Tufts

CS135
Introduction to Machine Learning

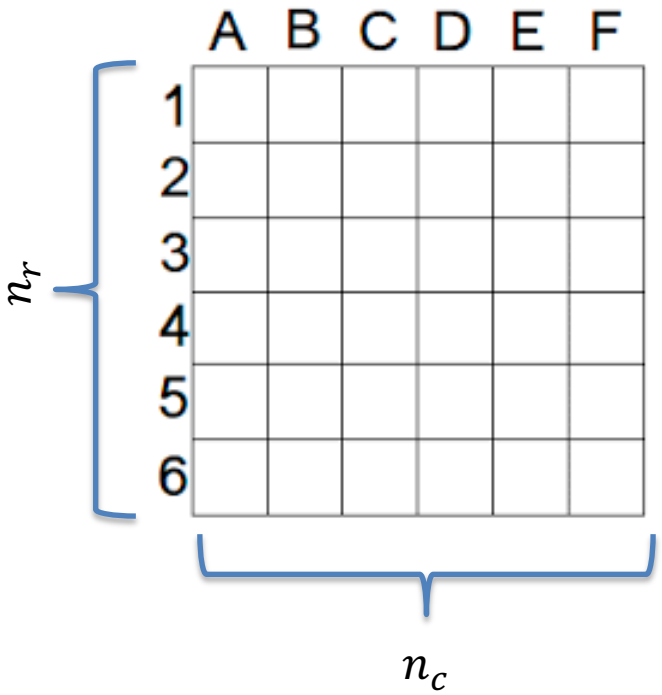
Lecture 3: Types of Data. What is Data?

Objectives

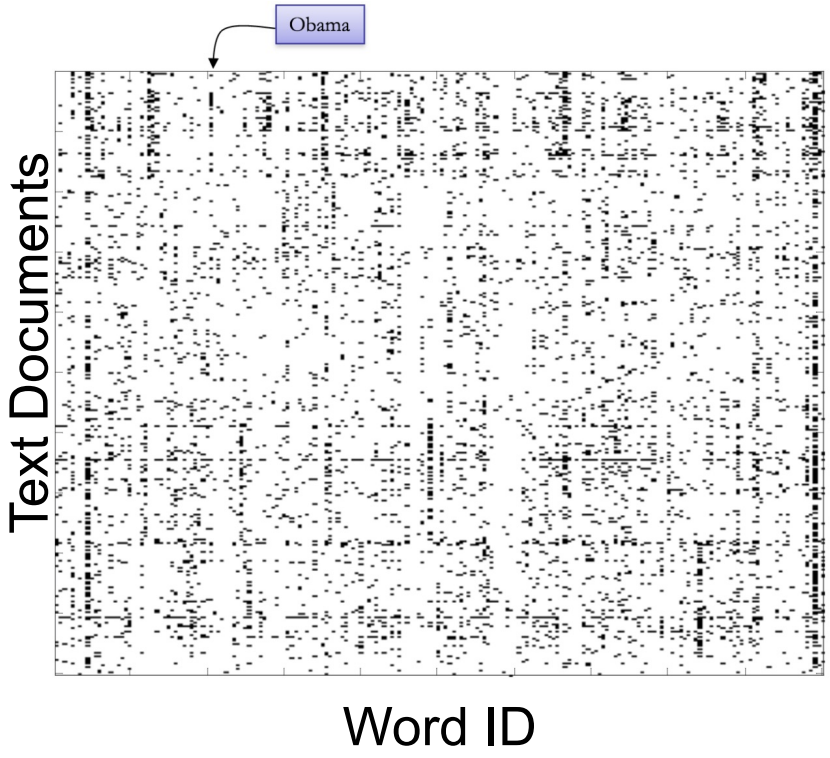
- Introduce different **Types of Data**
- Cover **Data Preparation** and where it fits in in the modeling process
- Discuss **Data Quality**
- Focus on a key part of data preparation
  - **Exploratory data analysis**
    - Identify data glitches and errors
    - Understanding the data
    - Identify possible transformations
- Handling **missing data**
- Review **tools** data preparation

Flat File Data (aka Tabular)

- Rows are objects
- Columns are measurements or attributes of objects
- n_r and n_c can grow really large



Text Data



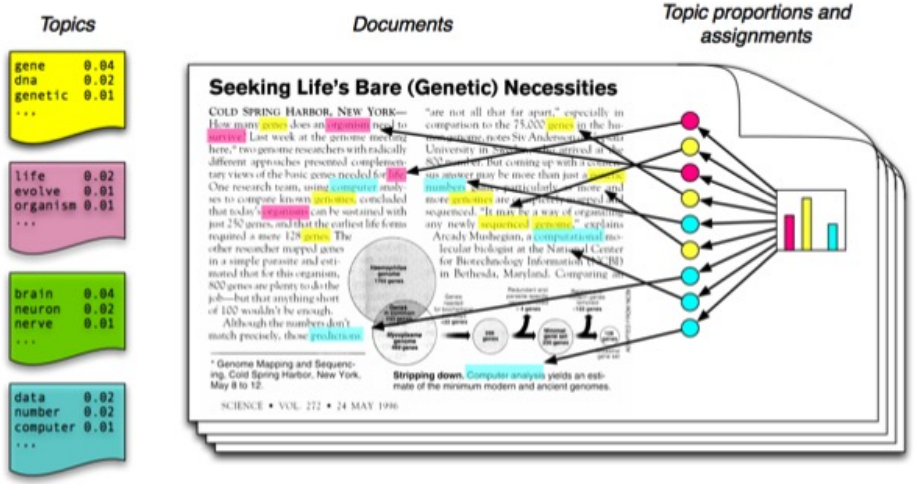
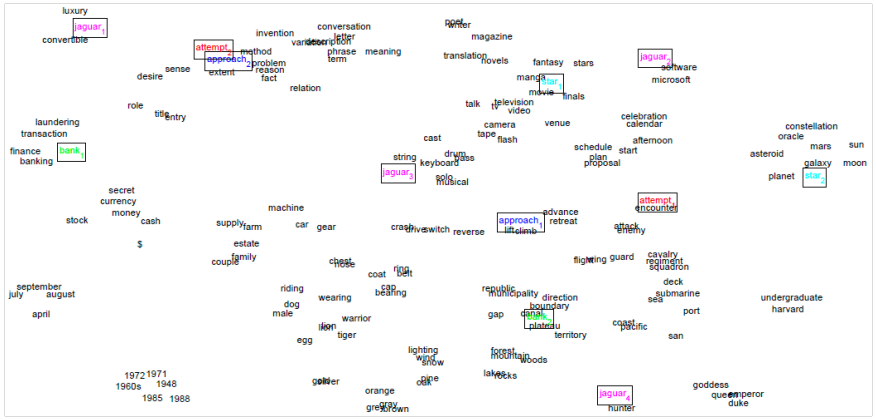


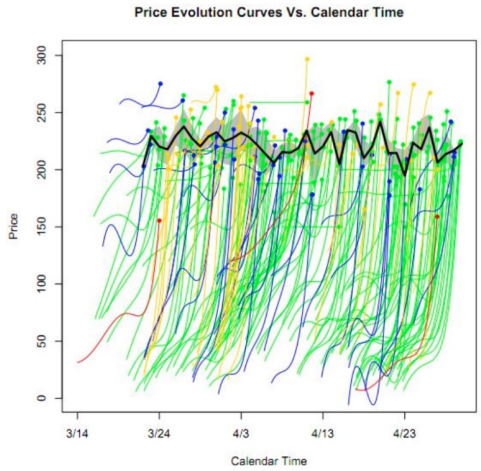
Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.



