Lecture Notes for Machine Learning in Python

Professor Eric Larson

Week Two, Lecture Three

Class Logistics and Agenda

- Participation for Distance
- Look at **Lab One**!
- Agenda
  - Pandas Demo with Imputation
  - Data Exploration
  - Data Preprocessing
  - Data Visualization

Demo

Pandas

- Make Up Demo
  - DataFrames
  - Loading
  - Indexing
  - Imputing

Data Exploration

What is data exploration?

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

- Key motivations of data exploration include
  - Helping to **select** the **right tool** for preprocessing or analysis
  - Making use of **humans' abilities** to recognize **patterns**
    - **People** can **recognize patterns** not captured by data analysis tools, Visualization is key

Techniques Used In Data Exploration

- In Exploratory Data Analysis, as originally defined by Dr. John Tukey:
  - The focus was visualization
  - Clustering and anomaly detection were viewed as exploratory techniques
- In our discussion of data exploration, we focus on
  - Summary statistics
  - Visualization

Summary Statistics

- **Earth shattering definition:**
- Summary statistics are numbers that summarize properties of the data

  - Summarized properties include 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

Measures of Location: Mean and Median

- The mean is the most common measure of the location of a set of points.
- However, the mean is very sensitive to outliers.
  - Solution: median or a trimmed mean

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

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

- 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.

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

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

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

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

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

Higher order statistics

- A comparison of the tails of a distribution

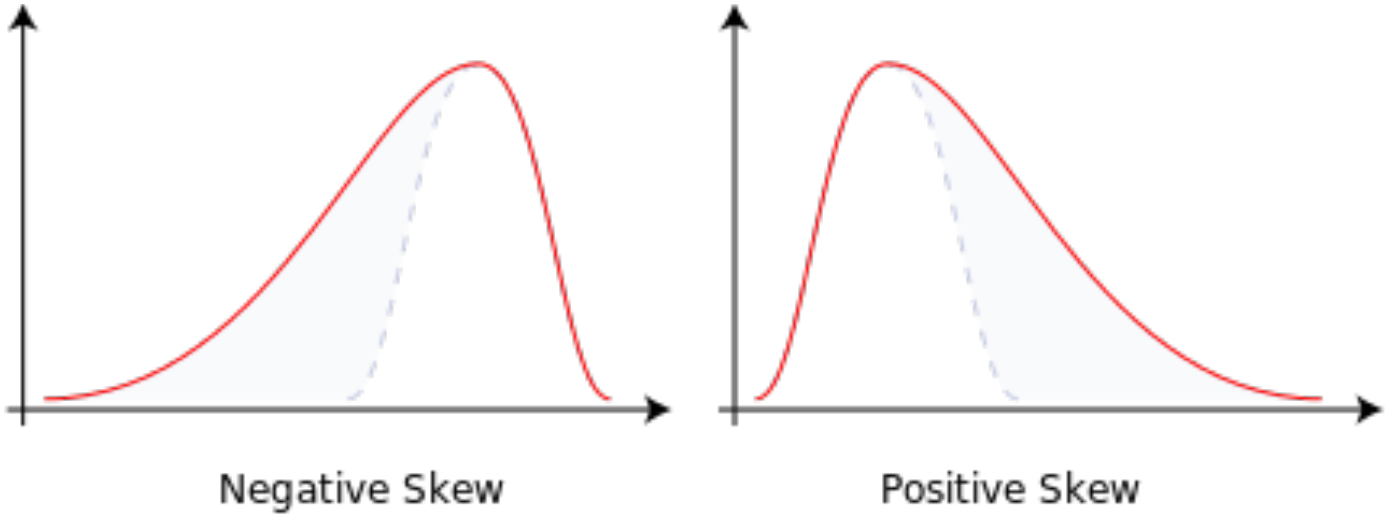


image: wikipedia

$$skewness(x) = \frac{1}{N}\sum_i \left(\frac{x_i - \bar{x}}{\sigma}\right)^3$$

$$kurtosis(x) = \frac{1}{N}\sum_i \left(\frac{x_i - \bar{x}}{\sigma}\right)^4$$

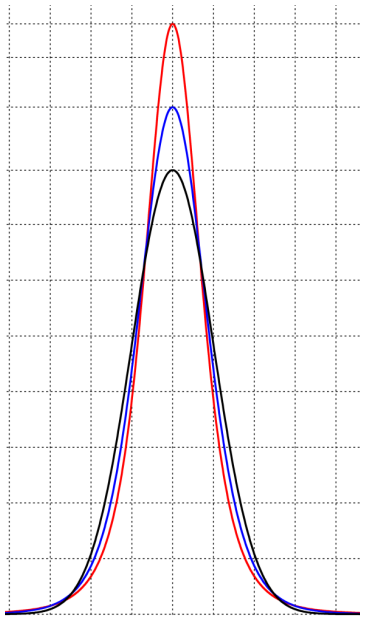
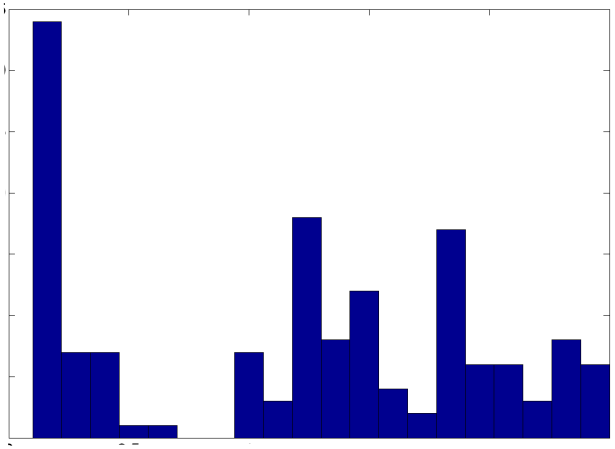


image: wikipedia

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

covered in stats

Data Preprocessing

- Aggregation
- Quantization: Making Discrete or Binary
- Attribute Transformation
- *Dimensionality Reduction*
  - *PCA and LDA* (*look at separately, next time*)

covered in stats

Aggregation

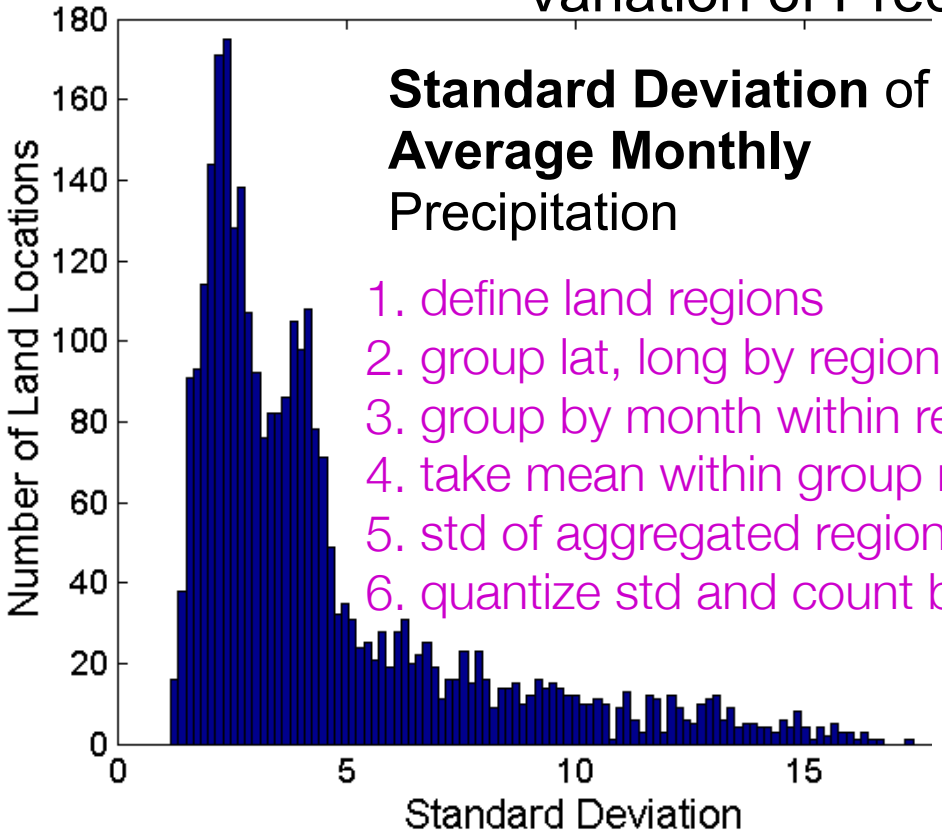
- Combining two or more attributes (or objects) into a single attribute (or object)

