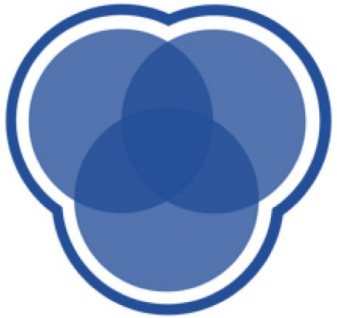
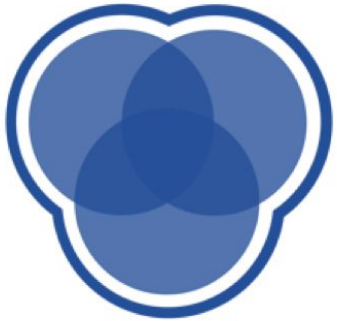


CDS Education

We explore, learn, and educate big minds.



CDS Education

Introduction to Machine Learning for Python

Manipulation Techniques & Visualization

Sanity Check

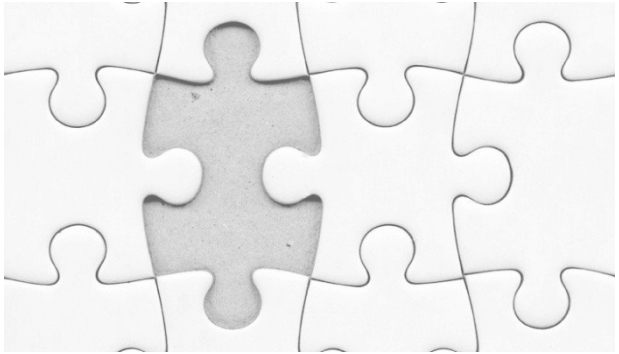
- Have you looked at the notes and started the quiz?
- Are you getting email notifications from Piazza?
- Did you enroll yourself on the Student Center?
- Are you in a group of 3-4 people for the project?
  - If not, post on Piazza or we can randomly assign groups

Dealing with Missing Data

Datasets are usually incomplete. We can handle this by:

Leaving out missing samples

Data imputation



NaN Values

- NaN values are “Undefined”
- Variety of uses
  - Error in collecting data
  - Feature is only present/ measurable among a subset data samples
- Can often be filled be a 0 or "None"

Removing Rows or Columns

- You can remove NaN values by removing specific samples or entire features
- Beware not to remove too many samples or features
  - Information about the dataset is lost each time you do this
  - Could lead to biased model
- How much is too much?

```
import numpy as np
import functools
def conjunction(*conditions):
    return functools.reduce(np.logical_and, conditions)
```

```
c1 = data == True
```

```
data_filtered = data[conjunction(c1)]
```

```
data_filtered[np.isfinite(data_filtered)]
```

	0	1	2	3	4	5	6	7	8	9	...	990	991	992	993	994	995	996	997	998	999
0	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	NaN	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	NaN	NaN	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	NaN	NaN	NaN	1	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Randomly Replacing NaNs

- This is not done - don’t do it
- Replacing NaNs with random values adds unwanted and unstructured noise
    - Not useful for data imputation

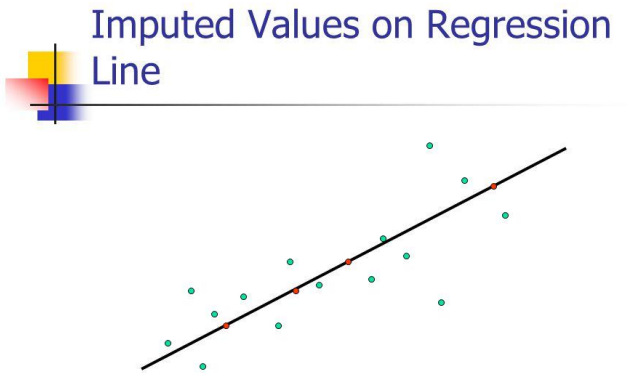


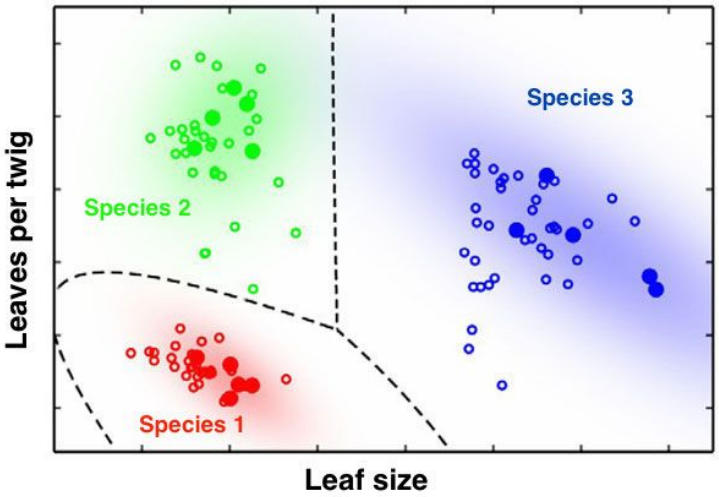
Summary Statistic Imputation

- Can replace missing values with an average value
  - Won't change the average of the data
- If numerical, use the median or mean
  - Check if the data is normal for the mean - may be better to do median
- If categorical, use the mode
- Still can add noise, but not as much

Regression or Clustering

- Use other variables to predict the missing values
  - Through either regression or clustering model
- Doesn't include an error term, so it's not clear how confident the prediction is





