DEPAUL UNIVERSITY

DSC478: Programming Machine Learning Applications

Roselyne Tchoua

rtchoua@depaul.edu

School of Computing, CDM, DePaul University

Understanding Your Data

Before doing any machine learning, you should have an idea of what your data looks like

- What type of data is it?
- What are the features? What types are they? Do they need to be converted into different types? e.g., categorical to numeric
- Are some features correlated?
- Is the data skewed?
- Are there outliers?
- If you have class labels, is it imbalanced? e.g., a few transactions labeled as fraud in a sea of “normal” transactions
- Part of visualization is about whether you can easily digest and convey the characteristics of your data

Types of Dataset

- **Record Data**
  - Relational records
  - **Data matrix, similar to records but all numeric data**
  - Document data: text documents: term-frequency vector
  - Transaction data
- **Graph and network**
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- **Ordered**
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data
- **Spatial and multimedia**:
  - Spatial data: maps
  - Image + Video + Text data…

Data Objects

- Data sets are made up of data objects.
- A **data object** represents an entity.
- Examples:
  - sales database: object → customers, store items, sales
  - medical database: object → patients, treatments
  - university database: object → students, professors, courses
- Also called *samples , examples, instances, data points, objects, tuples, vectors*.
- Data objects are described by **attributes**.
- Database rows → data objects; columns → attributes.

Attributes

- **Attribute (**or **dimensions, features, variables**): a data field representing a characteristic or property of a data object
  - *E.g., customer _ID, name, address, income, GPA, ….*
- Types:
  - Nominal (Categorical)
    - e.g., Gender (M,F), Movie Genre (Action, Drama, Comedy etc.)
  - Ordinal (Ordered categories)
    - e.g., Army ranks, Age groups, Grades
  - Numeric: quantitative
    - Interval-scaled: e.g., dates (can add/subtract but cannot multiply)
    - Ratio-scaled: e.g., length, counts (distinct, ordered and can +, - , * and /)

Data Objects and Attributes

Attributes

Objects

ID	Outlook	Temperature	Humidity	Windy
1	sunny	85	85	FALSE
2	sunny	80	90	TRUE
3	overcast	83	78	FALSE
4	rain	70	96	FALSE
5	rain	68	80	FALSE
6	rain	65	70	TRUE
7	overcast	58	65	TRUE
8	sunny	72	95	FALSE
9	sunny	69	70	FALSE
10	rain	71	80	FALSE
11	sunny	75	70	TRUE
12	overcast	73	90	TRUE
13	overcast	81	75	FALSE
14	rain	75	80	TRUE

Example of Record Data

Data Matrices: Bag-of-Word

- Data is an m by n matrix, where there are m rows (objects), one for each object, and n columns, one for each attribute
- For text data, a common way to represent a document (text message, email, document) is a vector of the frequencies of the words it contains (0 for words in the corpus that do not occur in that document)

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Feature vector

Text Data Matrices

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute
- Data is an m by n matrix, where there are m rows (objects), one for each object, and n columns, one for each attribute

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Feature vector

Basic Statistical Description of Data

Before deeper analysis, it's important to **explore** the basic characteristics and relationships in the data set

- **Descriptive Statistics**
  - To better understand the characteristics of attributes and fields: central tendency, variation, spread, etc.
  - To get a feel for general patterns or relationships among variables: e.g., correlation, covariance, etc.
- **Data Visualization**
  - Visual examination of data distributions often help in uncovering important patterns and guide further investigation or decision making

Measuring the Central Tendency

- Mean (algebraic measure) (sample vs. population):

  $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N}$$

  Note: *n* is sample size and *N* is population size.

