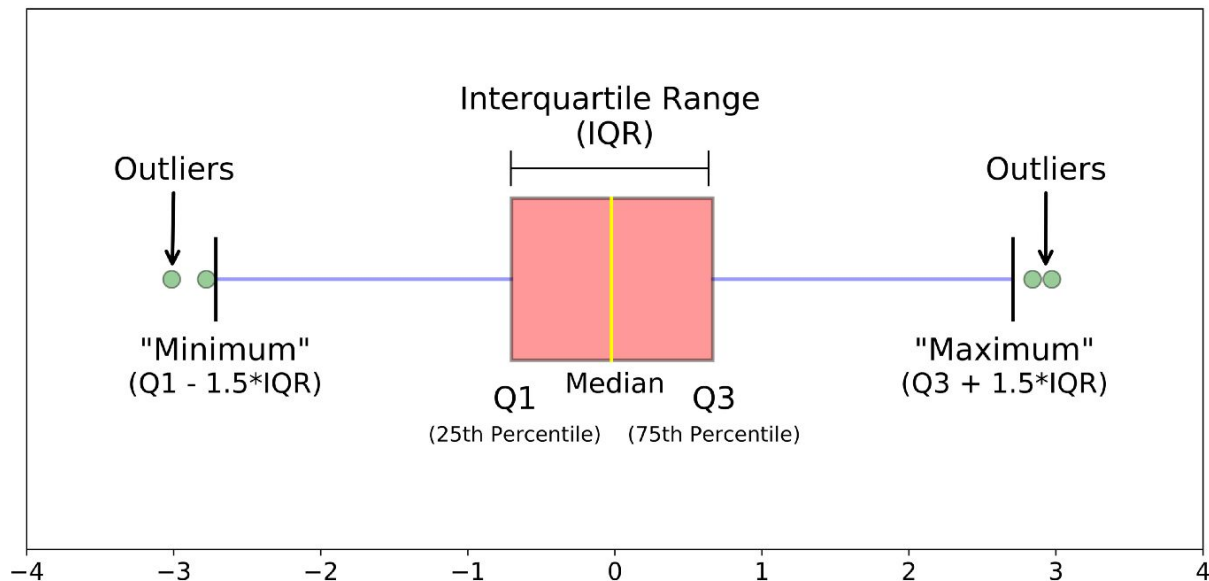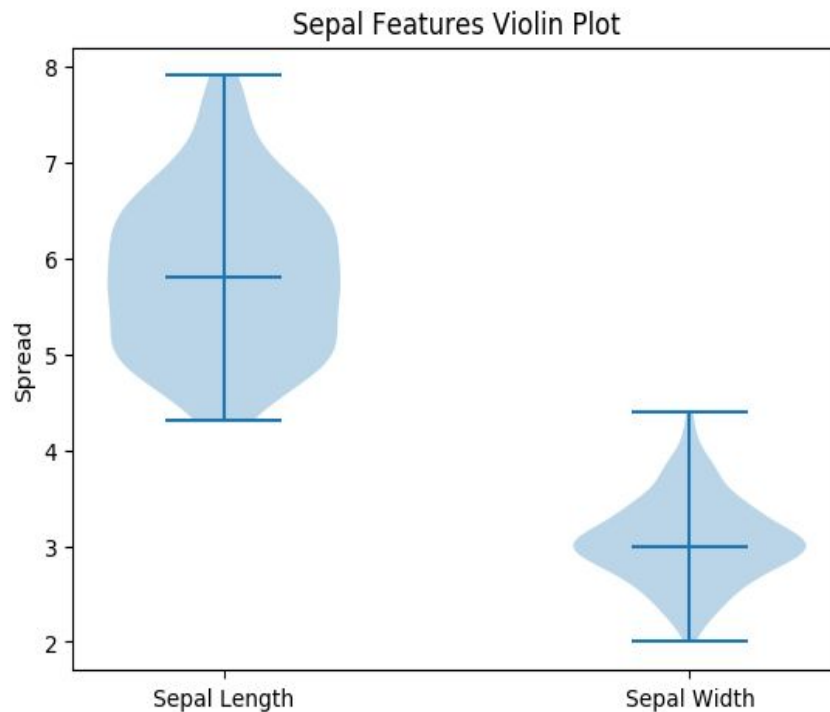
Boxplot (a.k.a box and whisker plot)