Data Imputation Example

Go to the course website to follow along with the code

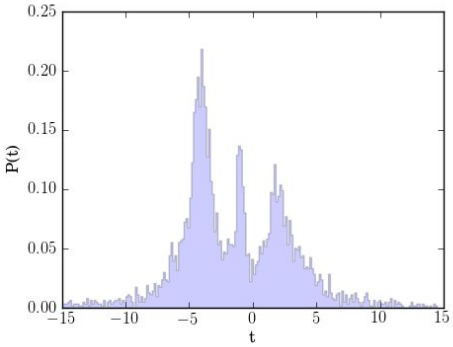
Techniques for Data Manipulation

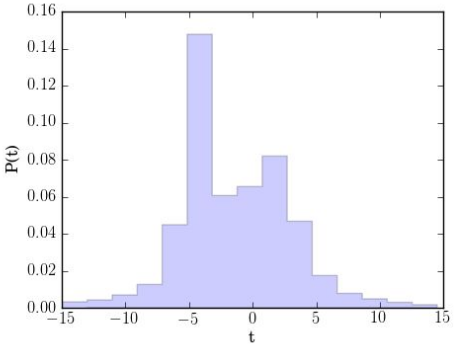
Formatting the shape of our data

Changing the actual content of the data

Technique: Binning

What it does	Makes continuous data categorical by lumping ranges of data into discrete “levels”
Why?	Applicable to problems like (third-degree) price discrimination



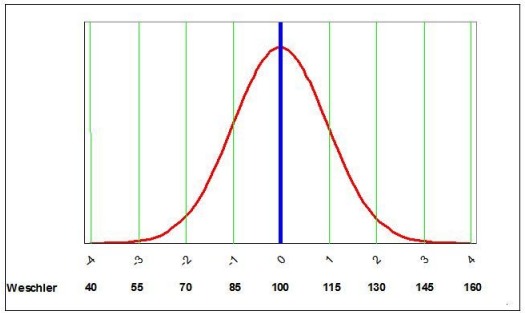


Source

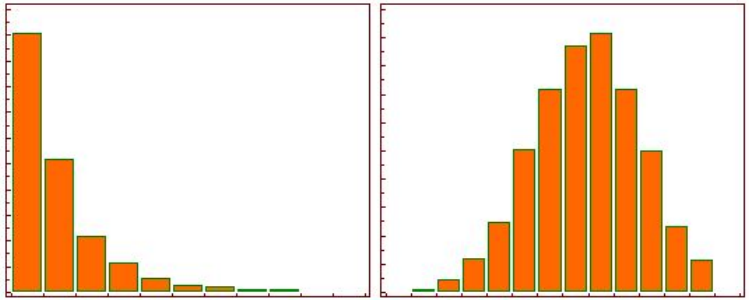
Technique: Normalizing

What it does	Turns the data into a bell curve (Gaussian) shape by standard, log, or another transformation
Why use it	Meet model assumptions of normal data; act as a benchmark since the majority of data is normal; wreck GPAs

Standardizing



Log transformation



Others include square root, cubic root, reciprocal, square, cube...

Source Source

Technique: Ordering

What it does	Why?	Example
Converts categorical data that is inherently ordered into a numerical scale	Numerical inputs often facilitate analysis	January → 1 February → 2 March → 3 …

Technique: Dummy Variables

What it does

Creates a binary variable for each category in a categorical variable

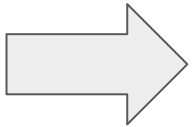
plant	is a tree
aspen	1
poison ivy	0
grass	0
oak	1
corn	0

Technique: Feature Engineering

What it does	Generates new features which may provide additional information to the user and to the model
How to do it	You may add new columns of your own design using the assign function in pandas

tab ->

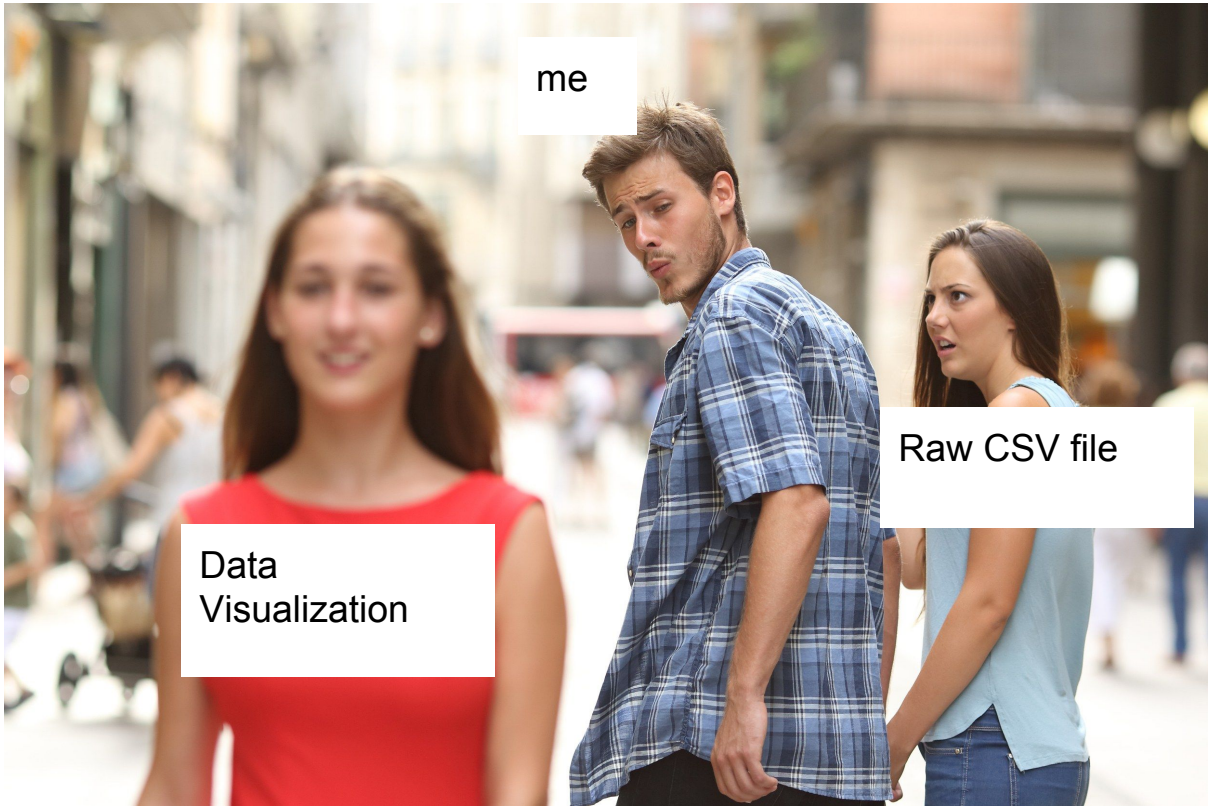
ID	Num
0001	2
0002	4
0003	6



ID	Num	Half	SQ
0001	2	1	4
0002	4	2	16
0003	6	3	36

```
tab.assign(SQ=arr['Num']**2, Half=0.5 * arr['Num'])
```