  - Weighted arithmetic mean:

    $$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

  - Trimmed mean: chopping extreme values

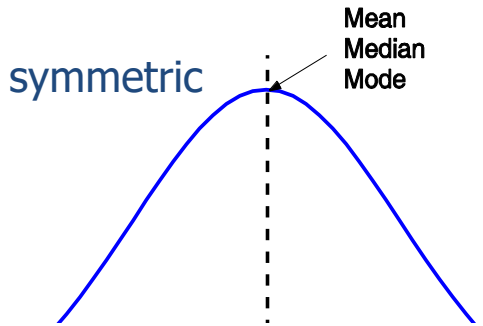
- Median:
  - Middle value if odd number of values, or average of the middle two values otherwise
  - Estimated by interpolation (for *grouped data*):

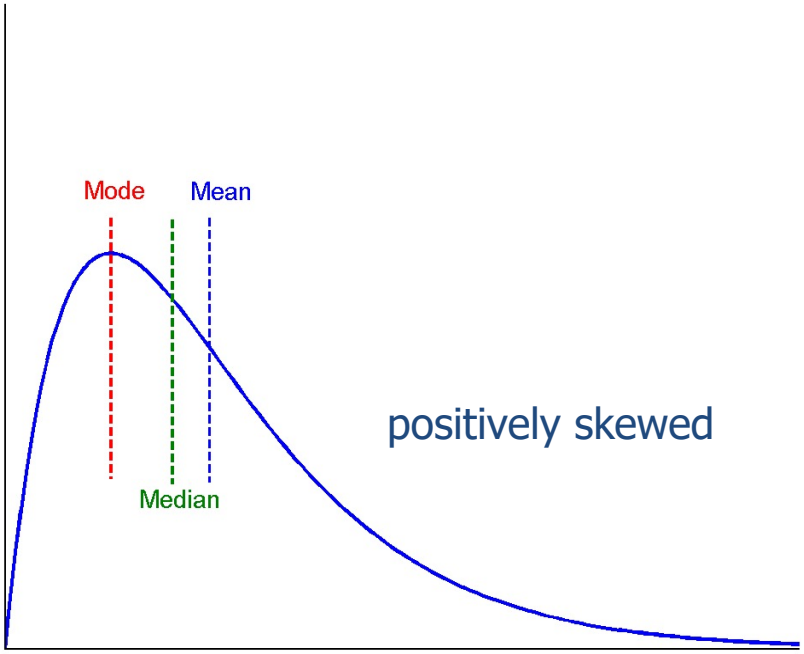
	age	*frequency*
	1–5	200
	6–15	450
	16–20	300
Median interval →	21–50	1500
	51–80	700
	81–110	44

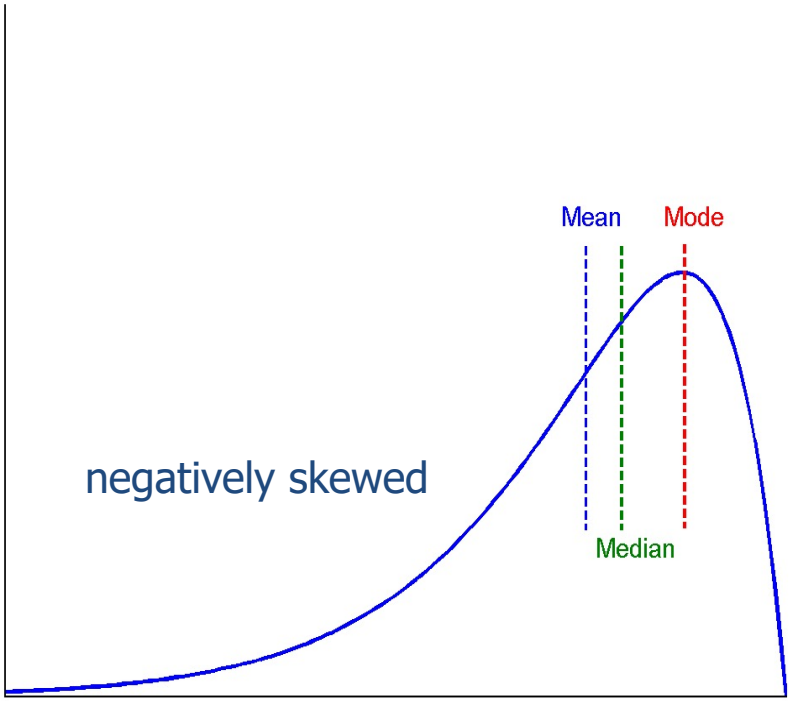
- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula: $mean - mode = 3 \times (mean - median)$
- Midrange: $\text{Midrange} = \frac{\text{Max}(x) + \text{Min}(x)}{2}$ (susceptible to outliers)

Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data







Measuring the Dispersion of Data

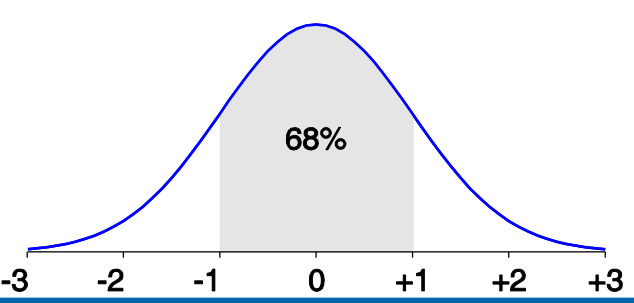
- Quartiles, outliers and boxplots
  - Quartiles: Q1 (25th percentile), Q3 (75th percentile)
  - Inter-quartile range: **IQR = Q3 – Q1**
  - Five number summary: **min, Q1, median, Q3, max**
  - Boxplot: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
  - Outlier: usually, a value **lower than Q1 - 1.5 x IQR and higher than Q3 + 1.5 x IQR**
- Variance and standard deviation
  - Variance or s^2: (algebraic, scalable computation)

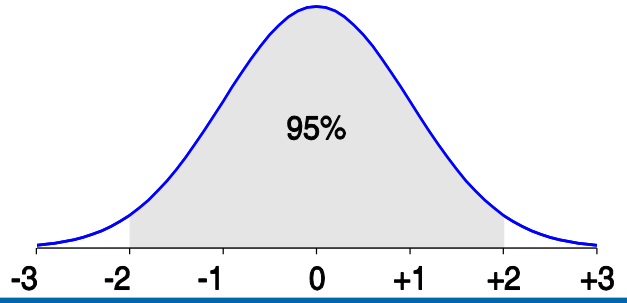
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2] \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n} x_i^2 - \mu^2$$

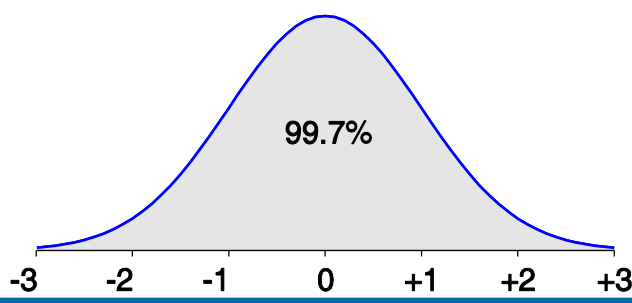
  - Standard deviation is the square root of variance

Properties of Normal Distribution Curve

- The normal (distribution) curve
- From μ–σ to μ+σ: contains about 68% of the measurements (μ: mean, σ: standard deviation)
- From μ–2σ to μ+2σ: contains about 95% of it
- From μ–3σ to μ+3σ: contains about 99.7% of it
- Depending on problem 3σ *may be high*

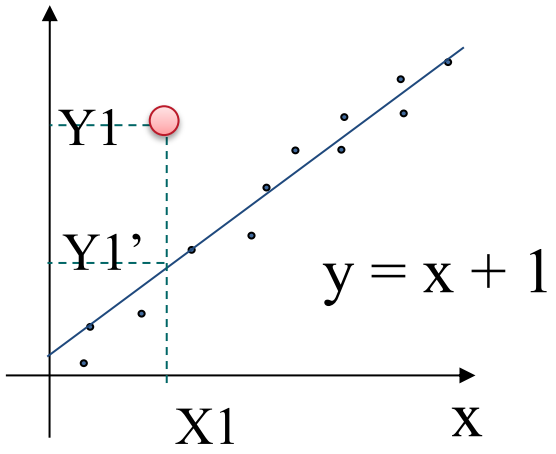


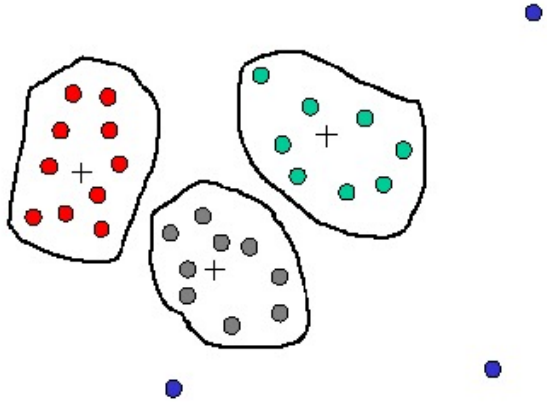




Outlier Detection

- Univariate case: use IQR or StanDev
- Multivariate case, use:
  - Regression: Errors $\geq$ some threshold
  - Clustering: Points that do not belong to any group



