

Lecture Notes for **Machine Learning in Python**

Professor Eric Larson
Preprocessing and Visualization

Class Logistics and Agenda

- Participation/Teams
- Be sure you look at **Lab One!**
- Dataset Selection Questions?
- Agenda
 - Finish Pandas Demo with Imputation, *if needed*
 - Data Exploration
 - Data Preprocessing
 - Data Visualization

Last Time

- Datatypes
- Imputation
- Document Features

Feature Type Representation Review

	Attribute	Representation Transformation	Comments
Discrete	Nominal	Any permutation of values one hot encoding	If all employee ID numbers were reassigned, would it make any difference?
	Ordinal	An order-preserving change of values, i.e., $\text{new_value} = f(\text{old_value})$ where f is a monotonic function. integer	An attribute encompassing the notion of good, better, best can be represented equally well by the values $\{1, 2, 8\}$ or by $\{0.5, 1, 10\}$.
Continuous	Interval	$\text{new_value} = a * \text{old_value} + b$ where a and b are constants. float	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
	Ratio	$\text{new_value} = a * \text{old_value}$ float	Length can be measured in meters or feet.

K-Nearest Neighbors Imputation

ID	Pregnant	SMV	Age	Diabetes
1	Y	23.0	21-30	positive
2	N	90.0	31-40	negative
3	Y	33.3	?	positive
4	?	38.1	21-30	negative
5	N	23.1	31-40	positive
6	Y	90.0	21-30	negative
7	Y	0.0	21-30	positive
8	Y	35.1	?	negative
9	N	30.1	31-40	positive
10	Y	27.5	31-40	positive

For $K=3$, find 3 closest neighbors

ID	Pregnant	SMV	Age	Diabetes	Dist
3	Y	23.3	?	positive	C
6	Y	26.6	21-30	negative	$\sqrt{(1-2)^2 + (1-0)^2}$
9	N	90.0	31-40	negative	$\sqrt{(1-0)^2 + (1-1)^2}$
4	?	38.1	21-30	negative	$\sqrt{(4-2)^2 + (1-0)^2}$

How to calculate distance?

- Difference for valid features only
- May need to normalize ranges
- Or weight neighbors differently
- Or have min # of valid features
- Euclidean, city-block, etc.

Demo

Start
Pandas demo
DataFrames
Loading
Indexing
Imputing



“Start” Pandas demo

DataFrames
Loading
Indexing
Imputing



03.Data Visualization.ipynb

Data Exploration



What is data exploration?

A preliminary exploration of the data to better understand its characteristics.

- Help **select** the **right tool** for preprocessing or analysis
- Exploratory Data Analysis (EDA) by Dr. John Tukey:
 - The focus was visualization
 - Clustering and anomaly detection were viewed as exploratory techniques
- In our discussion,
 - Summary statistics, aggregations
 - Visualizing summaries



Summary Statistics

- frequency, location, and spread
 - Examples: location by **mean**
spread by **standard deviation**
- Most summary statistics can be calculated in a single pass through the data

$$\text{sample mean}(x) = \bar{x} = \frac{1}{m} \sum_{i=1}^m x_i$$

$$\text{sample median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r + 1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

- For nominal data, mode or frequency is most common

Measures of Spread

- **Range** is the difference between the max and min
- The **variance** or standard deviation is the most common measure of the spread of a set of points.

$$\text{sample variance}(x) = s_x^2 = \frac{1}{m-1} \sum_{i=1}^m (x_i - \bar{x})^2$$

- However, this is also sensitive to outliers, so that other measures are often used.

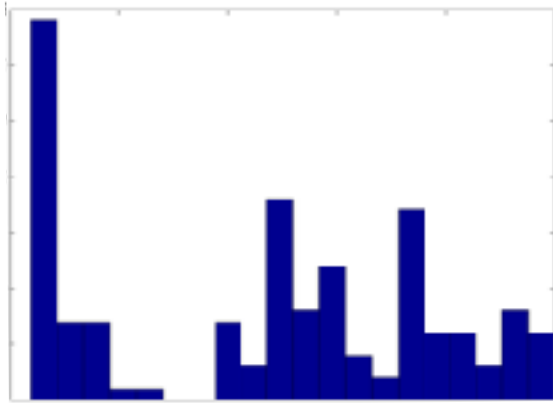
$$\text{AAD}(x) = \frac{1}{m} \sum_{i=1}^m |x_i - \bar{x}|$$

$$\text{MAD}(x) = \text{median}\left(\{|x_1 - \bar{x}|, \dots, |x_m - \bar{x}|\}\right)$$

$$\text{interquartile range}(x) = x_{75\%} - x_{25\%}$$

Self Test 2a.1

What measure of **spread** is **most appropriate** for the data in the histogram below?