Data Visualization

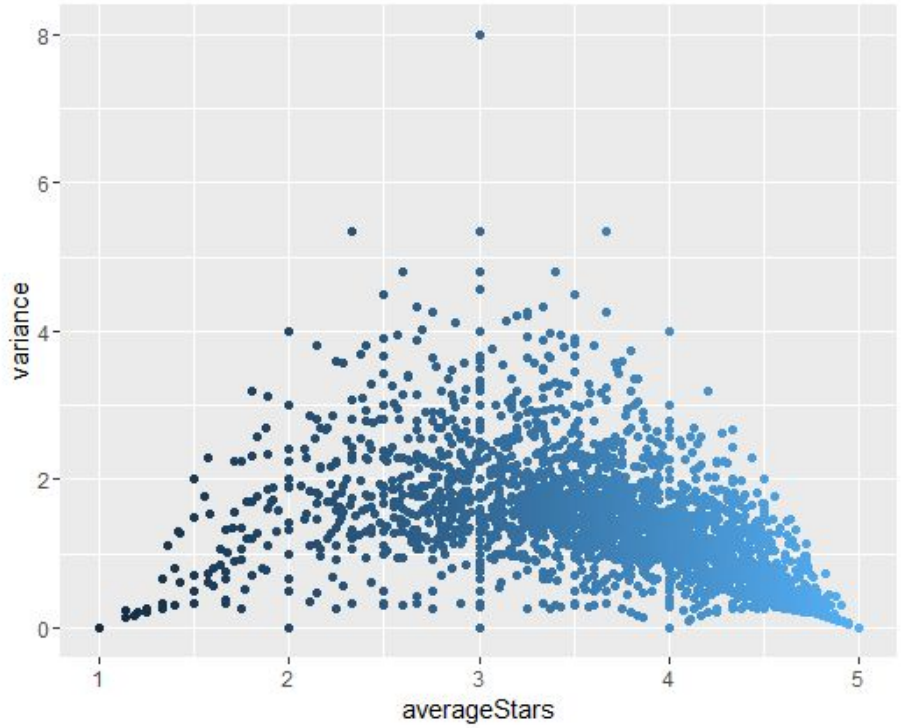


Source

Data Visualization Simple Example: Yelp

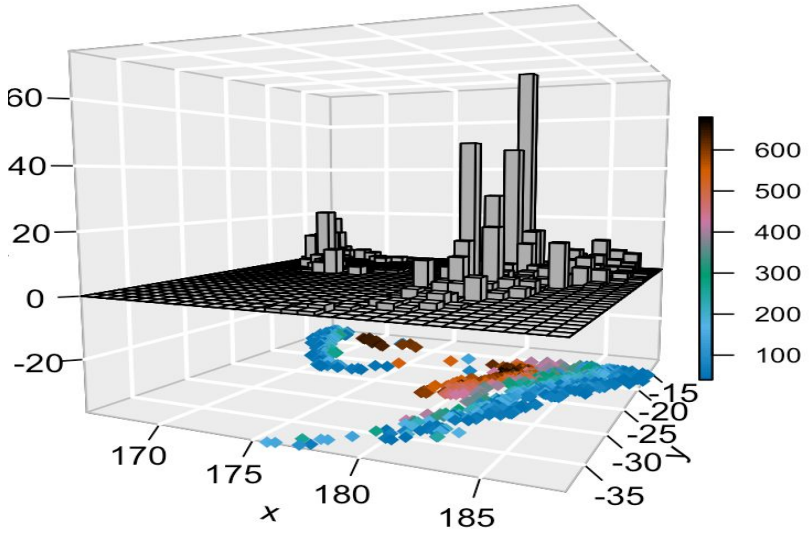
	AVG(stars)	var
AVG(stars)	1.00	-0.43
var	-0.43	1.00

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



Why Data Visualization?

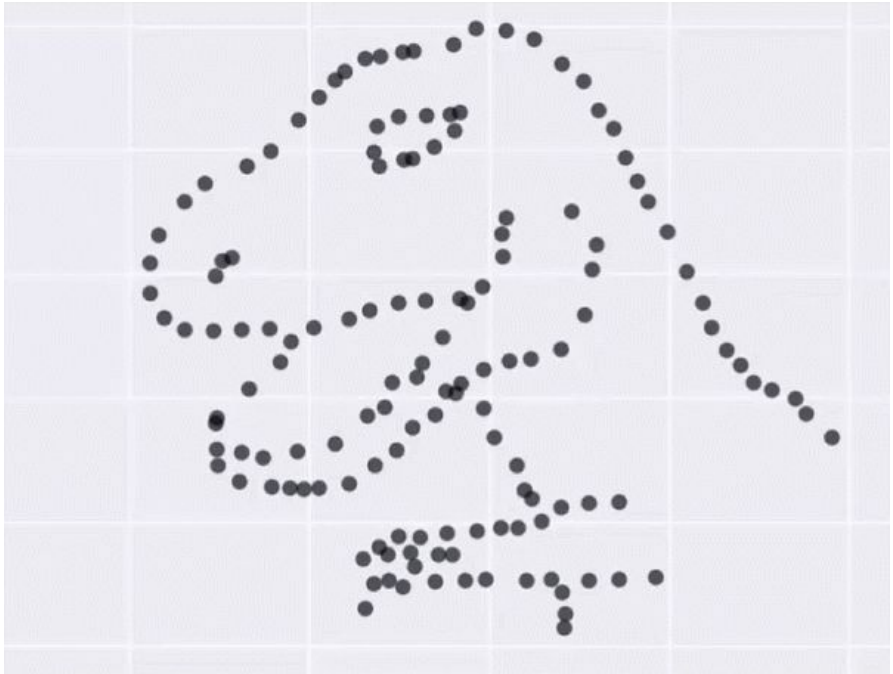
- Understanding a dataset
- Communication of knowledge to an audience