Word Embedding

Tufts

Time Series Data



Jank, Shmueli, et al (2005)

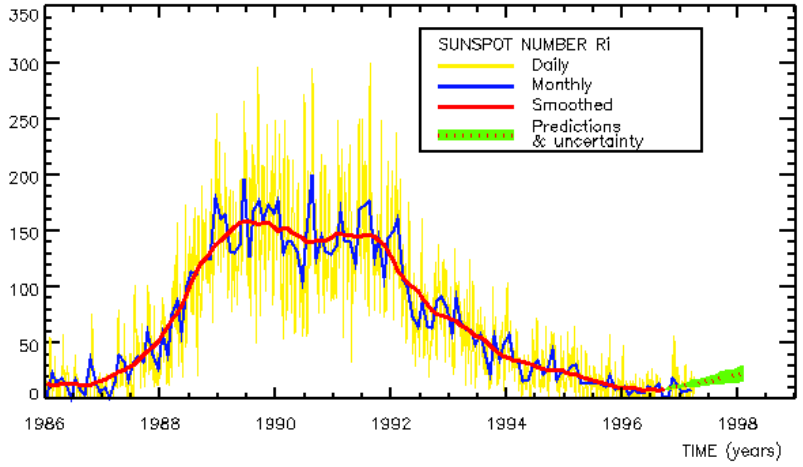
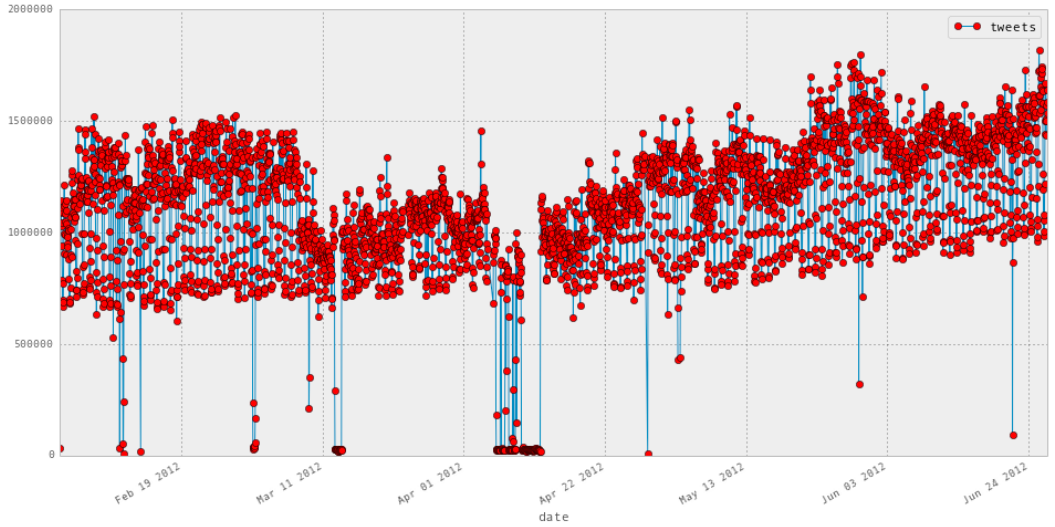
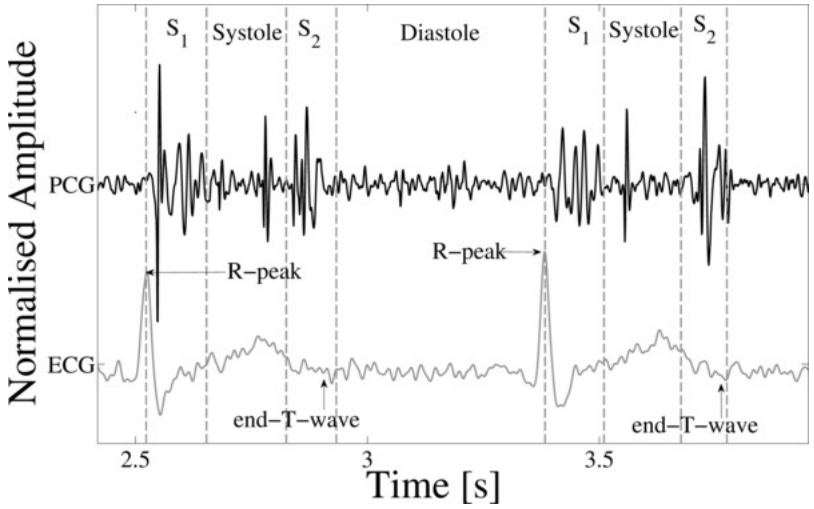


Fig. 10. Rug Plot displaying the price evolution (y-axis) of 217 online auctions over calendar time (x-axis) during a 3-month period. The colored lines show the price path of each auction with color indicating auction length (yellow = 3-day; blue=5-day; green = 7-day; red = 10-day). The dot at the end of each line indicates the final price of the auction. The black line represents the average of the daily closing price , and the gray band is the inter-quartile range.



Tufts

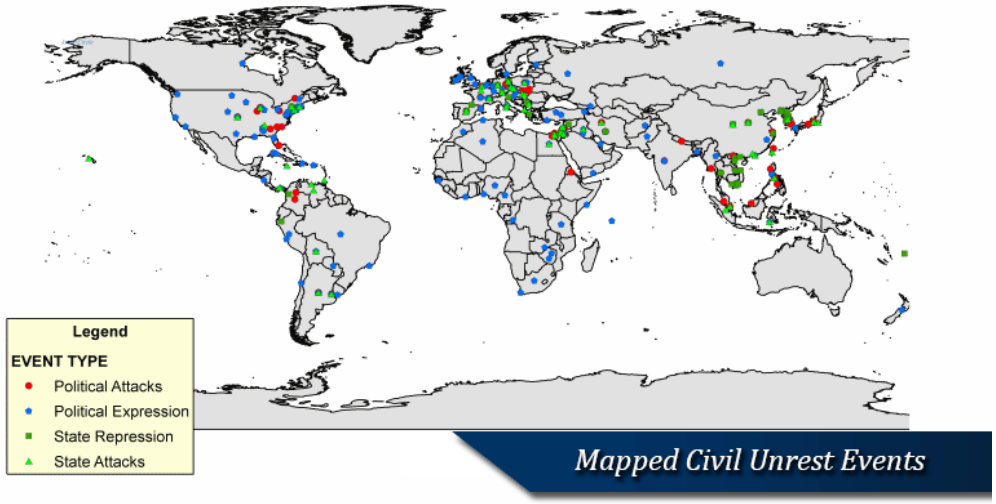
Sound Data[1]