- Purpose
  - Data reduction
    - Reduce the number of attributes or objects
  - Change of scale
    - Cities aggregated into regions, states, countries, etc
  - More “stable” data
    - Aggregated data tends to have less variability

covered in stats

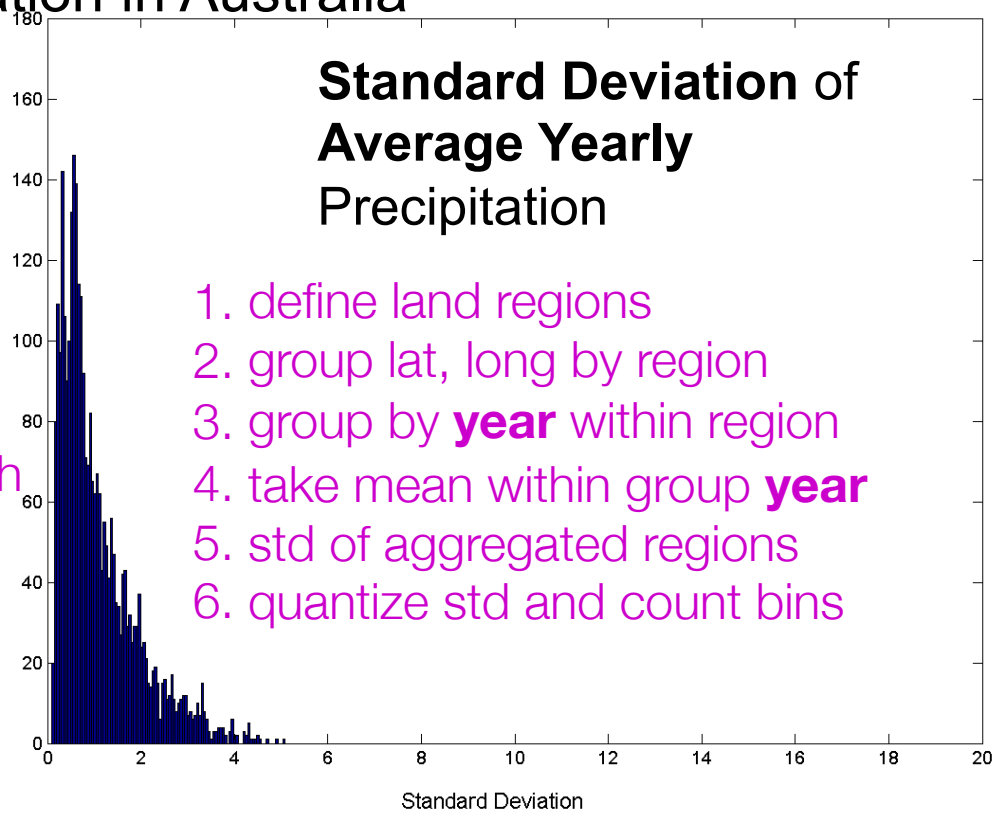
Aggregation

Variation of Precipitation in Australia



Standard Deviation of **Average Monthly** Precipitation

1. define land regions
2. group lat, long by region
3. group by month within region
4. take mean within group month
5. std of aggregated regions
6. quantize std and count bins



Standard Deviation of **Average Yearly** Precipitation

1. define land regions
2. group lat, long by region
3. group by **year** within region
4. take mean within group **year**
5. std of aggregated regions
6. quantize std and count bins

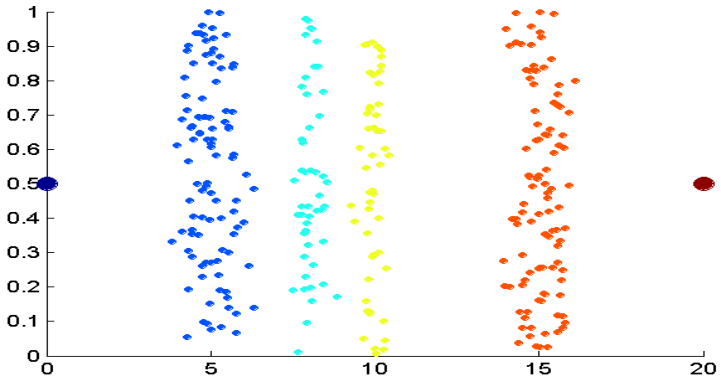
How has aggregation has been used to create these plots?

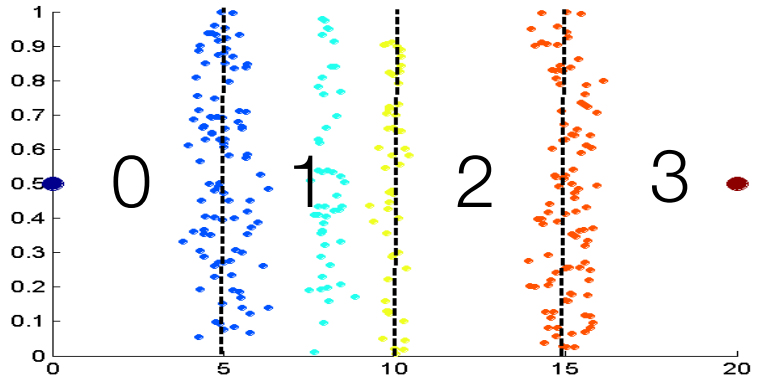
TID	Location	time	measured rainfall
1	*lat, long*	*measured daily*	*X.XX cm*

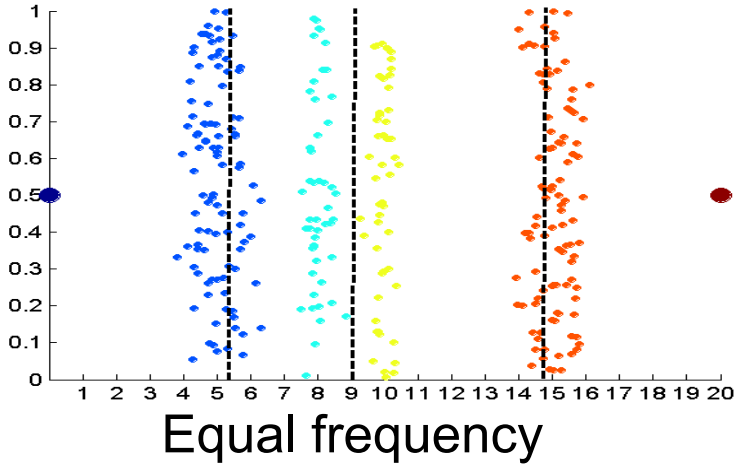
covered in stats

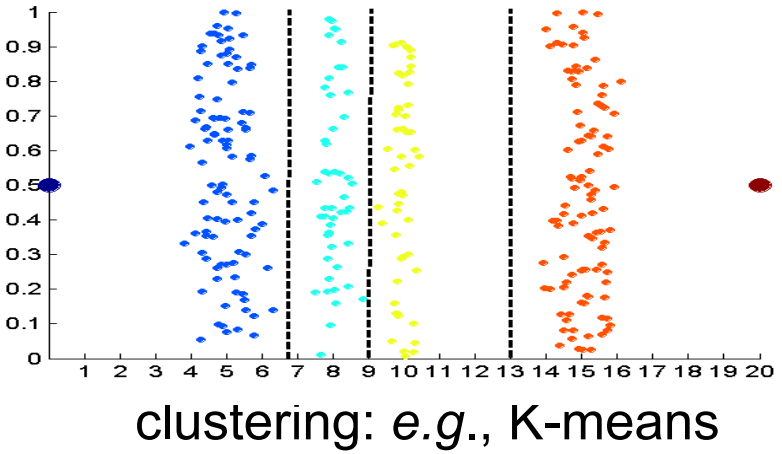
Feature quantization: make ordinal

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










```
num_quantiles = 4
pandas.qcut(dataframe.var, num_quantiles)
```