4D Plot For Earthquake Data

Why Data Visualization is Important

- **All Different Datasets**
  They all have same mean, median, mode, variance, line of best fit
- **Same Summary Stat**
  But we need to see how the **actual** data looks



Source

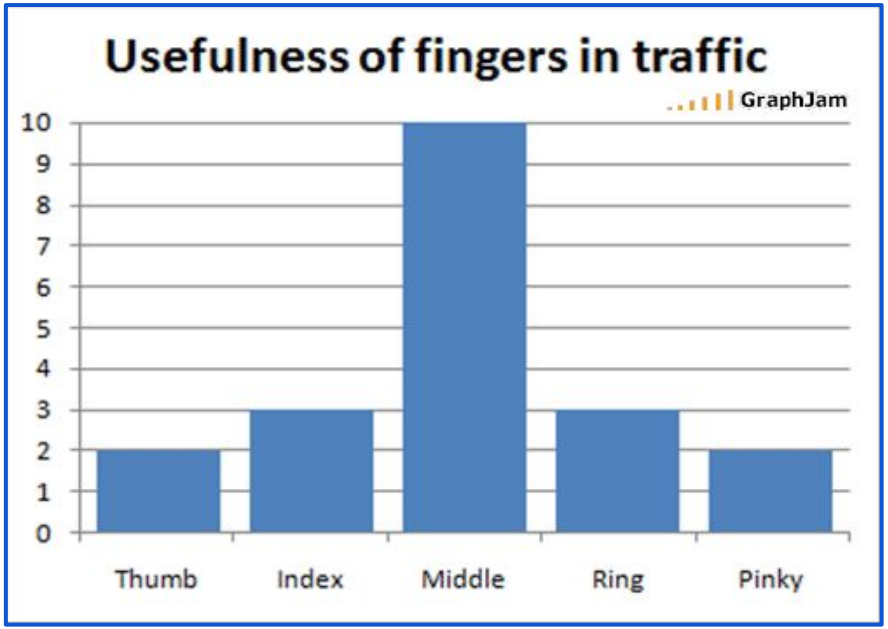
What is matplotlib?

- **Python data visualization package**
  - Capable of handling most data visualization needs
  - Simple object-oriented library inspired from MATLAB
  - Cheatsheet

matplotlib

Let's start with an easy one... a bar graph!

- Represent **magnitude** or **frequency**
- Allows us to compare features



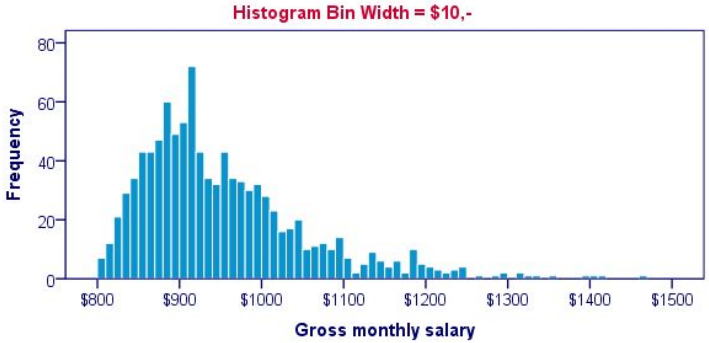
Source

Histograms

Source

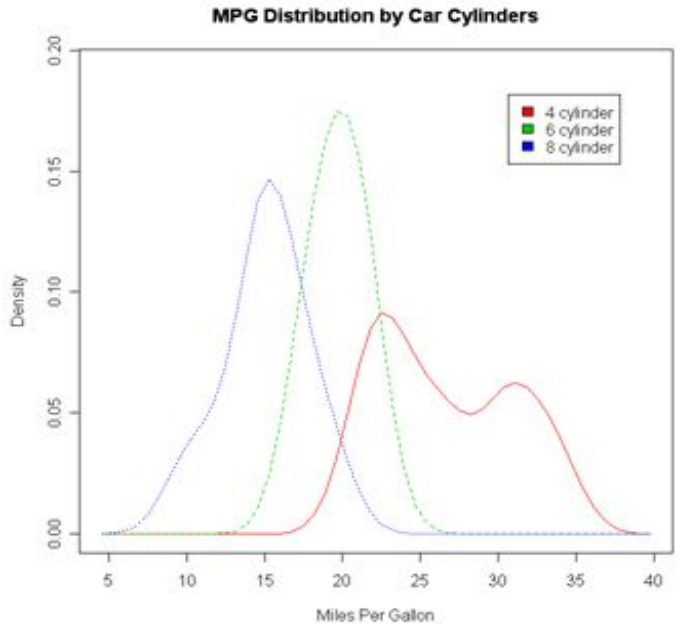
- Used to observe **frequency distribution** of numerical data
- Data split into **bins**