- 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

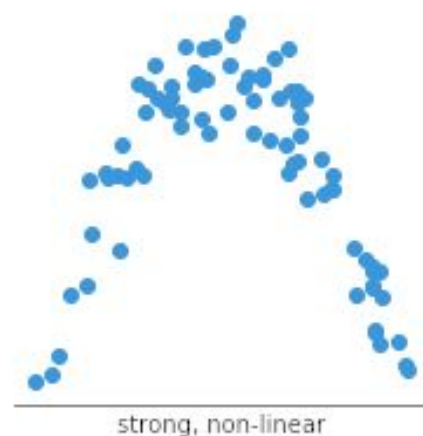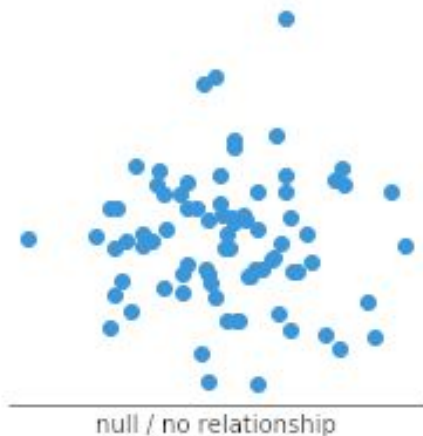- Can show whether the data is **normal** (i.e. is distributed normally)
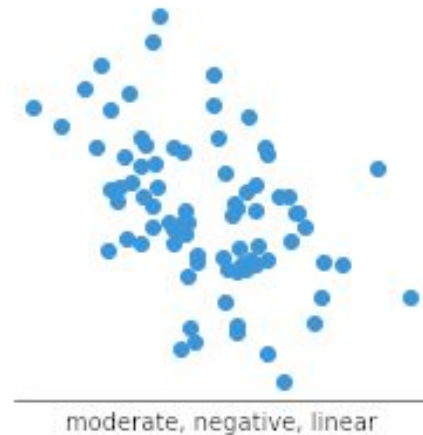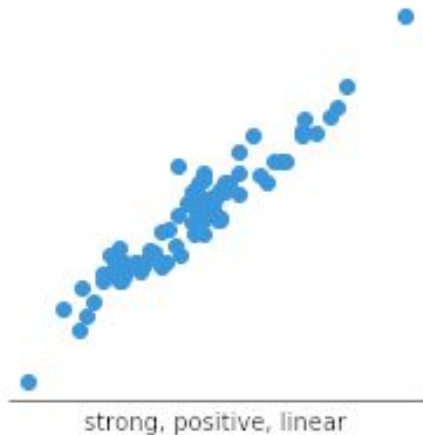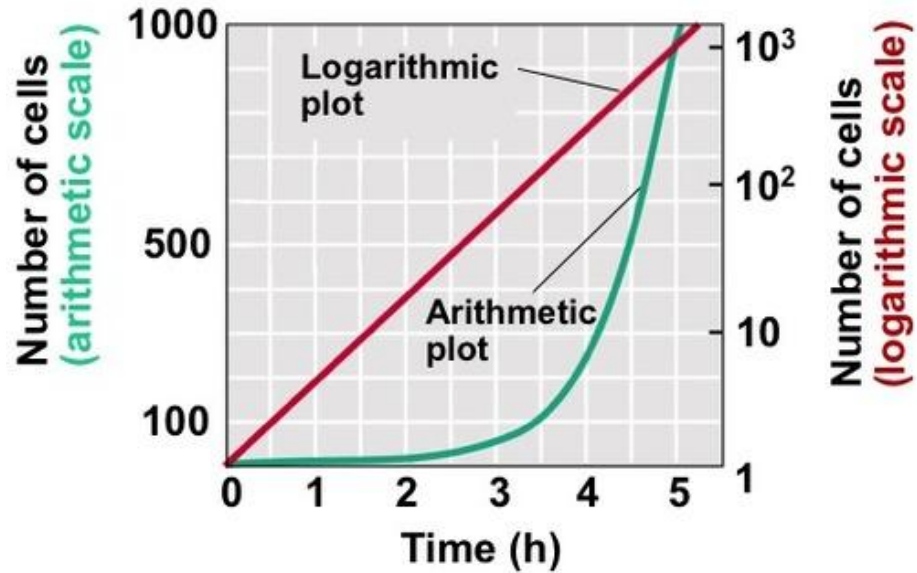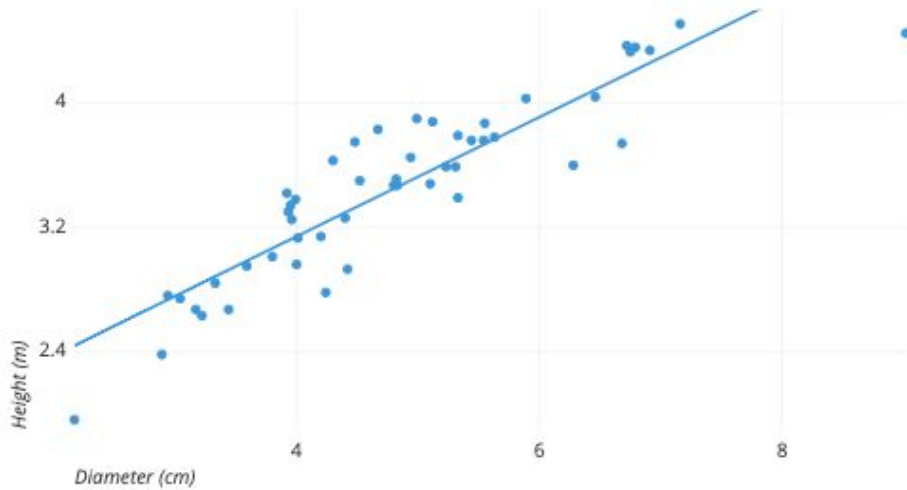

