Lecture 3: Data Visualization

INFO 1998: Introduction to Machine Learning

CDS Education

Agenda

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

Announcements

1. A1 Due Last Wednesday!
    - Please post on Ed about extenuating circumstances that prevented submission
2. A2 released!
    - Due Monday, 03/03/25
3. Ed > Emails

Coming Up

Assignment 2: Due tonight at 11:59pm!

Assignment 3: Due next Wednesday (10/02) at 11:59 PM

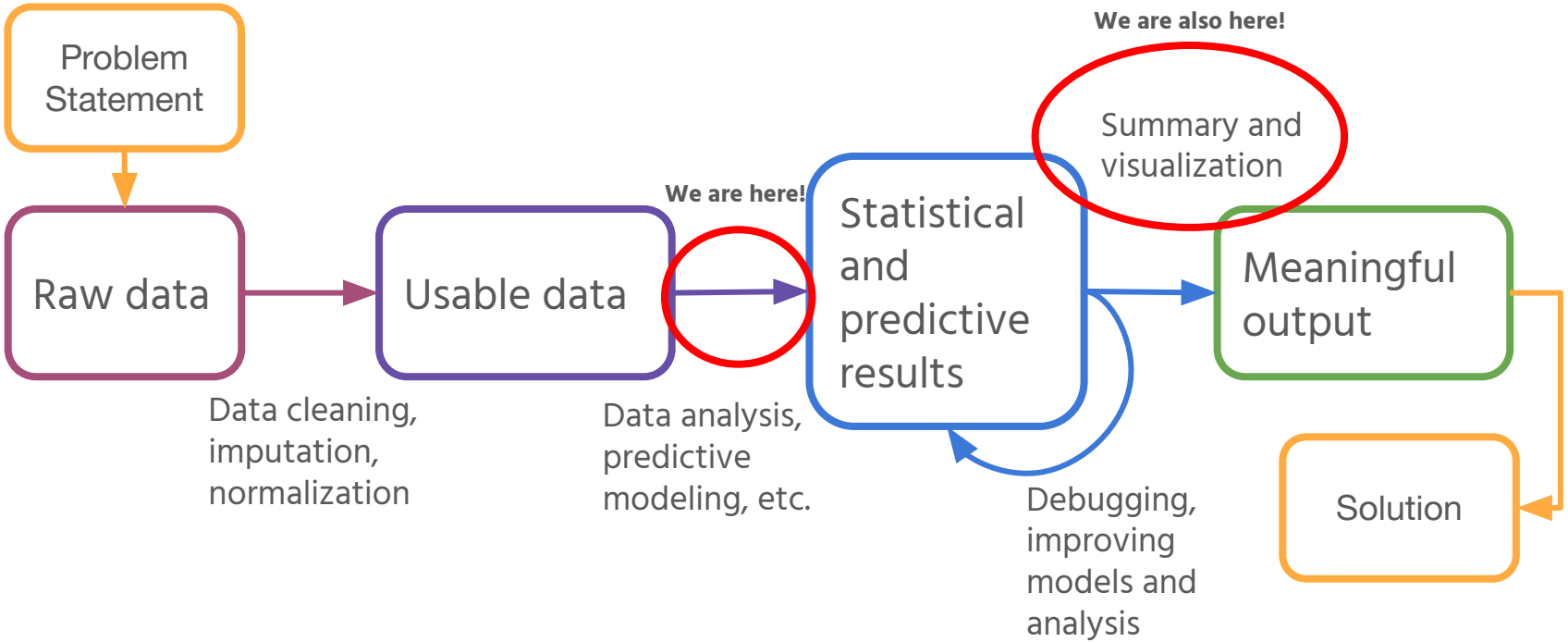
Next Lecture: Fundamentals of Machine Learning

Web Scraping Workshop 👀

Check ED before writing emails! Post Questions on ED!

CDS Education

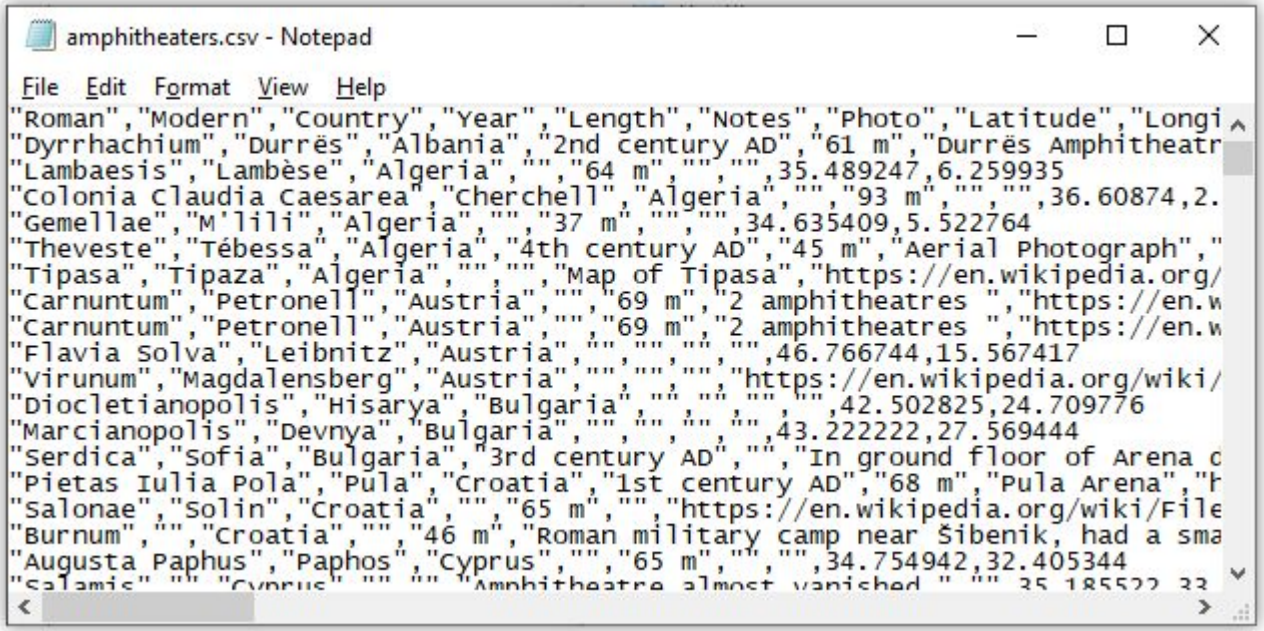
The Data Pipeline



https://towardsdatascience.com/5-steps-of-a-data-science-project-lifecycle-26c50372b492

Why is Data Visualization Important?

We don’t understand this.



https://manifold.net/doc/mfd9/images/eg_formats_csv01_01.png

Why is Data Visualization Important?

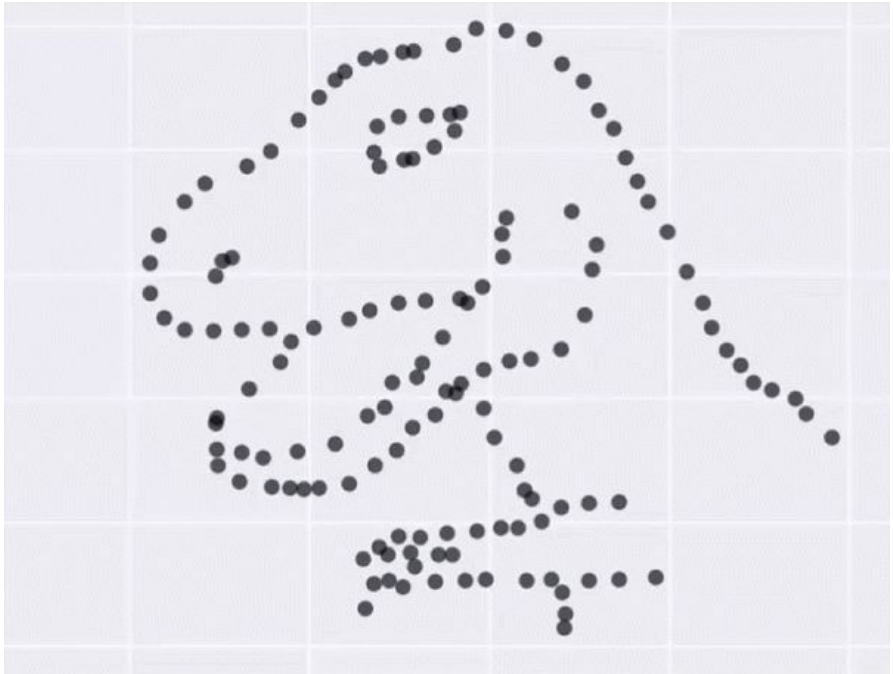
- Informative
- Appealing
- Universal
- Predictive