Histograms



Source

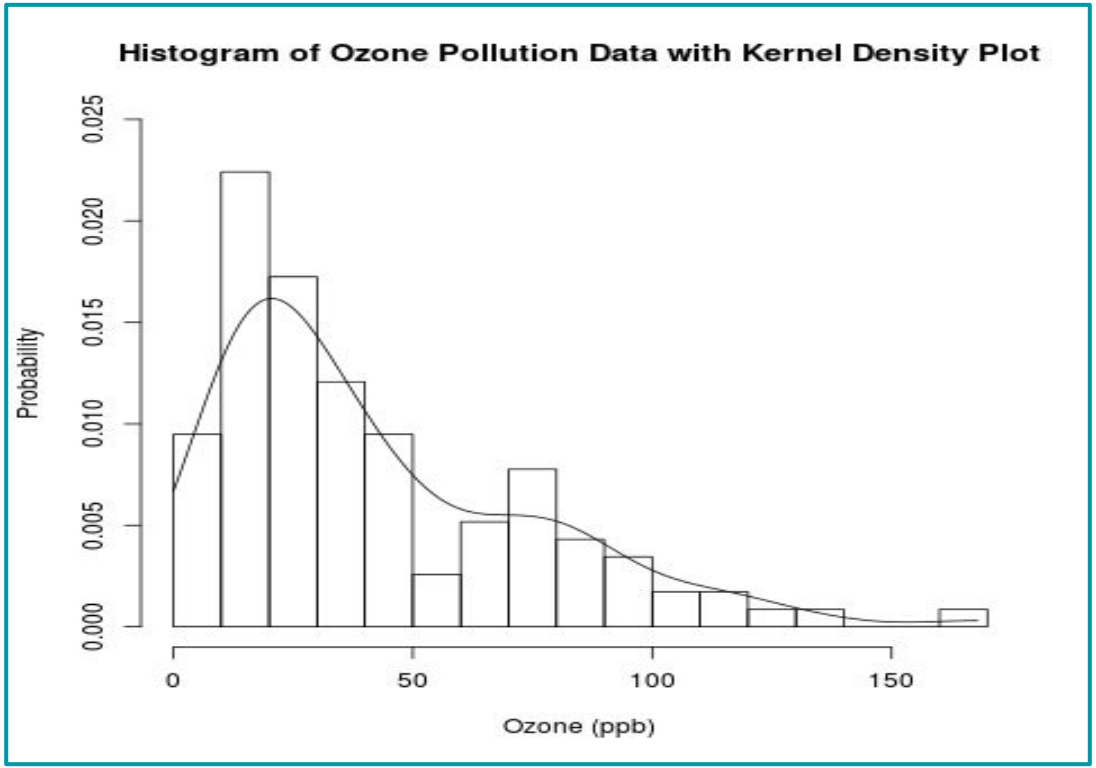
Density Plot



- Like a histogram, but **smooths** the shape of the distribution
- Why is Density Plot important?

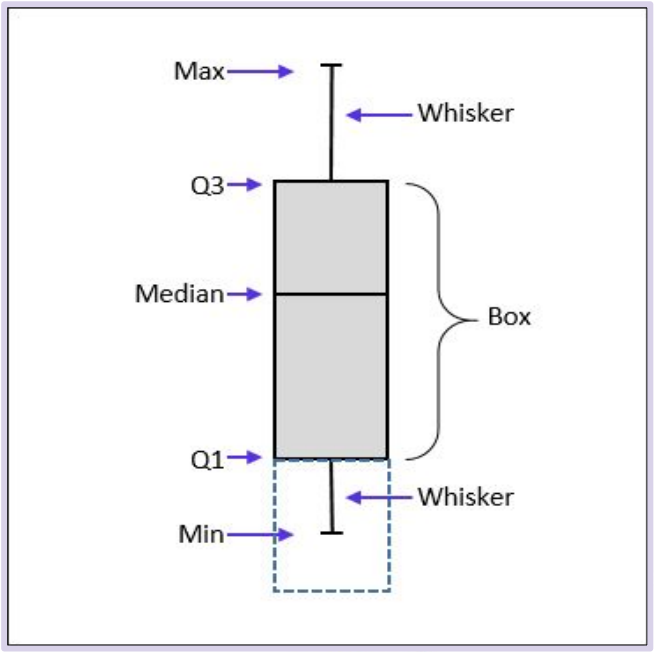
Source

Histogram vs. Density Plot



Source

Boxplot (a.k.a Box-and-whisker plot)

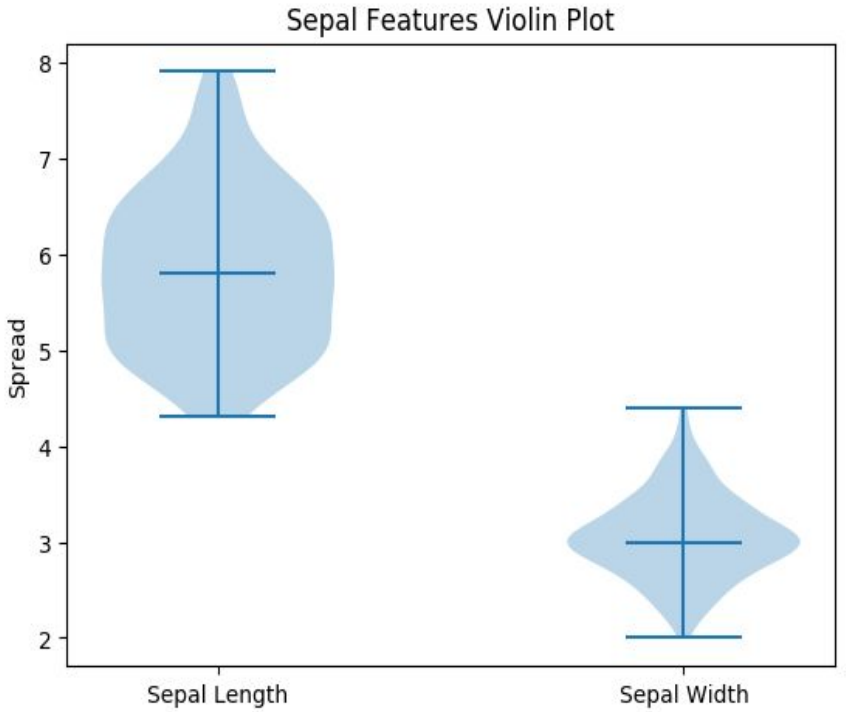


- Summary of data
- Shows **spread** of data
- Gives range, interquartile range, median, and outlier information