Sepal Features Violin Plot

Advanced Data Visualizations

Scatterplot

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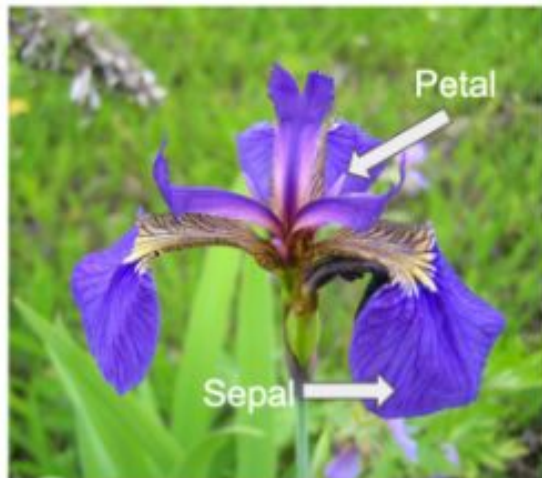
- Can be useful for **extrapolating** information

strong, positive, linear

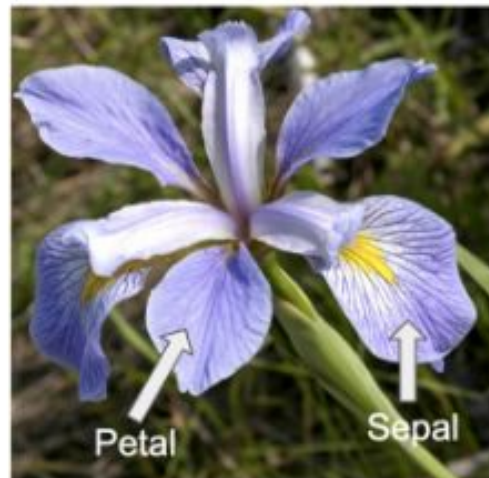moderate, negative, linear

null / no relationship

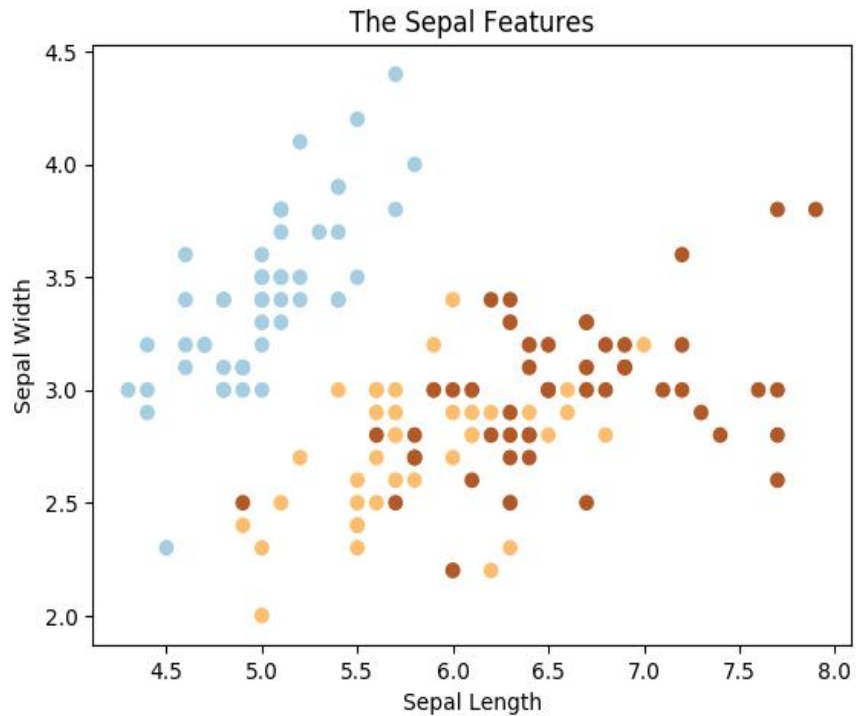strong, non-linear

Scatterplot – more ways

- Line of best fit

Iris setosa — Petal, Sepal