covered in stats

Attribute Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values
  - Simple functions: x^k, log(x), e^x, |x|
  - Standardization and Normalization
  - Polynomial and Interaction Variables, x[:,1], x[:,1]*x[:,2]

covered in stats

Attribute Transformation in Python

Standardization and Normalization

```
>>> from sklearn import preprocessing
>>> import numpy as np
>>> X = np.array([[ 1., -1.,  2.],
...               [ 2.,  0.,  0.],
...               [ 0.,  1., -1.]])
>>> X_scaled = preprocessing.scale(X)
>>> X_scaled
array([[ 0.  ..., -1.22...,  1.33...],
       [ 1.22...,  0.  ..., -0.26...],
       [-1.22...,  1.22..., -1.06...]])
```

```
>>> import pandas
>>> df_normalized = (df-df.mean())/(df.std())
```

```
>>> scaler = preprocessing.StandardScaler().fit(X)
>>> scaler
StandardScaler(copy=True, with_mean=True, with_std=True)

>>> scaler.mean_
array([ 1. ...,  0. ...,  0.33...])

>>> scaler.std_
array([ 0.81...,  0.81...,  1.24...])

>>> scaler.transform(X)
array([[ 0.  ..., -1.22...,  1.33...],
       [ 1.22...,  0.  ..., -0.26...],
       [-1.22...,  1.22..., -1.06...]])
```


covered in stats

Data Visualization

Visualization

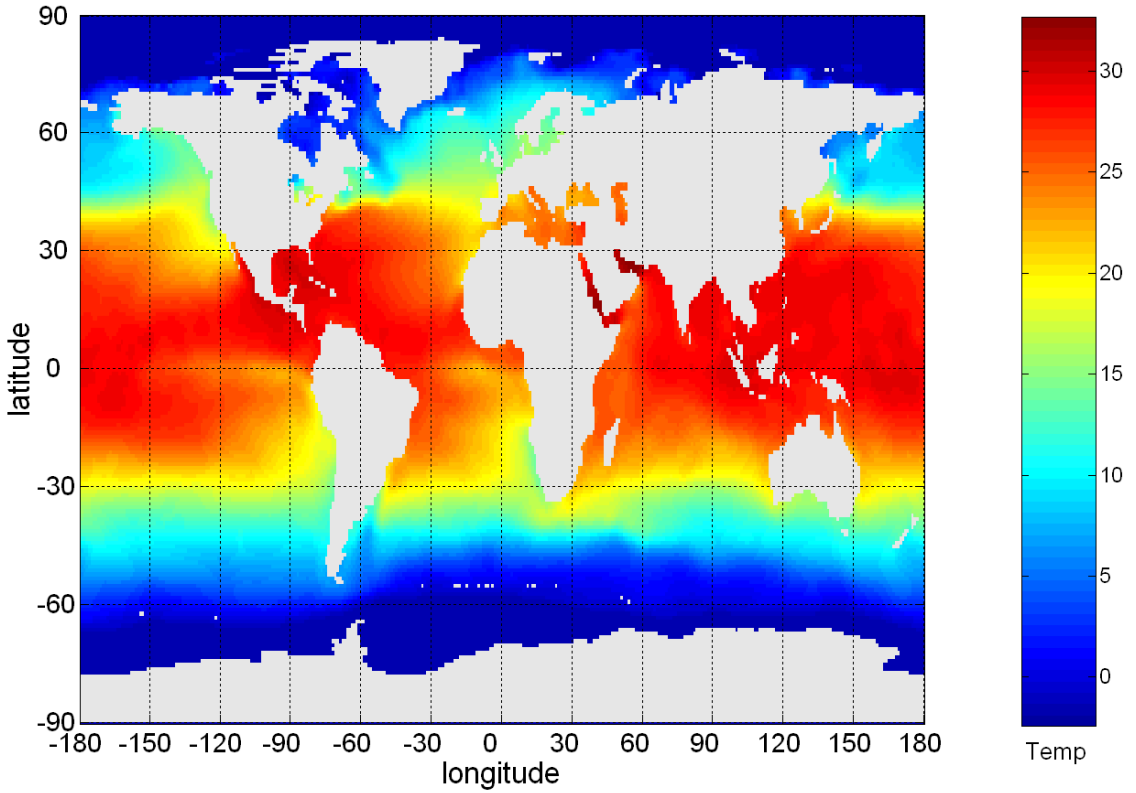
Visualization is the conversion of data into a **visual** or **tabular** format so that the **characteristics** of the data and the **relationships** among data items or attributes can be **analyzed** or reported.

- Visualization of data is one of the most powerful and appealing techniques for data exploration.
  - Humans have a well developed ability to analyze large amounts of information that is presented visually
  - Can detect general patterns and trends
  - Can detect outliers and unusual patterns

section 3: moved up

Example: Sea Surface Temperature

- The following shows the Sea Surface Temperature (SST) for July 1982
  - Tens of thousands of data points are summarized in a single figure



section 3: moved up

Arrangement is important for humans!

- Can make a large difference in how easy it is to understand the data
- Example:

	1	2	3	4	5	6
1	0	1	0	1	1	0
2	1	0	1	0	0	1
3	0	1	0	1	1	0
4	1	0	1	0	0	1
5	0	1	0	1	1	0
6	1	0	1	0	0	1
7	0	1	0	1	1	0
8	1	0	1	0	0	1
9	0	1	0	1	1	0

	6	1	3	2	5	4
4	1	1	1	0	0	0
2	1	1	1	0	0	0
6	1	1	1	0	0	0
8	1	1	1	0	0	0
5	0	0	0	1	1	1
3	0	0	0	1	1	1
9	0	0	0	1	1	1
1	0	0	0	1	1	1
7	0	0	0	1	1	1

section 3: moved up

Selection (people do not think > 3D)

- Need to eliminate or de-emphasize too much data
- Selection may involve choosing a subset of instances
  - A region of the screen can only show so many points
  - Can sample, but want to preserve points in sparse areas
- Selection may involve choosing a subset of attributes
  - Dimensionality reduction is often used to reduce the number of dimensions to two or three
  - Alternatively, pairs of attributes can be aggregated

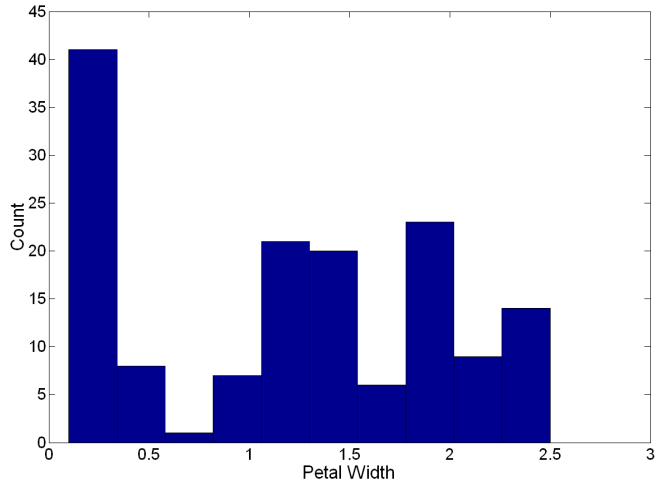
Let's look some graphs

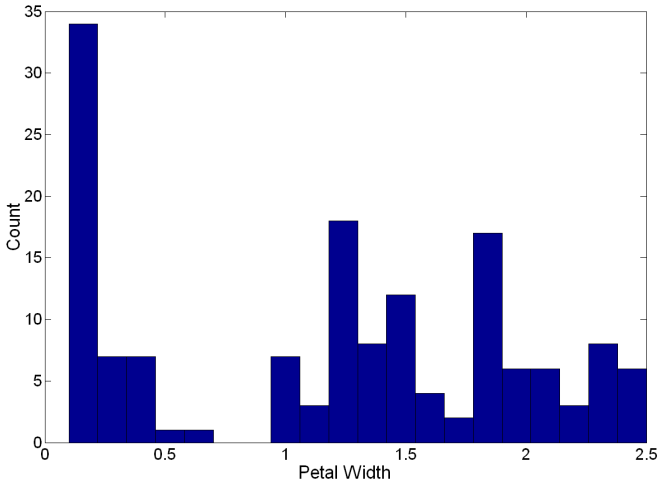
WAKE UP (*please*)