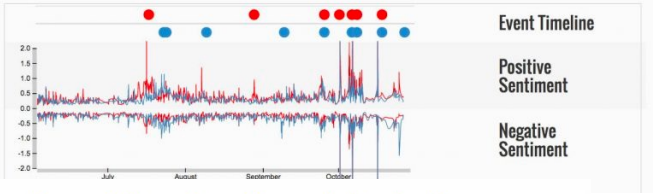
1. Often handled as time-series

Tufts

Event Data

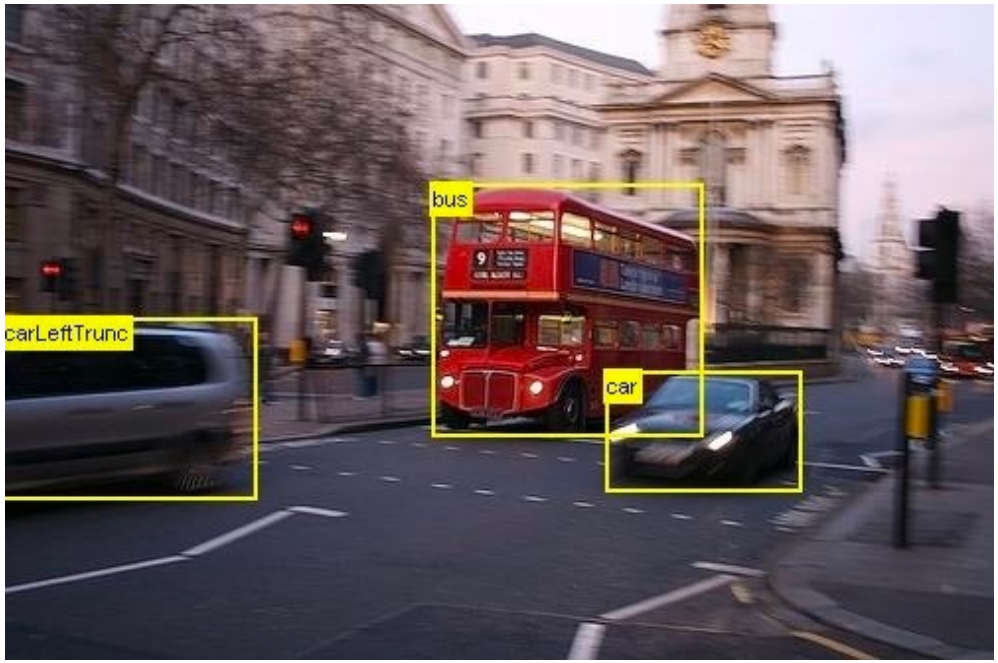


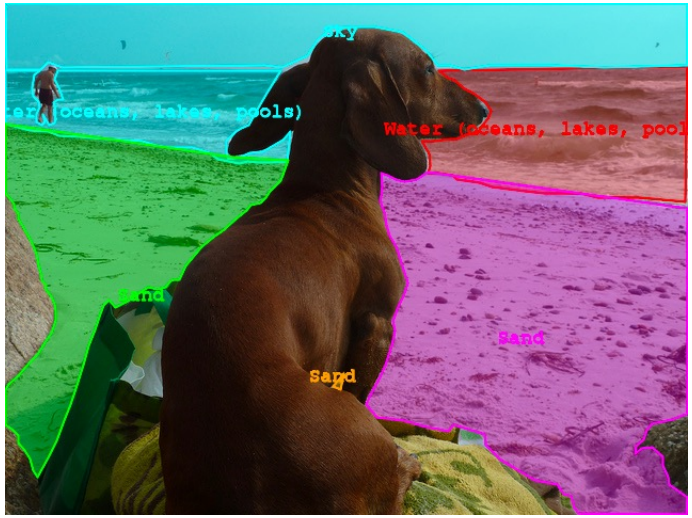
Tweets during Election

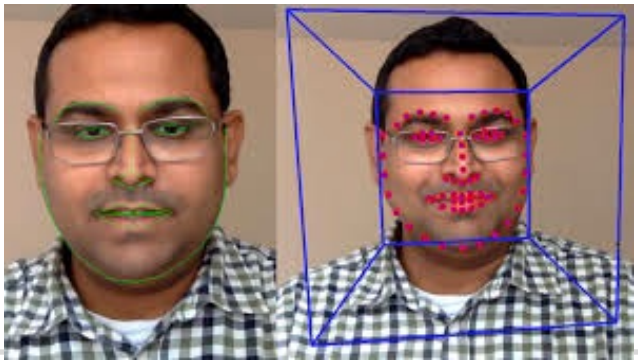


Tufts

Image Data

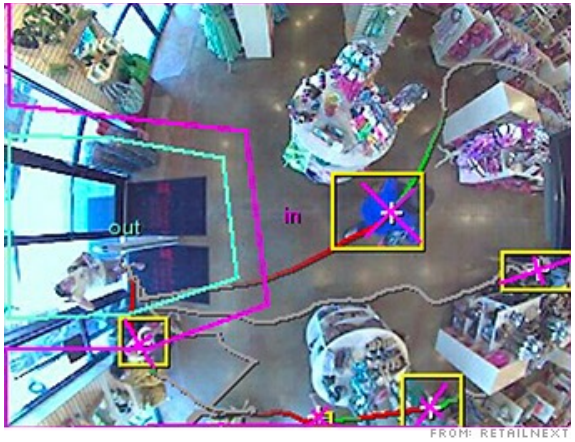