- A) Standard Deviation
- B) Interquartile Range
- C) Median Absolute Difference
- D) None of these

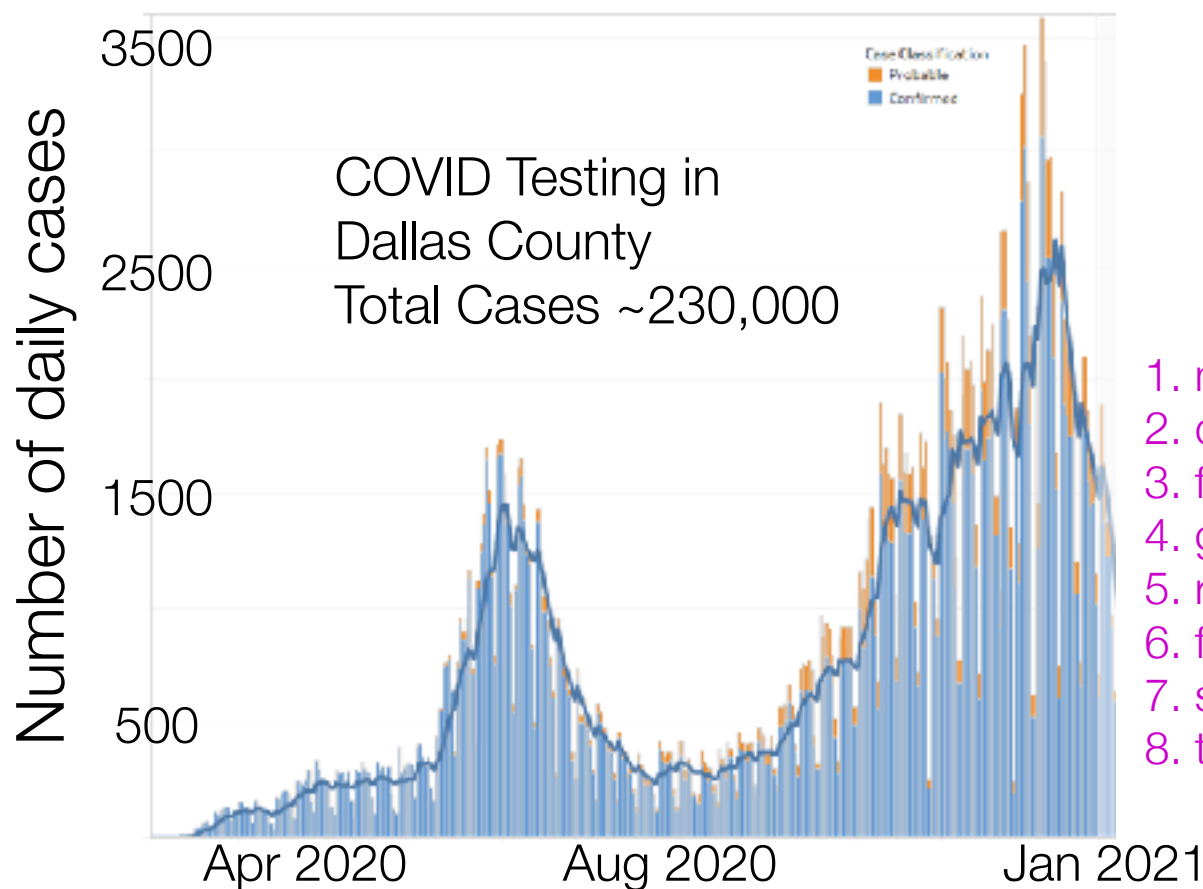
Data Preprocessing



Preprocessing

- Common preprocessing techniques:
 - **Aggregation: Combine features/samples**
 - ◆ Reduce the number of attributes or objects
 - ◆ Aggregated data tends to be more stable
 - **Transformation: Change of scale**
 - ◆ Normalize dynamic ranges
 - ◆ More numerically stable when combining
 - **Quantization: Make discrete**
 - ◆ More stable
 - ◆ More semantically meaningful

Preprocessing: Aggregation



Steps:

1. map test site to address
2. define address by county
3. filter to only Dallas county
4. group "result" by day w/n county
5. repeat for "probable"
6. filter out repeated entries, if any
7. sum up results for county in group
8. take rolling mean within group

How has aggregation has been used to create these plots?