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

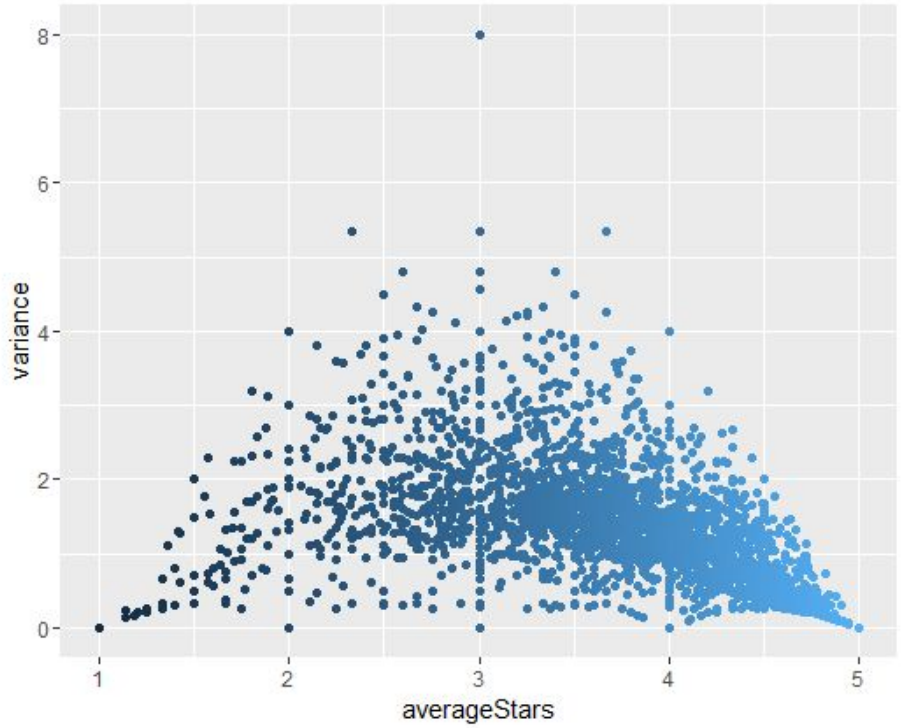


Source

Data Visualization Simple Example: *Ratings on Yelp*

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

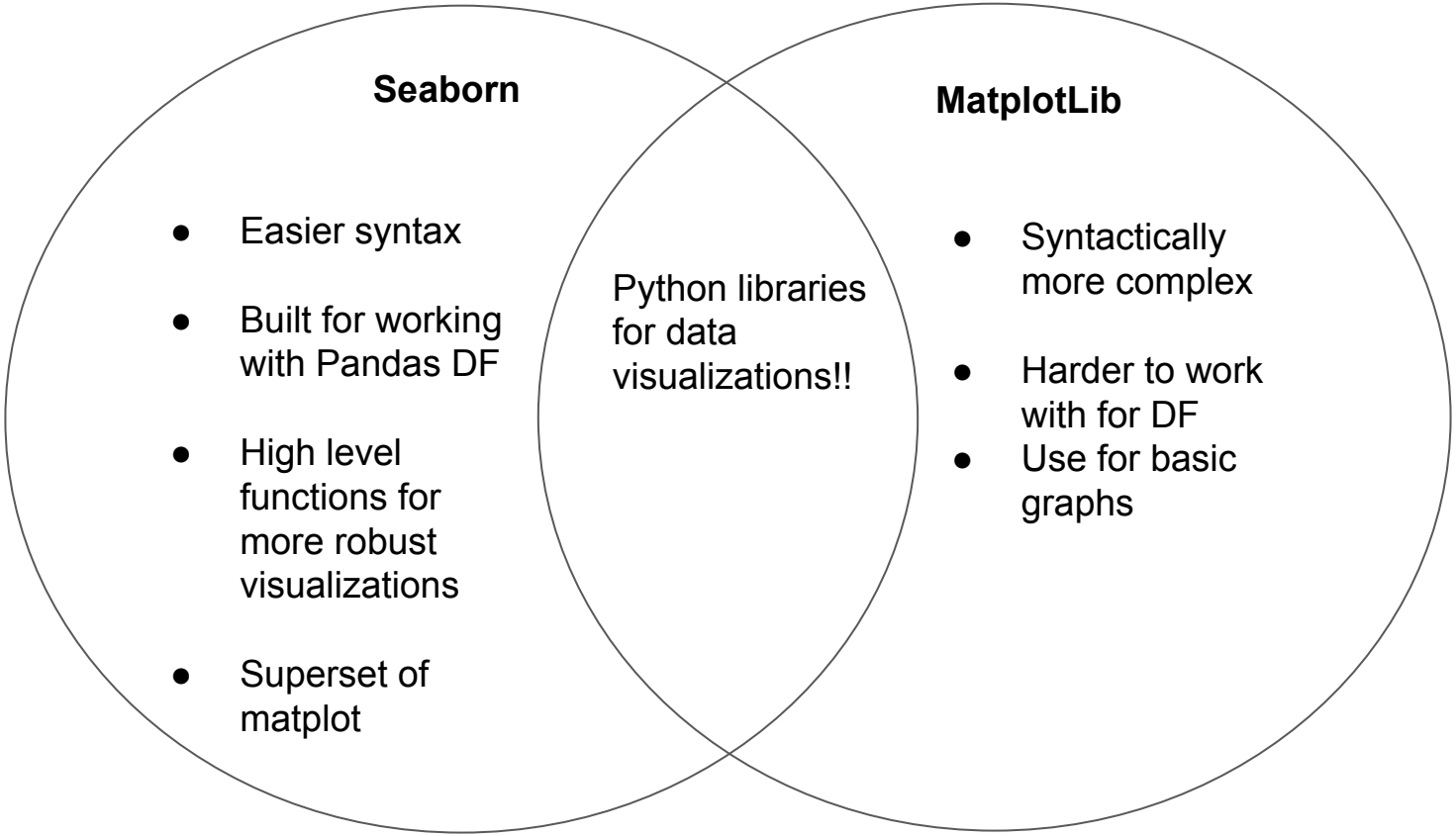
Question: What do you notice? What trends do you see?



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

Seaborn vs MatplotLib



Lecture 3: Data Visualization

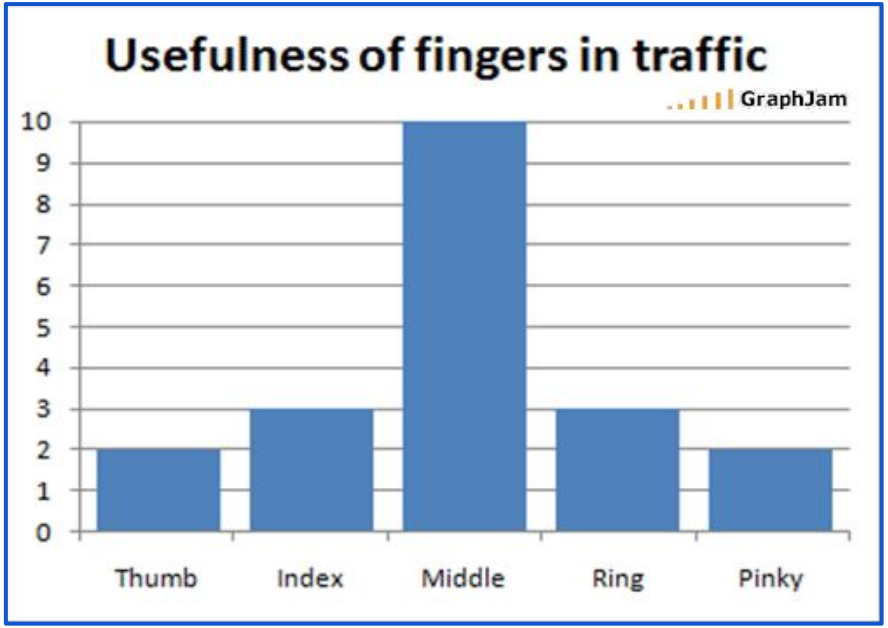
INFO 1998: Introduction to Machine Learning

Attendance Form!