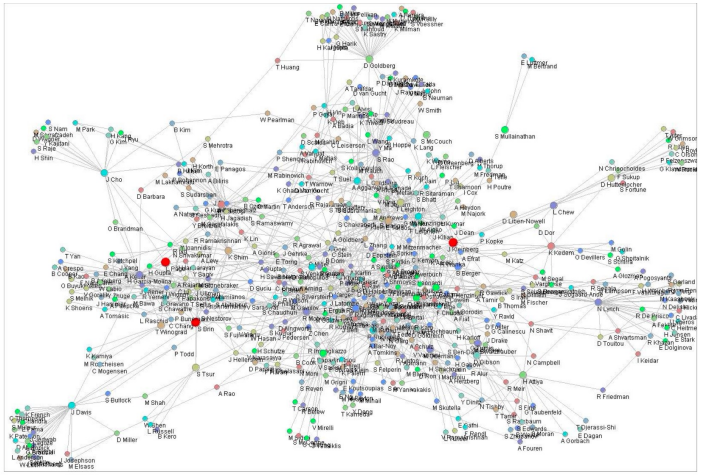
Tufts

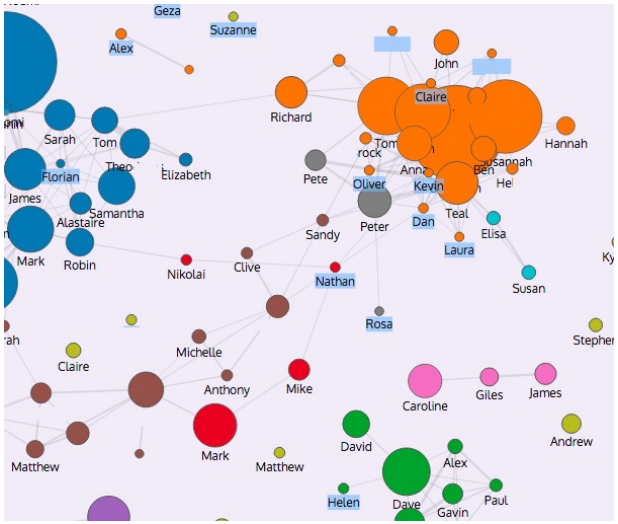
Video Data



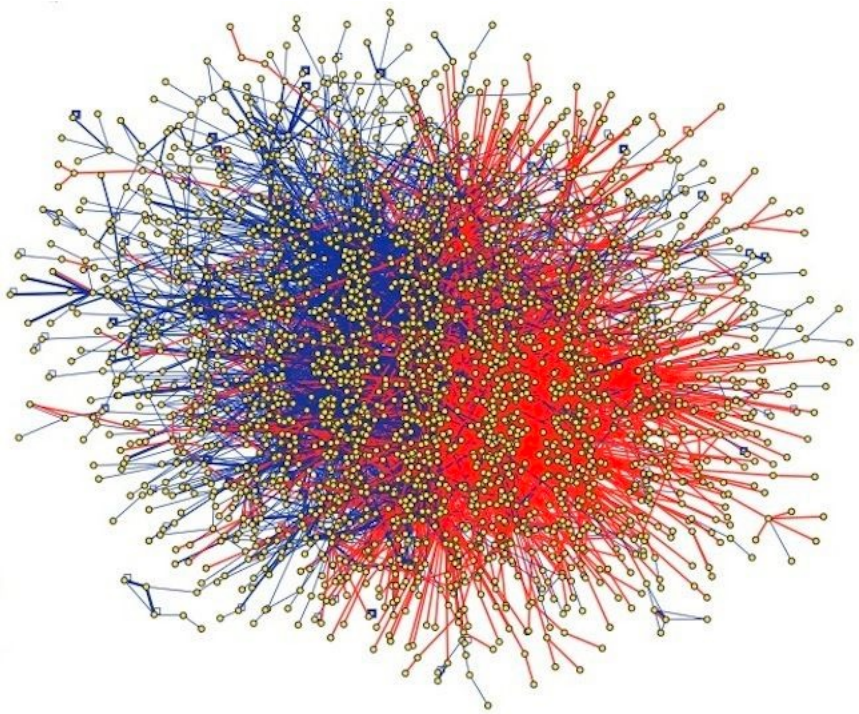
Tufts

Network Data





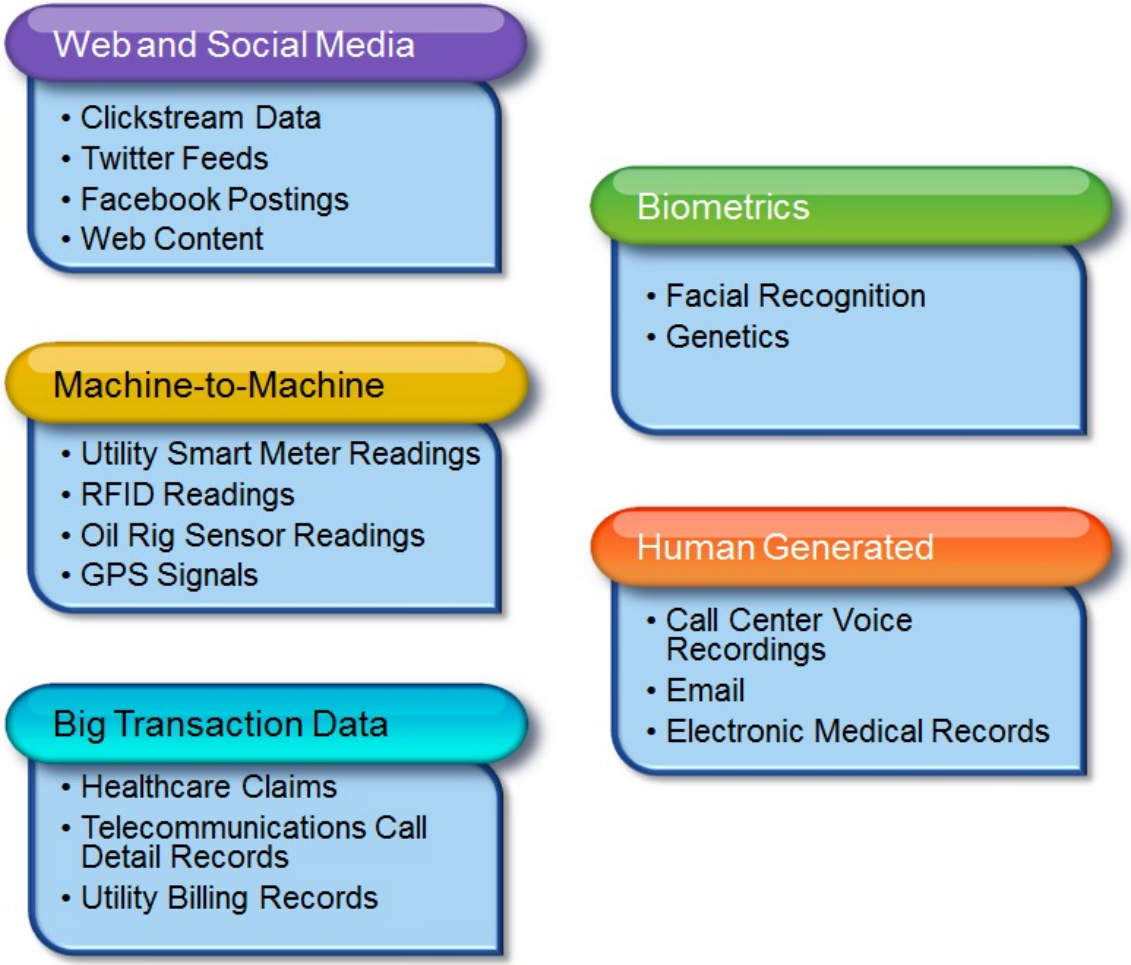
Algorithms for estimating relative importance in networks
S. White and P. Smyth, *ACM SIGKDD*, 2003.