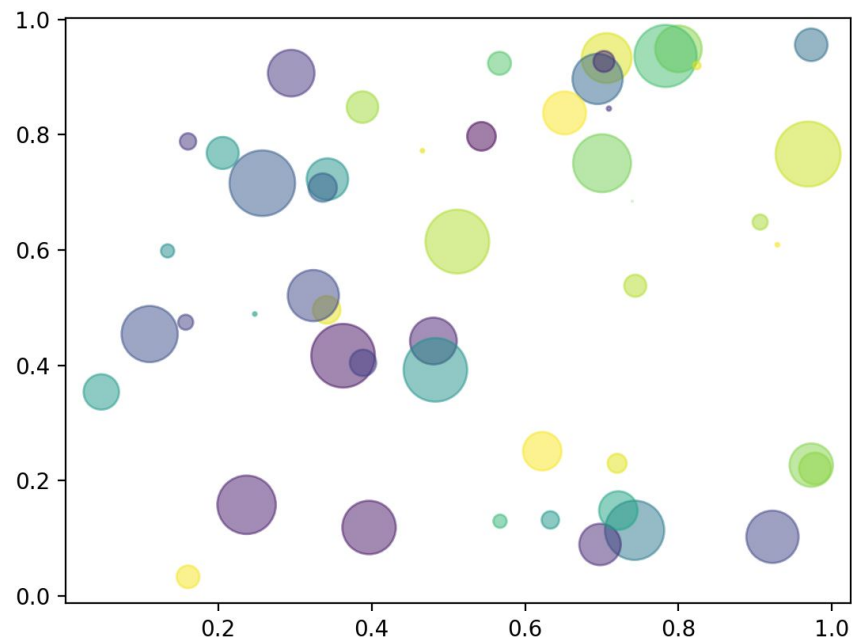*Iris versicolor* — Petal, Sepal
Iris virginica — Petal, Sepal
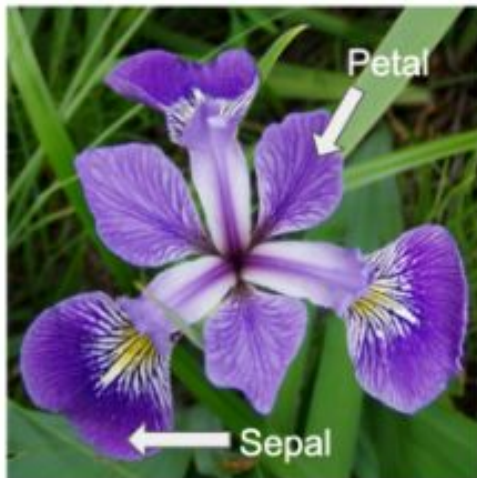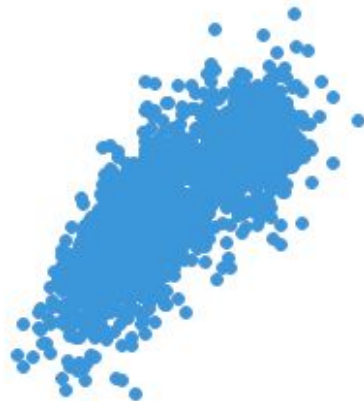
Scatterplot – more ways

- Line of best fit
- Demonstrate clusters
- Bubble chart

Scatterplot - Overplotting

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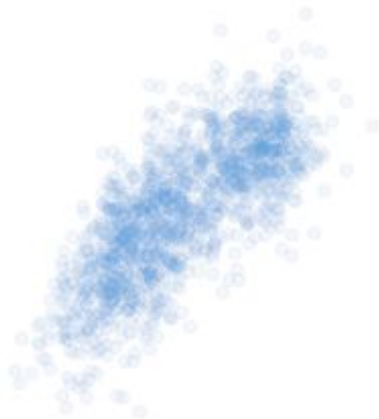- Change dot form (eg. add transparency)
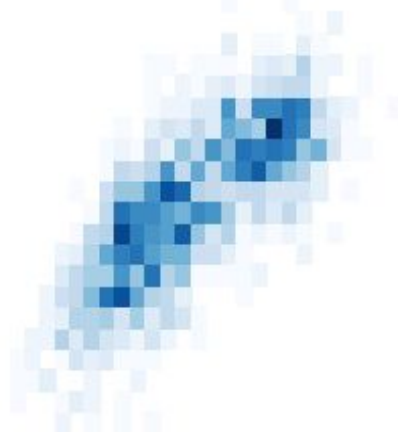- Use heatmap