Source

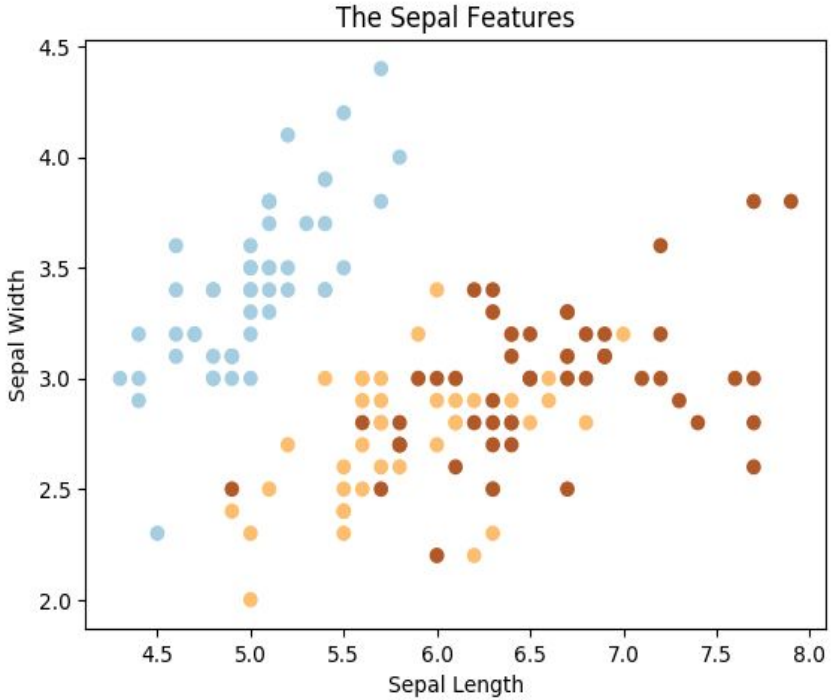
Violin Plot

- Combination of **boxplot** and **density plot** to show the **spread** and **shape** of the data
- Can show whether the data is **normal**

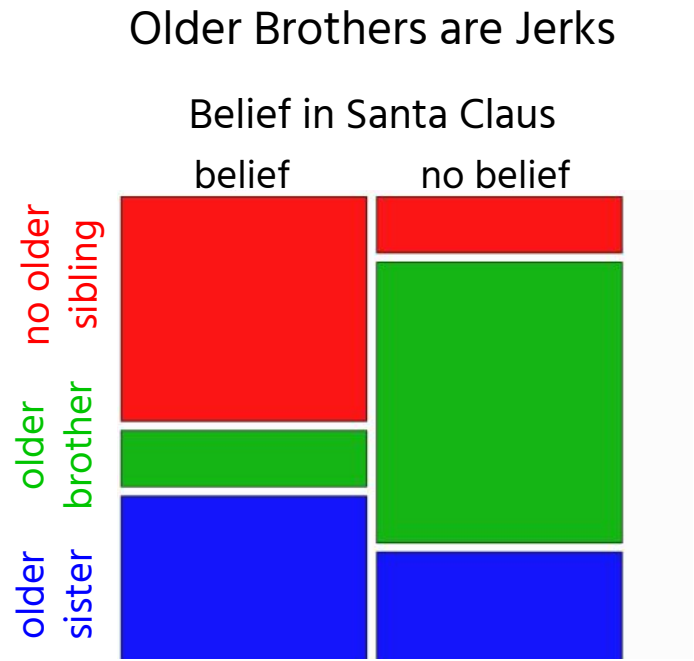


Scatterplot

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



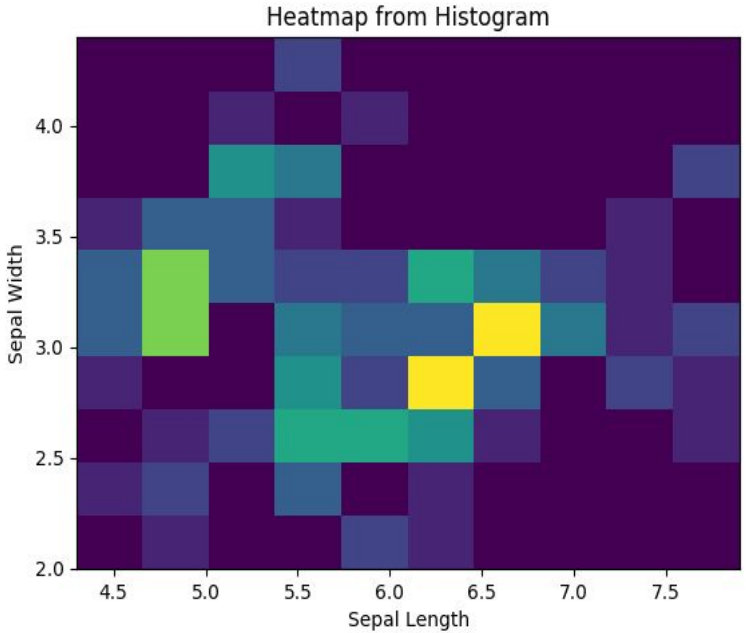
Mosaic Plot



Source

- Represents **two-way frequency**
- Horizontal dimension represents the frequency of one variable while the vertical dimension represents the other

Heatmaps

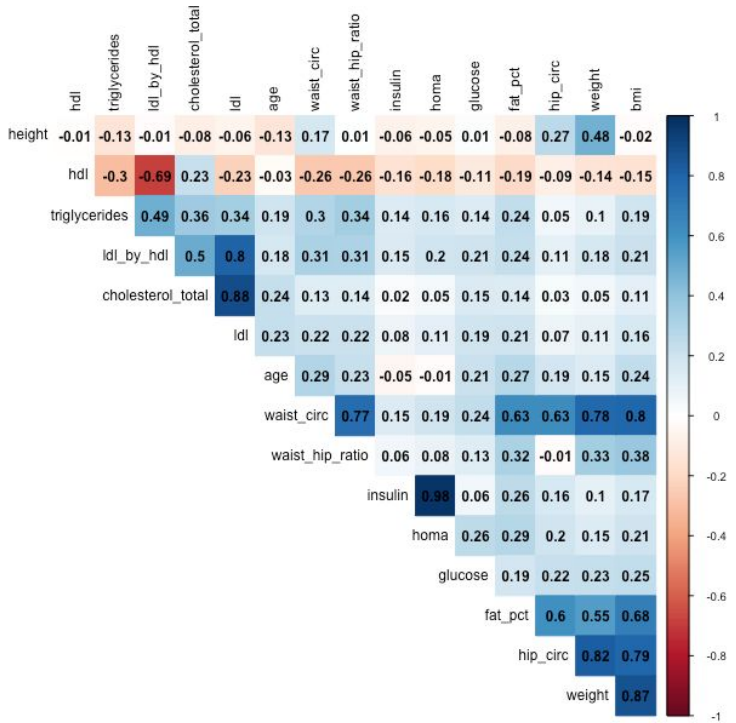


- Varying degrees of one metric are represented using **color**[1]
- Especially useful in the context of **maps** to show geographical variation

