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

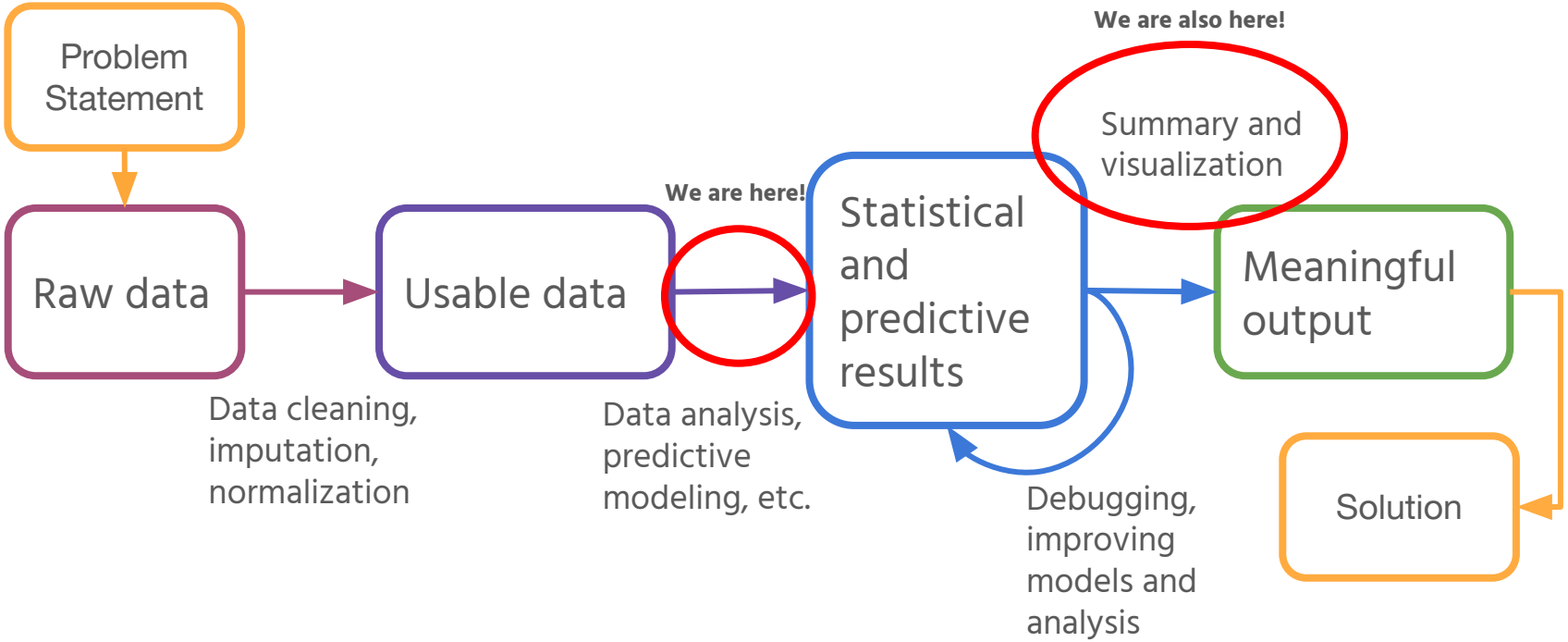
CDS Education

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



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

This!!!

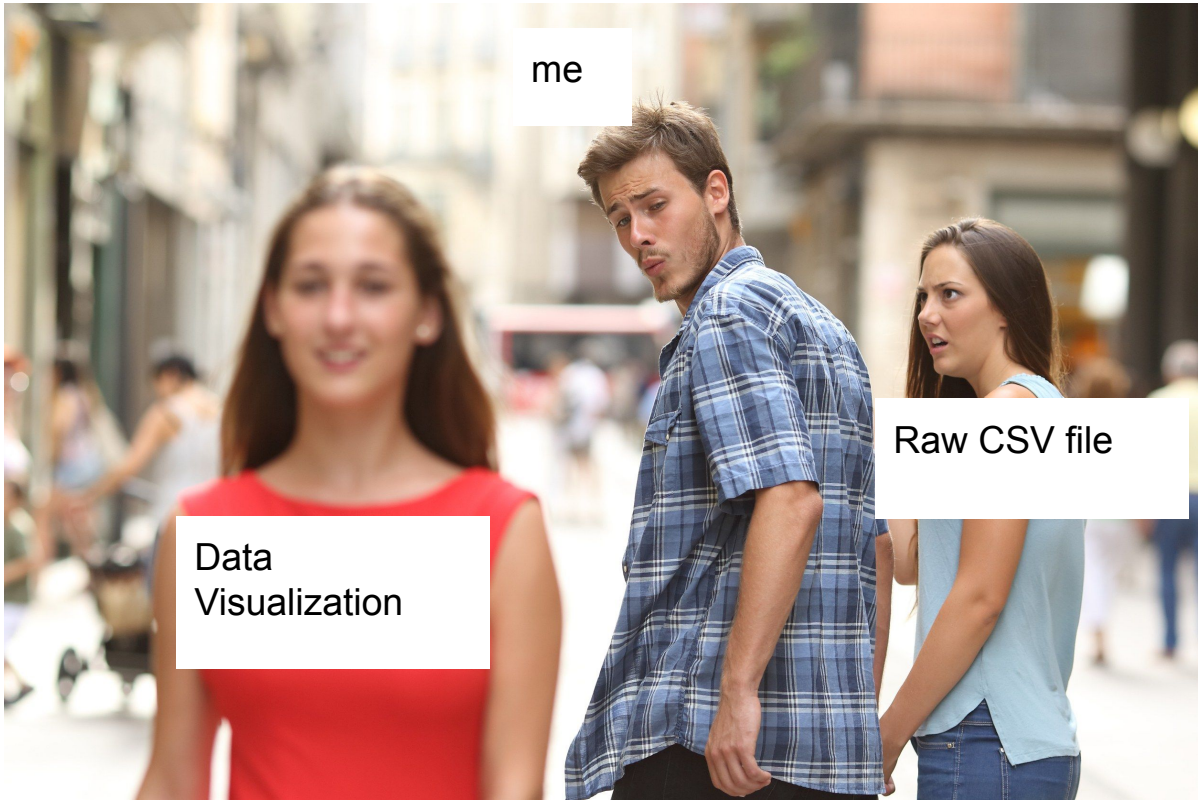
amphitheaters.csv - Notepad

File Edit Format View Help

```
"Roman","Modern","Country","Year","Length","Notes","Photo","Latitude","Longi
"Dyrrhachium","Durrës","Albania","2nd century AD","61 m","Durrës Amphitheatr
"Lambaesis","Lambèse","Algeria","","64 m","","",35.489247,6.259935
"Colonia Claudia Caesarea","Cherchell","Algeria","","93 m","","",36.60874,2.
"Gemellae","M'lili","Algeria","","37 m","","",34.635409,5.522764
"Theveste","Tébessa","Algeria","4th century AD","45 m","Aerial Photograph","
"Tipasa","Tipaza","Algeria","","","Map of Tipasa","https://en.wikipedia.org/
"Carnuntum","Petronell","Austria","","69 m","2 amphitheatres ","https://en.w
"Carnuntum","Petronell","Austria","","69 m","2 amphitheatres ","https://en.w
"Flavia Solva","Leibnitz","Austria","","","","",46.766744,15.567417
"Virunum","Magdalensberg","Austria","","","","","https://en.wikipedia.org/wiki/
"Diocletianopolis","Hisarya","Bulgaria","","","","","",42.502825,24.709776
"Marcianopolis","Devnya","Bulgaria","","","","","",43.222222,27.569444
"Serdica","Sofia","Bulgaria","3rd century AD","","In ground floor of Arena c
"Pietas Iulia Pola","Pula","Croatia","1st century AD","68 m","Pula Arena","h
"Salonae","Solin","Croatia","","65 m","","https://en.wikipedia.org/wiki/File
"Burnum","","Croatia","","46 m","Roman military camp near Šibenik, had a sma
"Augusta Paphus","Paphos","Cyprus","","65 m","","",34.754942,32.405344
"Salamis","","Cyprus","","","Amphitheatre almost vanished ","",35.185522,33
```

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

Why is Data Visualization Important?