<i>TID</i>	<i>Location</i>	<i>time</i>	<i>test</i>	<i>Probable?</i>
1	<i>test site name</i>	<i>day and hour</i>	<i>test result</i>	<i>yes/no</i>

<https://www.dallascounty.org/covid-19/>

Preprocessing: Transformation

- Monotonically map one set of values to a set of replacement values

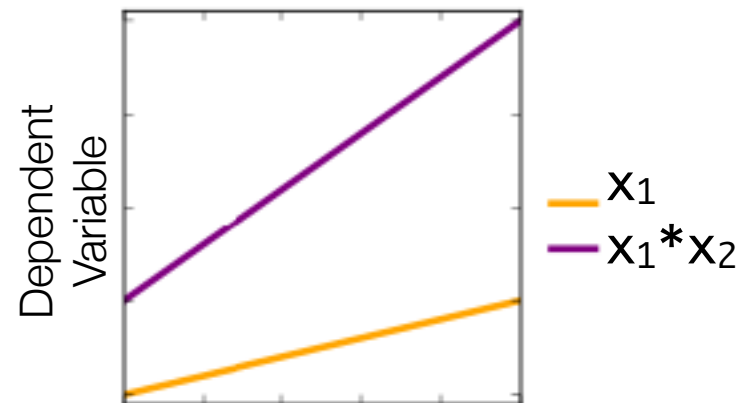
- **Standardization and Normalization**

- min/max, z-scores

```
df_normalized = (df-df.mean())/(df.std())
```

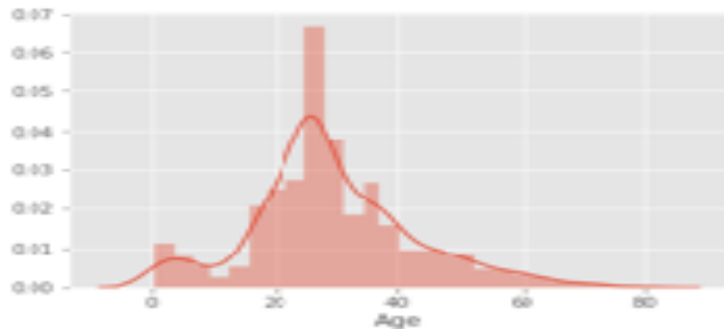
- **Polynomial and Interaction Variables**

- $x[:, \text{col}=1]**2$
 - $x[:, \text{col}=1]*x[:, \text{col}=2]$

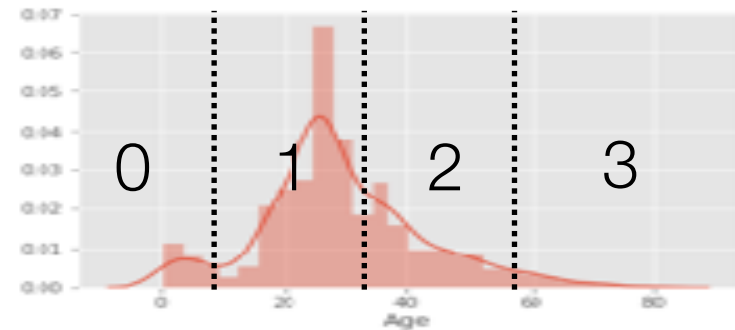


Preprocessing: Quantization

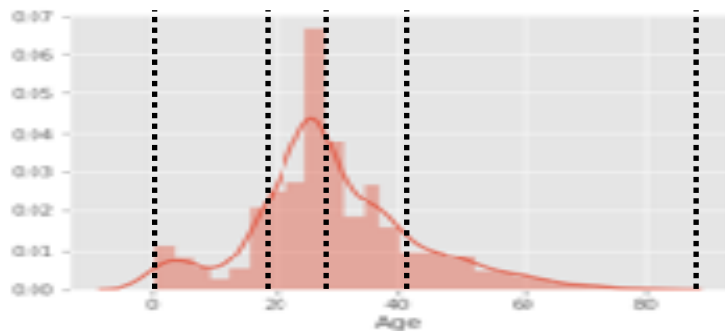
```
pandas.cut(dataframe.var, [5,10,15])
```