1 Defined by https://www.marketingterms.com/dictionary/heatmap/

Correlation Plot

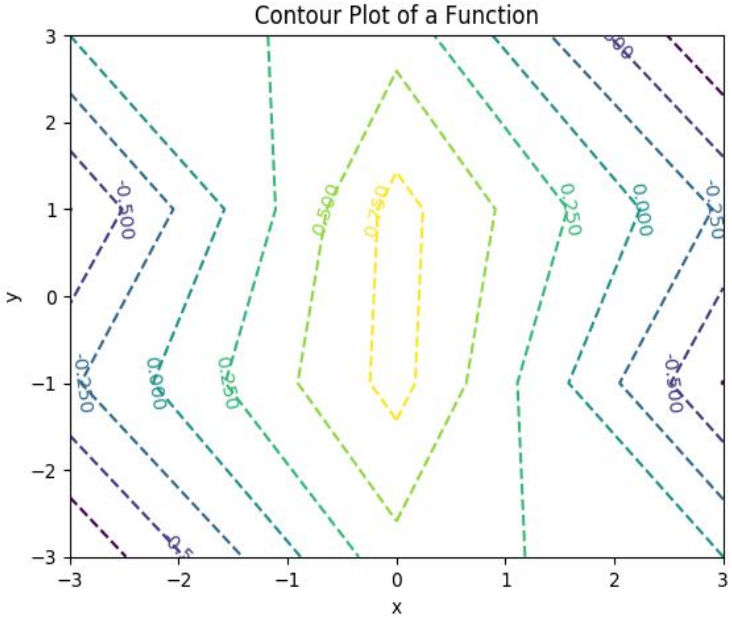
- 2D matrix with all variables on each axis
- Entries represent the **correlation coefficients** between each pair of variables




```
[[ 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.        ]]
```

Source

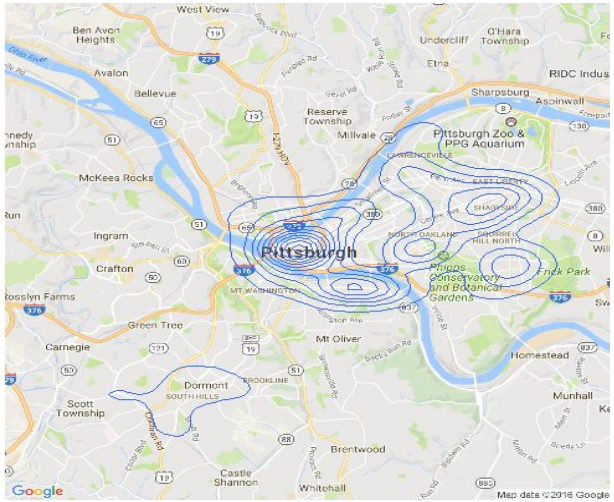
Contours

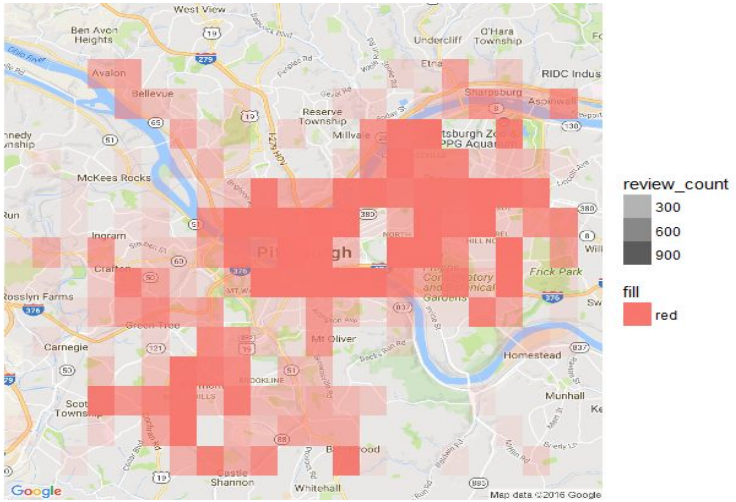


- Used to show **distribution** of the data or a function
- Observe variation among portions of data
- In maps, they indicate the shape of the land