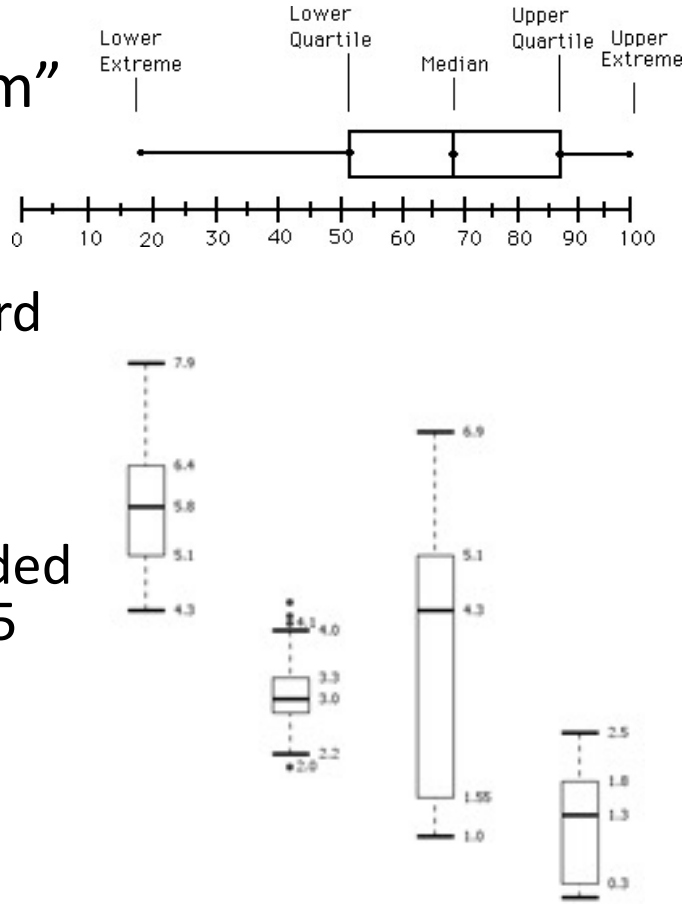


Graphic Display of Basic Stat. Description

- **Boxplot**: graphic display of five-number summary
- **Histogram**: x-axis ➔ values, y-axis ➔ frequencies
- **Quantile** plot: each value x_i is paired with f_i indicating that ~ 100 f_i % of data are $\leq x_i$
- **Quantile-quantile** (**q-q**) plot graphs the quantiles of one univariant distribution against the corresponding quantiles of another
  - Giving you an at-a-glance idea whether two vars may come from similar distribution
- **Scatter plot**: each pair of values is a pair of coordinates and plotted as points in the plane

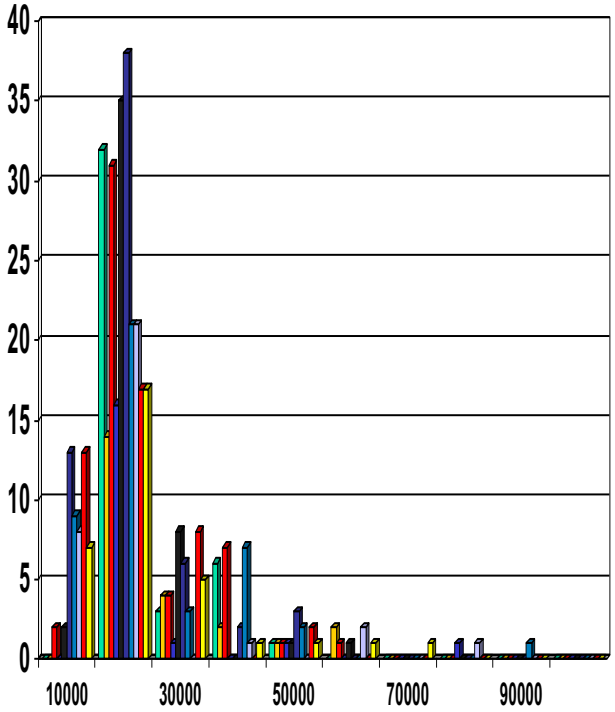
Boxplot

- Five-number summary of a distribution
  "Minimum", Q1, Median, Q3, "Maximum"
- Boxplot
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extended to Minimum and Maximum (within [Q1-1.5 IQR, Q3+1.5IQR])
  - Outliers: points beyond a specified outlier threshold, plotted individually