- You tell me what conclusions we are getting from these graphs

Visualization Techniques: Histograms

- Histogram
  - Usually shows the distribution of values of a single variable
  - Divide the values into bins and show a bar plot of the number of objects in each bin.
  - The height of each bar indicates the number of objects
  - Shape of histogram depends on the number of bins
- Example: Petal Width (10 and 20 bins, respectively)

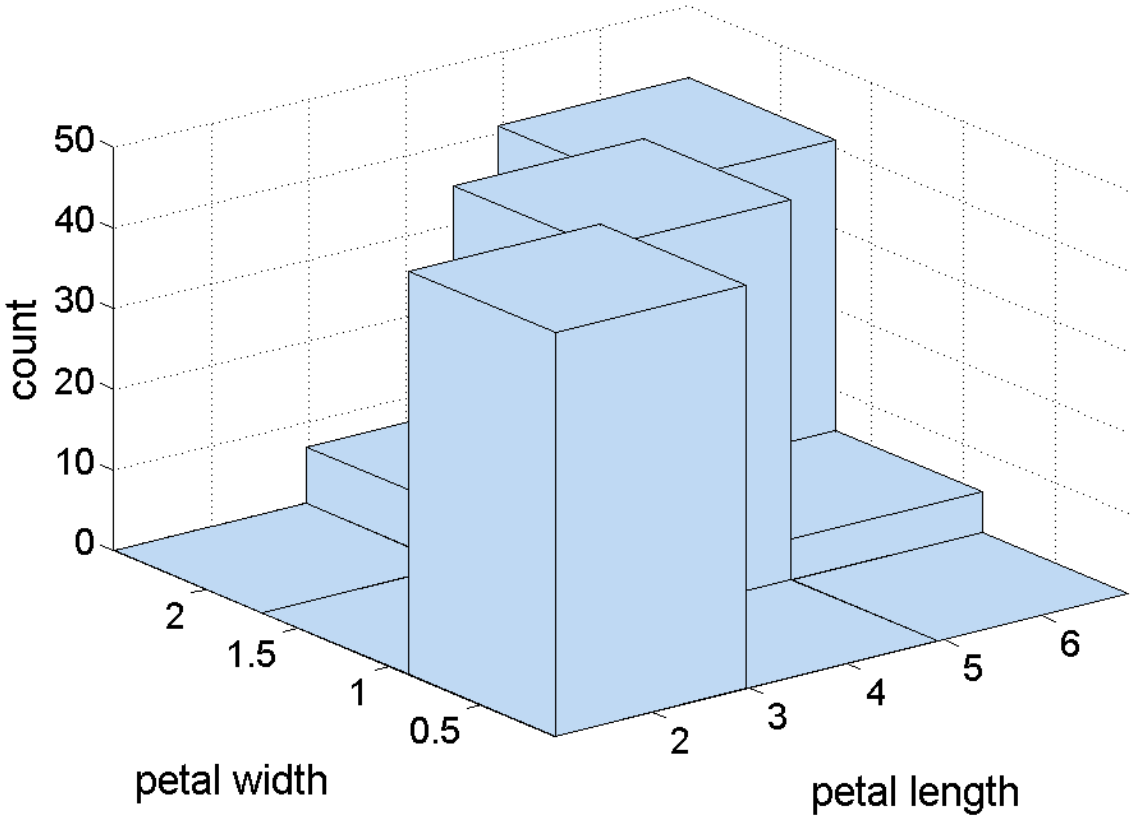




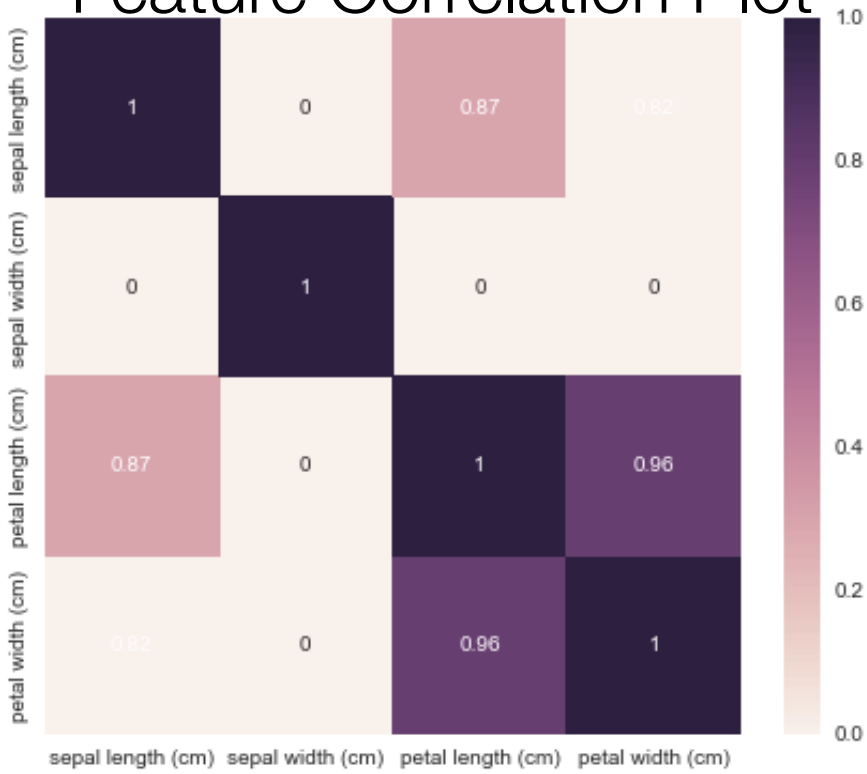
section 3: moved up

Two-Dimensional Histograms

- Estimate the joint distribution of the values of two attributes
- Example: petal width and petal length
  - What does this tell us?