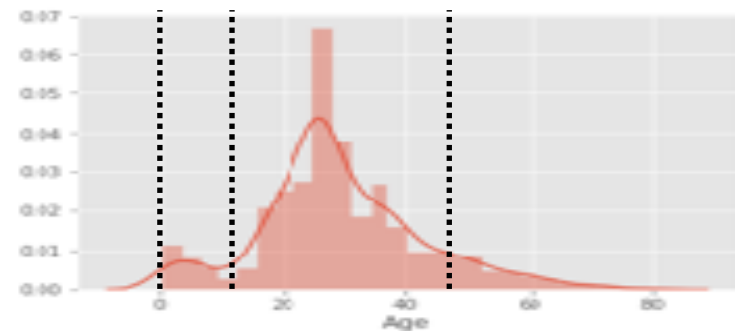
Data



Equal interval width



Equal frequency

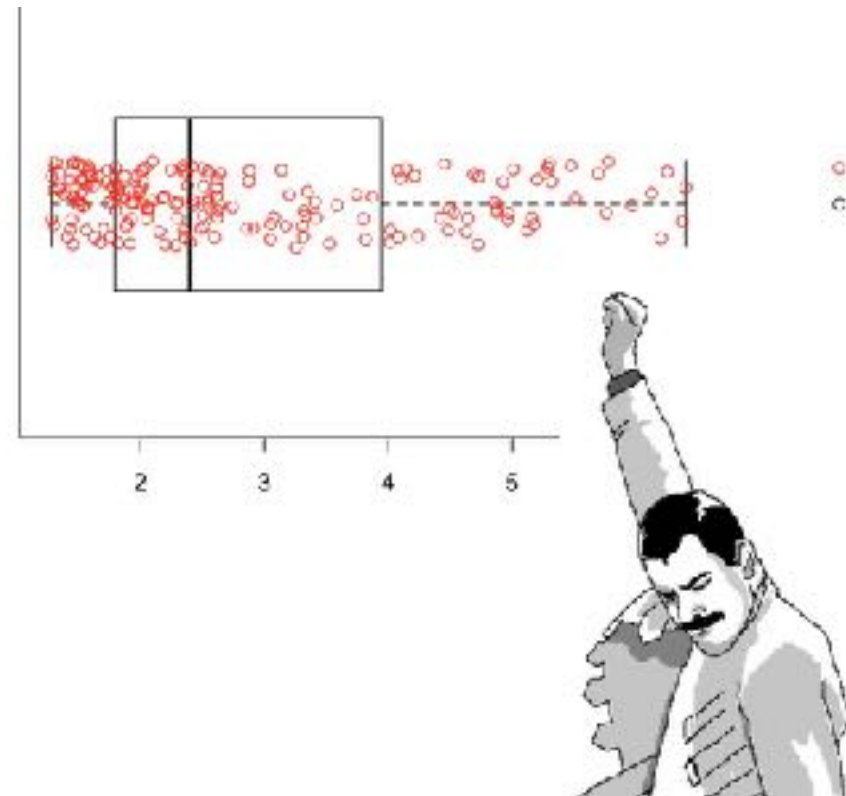
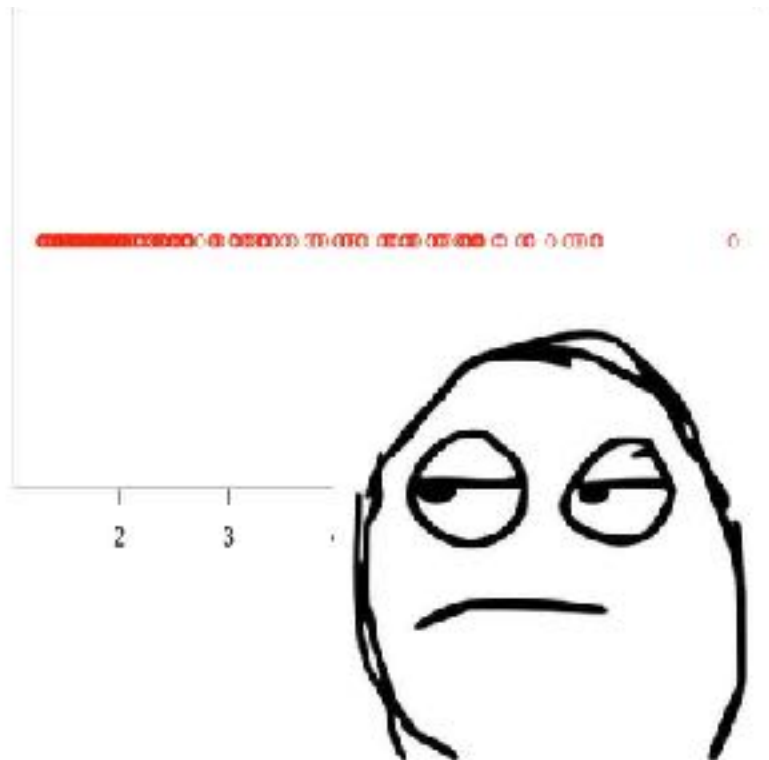