Data Visualizations

Bar Graph

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



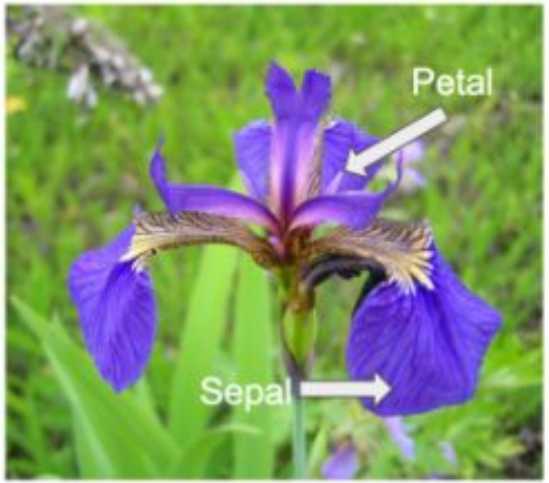
Source

Histograms

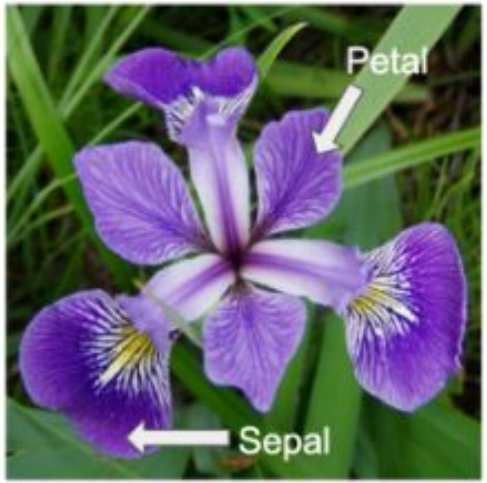
Source

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

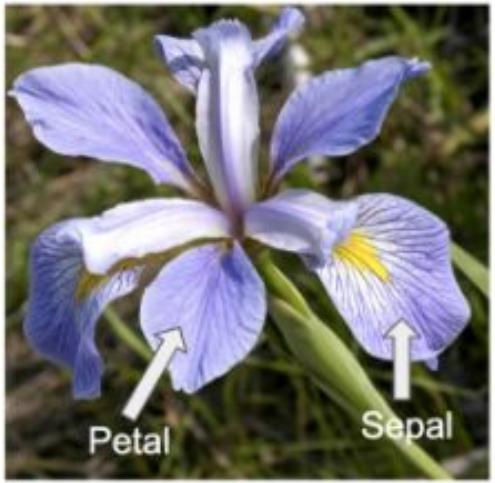
Iris setosa



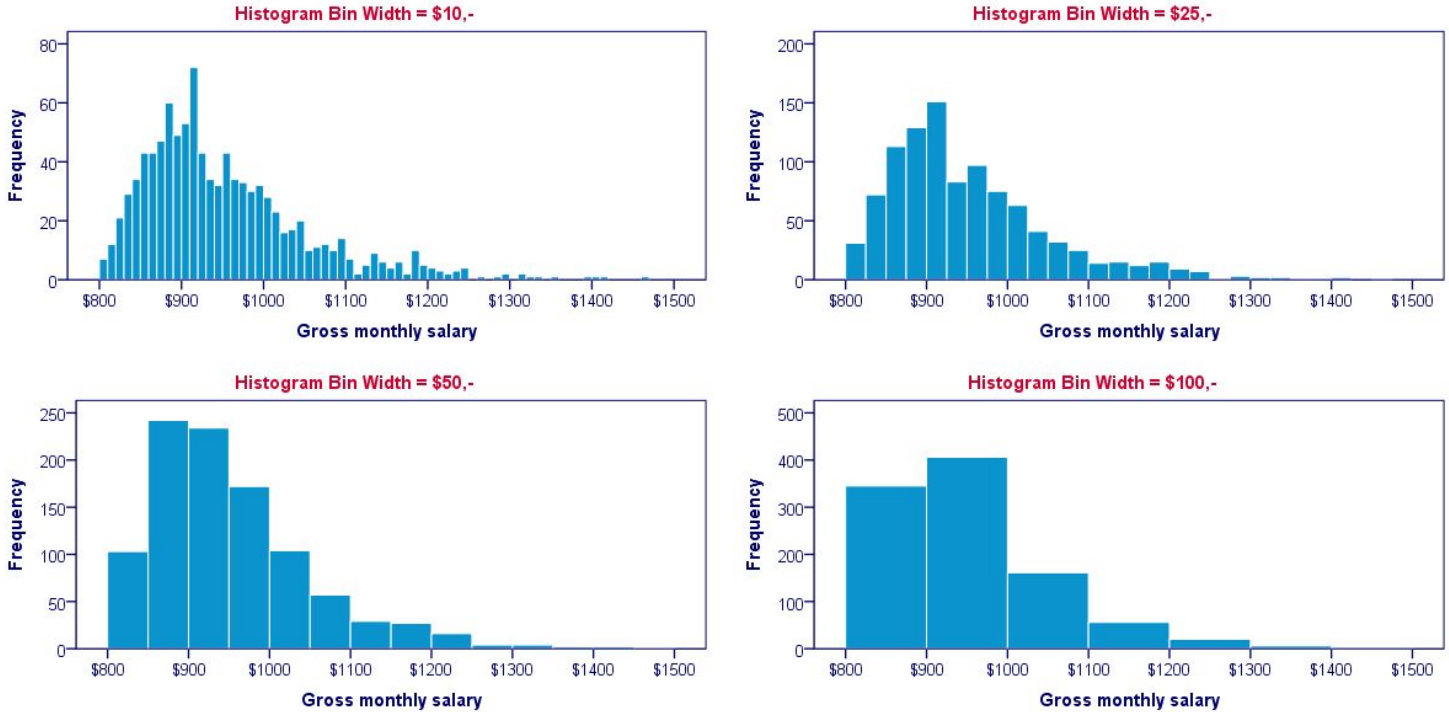
Iris versicolor



Iris virginica

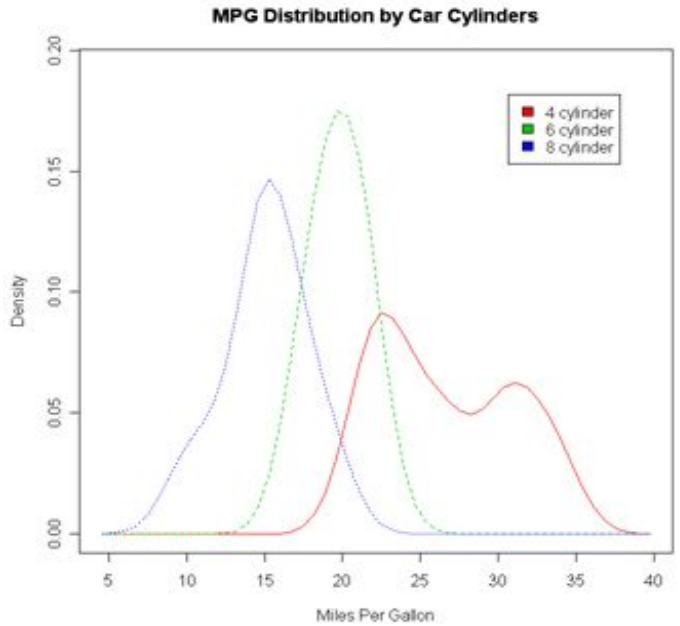


Histograms: Different Bin Sizes



Source

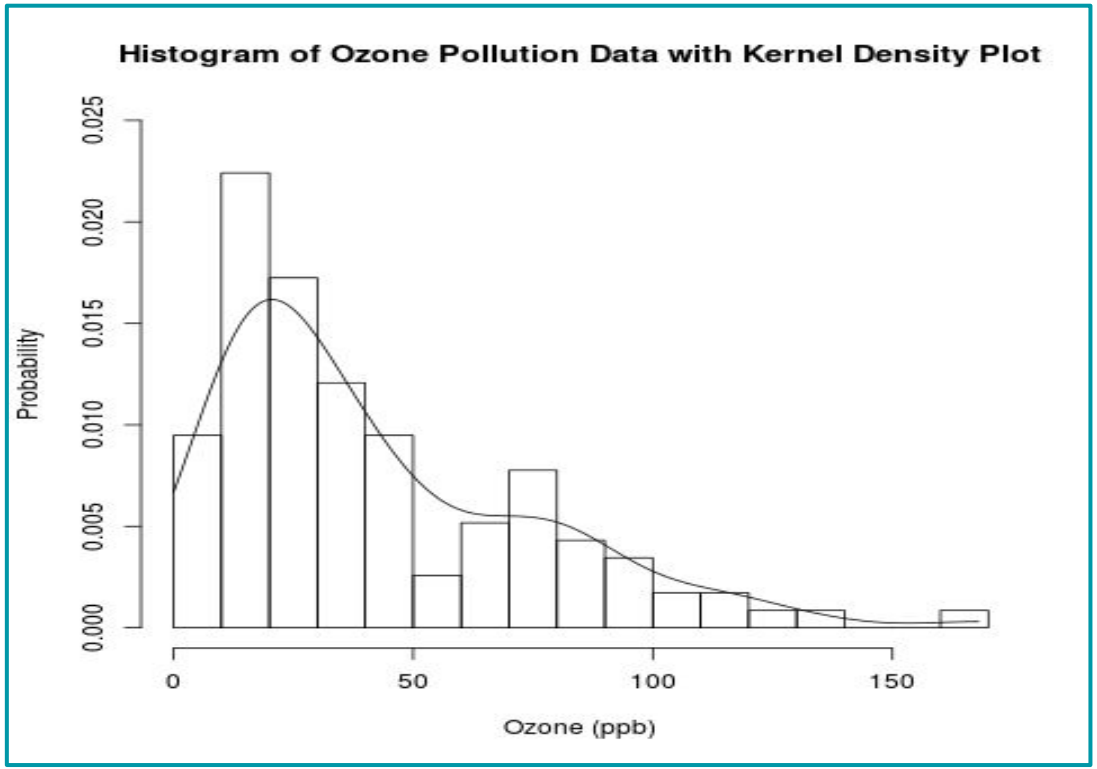