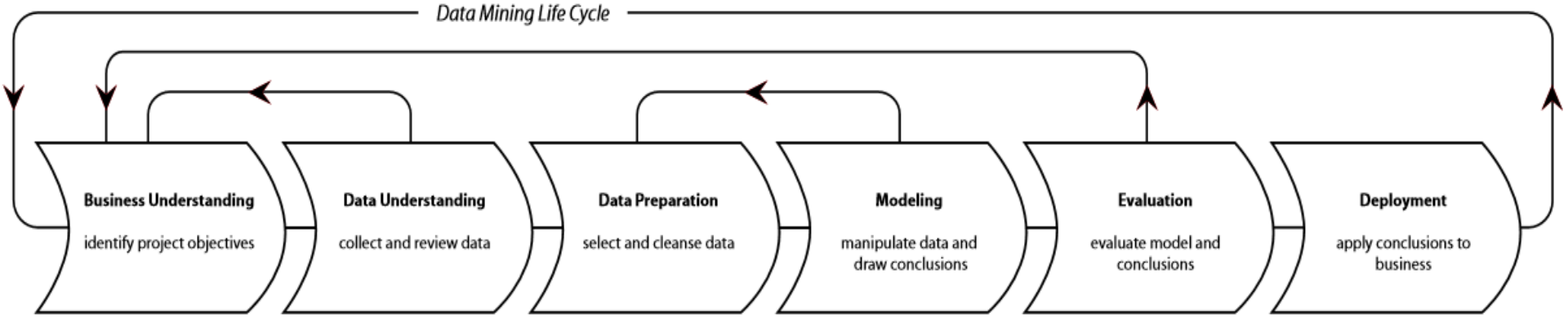
LinkedIn Network

Tufts

Other Types (per category)



Phases



Generic Tasks
Specialized Tasks
(Process Instances)

a visual guide to CRISP-DM methodology

SOURCE CRISP-DM 1.0
http://www.crisp-dm.org/download.htm
DESIGN Nicole Leaper
http://www.nicoleleaper.com

Data in Data Analytics

Data in Data Analytics

- Researchers classify *variables* according to the extent to which the values measure the intended characteristics.

- What option precisely measures intended characteristic of age?

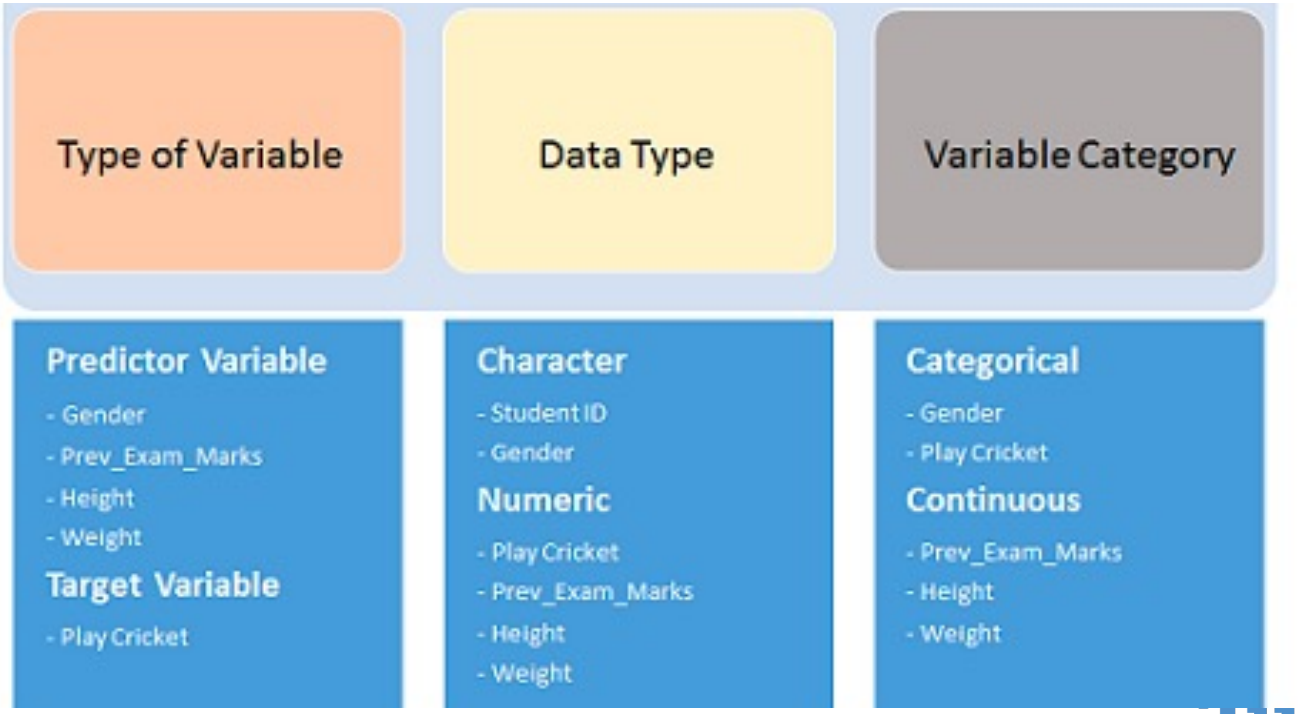
Option A	Option B	Option C	Option D
Young Middle-Age Old	Under 20 20 – 29 30 – 39 40 – 49 50 – 59 60 or above	How old are you?	What is your birth date?

- Clearly *Option D* is the most precise measure of age.
- **Note**: Variable age can be measured with different levels of precision.

Variable Identification

First, identify **Predictor** (Input) and **Target** (output) variables. Next, identify the data type and category of the variables.

Student_ID	Gender	Prev_Exam_Marks	Height (cm)	Weight Caregory (kgs)	Play Cricket
S001	M	65	178	61	1
S002	F	75	174	56	0
S003	M	45	163	62	1
S004	M	57	175	70	0
S005	F	59	162	67	0



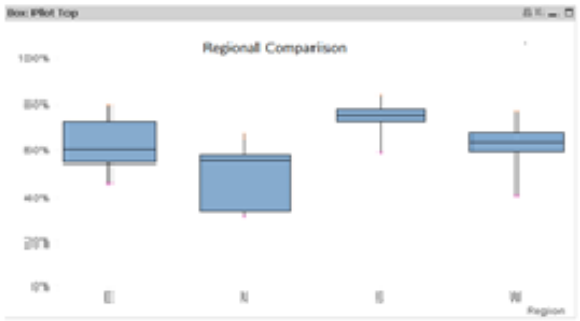
Univariant Analysis

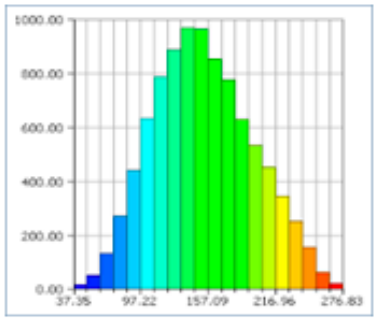
At this stage, we explore variables one by one. Method to perform uni-variate analysis will depend on whether the variable type is categorical or continuous.