Source

Why is Data Visualization Important?

Informative

Appealing

Universal

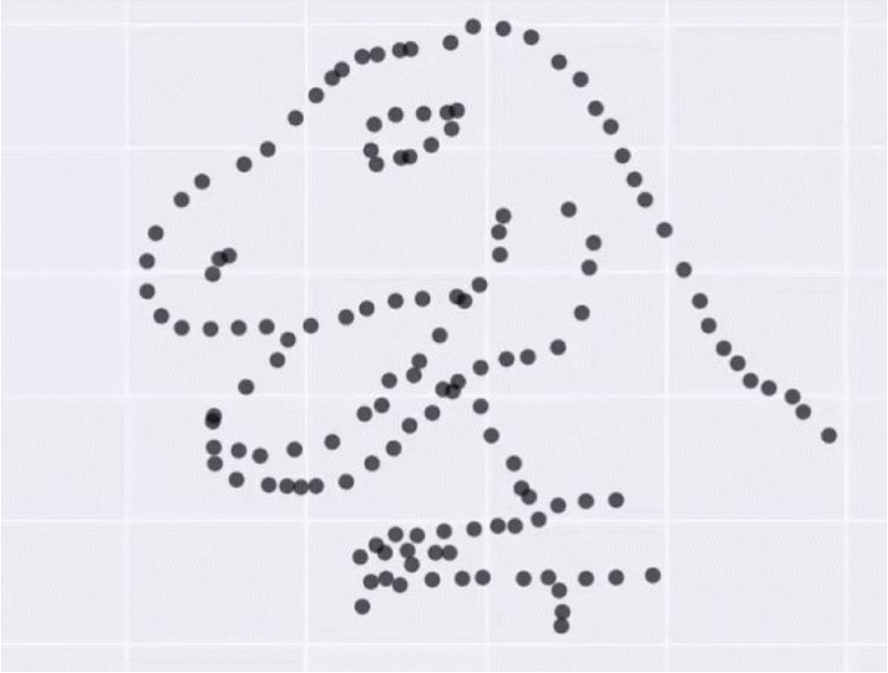
Predictive

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

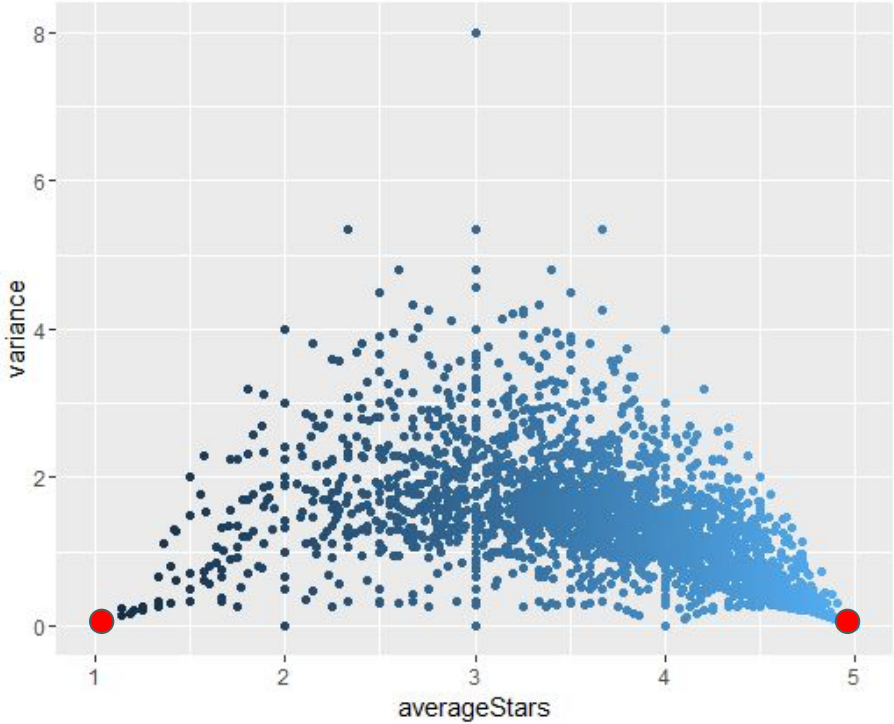


Source

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

Seaborn vs Matplot

Seaborn

- Easier syntax
- Built for working with Pandas DF
- High level functions for more robust visualizations
- Superset of matplot

Python libraries for data visualizations!!