Density Plot



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

Source

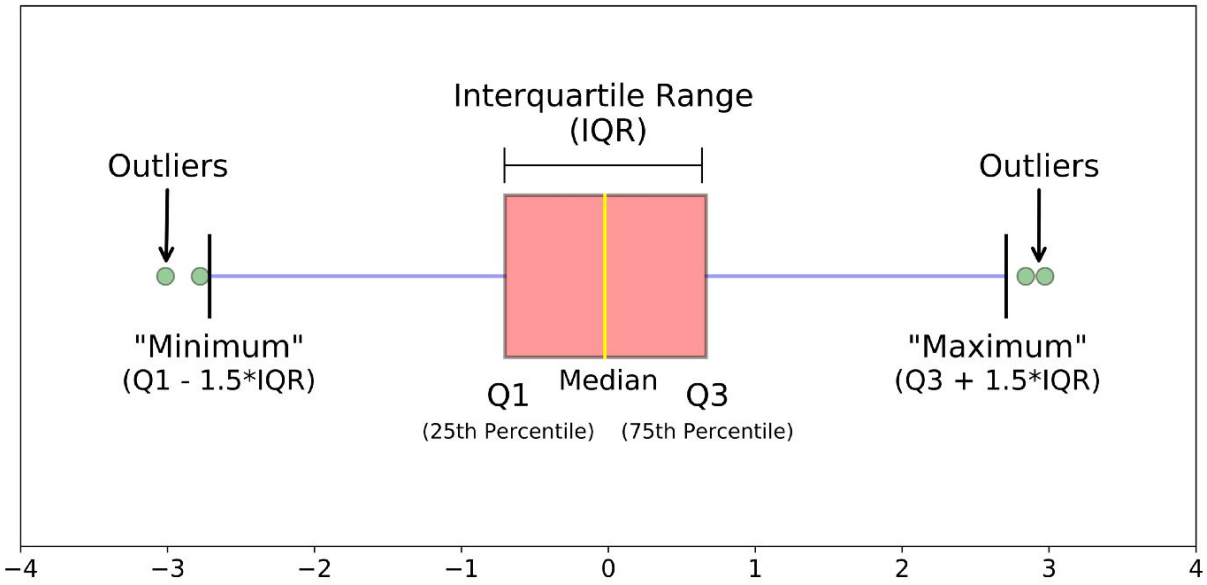
Histogram vs Density Plot



Source

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

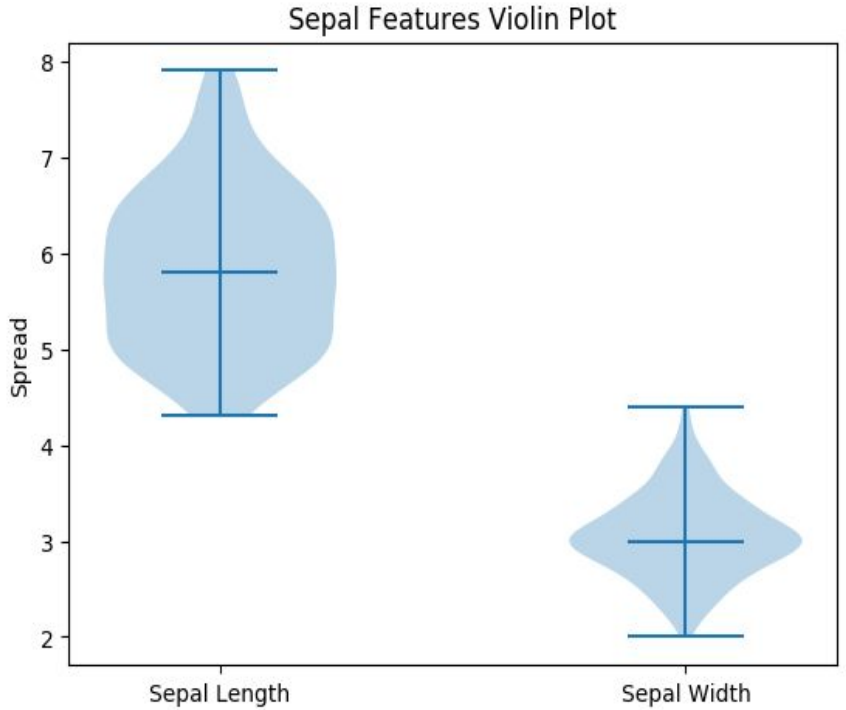
- Summary of data
- Shows **spread** of data
- Gives range, interquartile range, median, and outlier information



Source

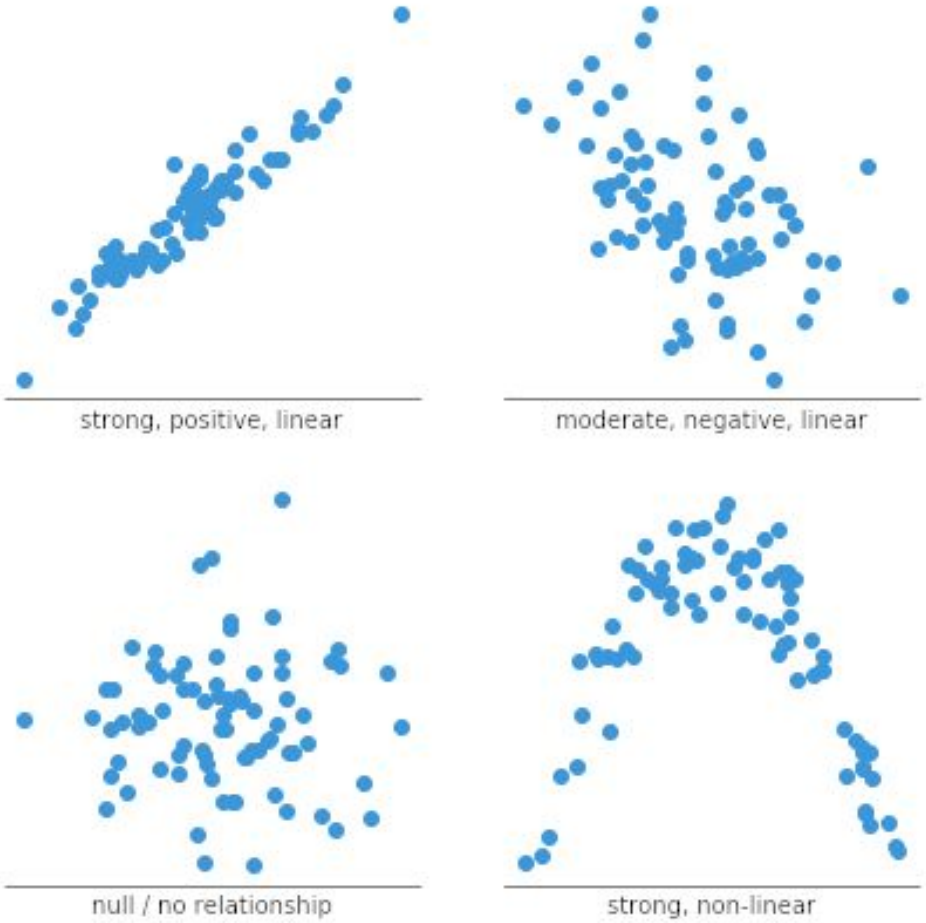
Violin Plot

- 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)