Continuous Variables: Need to understand the variable's central tendency and spread. These are measured using various statistical metrics visualization methods:

Categorical Variables: Use frequency table to understand distribution per category. Also, read as percentage of values per category. It can be measured using two metrics, **Count** and **Count%** per category. Bar chart can be used as visualization.

Central Tendency	Measure of Dispersion	Visualization Methods
Mean	Range	Histogram
Median	Quartile	Box Plot
Mode	IQR	
Min	Variance	
Max	Standard Deviation	
	Skewness and Kurtosis	





Tufts

Bivariant Analysis

Finds relationship between 2 variables. Look for association and disassociation between variables at a pre-defined significance level.

Bi-variate analysis for any combination of categorical and continuous variables.

Combinations can be: Categorical & Categorical, Categorical & Continuous and Continuous & Continuous.

Different methods are used to tackle these combinations during analysis process.

Bivariant Analysis (More on this later)

Finds relationship between 2 variables. Look for association and disassociation between variables at a pre-defined significance level.

Bi-variate analysis for any combination of categorical and continuous variables.

Combinations can be: Categorical & Categorical, Categorical & Continuous and Continuous & Continuous.

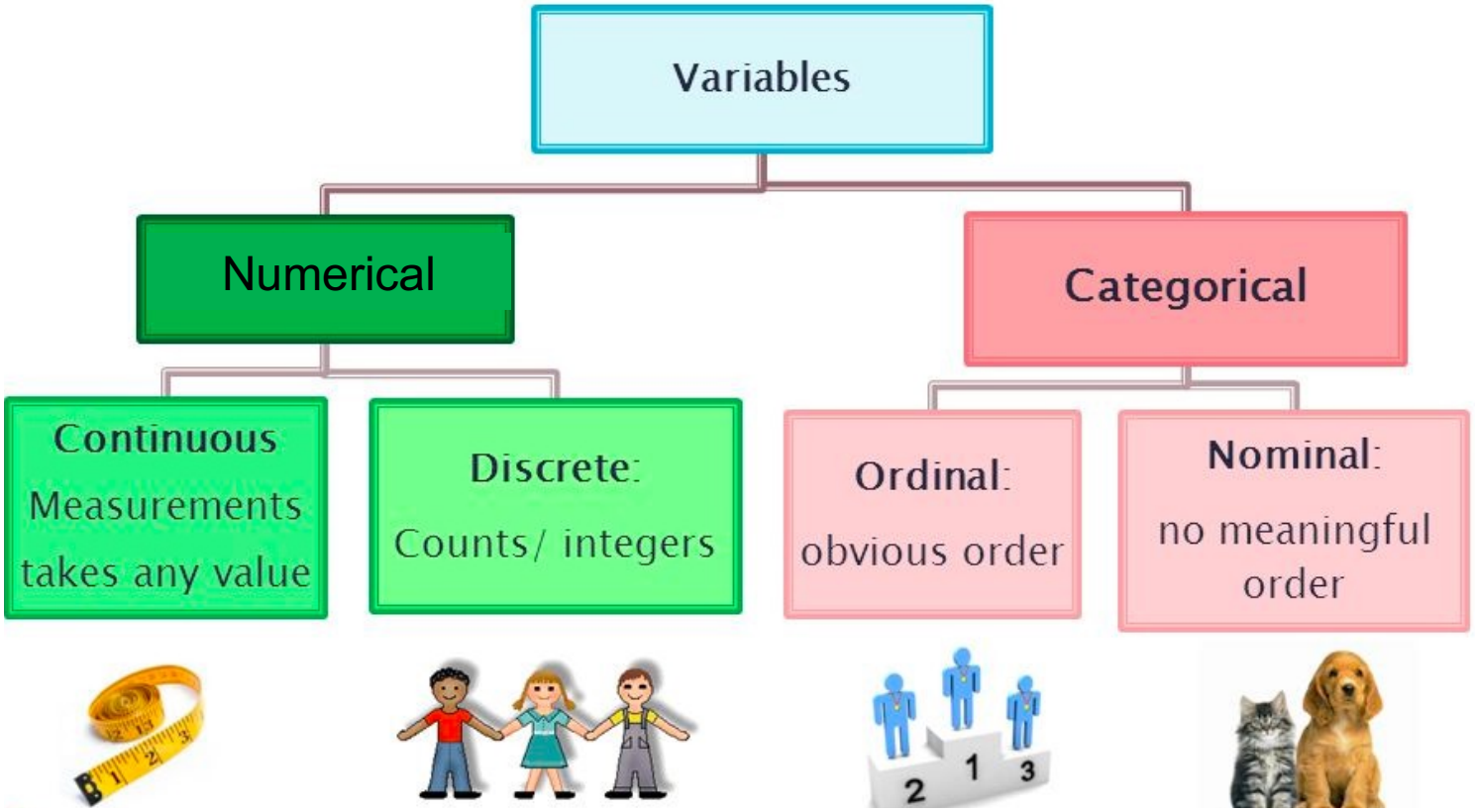
Different methods are used to tackle these combinations during analysis process.

Levels of Measurement

- L*evels of measurement* refers to the precision a variable measures intended empirical characteristic.
  - Different *levels* correspond to how the math is handled.
- Knowledge of *levels of measurement* helps determine statistical methods to apply to the analysis of data.
  - Also affecting the conclusion that can be drawn from the data.

The Four Levels of Measurement

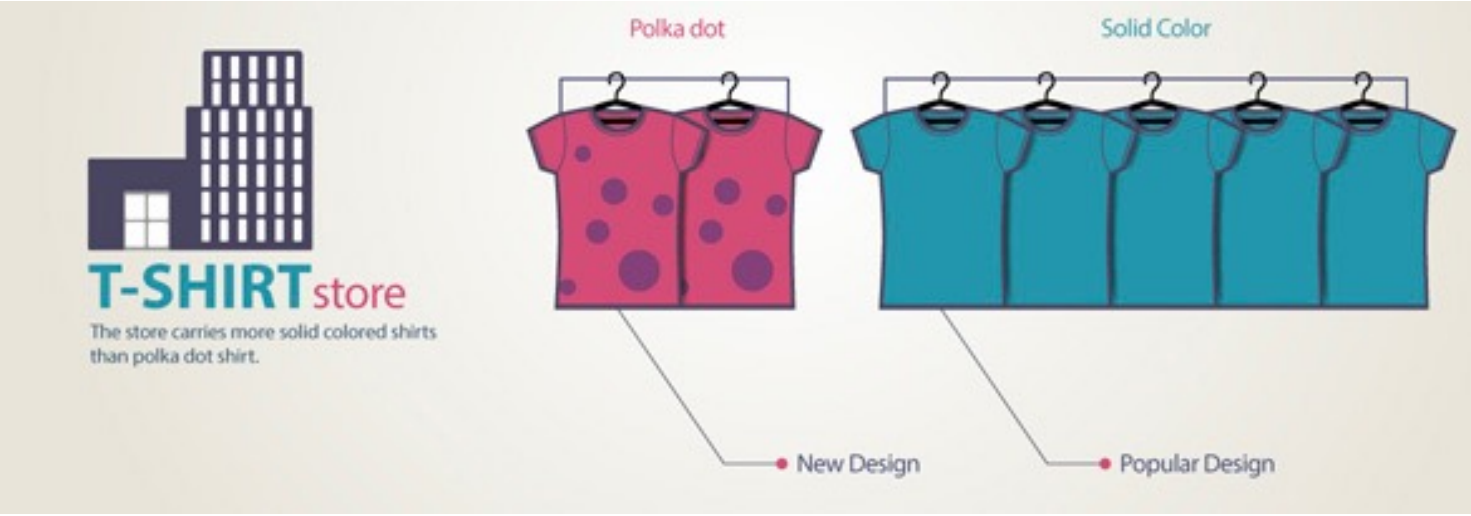
- 4 levels of measurement are *nominal, ordinal, interval*, and *ratio*.
  - Nominal is the lowest of 4 levels, followed by ordinal, interval, and ratio.