Feature Correlation Plot



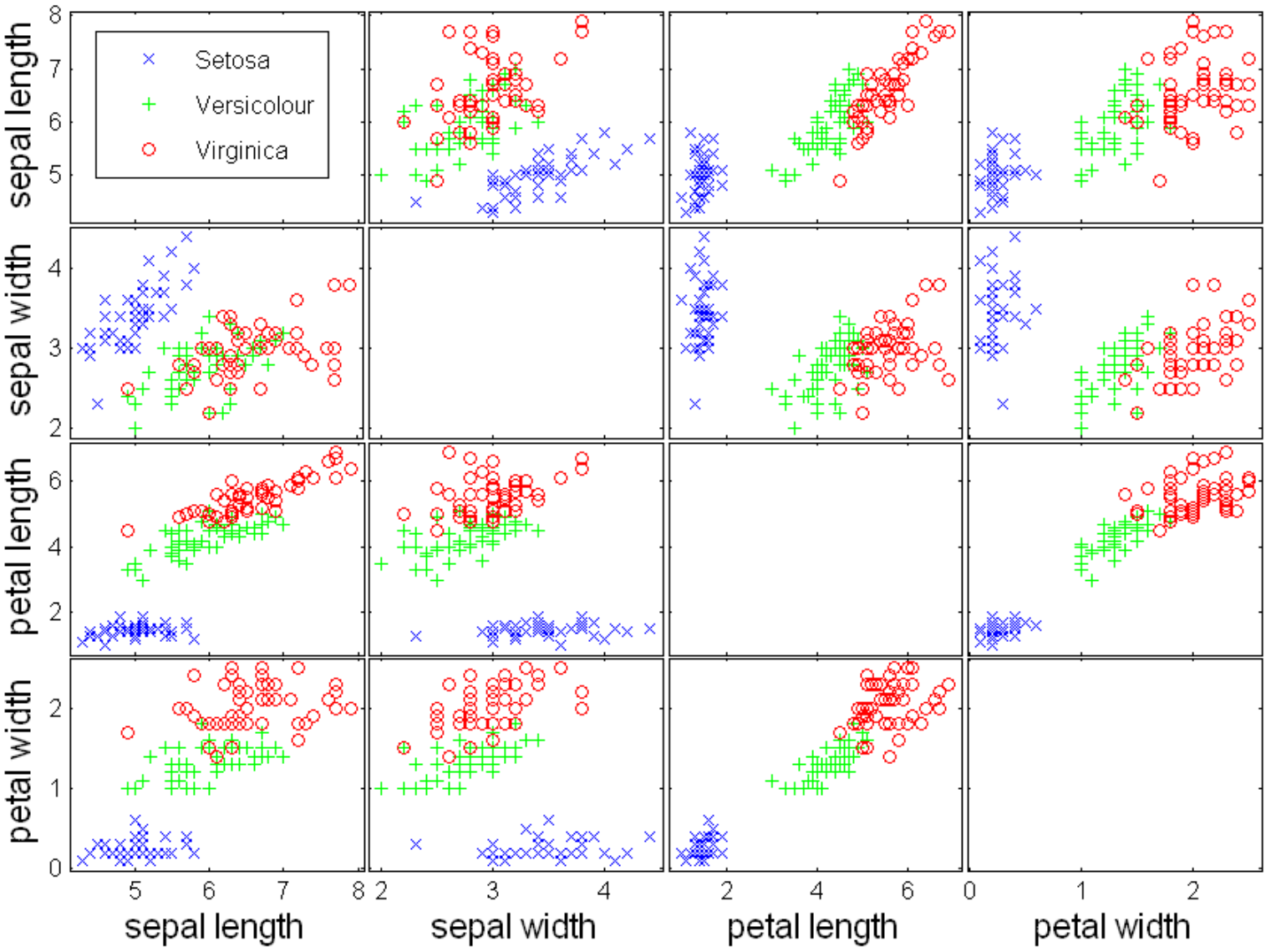
section 3: moved up

Visualization Techniques: Scatter Plots

- Scatter plots
  - Two-dimensional scatter plots most common, but can have three-dimensional scatter plots
  - Additional attributes can be displayed by using the size, shape, and color of the markers that represent the objects
  - It is useful to have arrays of scatter plots can compactly summarize the relationships of several pairs of attributes
  - Good for numeric data, but needs jitter for categorical data

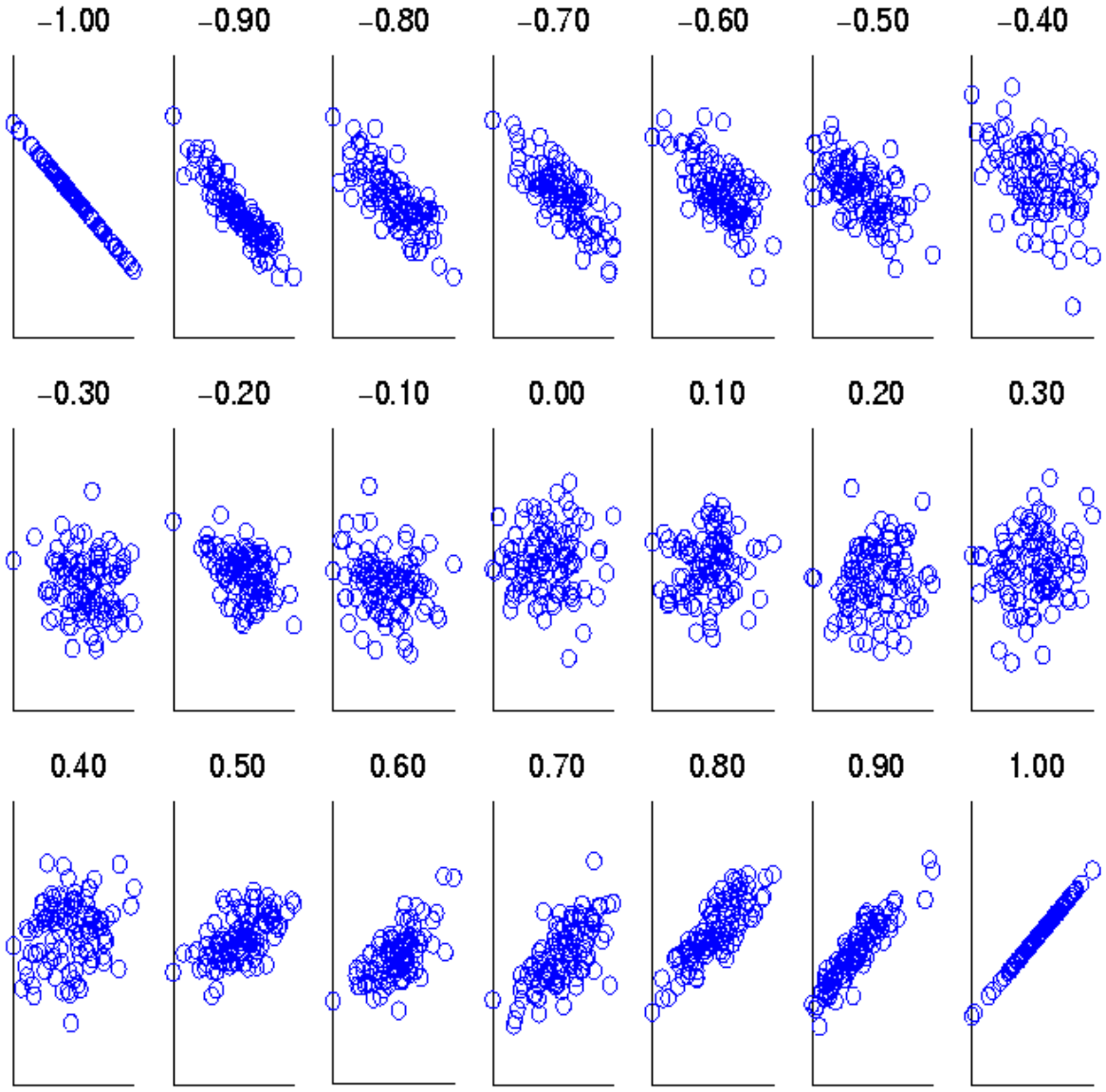
section 3: moved up

Scatter Plot Matrix of Iris Attributes



section 3: moved up

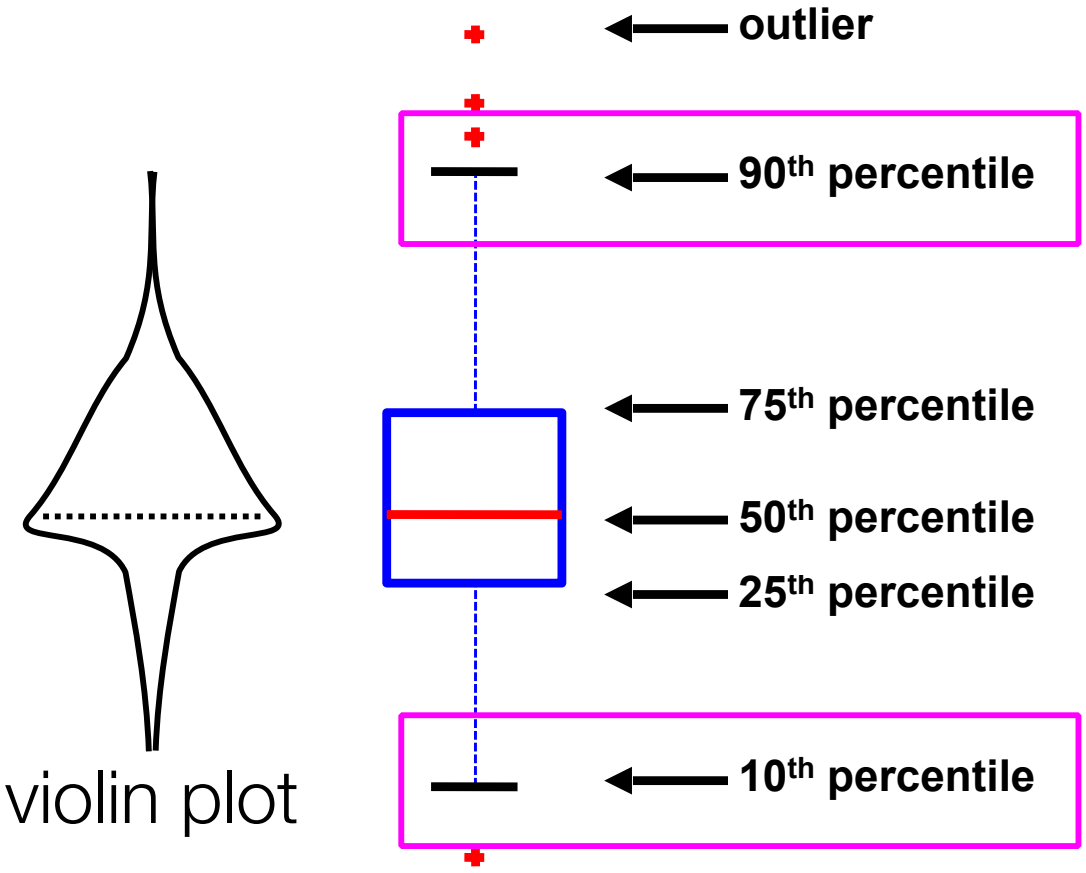
Visually Evaluating Correlation



Scatter plots showing the similarity from –1 to 1.