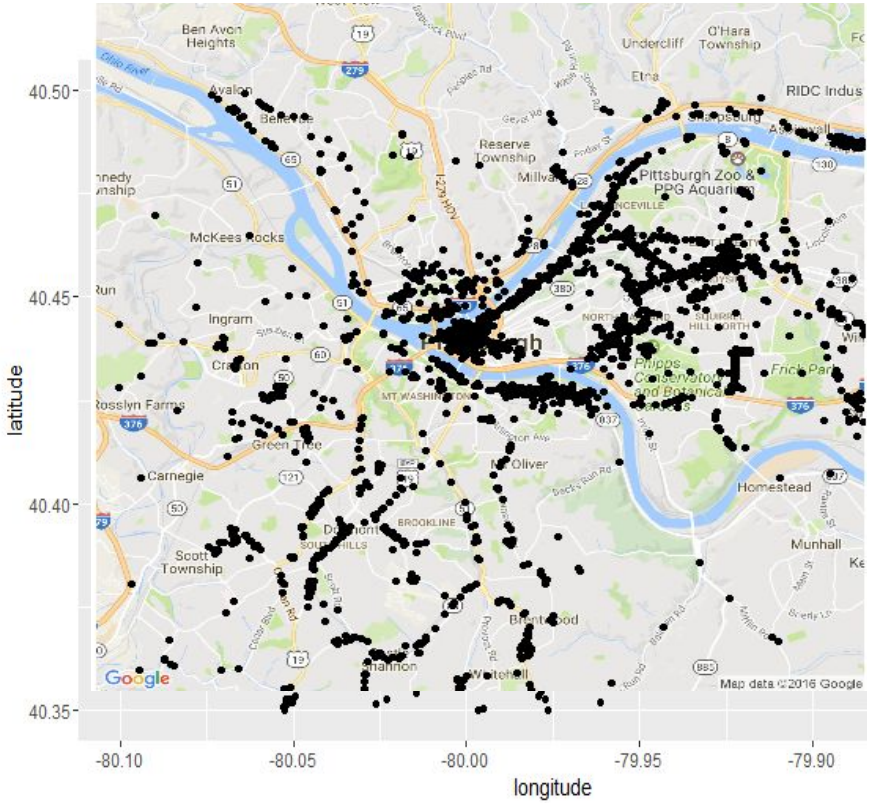
Using Maps

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





Example: Pittsburgh Data



Challenges of Visualization

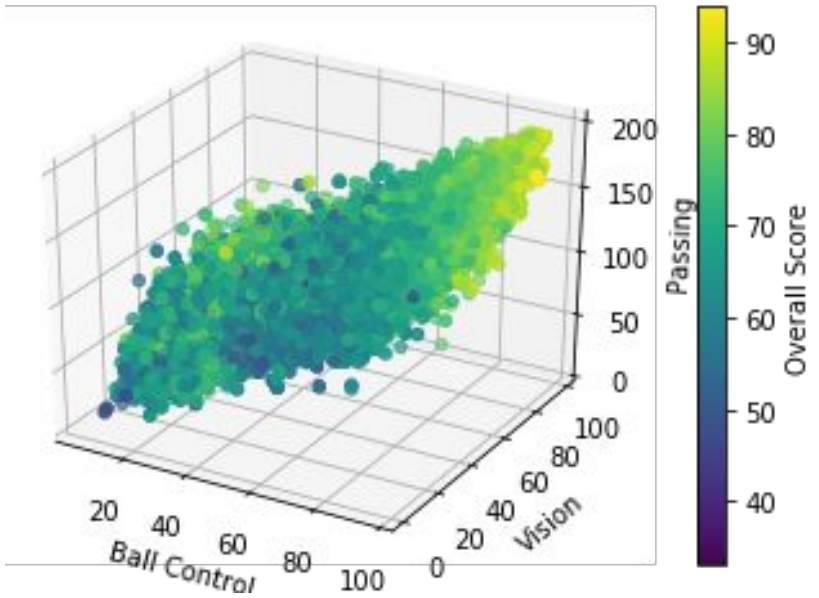
Higher Dimension

Non-Trivial

Time Consuming

Hard to Show Uncertainty

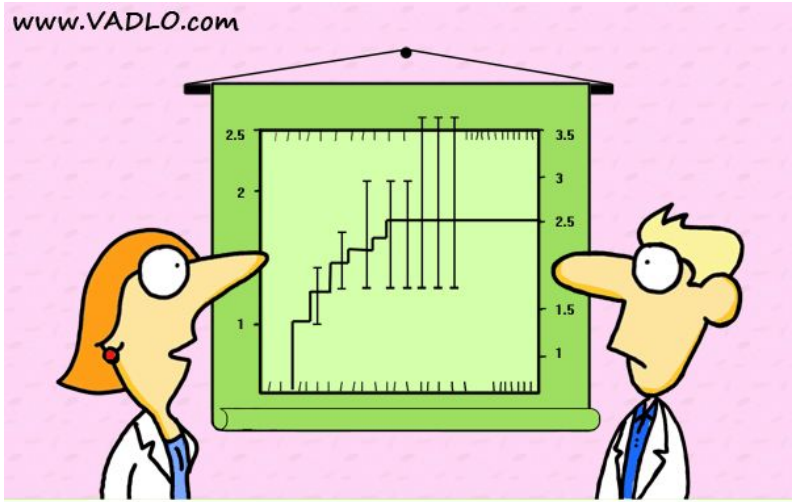
Higher Dimensional Data



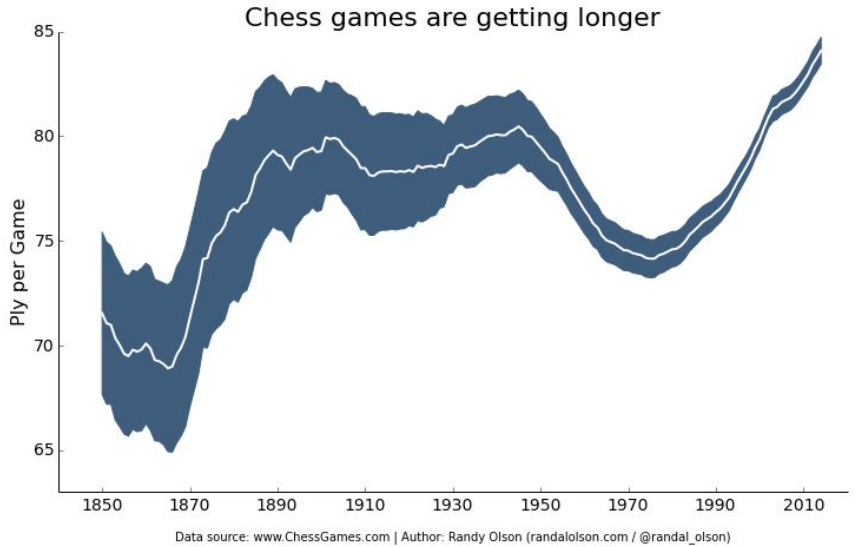
- **Color**, **time animations**, or **point shape** can be used for higher dimensions
- There is a limit to the number of features that can be displayed

Error Bars

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



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



Source

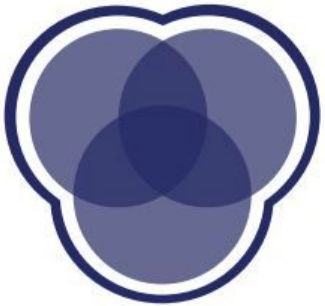
Coming Up

Your assignment: Finish quiz and start project A

Due dates: Quiz due 2/25 & Project A due 3/6

Next week: Introduction to Supervised Learning

See you then!



CDS
Cornell Data Science