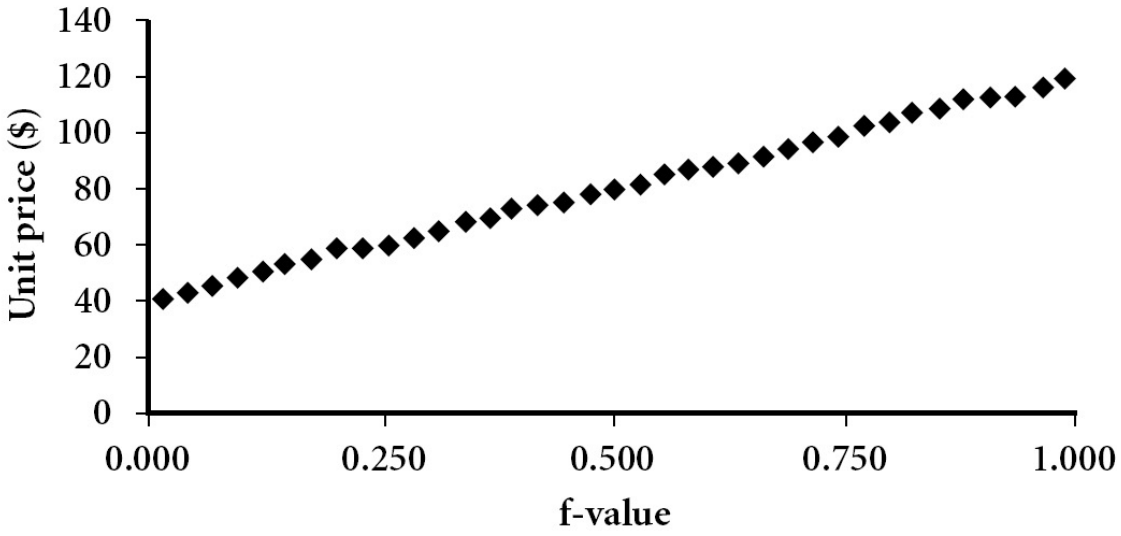
Histograms

- Histogram: displays the frequency distribution of continuous data. It indicates the number of observations that lie in-between a range of values (bin)
- The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent



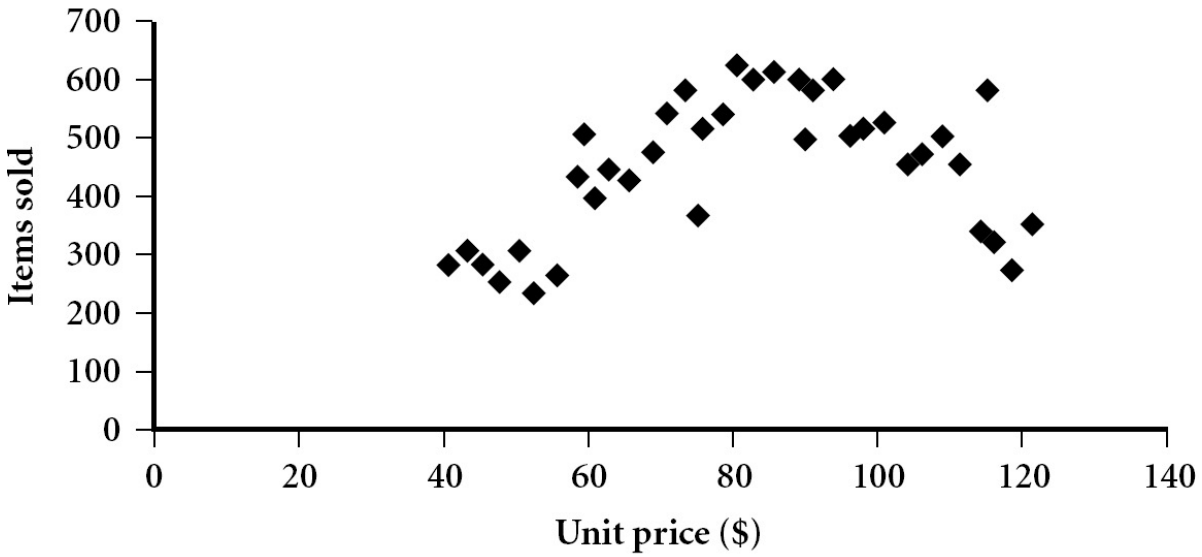
Quantile Plot

- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots quantile information
- For a data x_i data sorted in increasing order, f_i indicates that approximately 100 fi% of the data are below or equal to the value x_i



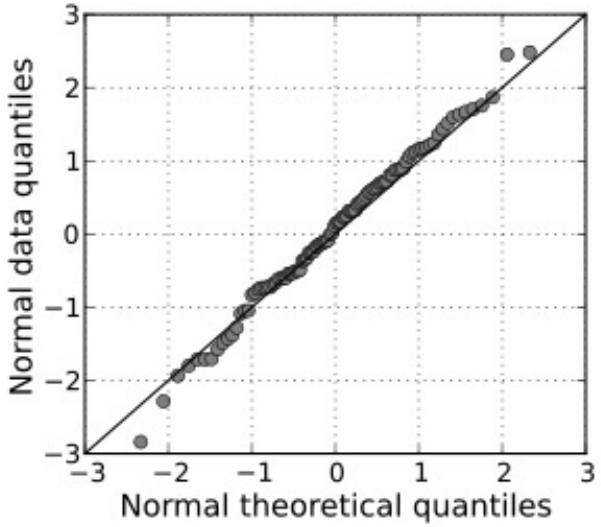
Scatter Plot

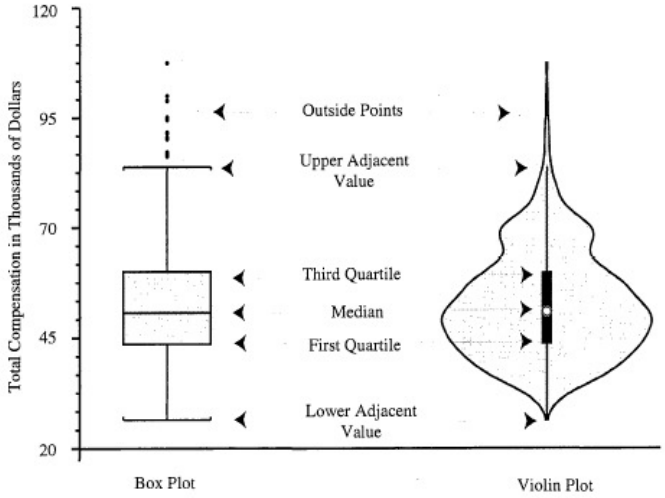
- Provides a first look at bivariate data to see clusters of points, outliers, etc.
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



Other Useful Stat. Plots for Projects

- Violin plot (to see what boxplots maybe hiding)
- QQ plots





https://towardsdatascience.com/violin-plots-explained-fb1d115e023d

Correlations (Categorical)

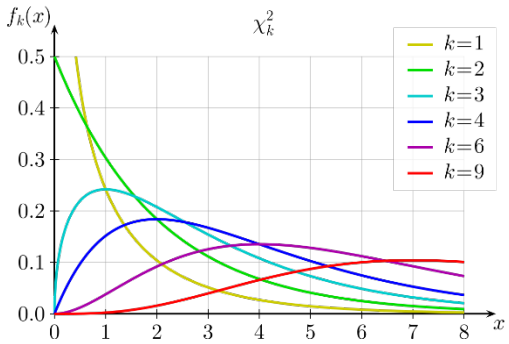
- $Chi-square\ (\chi^2) test$

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- Chi-squared test is used to determine whether there is a statistically significant difference between the expected frequencies and the observed frequencies in one or more categories of a contingency table
- The larger the χ^2 value, the more likely the variables are related
- The cells that contribute the most to the χ^2 value are those whose actual count is very different from the expected count
- **Correlation does not imply causality**
  - # of hospitals and # of car-theft in a city are correlated
  - Both are causally linked to the third variable: population