Tufts

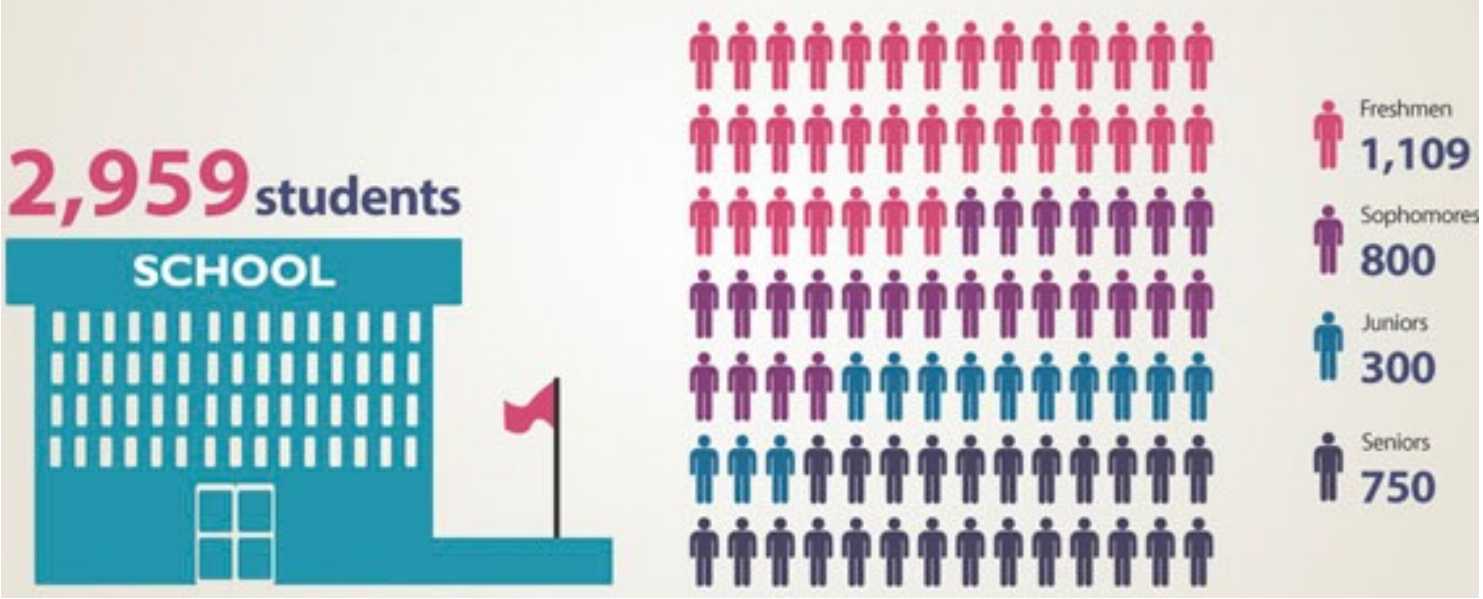
Nominal-level Variables

- Values or responses of variable only differ in **name**, **label**, or **category**
  - no intrinsic or obvious order in values or categories of the variable.
- Examples: gender (M, F), favorite color (purple, blue, red, etc.), religious affiliation (Catholic, Protestant, Judaism, etc.)
- Numbers are assigned on nominal scale are simply as identifiers or names
  - e.g., #'s on the back of baseball jerseys. For convenience and does not measure any quantity possessed by the object.



Ordinal-level Variables

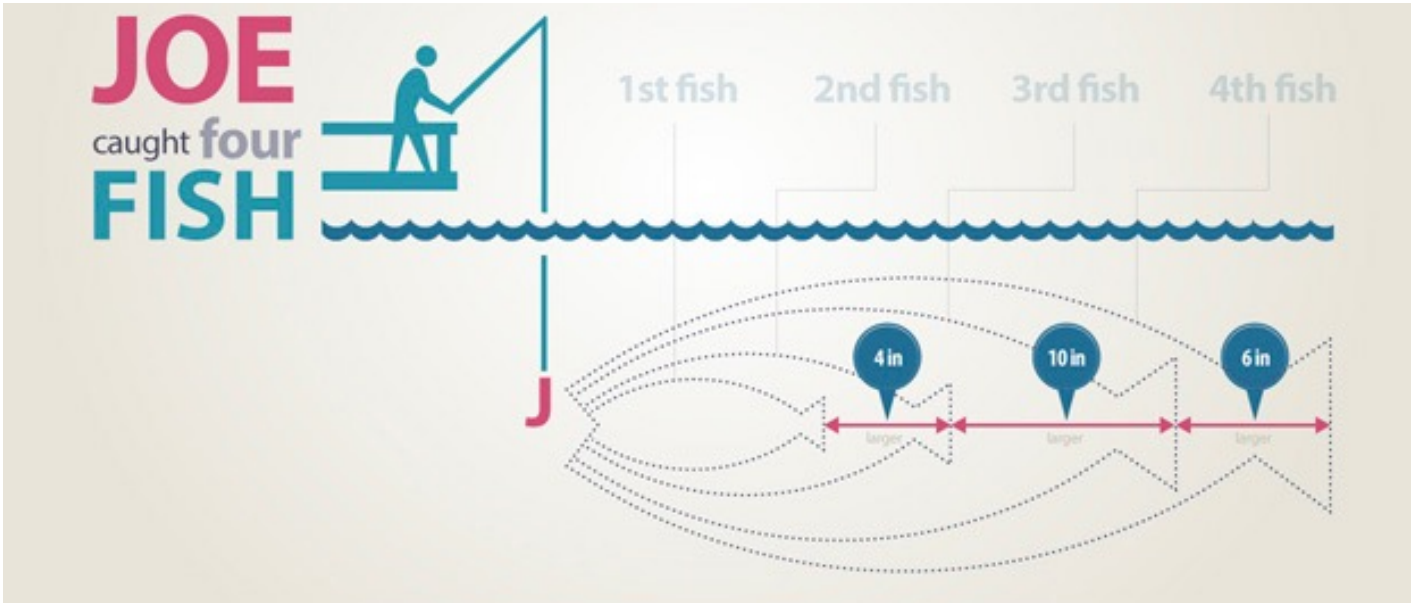
- Values or categories of the variable can be meaningfully or logically *listed in order* and there is *no* intention to determine the numerical value.
  - E.g., student's class standing in high school: freshman, sophomore, junior, or senior (i.e., category names are rank ordered).
  - Other examples include letter grades (A, A-, B+,…, F), political philosophy (very liberal, liberal, …, conservative, very conservative), social class, etc.