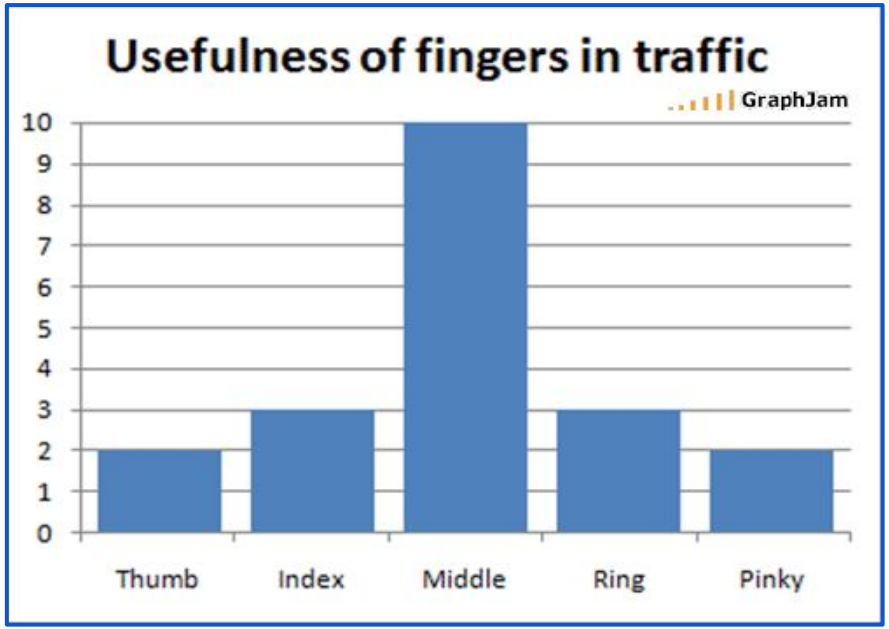
Matplot

- 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



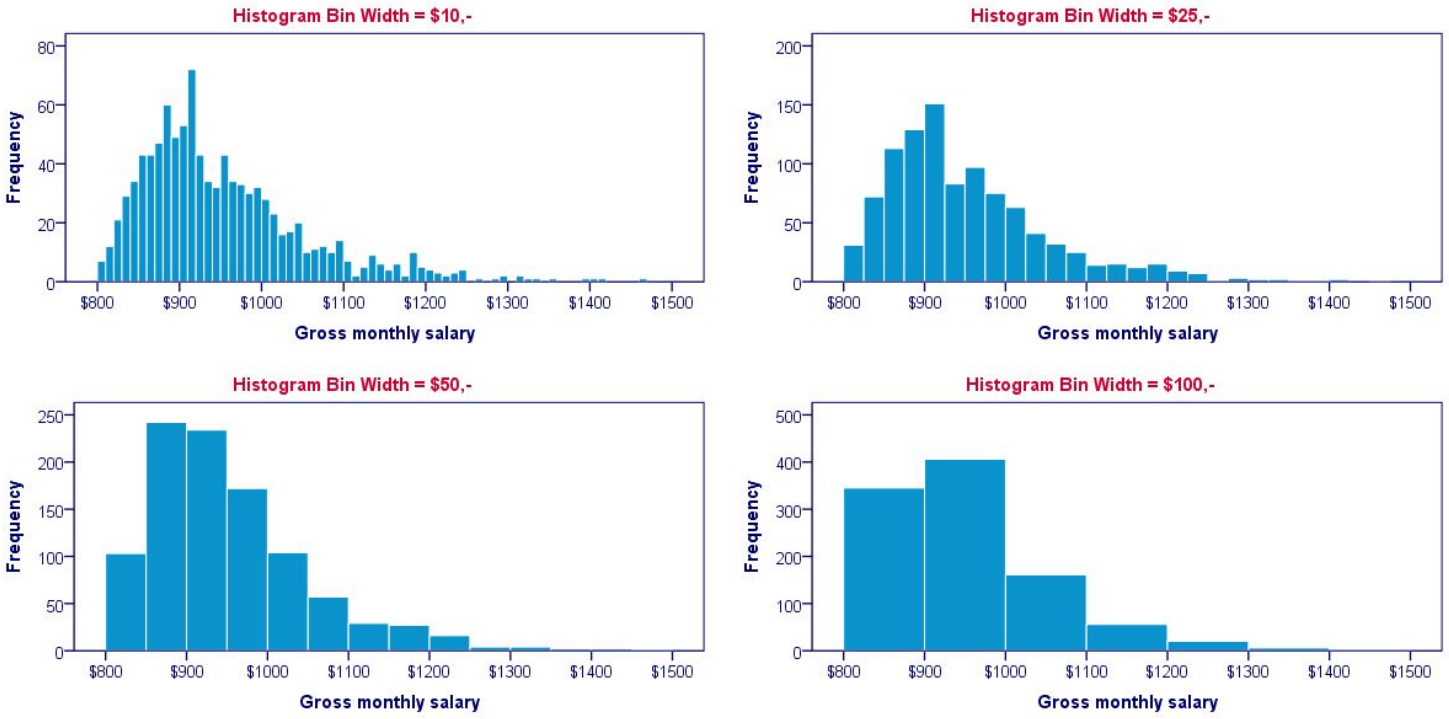
Source

Histograms

Source

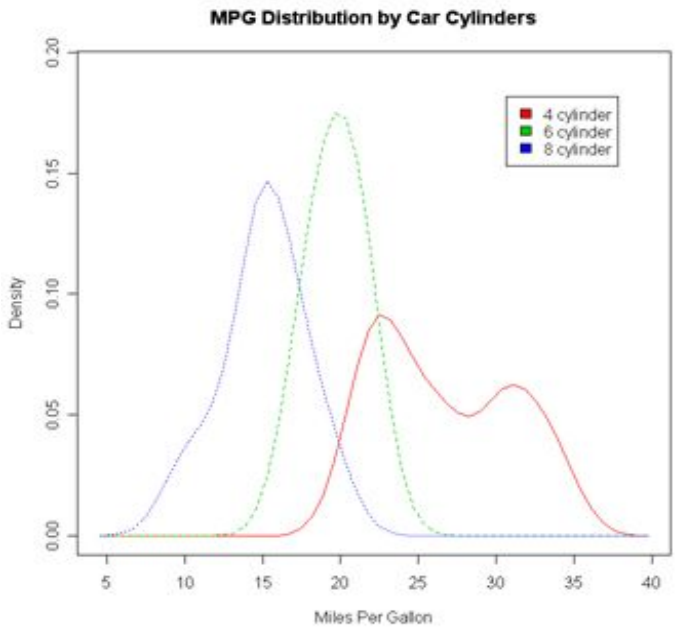
- Used to observe **frequency distribution** of continuous variables
- Data split into **bins**