Correlations (Categorical Example)

	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500



- χ^2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- It shows that like_science_fiction and play_chess are correlated in the group

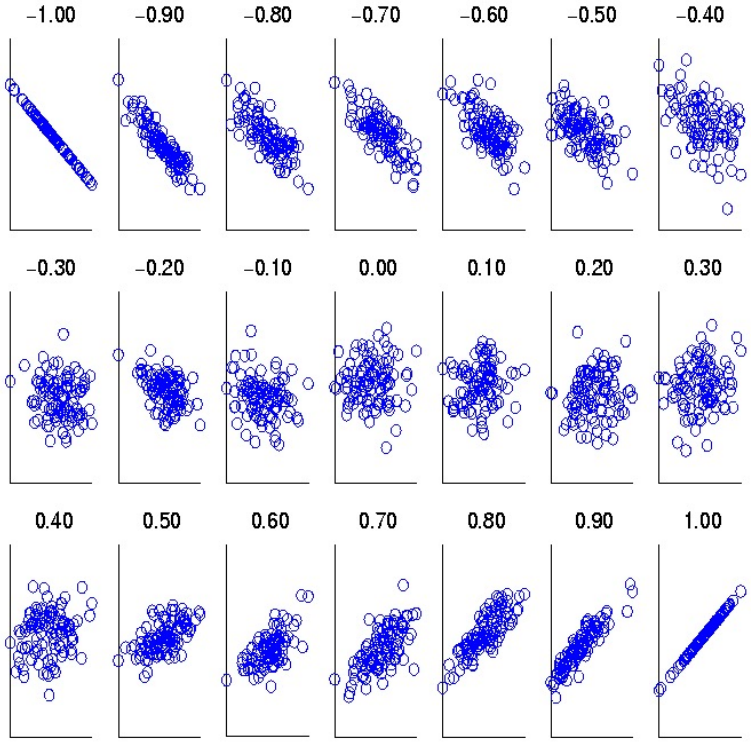
Correlations (Numeric)

- Correlation coefficient (also called Pearson's product moment coefficient)

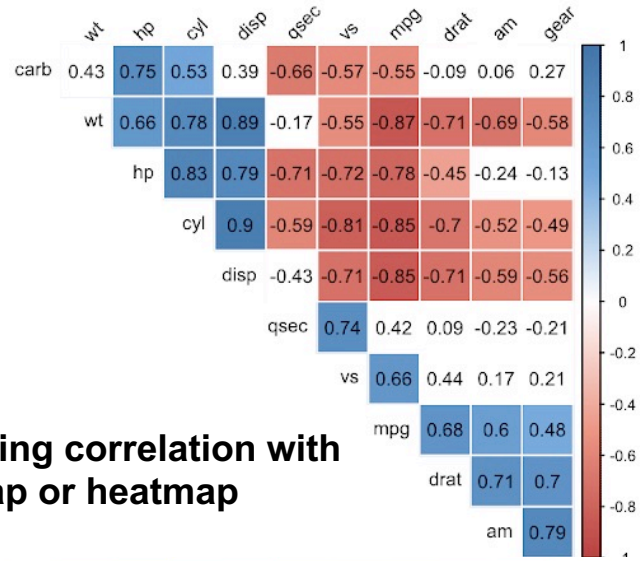
$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{(n-1)\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{(n-1)\sigma_A\sigma_B}$$

- where n is the number of tuples, and are the respective means of A and B, σA and σB are the respective standard deviation of A and B, and Σ(aibi) is the sum of the AB cross-product.
- If rA,B > 0, A and B are positively correlated (A's values increase as B's). The higher, the stronger correlation.
- rA,B = 0: independent; rAB < 0: negatively correlated