Scatterplot

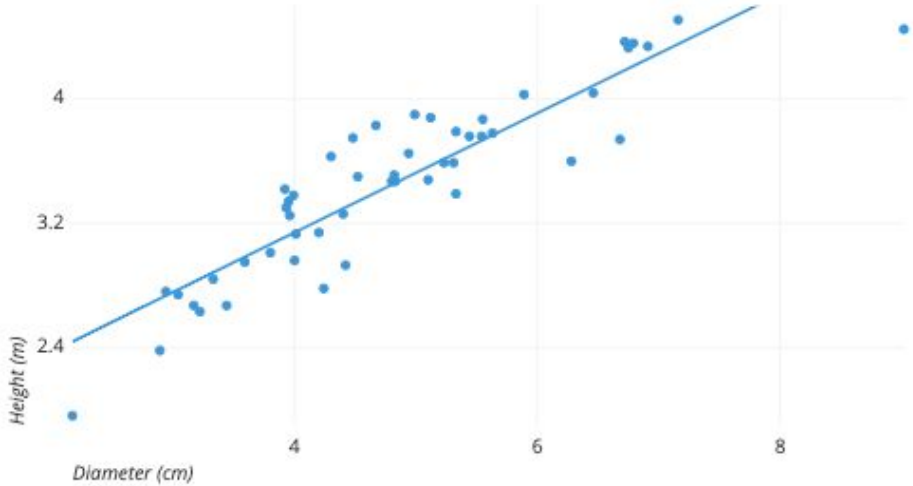
- See **relationship** between two features
- Can be useful for **extrapolating** information

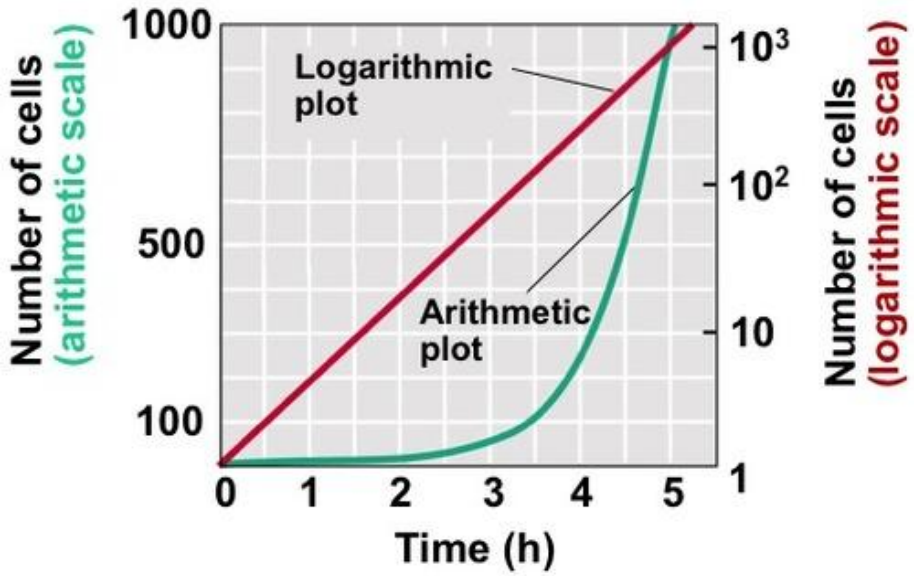


https://chartio.com/learn/charts/what-is-a-scatter-plot/

More Scatterplots!

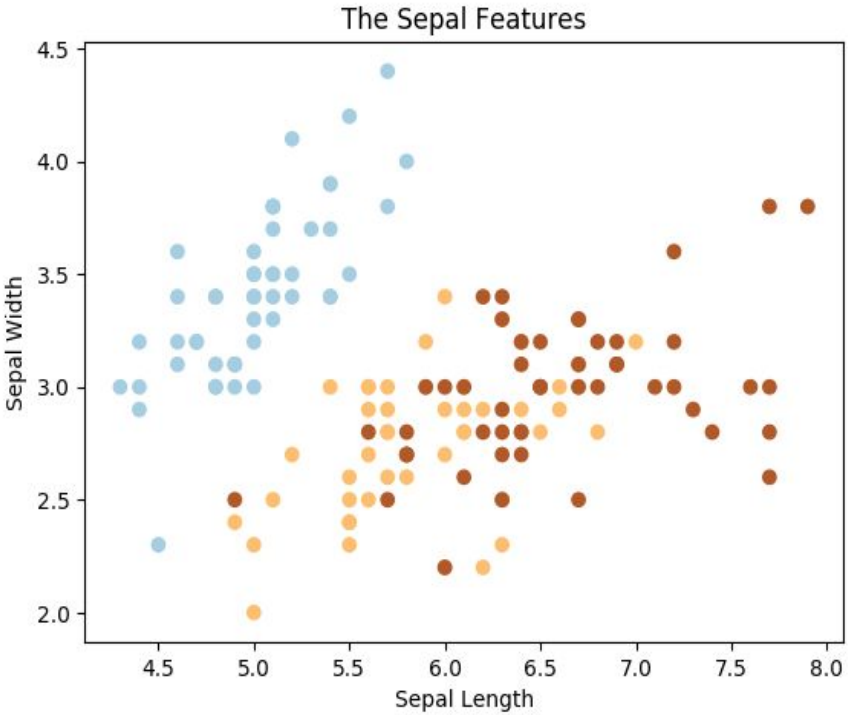
- Line of best fit

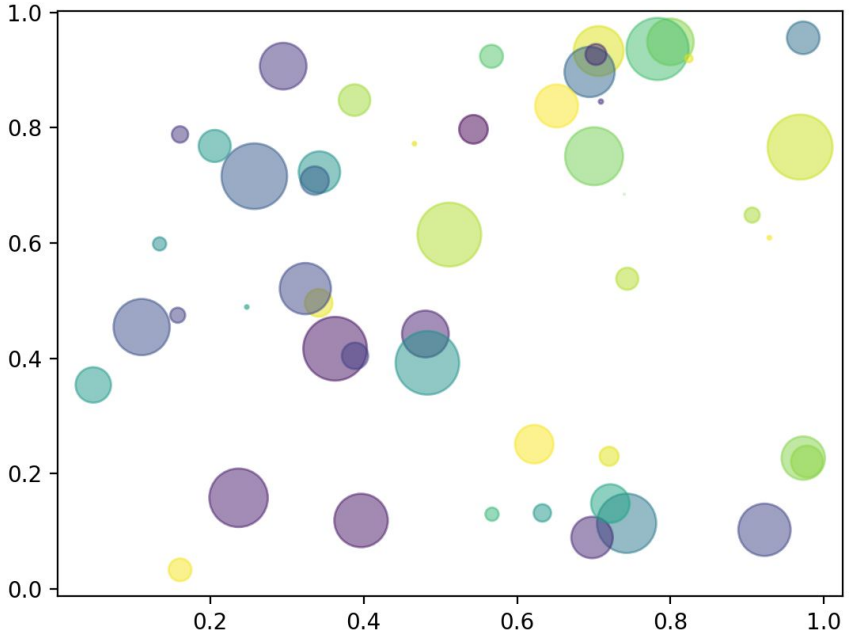




More Scatterplots!