Histograms: Different Bin Sizes



Source

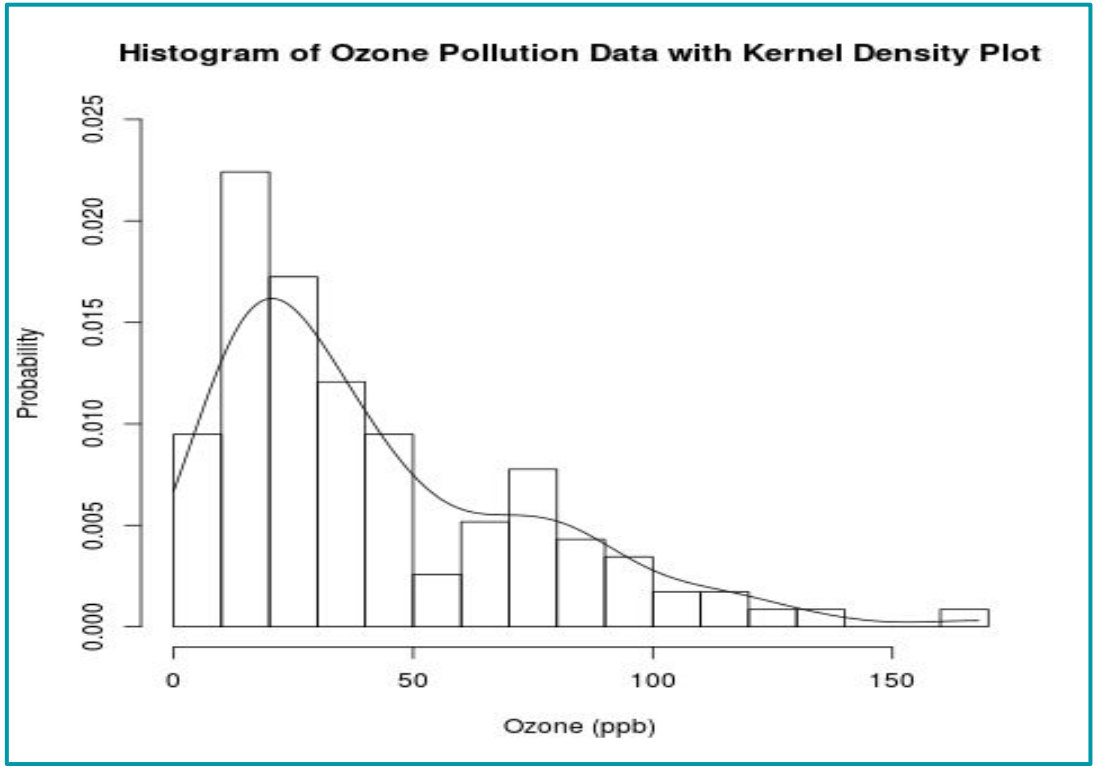
Density Plot



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

Source

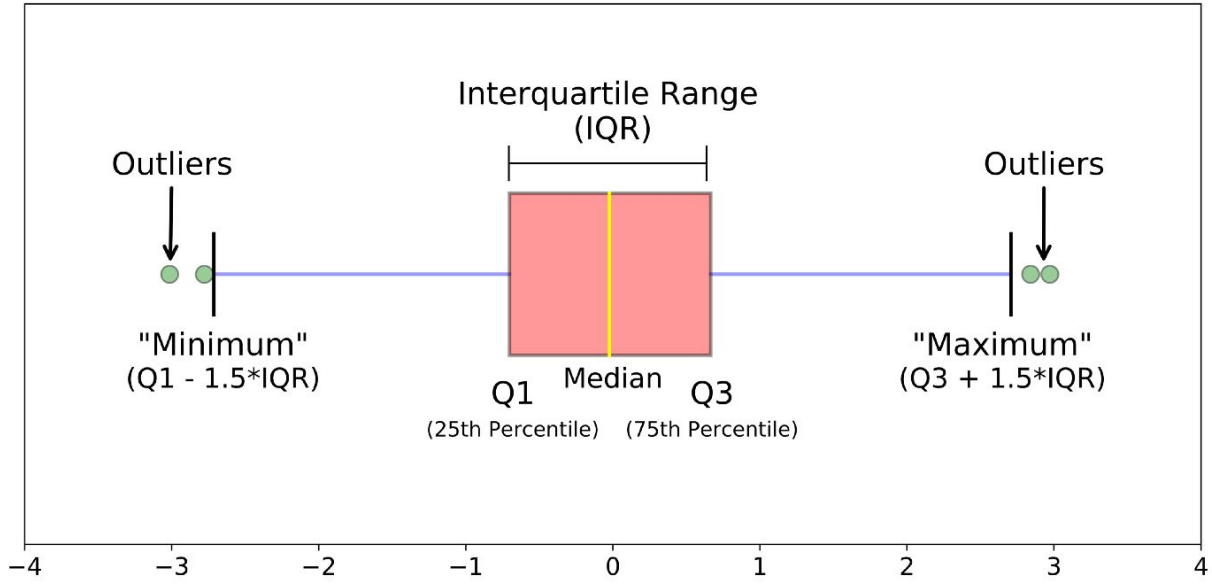
Histogram vs Density Plot



Source

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

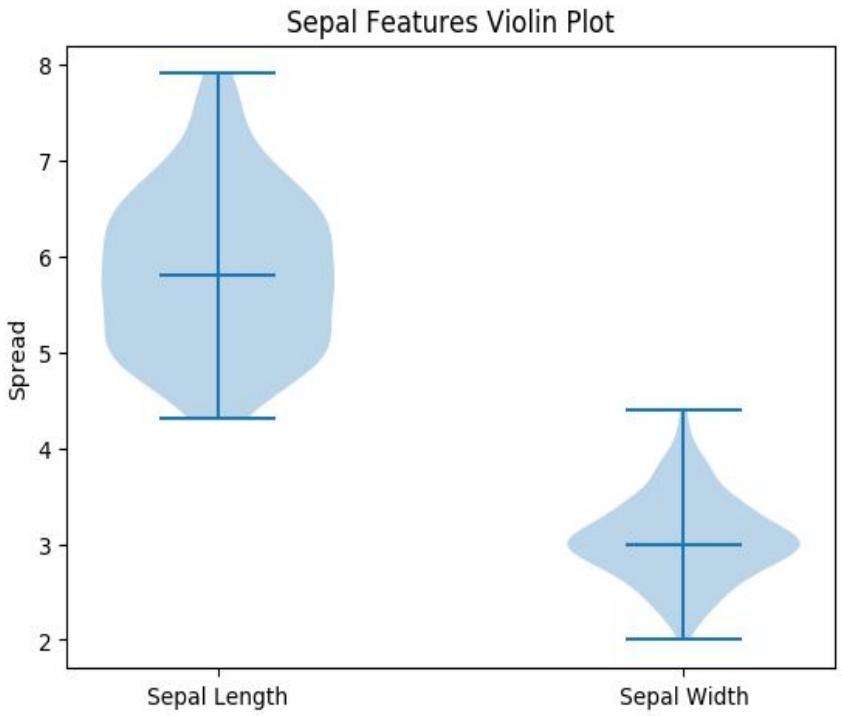
- Summary of data
- Shows **spread** of data
- Gives range, interquartile range, median, and outlier information



Source

Violin Plot

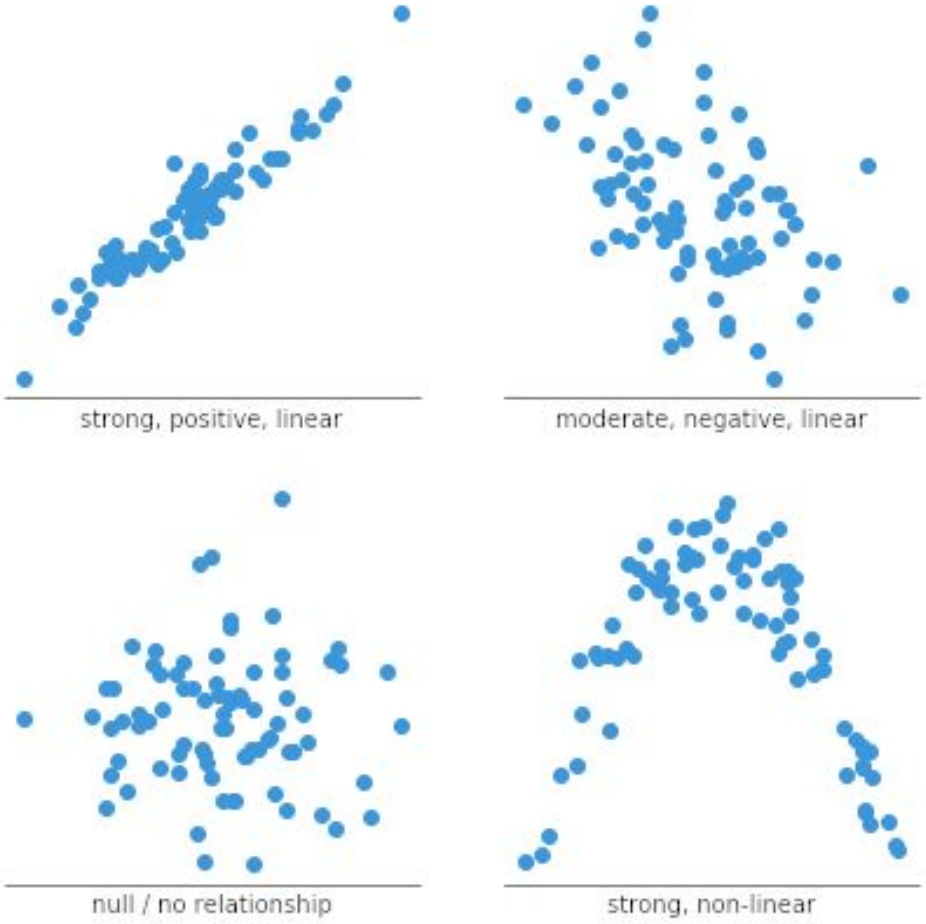
- Combination of **boxplot** and **density plot** to show the **spread** and **shape** of the data
- Can show whether the data is **normal** (i.e. is distributed normally)



Advanced Data Visualizations

Scatterplot

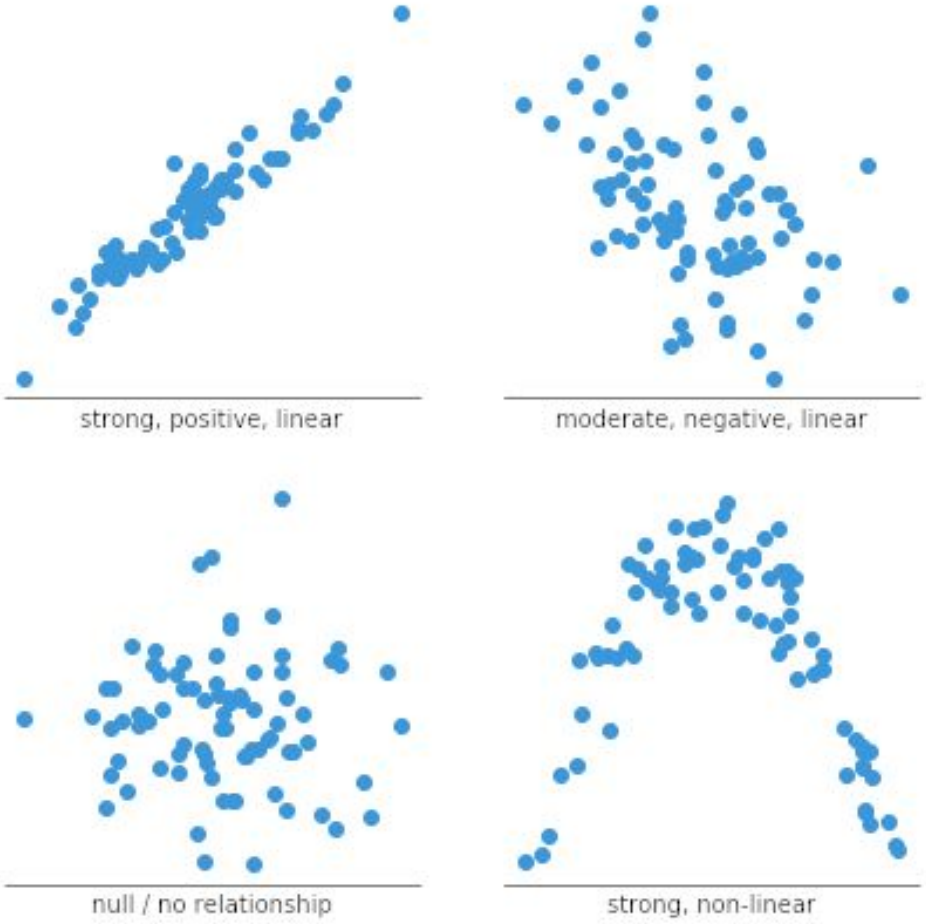
- See **relationship** between two features
- Can be useful for **extrapolating** information