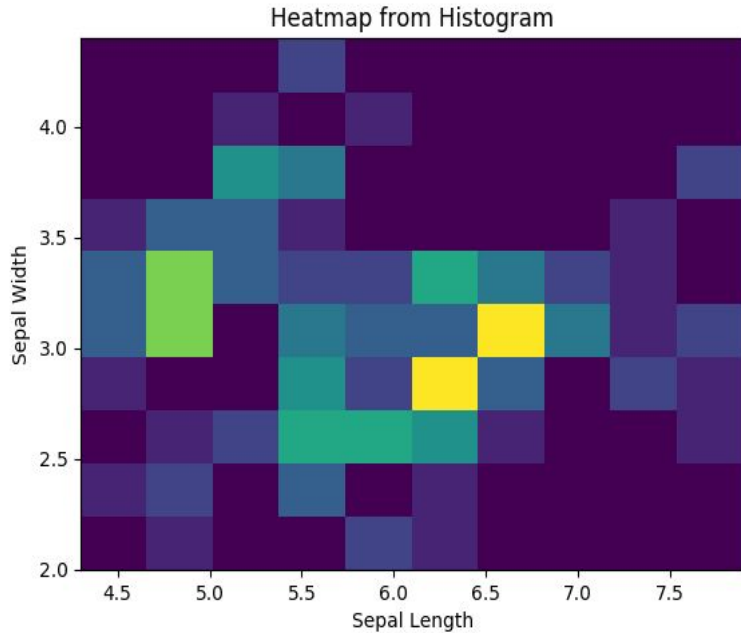Original data, 1500 points

Sampled data, 400 points

Plot w/ Transparency

Plot as 2-d histogram

Heatmap

Heatmap from Histogram
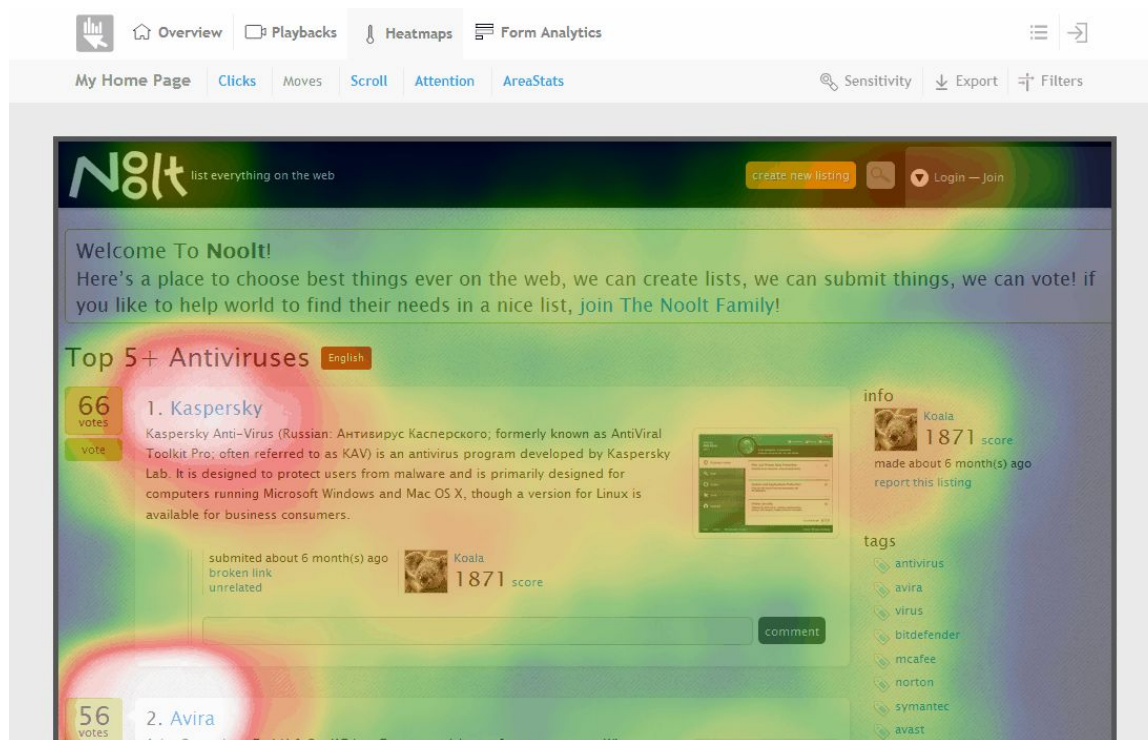
- Varying degrees of one metric are represented using **color**

- Especially useful in the context of **maps** to show geographical variation

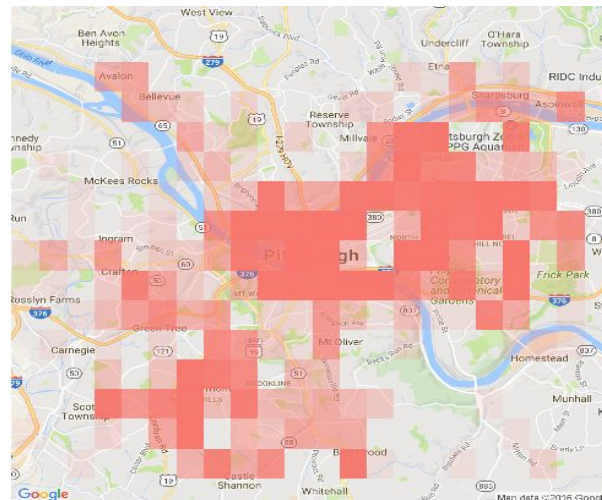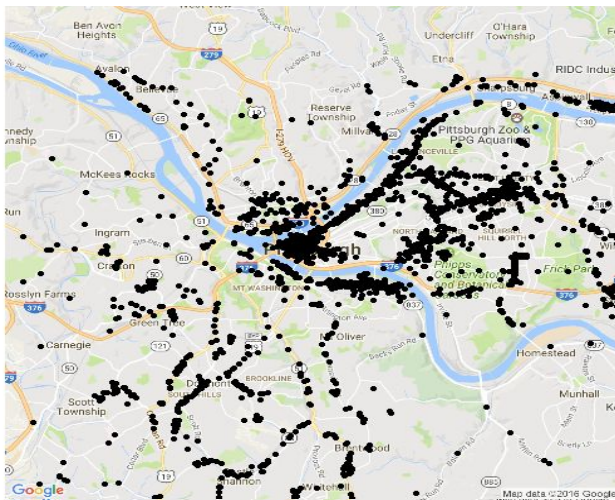# Heatmap - Click Density / Website Heatmaps