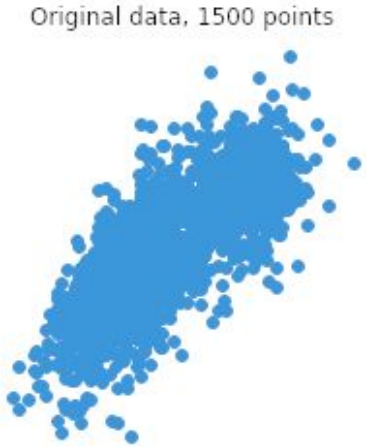
- Line of best fit
- Demonstrate clusters
- Bubble chart

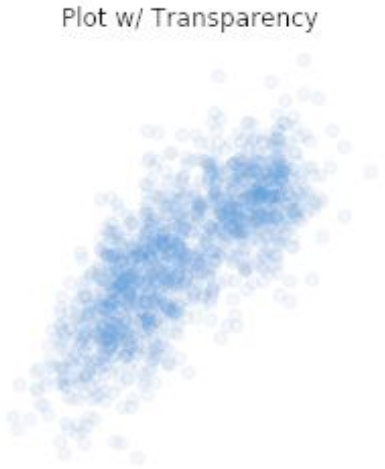


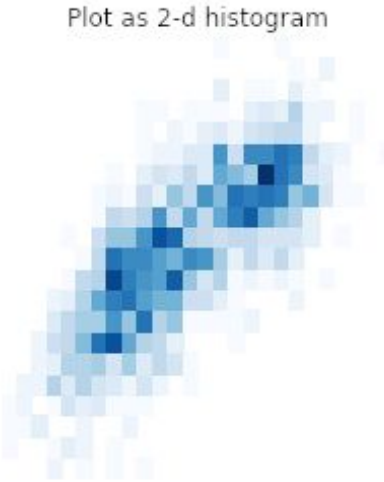


Scatterplot - Overplotting

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

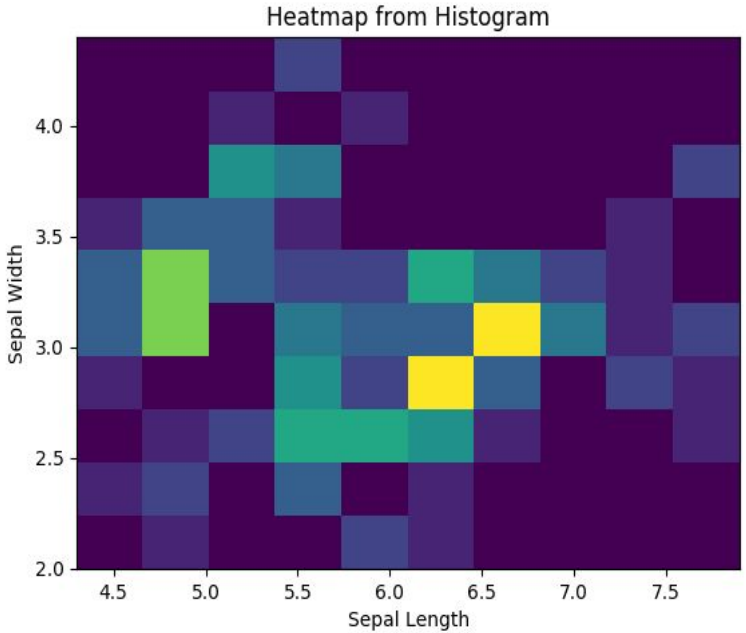






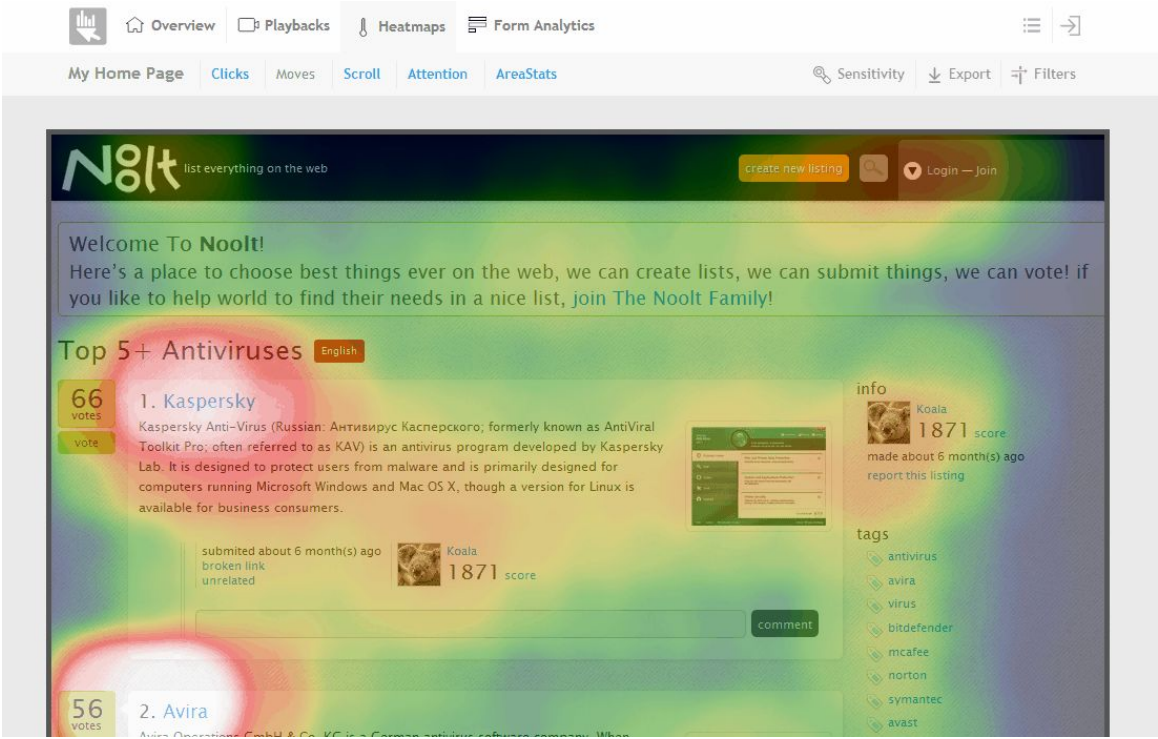
https://chartio.com/learn/charts/what-is-a-scatter-plot/

Heatmap



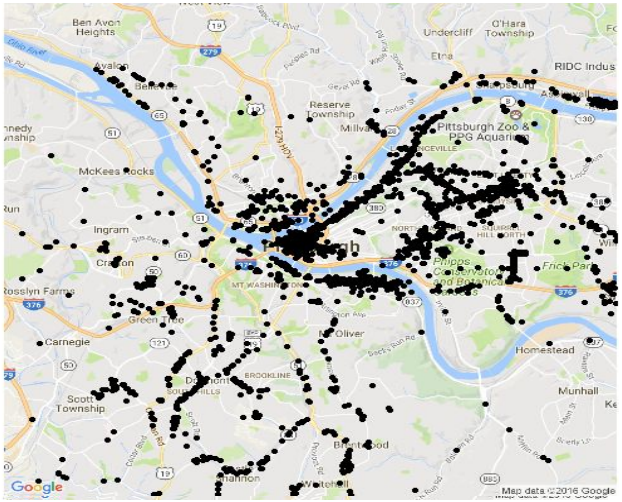
- Varying degrees of one metric are represented using **color**
- Especially useful in the context of **maps** to show geographical variation

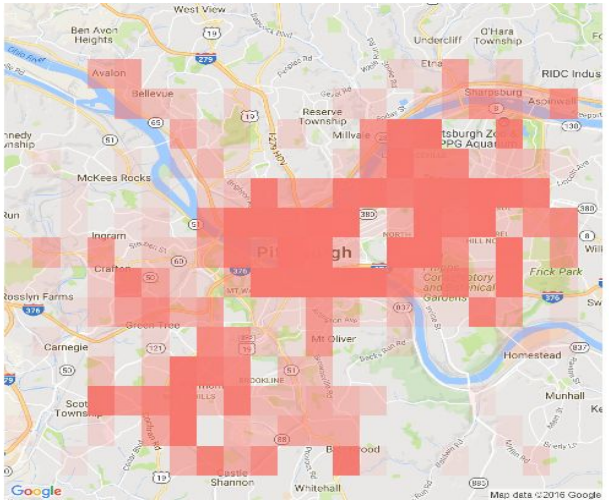
Heatmap - Click Density / Website Heatmaps