clustering: e.g., K-means

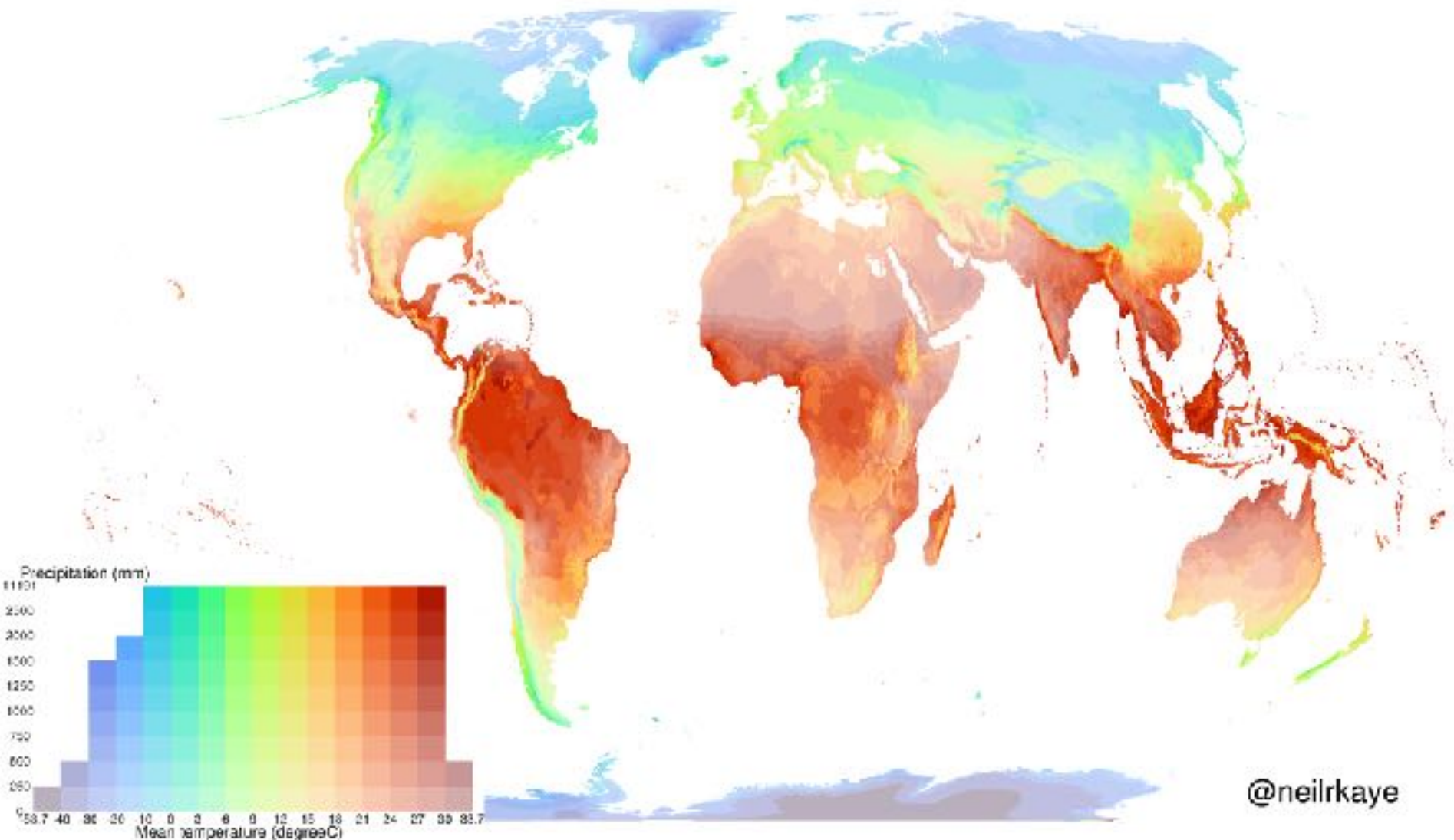
```
num_quantiles = 4
```

```
pandas.qcut(dataframe.var, num_quantiles)
```

Data Visualization



Annual mean temperature and precipitation totals (long term average)



Choosing How/What to Visualize?