Using Maps

➤ **Map visualization** → **contextual information**

- **Trends are not always apparent in the data itself**

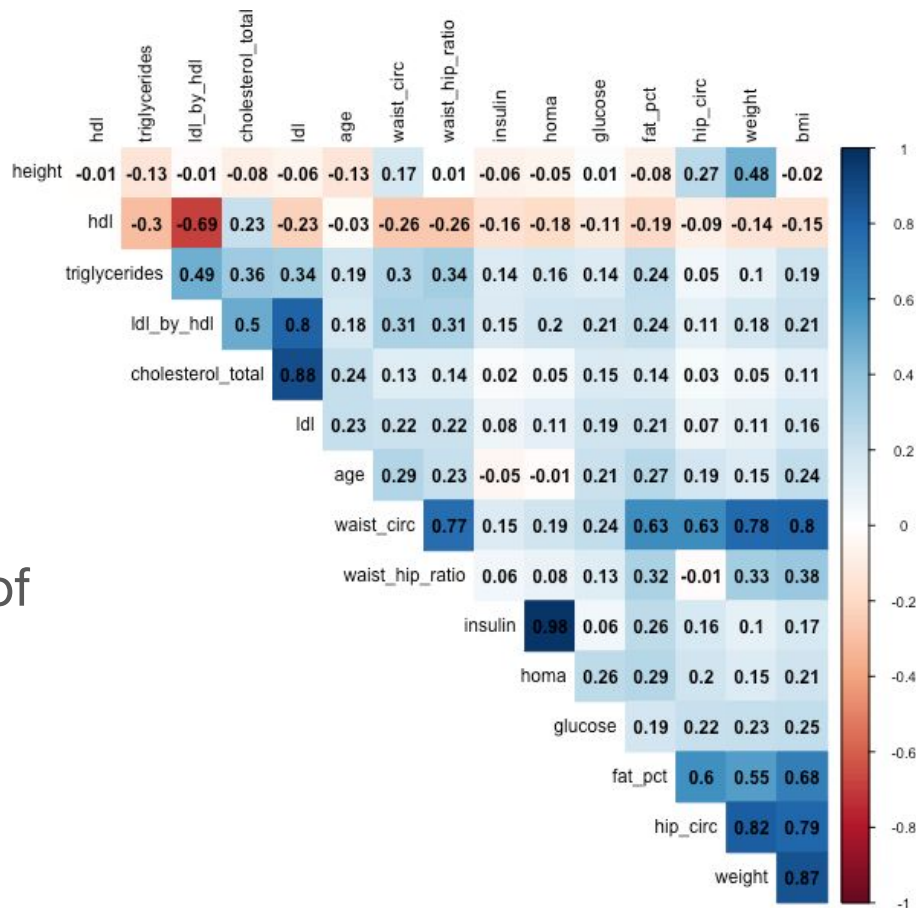- Eg. Longitudes + Latitudes → *Geographical Map*

Correlation Plots

- 2D matrix with all variables on each axis

- Entries represent the **correlation coefficients** between each pair of variables
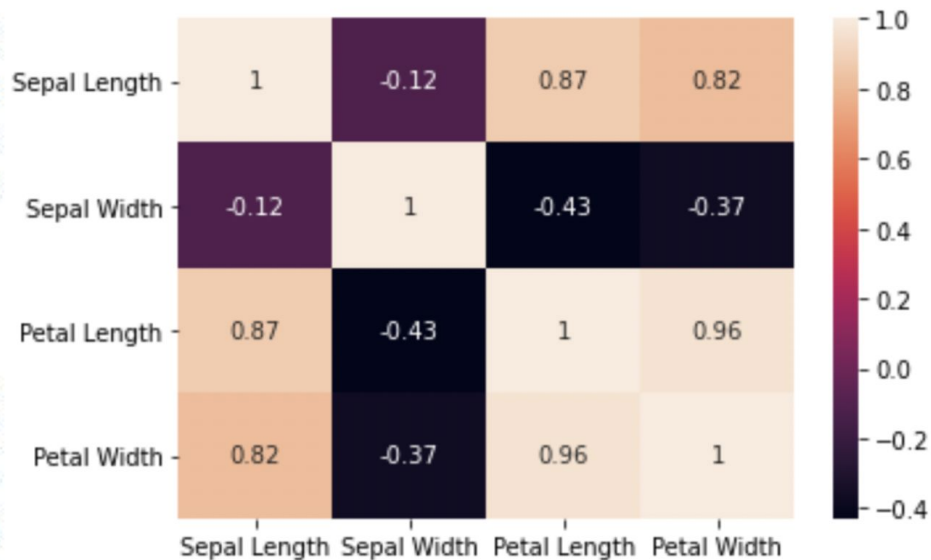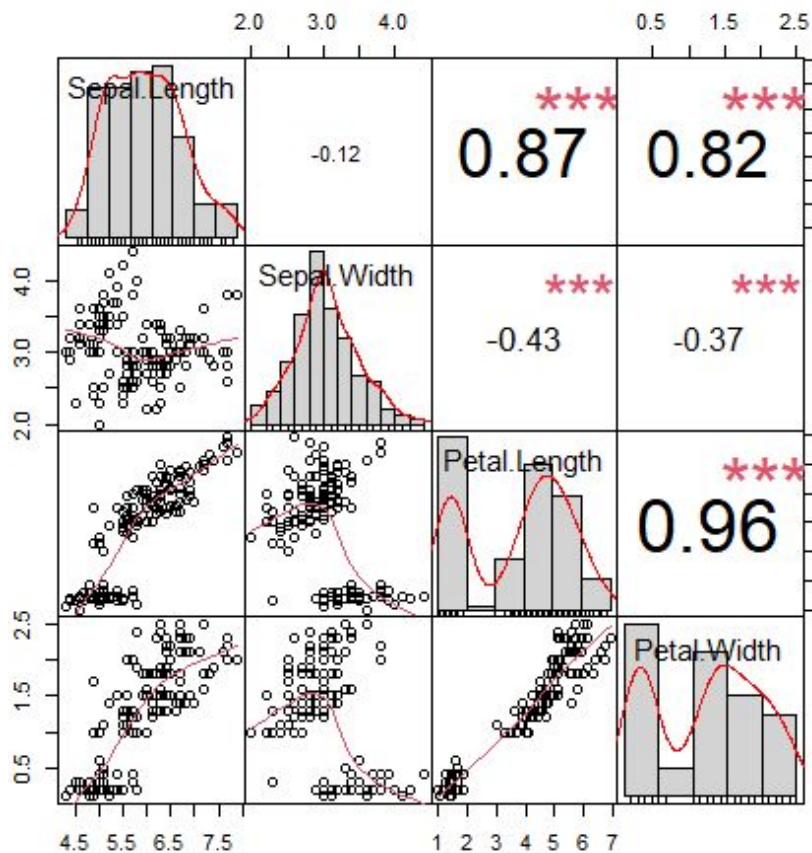
```
[[ 1.          -0.10936925   0.87175416   0.81795363]
 [-0.10936925   1.          -0.4205161   -0.35654409]
 [ 0.87175416  -0.4205161    1.           0.9627571 ]
 [ 0.81795363  -0.35654409   0.9627571    1.        ]]
```