Using Maps

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



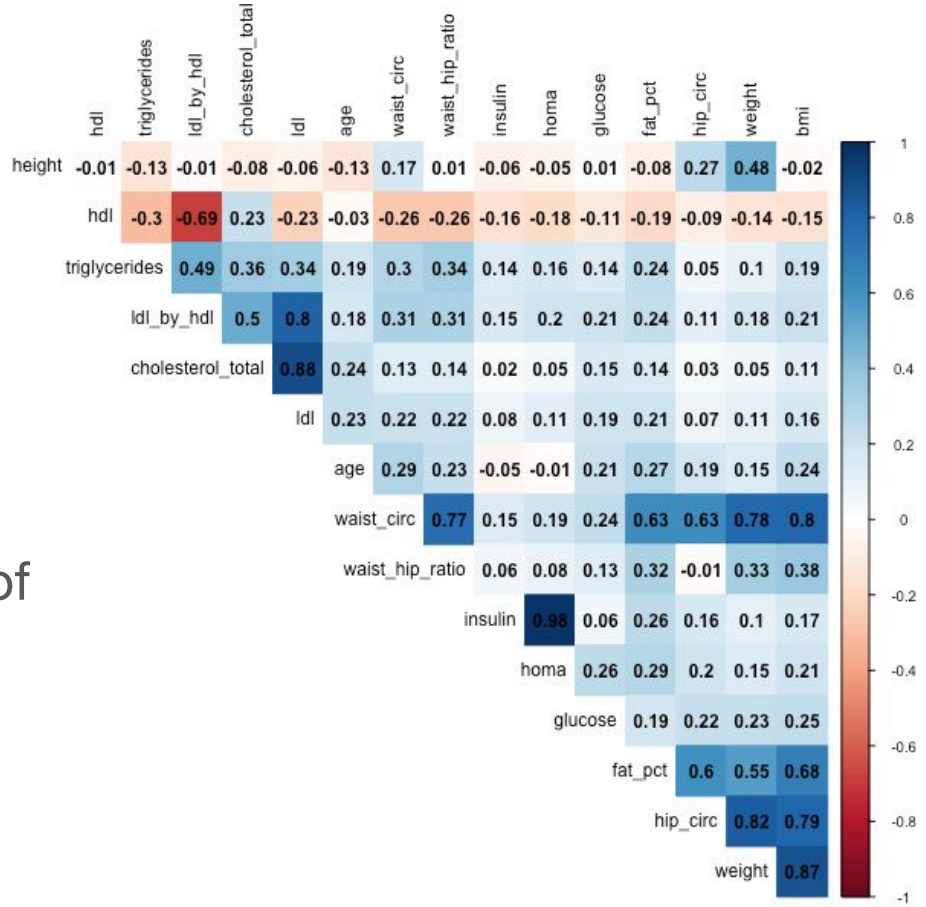


Correlation Plots

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

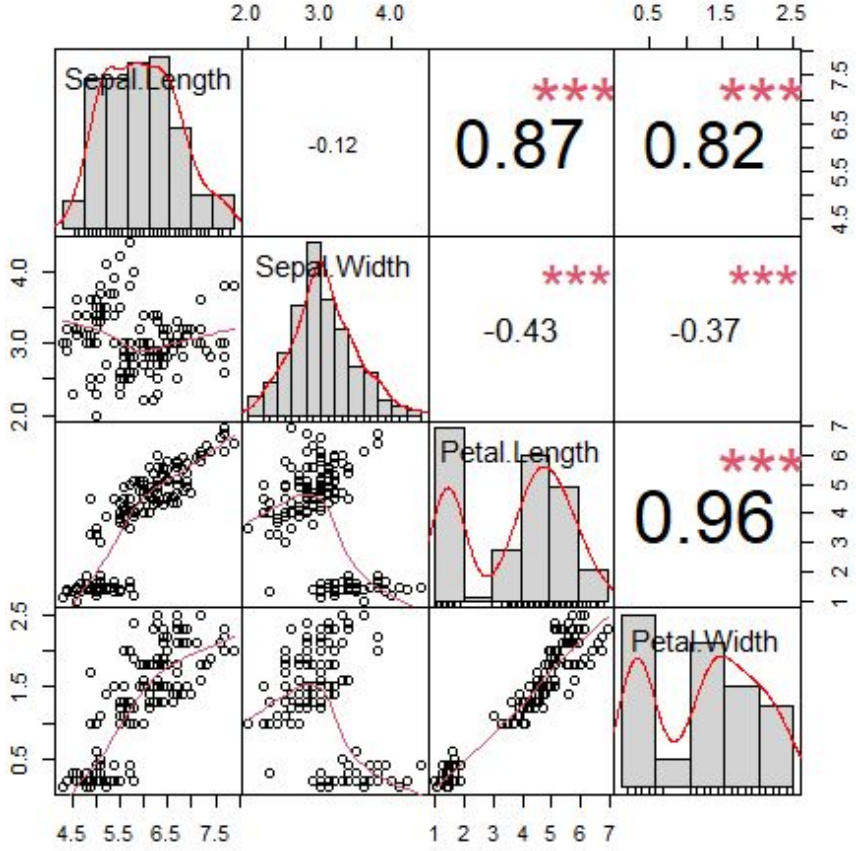
```
[[ 1.          -0.10936925  0.87175416  0.81795363]
 [-0.10936925  1.          -0.4205161  -0.35654409]
 [ 0.87175416 -0.4205161   1.           0.9627571 ]
 [ 0.81795363 -0.35654409  0.9627571    1.        ]]
```

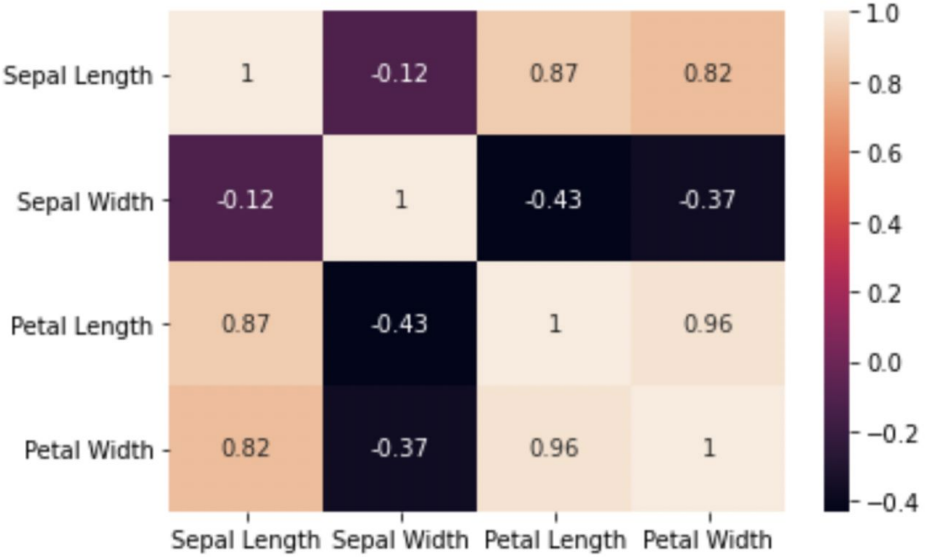
Why are all entries on the diagonal '1'?