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

Scatterplot

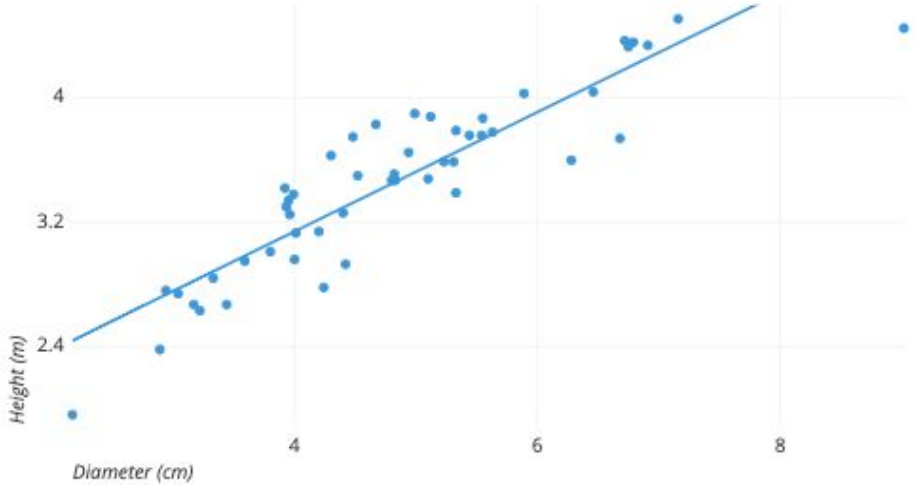
- See **relationship** between two features
- Can be useful for **extrapolating** information
- **Correlation ≠ Causation!**

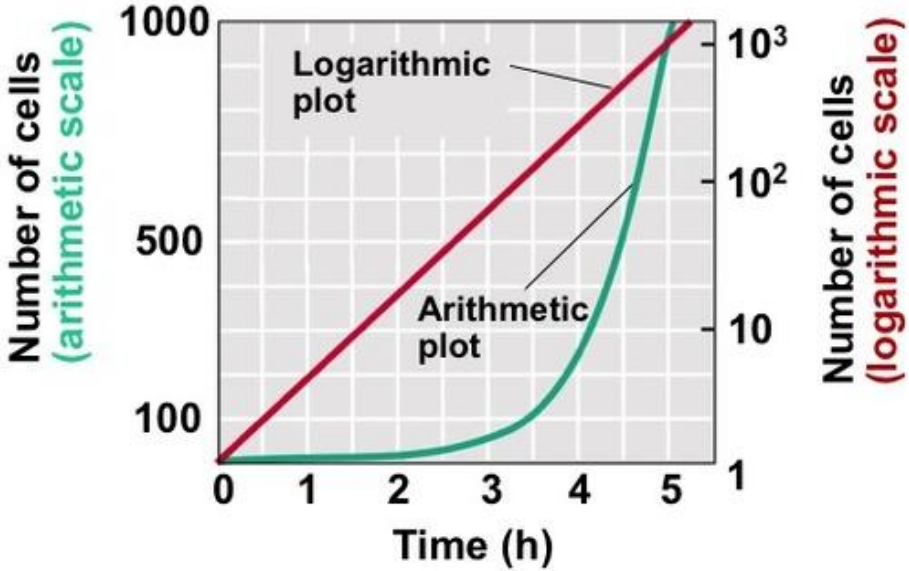


https://chartio.com/learn/charts/what-is-a-scatter-plot/

Scatterplot – more ways

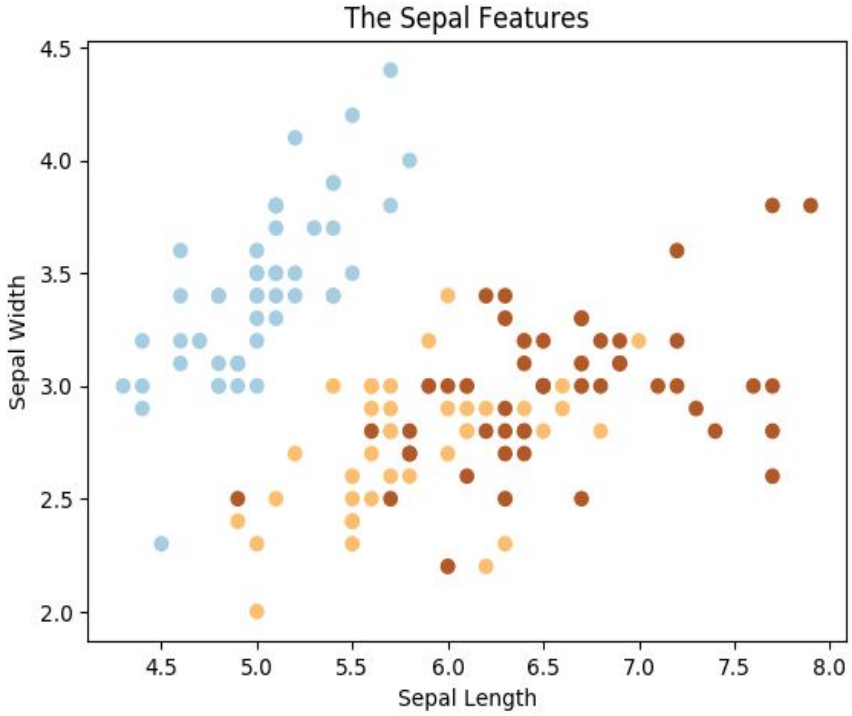
- Line of best fit

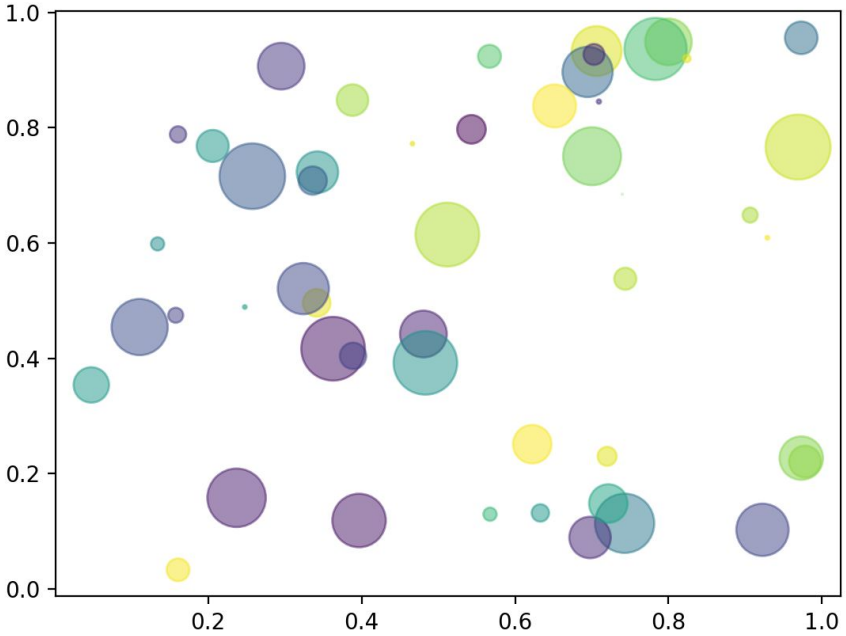




Scatterplot – more ways

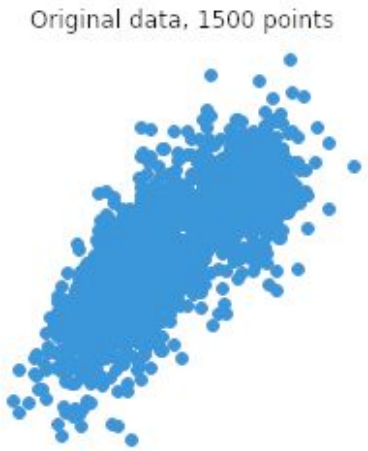
- Line of best fit
- Demonstrate clusters
- Bubble chart