Why are all entries on the diagonal '1'?

Correlation Plots

Demo

Challenges of Visualization
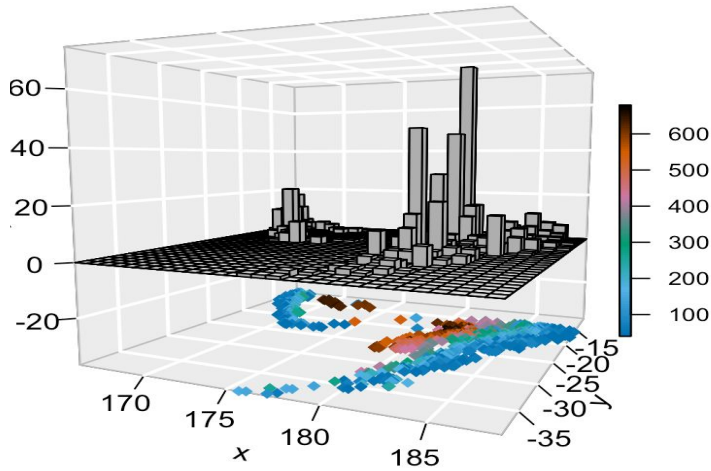
Higher Dimension

Non-Trivial

Time Consuming
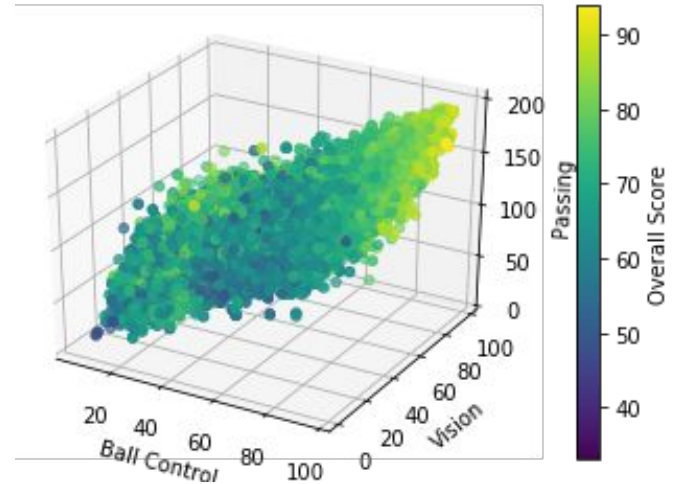
Hard to Show Uncertainty

High Dimensional Data

- Color, time animations, or point shape can be used for higher dimensions

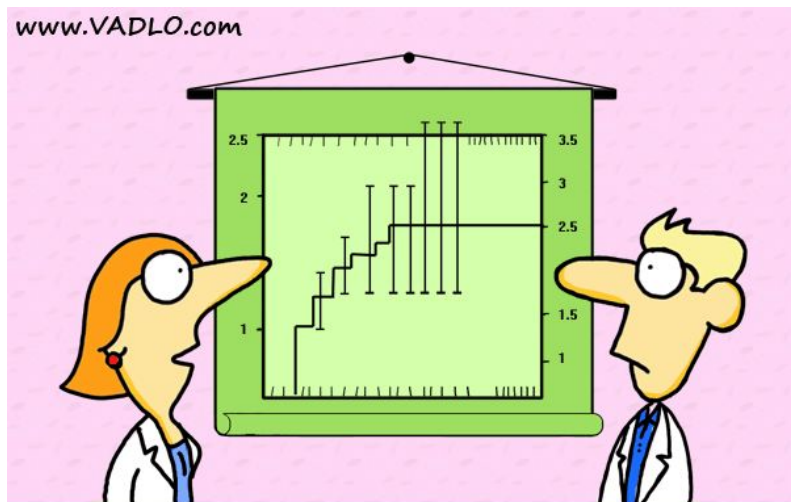- There is a limit to the number of features that can be displayed
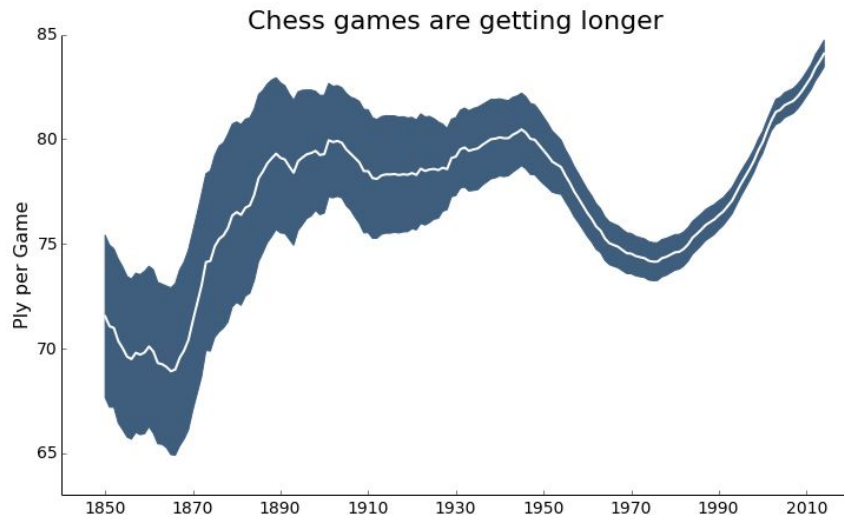

4D Plot For Earthquake Data

Error Bars

- Show uncertainty
- Usually display 95 percent confidence interval
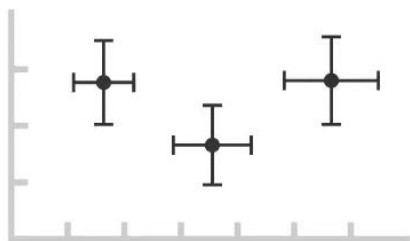

"Did you really have to show the error bars?"

Chess games are getting longer

Data source: www.ChessGames.com | Author: Randy Olson (randalolson.com / @randal_olson)