- Start with a question you want to understand
- Think about the **best plot** to answer the question
 - Do you have the **right data** for visualizing?
 - Do you need to **worry** about the **amount** of data in the plot (aliasing, low samples, etc.)?
 - Can your question be answered **reliably**?
- **Interpret** the visualization: Did it answer the question?
 - **No**: Think of another visual
 - **Kinda**: Ask a follow up question
 - **Yes**: No it didn't, think more critically

Matplotlib

- Python plotting utility
 - Has **low level plotting** functionality
 - Highly **similar to Matlab and R** for plotting
- Extended to be visually more beautiful by
 - **seaborn**: stanford data visualization group

John Hunter (1968-2012)

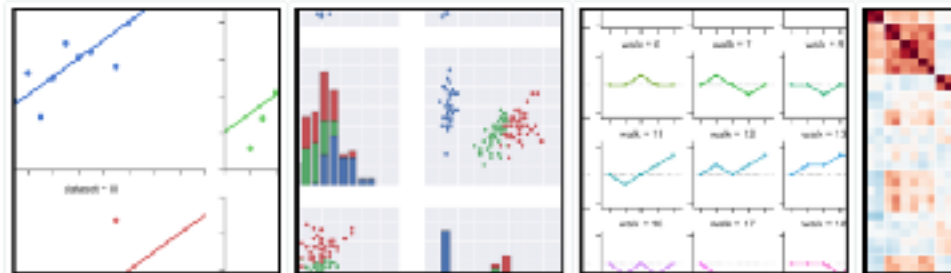


On August 28 2012, John D. Hunter, the creator of matplotlib, died from complications arising from cancer treatment, after a brief but intense battle with this terrible illness. John is survived by his wife Miriam, his three daughters Rahel, Ava and Clara, his sisters Layne and Mary, and his mother Sarah.

If you have benefited from John's many contributions, please say thanks in the way that would matter most to him. Please consider making a donation to the [John Hunter Memorial Fund](#).



Seaborn: statistical data visualization



- You tell me what conclusions we are getting from these graphs
 - Histogram
 - KDE
 - HeatMaps and Correlation
 - Scatter and Scatter Matrix
 - Box / Violin / Swarm



03.Data Visualization.ipynb

Matplotlib
Seaborn
Plotly

03.Data Visualization.ipynb

Other Tutorials:

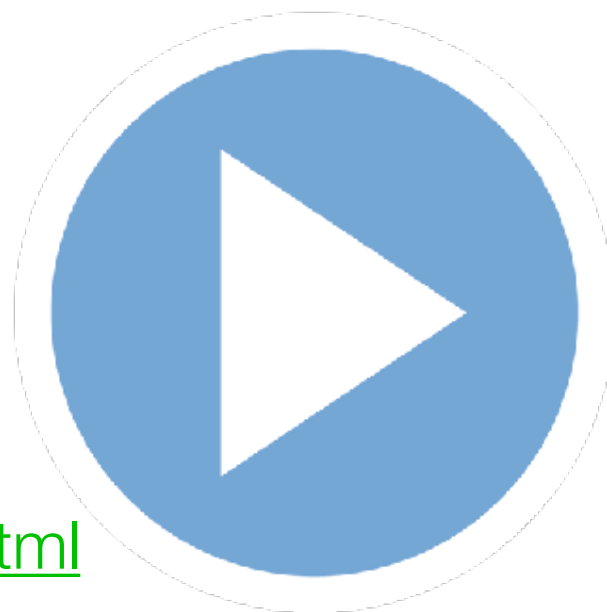
<https://t.co/zNzD8Q8w5E>

<http://matplotlib.org/examples/index.html>

<http://stanford.edu/~mwaskom/software/seaborn/index.html>

<http://pandas.pydata.org/pandas-docs/stable/visualization.html>

http://nbviewer.ipython.org/github/mwaskom/seaborn/blob/master/examples/plotting_distributions.ipynb



For Next Lecture

- Next Time:
 - Finish Visualization Demo
 - First Town Hall Meeting
- Look at chapter 5 of Python Machine Learning