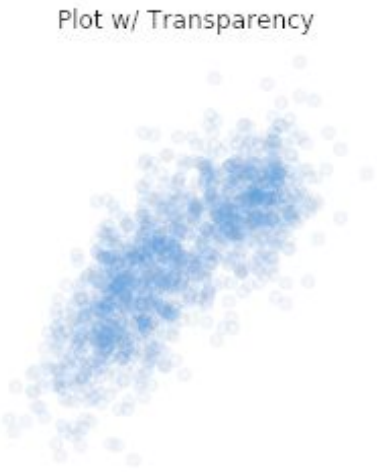


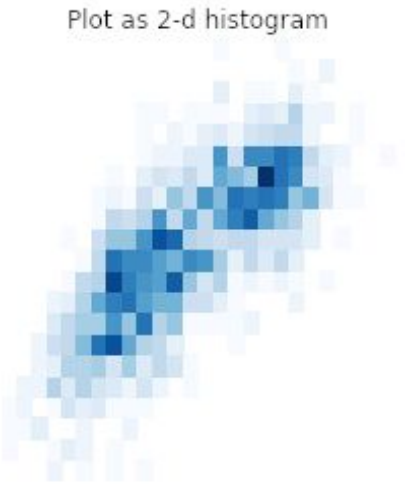


Scatterplot - Overplotting

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

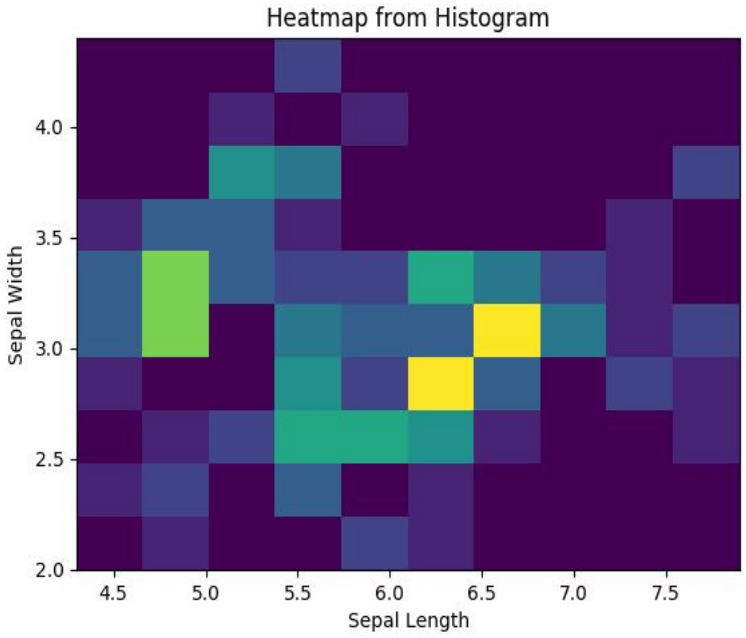






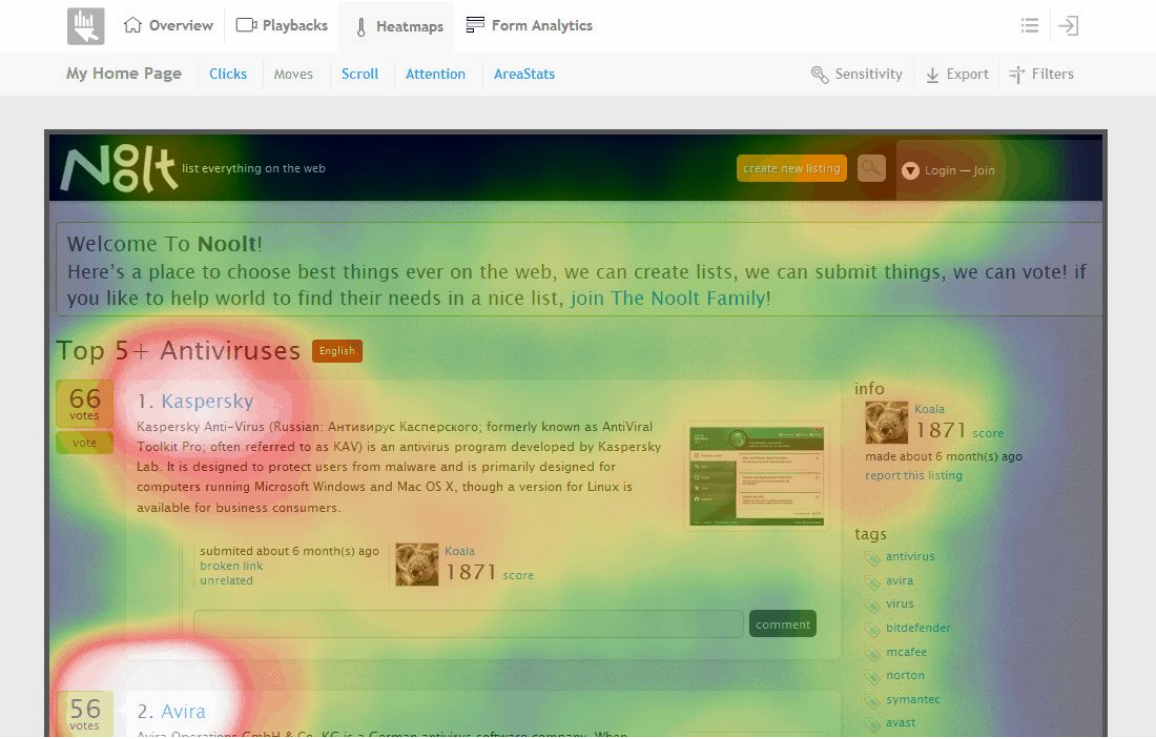
https://chartio.com/learn/charts/what-is-a-scatter-plot/

Heatmap



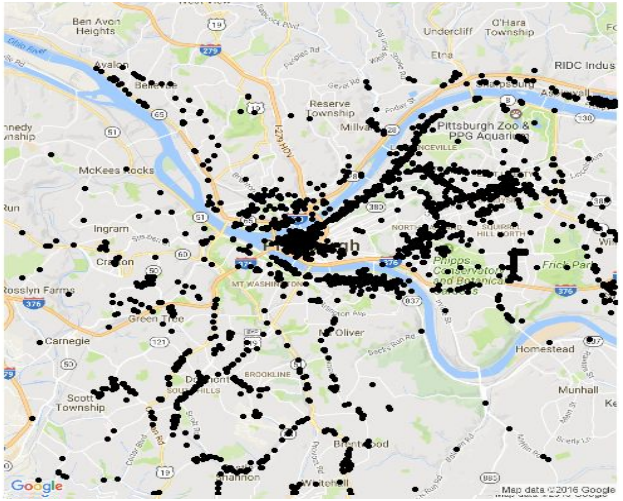
- Varying degrees of one metric are represented using **color**
- Especially useful in the context of **maps** to show geographical variation

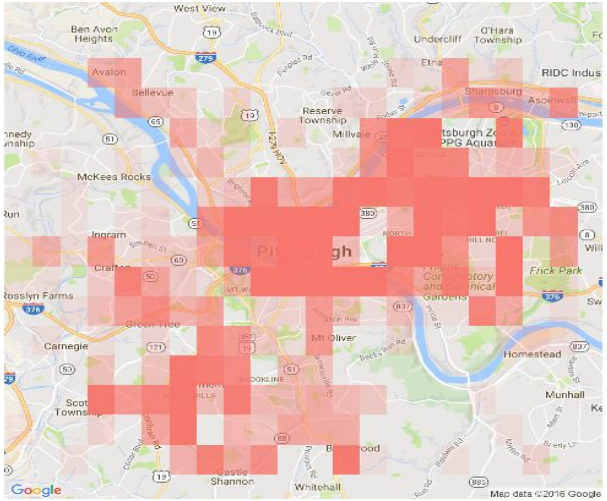
Heatmap - Click Density / Website Heatmaps