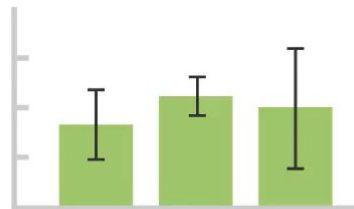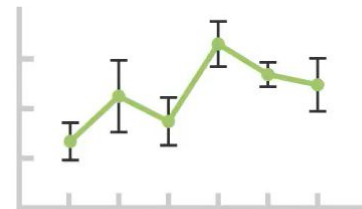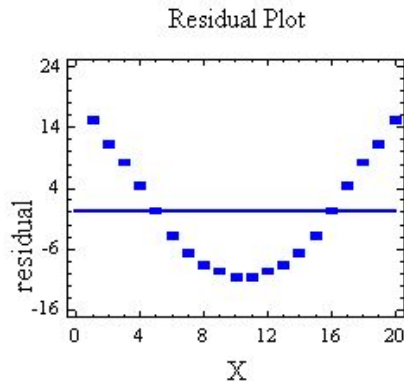# Error Bars
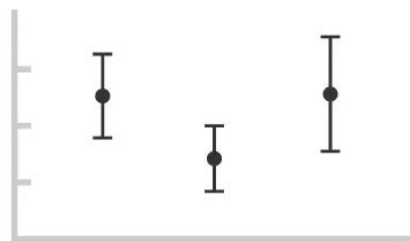
Residual Plot

- Values should be equally and randomly spaced on horizontal axis

- Regression line is called line of best fit

- Not optimal if data has outliers or is non-linear

Residual Plot

Projects!

For your visualizations..

- Choose the proper visualization

- Don't forget title, axis titles, etc.

2-3 people per project!

- Come up to the front to find someone

- **Partner finding on ED!**

CDS Education
We explore, learn, and educate big minds.

Coming Up

- **Assignment 3**: Due next Wednesday (03/06) at 11:59 PM

    ○ **Assignment 2:** Due this Friday (03/01) at 11:59pm

- **Next Lecture**: Fundamentals of Machine Learning

Check ED before writing emails! Post Questions on ED!

CDS Education
We explore, learn, and educate big minds.