Source

Correlation Plots





https://r-coder.com/correlation-plot-r/

Demo

Challenges of Visualization

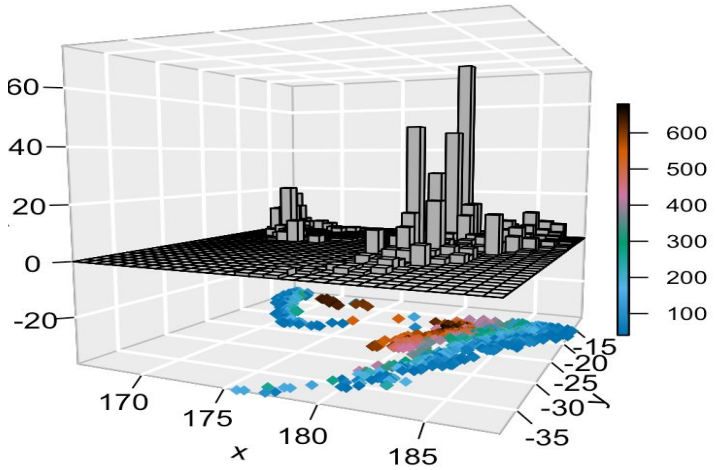
Higher Dimension

Non-Trivial

Time Consuming

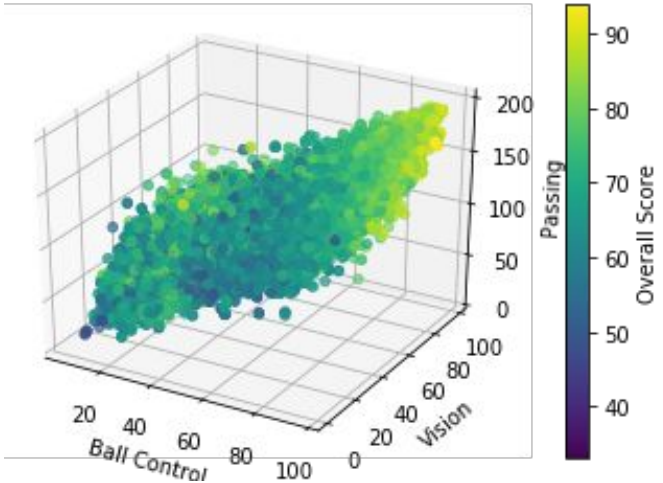
Hard to Show Uncertainty

High Dimensional Data



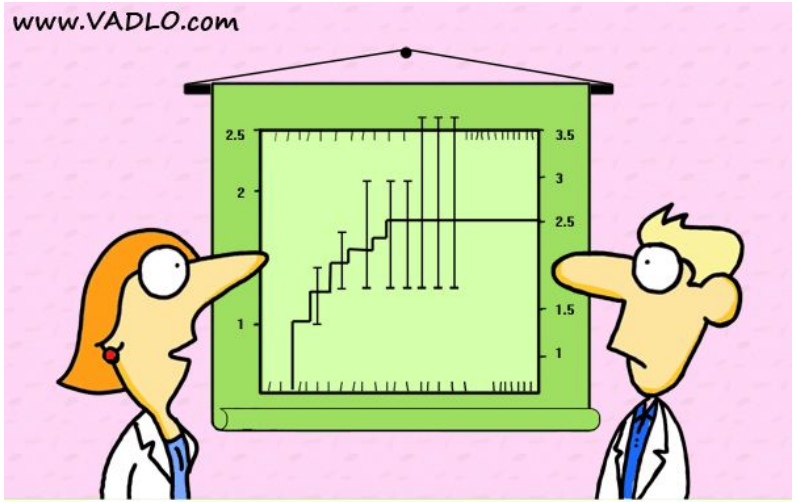
4D Plot For Earthquake Data

- 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

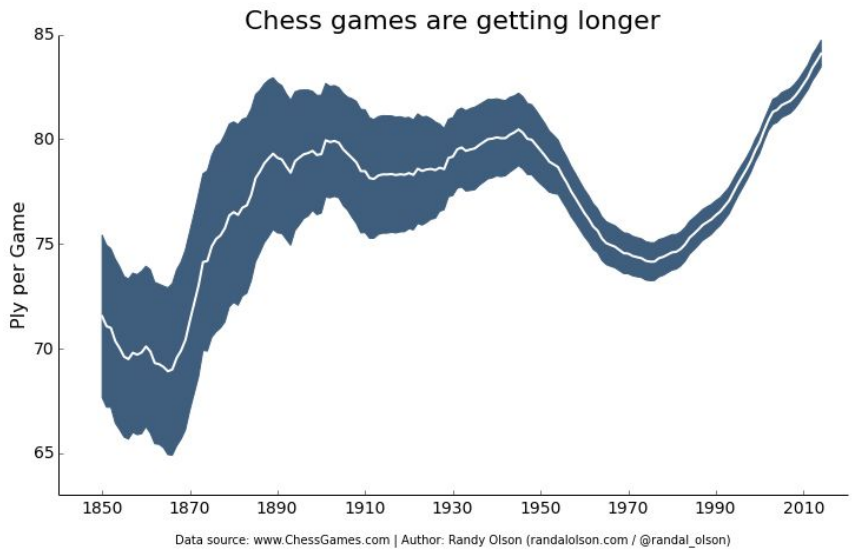


Error Bars

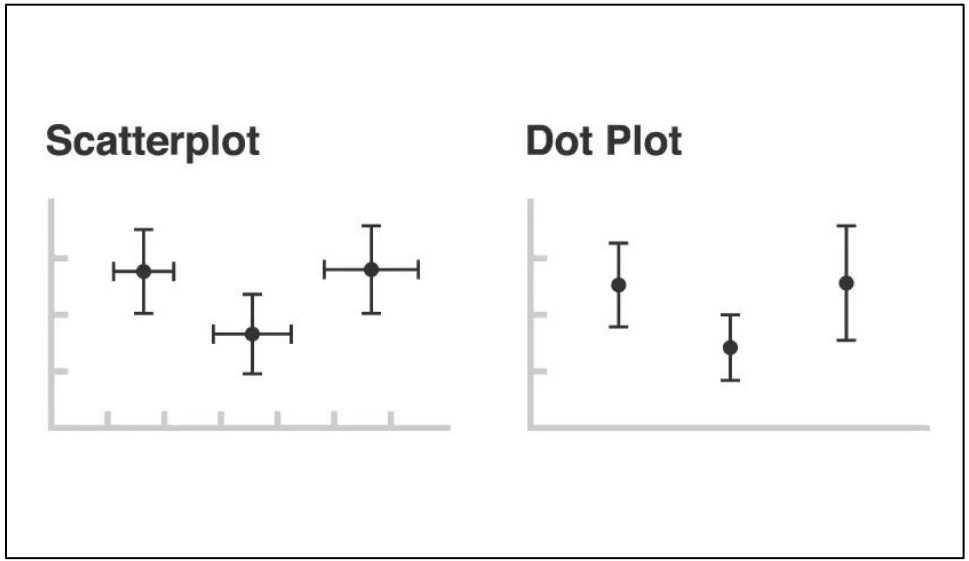
- Show uncertainty
- Usually display 95 percent confidence interval

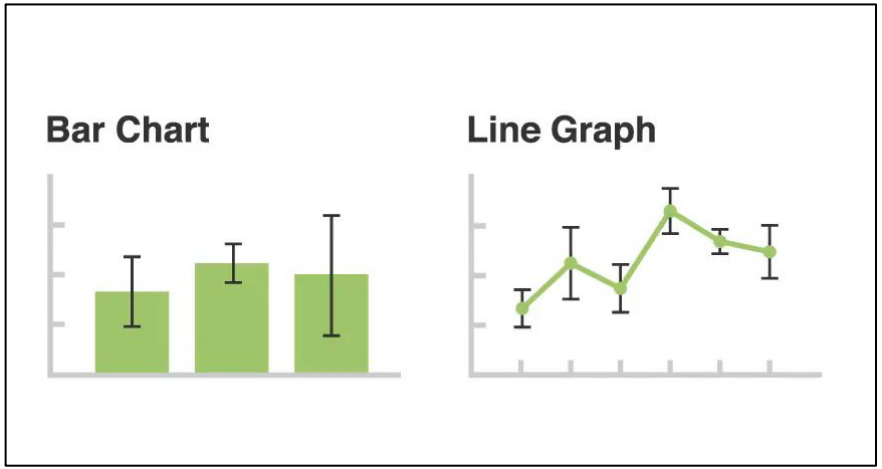


"Did you really have to show the error bars?"



Error Bars

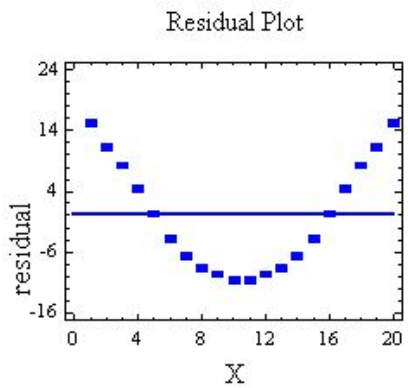




https://www.geeksforgeeks.org/errorbar-graph-in-python-using-matplotlib/

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



Projects!

For your visualizations..

- Choose the proper visualization
- Don’t forget title, axis titles, etc.

1-3 people per project!

- Partner finding on Ed Discussion!

Coming Up

Assignment 3: Due next Monday, 03/03

Next Lecture: Fundamentals of Machine Learning

Web Scraping Workshop 👀

CDS Education