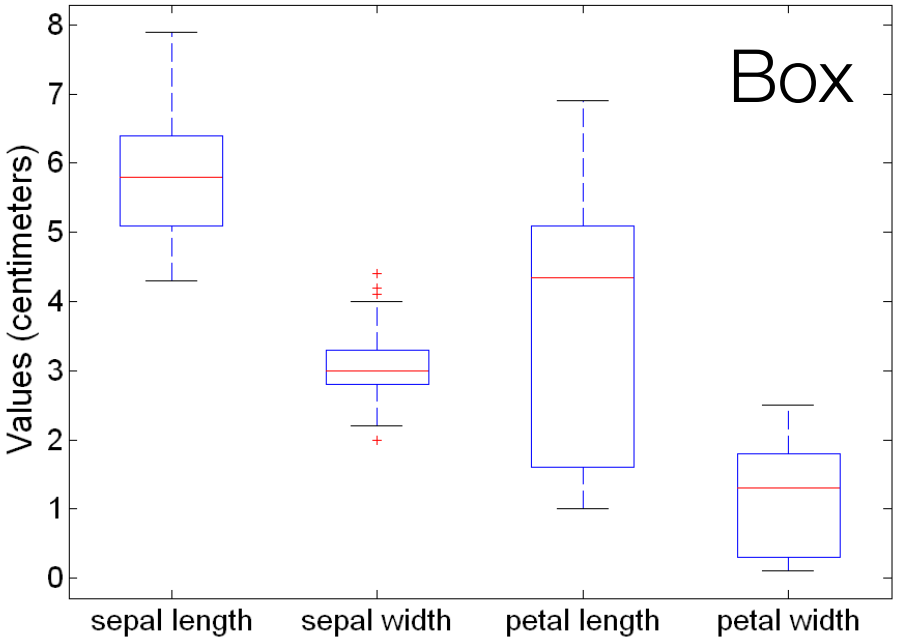
Visualization Techniques: Box Plots

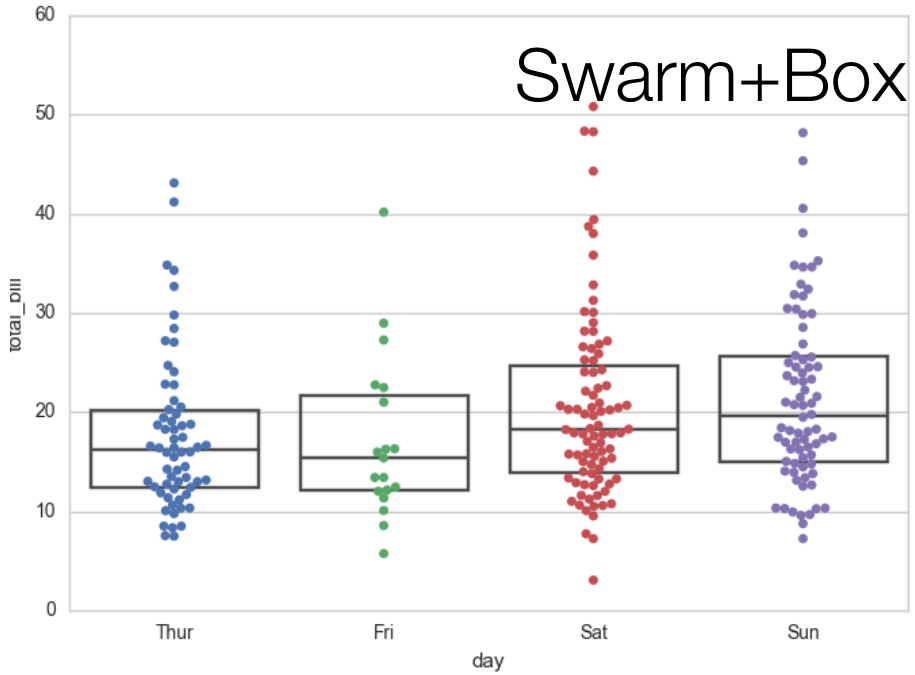
- Box Plots
  - Invented by J. Tukey
  - Another way of displaying the distribution of data
  - Following figure shows the basic part of a box plot

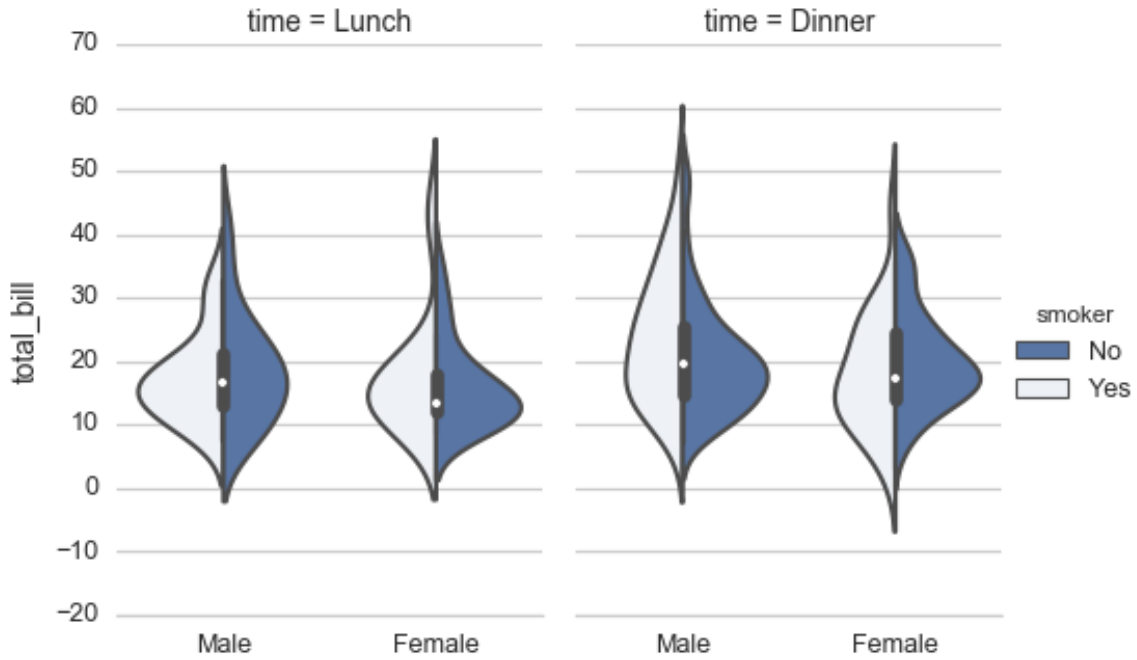


section 3: moved up

Example: Comparing Attributes





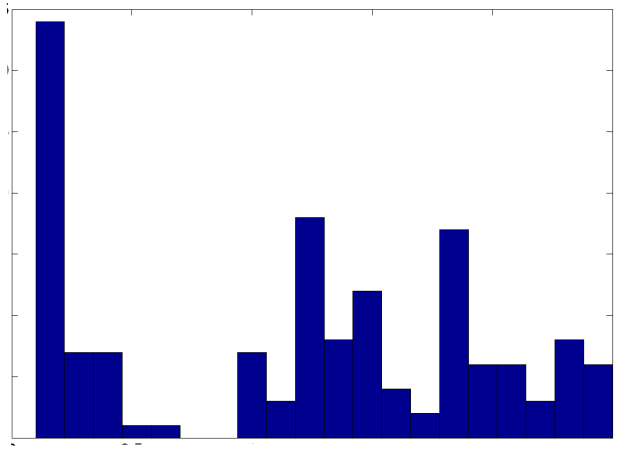


Mixed Violin + Box

section 3: moved up

Self Test 2a.2

What compact visualization is most appropriate for the data in the histogram below?