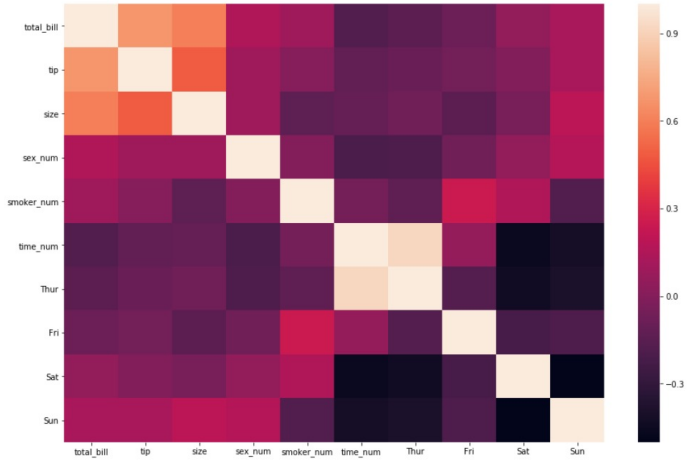
Visualizing Correlations



Scatter plots showing the correlations from –1 to 1.



Visualizing correlation with colormap or heatmap



Correlation viewed as a linear relationship

- Correlation measures the linear relationship between objects
- To compute correlation, we standardize data objects, A and B, and then take their dot product

$$a'_k = (a_k - mean(A)) / std(A)$$

$$b'_k = (b_k - mean(B)) / std(B)$$

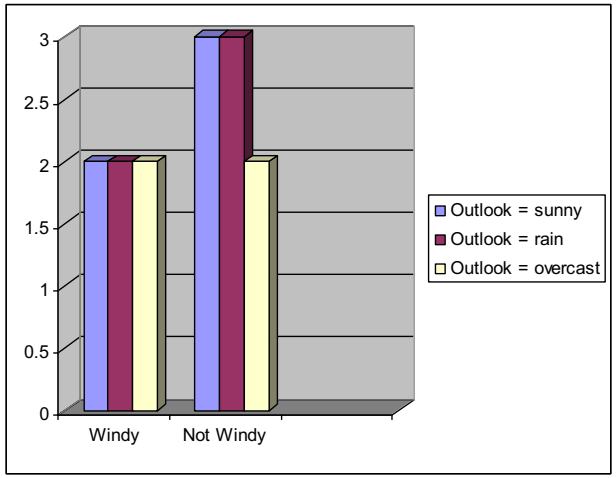
$$correlation(A, B) = A' \bullet B'$$

Visualizing Patterns Using Aggregation

Visualizing cross-tabulation (contingency table)

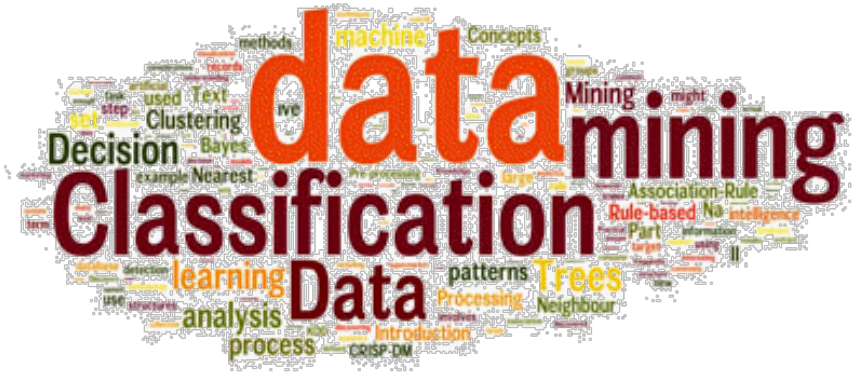
	Windy	Not Windy
Outlook = sunny	2	3
Outlook = rain	2	3
Outlook = overcast	2	2

ID	Outlook	Temperature	Humidity	Windy
1	sunny	85	85	FALSE
2	sunny	80	90	TRUE
3	overcast	83	78	FALSE
4	rain	70	96	FALSE
5	rain	68	80	FALSE
6	rain	65	70	TRUE
7	overcast	58	65	TRUE
8	sunny	72	95	FALSE
9	sunny	69	70	FALSE
10	rain	71	80	FALSE
11	sunny	75	70	TRUE
12	overcast	73	90	TRUE
13	overcast	81	75	FALSE
14	rain	75	80	TRUE



Other Types of Visualization

- Understanding Properties of Text
  - Zipf distribution
  - TF x IDF
  - Tag/Word Clouds



- Graph Visualization