Using Maps

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



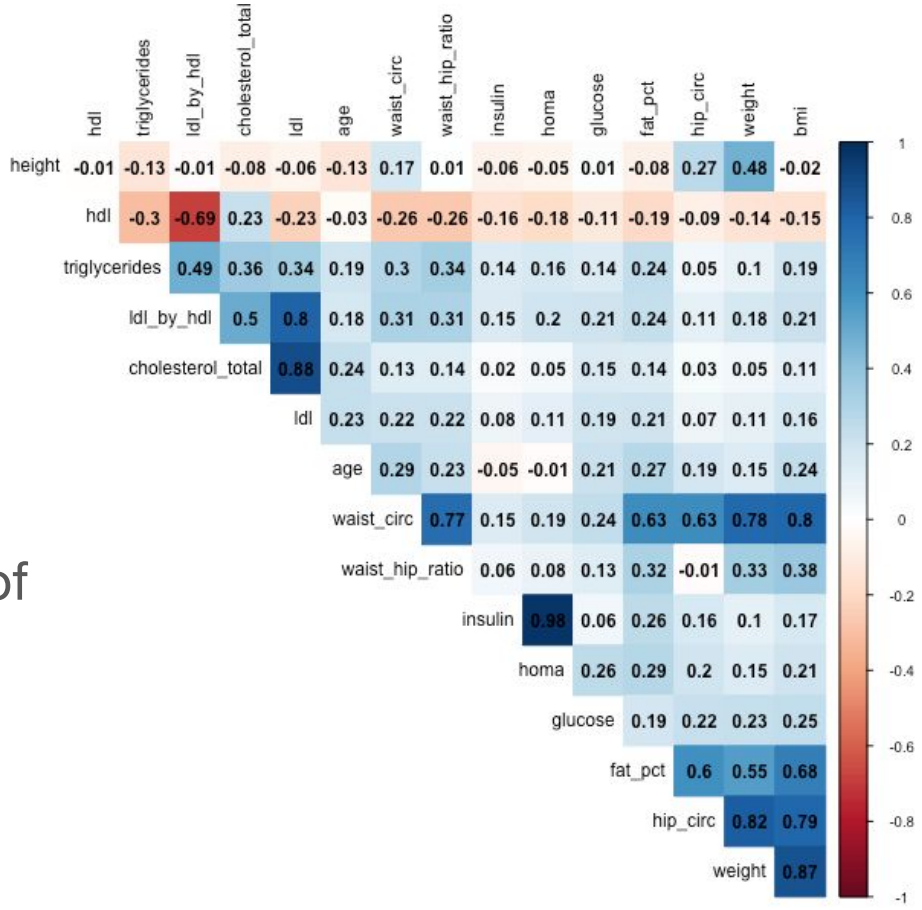


Correlation Plots

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

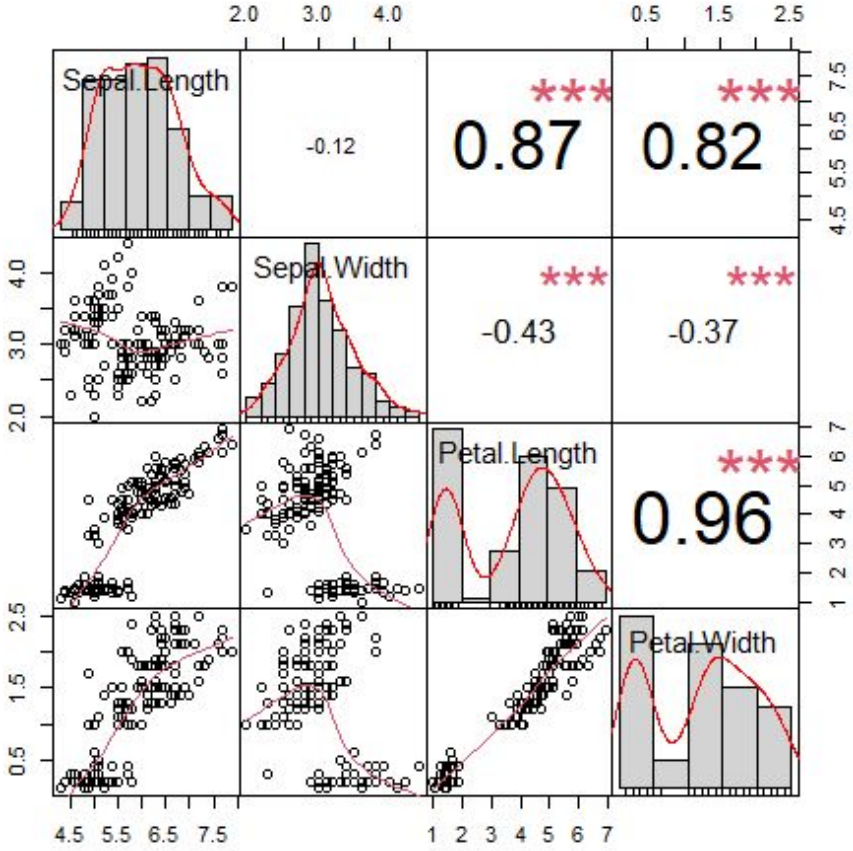
```
[[ 1.         -0.10936925  0.87175416  0.81795363]
 [-0.10936925  1.         -0.4205161  -0.35654409]
 [ 0.87175416 -0.4205161   1.          0.9627571 ]
 [ 0.81795363 -0.35654409  0.9627571   1.        ]]
```

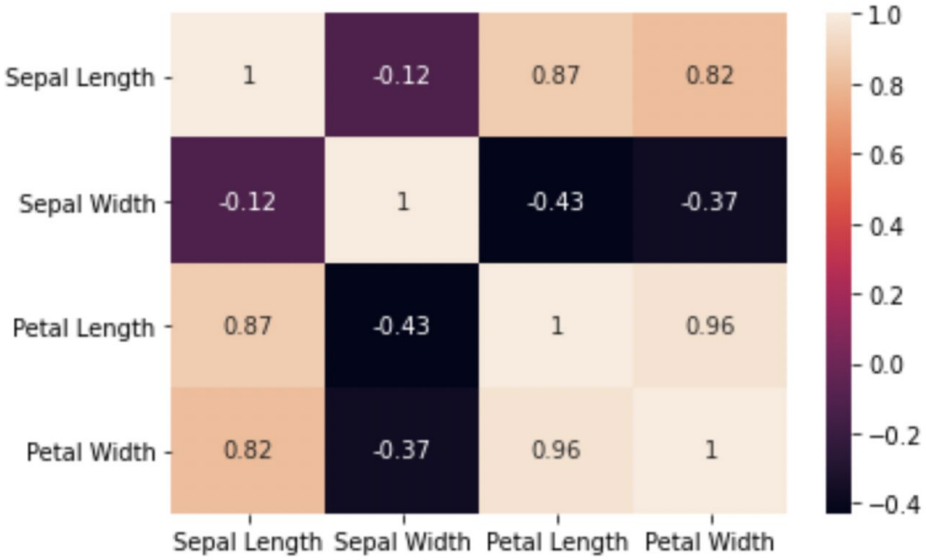
Why are all entries on the diagonal '1'?



Source

Correlation Plots





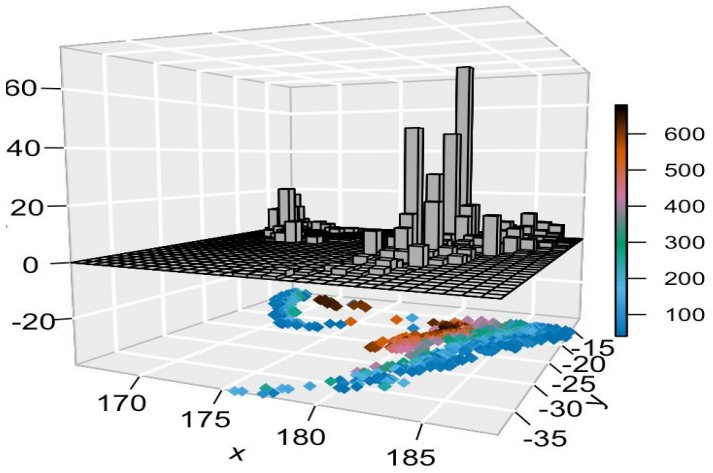
https://r-coder.com/correlation-plot-r/

Demo

Challenges of Visualization

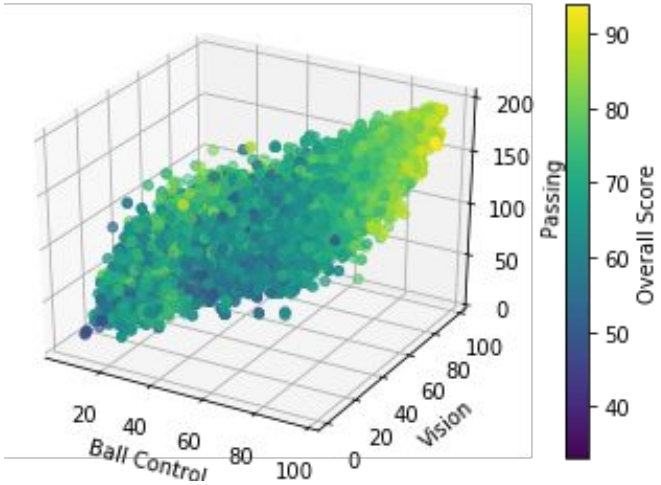
Higher Dimension	**Non-Trivial**
Time Consuming	**Hard to Show Uncertainty**