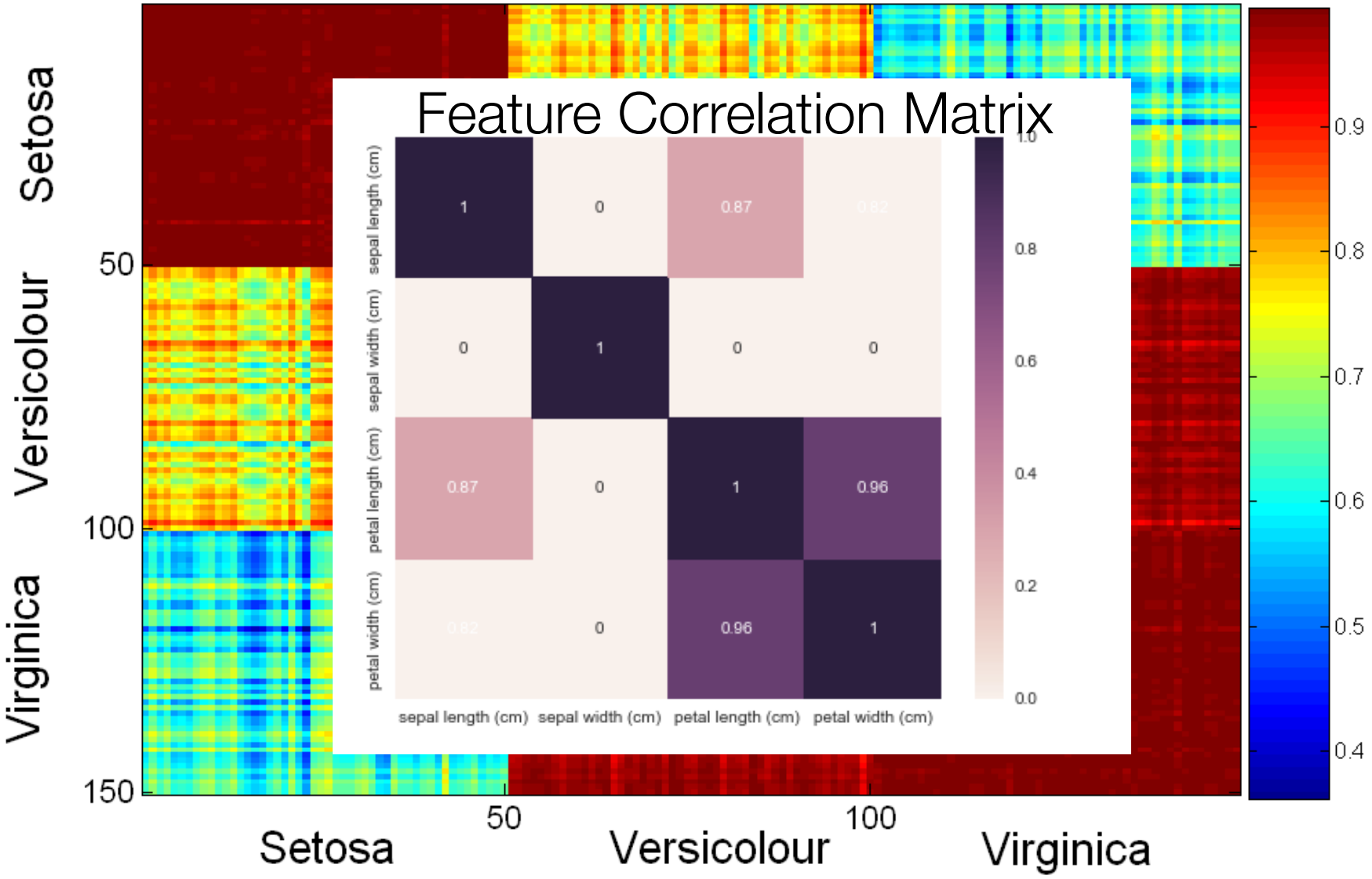
A) Box Plot
B) Violin Plot
C) Swarm Plot
D) None of these

Visualization Techniques: Matrix Plots

- Matrix plots (typically heatmaps)
  - Plot some data matrix
  - This can be useful when objects are sorted well
  - Typically, the attributes are normalized to prevent one attribute from dominating the plot
  - Plots of similarity or distance matrices can also be useful for visualizing the relationships between objects

section 3: moved up

Instance Correlation Matrix



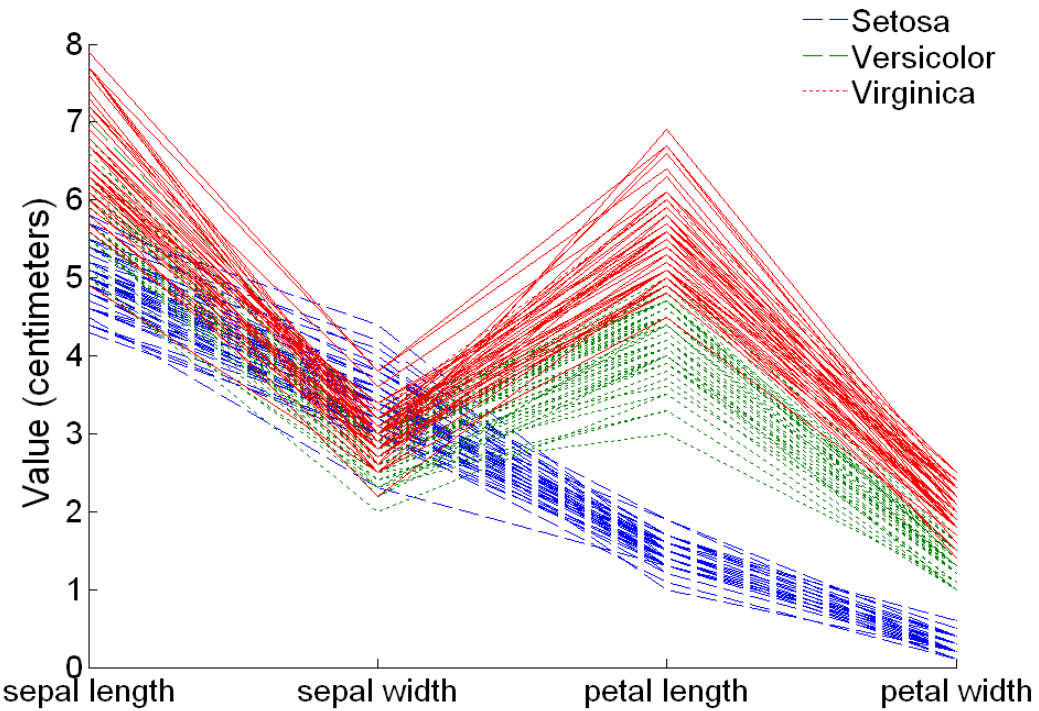
section 3: moved up

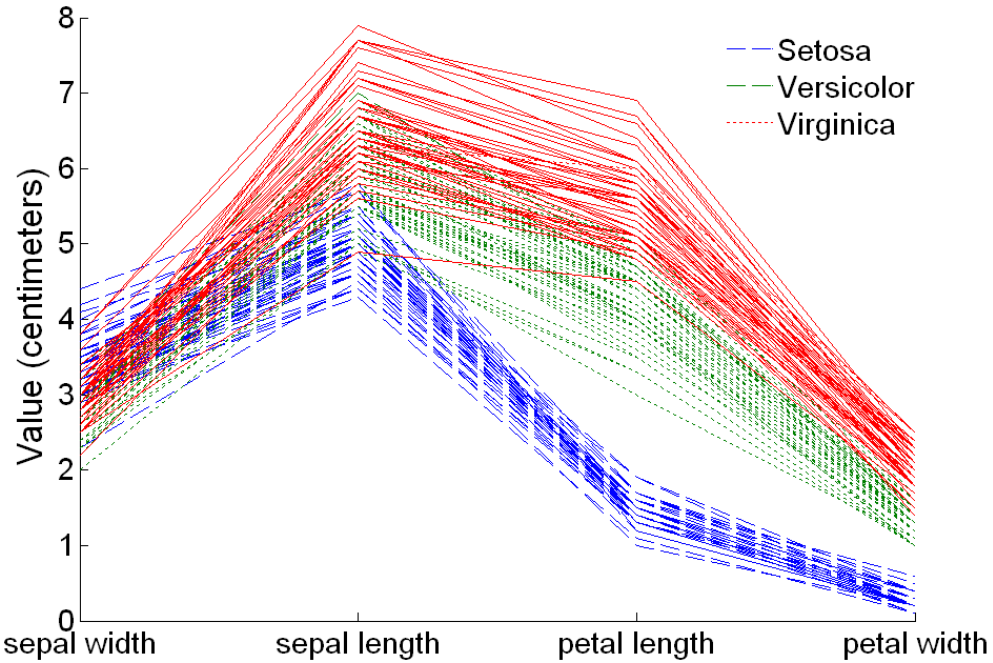
Visualization Techniques: Parallel Coordinates