High Dimensional Data



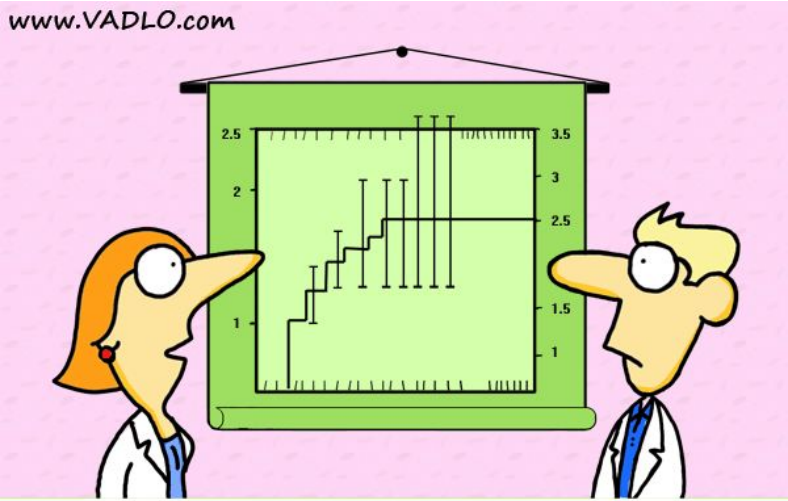
4D Plot For Earthquake Data

- 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

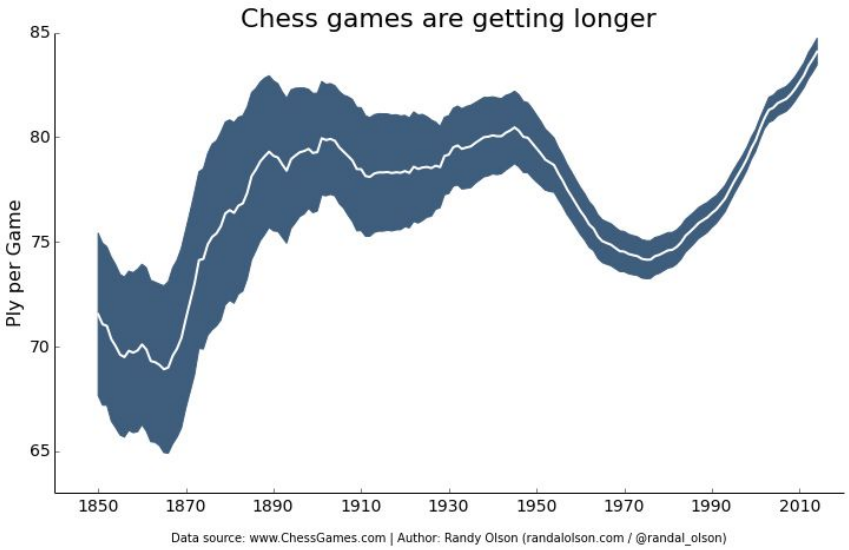


Error Bars

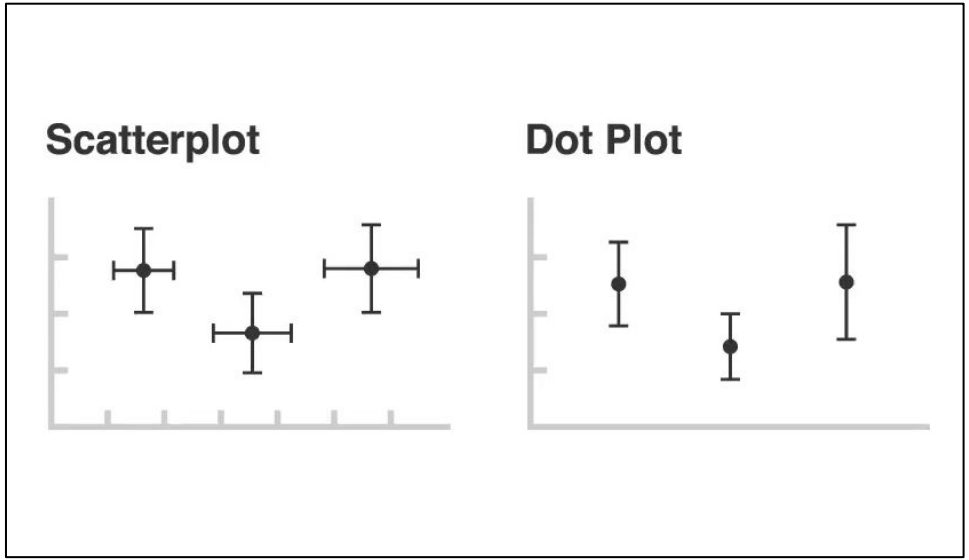
- Show uncertainty
- Usually display 95 percent confidence interval

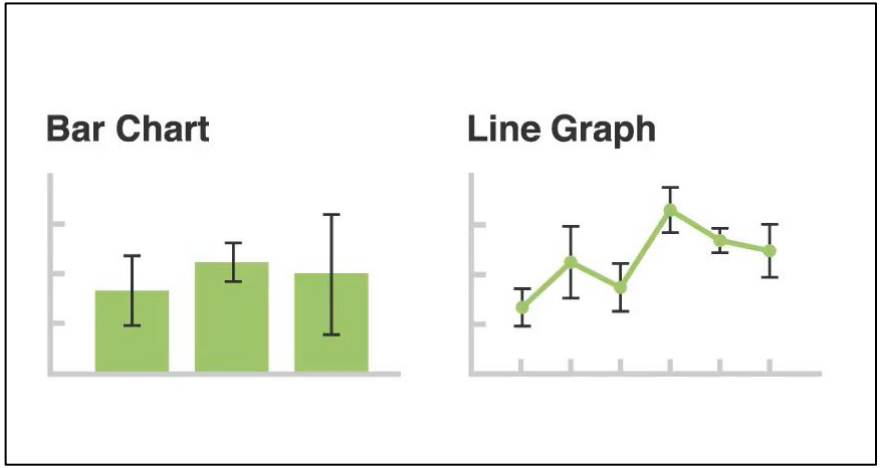


"Did you really have to show the error bars?"



Error Bars

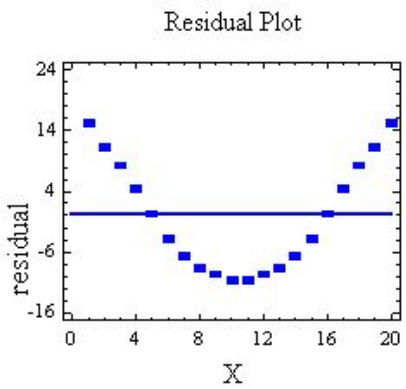




https://www.geeksforgeeks.org/errorbar-graph-in-python-using-matplotlib/

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



Coming Up

- **Assignment 2**: Due Tonight at 11:59 PM
- **Assignment 3**: Due next Wednesday at 11:59 PM
- **Next Lecture**: Fundamentals of Machine Learning

Check ED before writing emails! Post Questions on ED!

CDS Education

We explore, learn, and educate big minds.