- Parallel Coordinates
  - Used to plot the attribute values of multi-dimensional data
  - Instead of using perpendicular axes, use a set of parallel axes
  - The attribute values of each object are plotted as a point on each corresponding coordinate axis and the points are connected by a line
  - Thus, each object is represented as a line
  - Often, the lines representing a distinct class of objects group together, at least for some attributes
  - Ordering of attributes is important in seeing such groupings

section 3: moved up

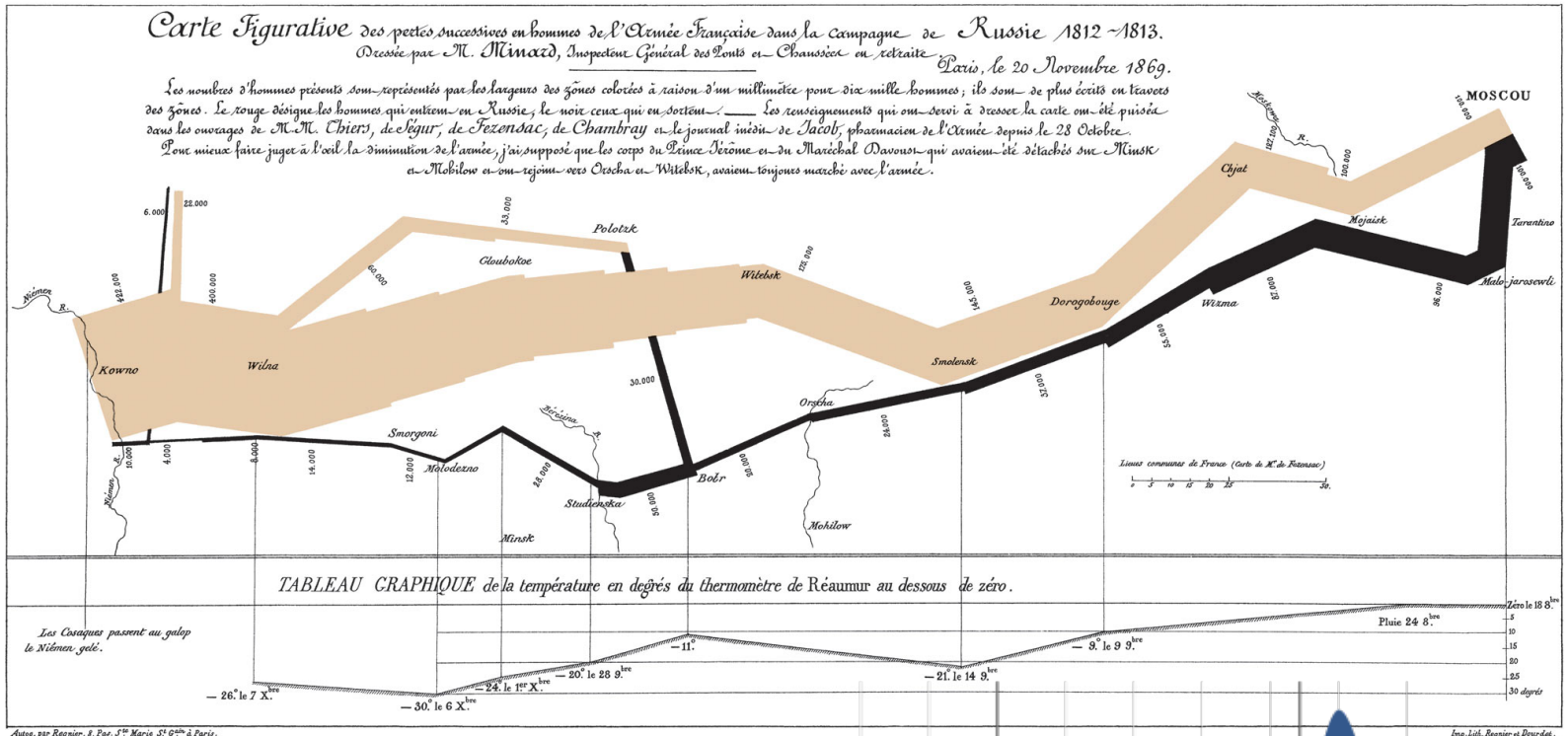
Parallel Coordinates Plots for Iris Data

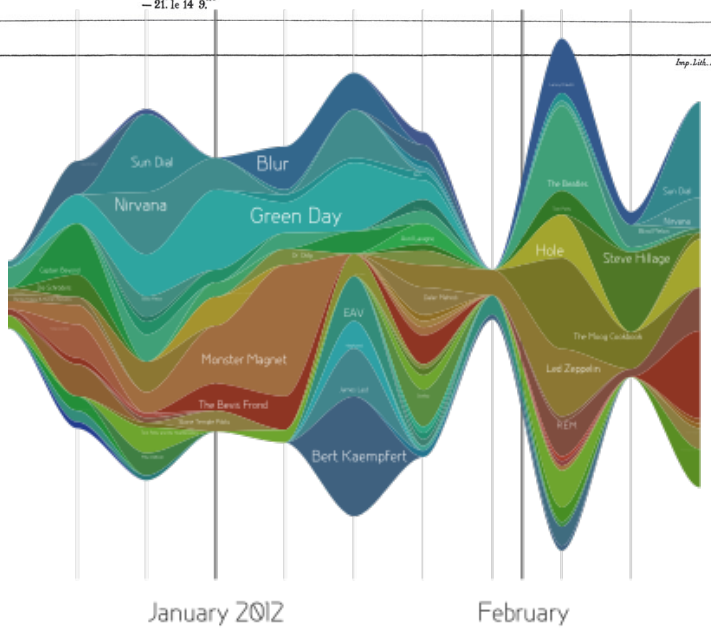




section 3: moved up

There are lots of plots out there!





section 3: moved up

Matplotlib

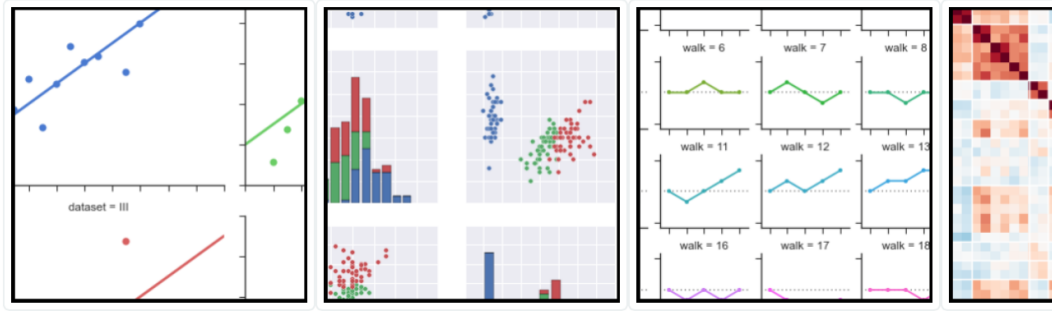
- Python plotting utility
  - Has low level plotting functionality
  - Highly similar to Matlab and R for plotting
- Extended for visually be more beautiful by
  - **seaborn**: stanford data visualization group

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



section 4: moved up

Demo

Visualization

- Matplotlib
- Seaborn
- Plotly (if time)

Other Tutorials:

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

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

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

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

section 4: moved up

For Next Lecture

- Next Time: Data Dimensionality Reduction
- Look at chapter 5 of Python Machine Learning

synchronous