Tufts

Interval-level Variables

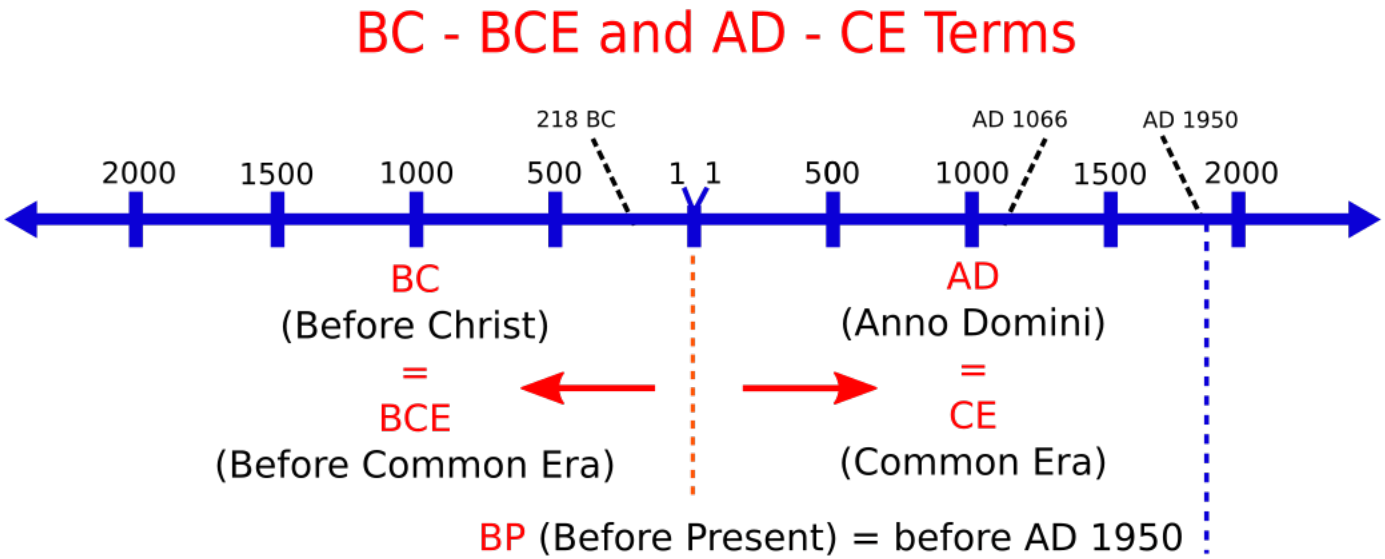
- Interval-level variables are numeric & more precise than nominal and ordinal.
- Values are **actual numerical amounts** of characteristics being measured.
- Think of interval-level variables as a number line with **equal intervals between values** and steps increase by the same amount.
- There is an **arbitrary starting point** (usually 0).

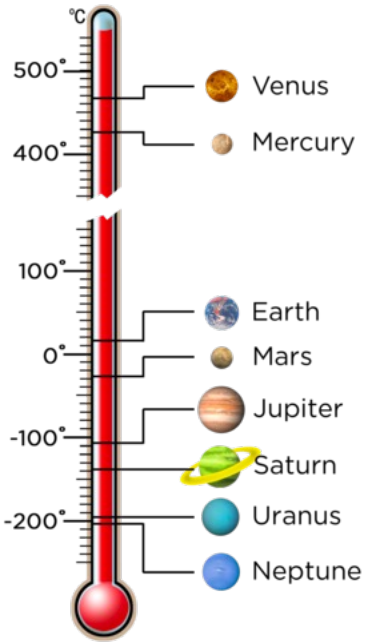


Tufts

Interval-level Variables

- Good example is temperature in degrees Celsius. The numbers correspond to levels of mercury in the thermometer and they are equal distance apart (1˚). 0 is an arbitrary value (i.e., it does not mean the absence of heat).
- Another is time in calendar years. The interval between categories is one year. 0 is an arbitrary point (between BC & AD). Time still existed at time 0.





Tufts

Ratio-level Variables

- Objects on the ratio scale possess all the properties of the interval scale AND have an *absolute zero point*.
- Scale usually used in physical sciences (eg distance, weights, heights, age, etc)
- Can use *division* to compare qualities being measured & make a statement about "how many times" one object is greater than or less than another.
- Example: A student who has completed 30 credits has twice as many credits as a student who has completed 15 credits.

Numerical vs Categorical Variables

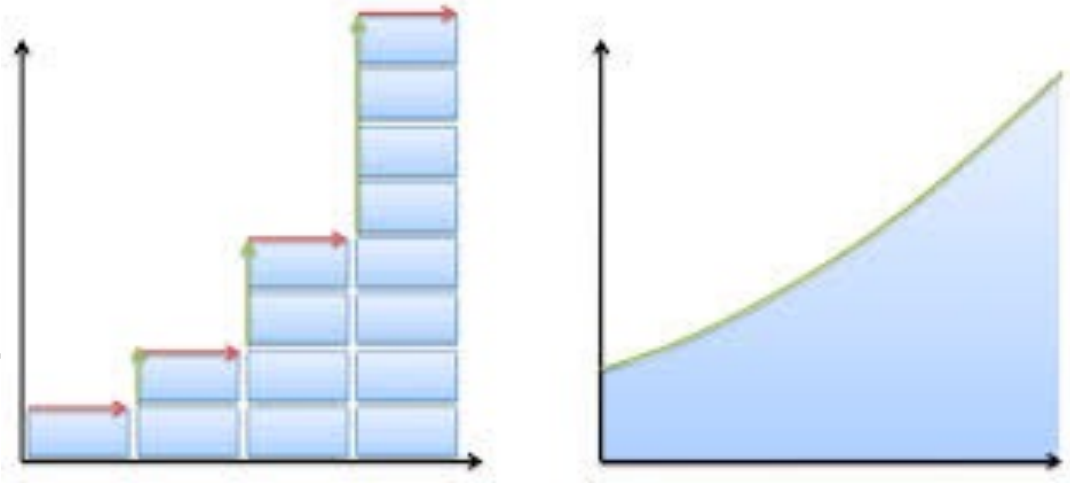
- Numerical variables are just as they sound (i.e., numbers).
  - Consist of numbers and can be interpreted as numbers.
- Interval & ratio variables are numerical + provide a *quantity* or a *dimension*.
- Note: Don't get confused. Just because something is a number doesn't automatically make it a numerical variable: eg given 1=male and 2=female doesn't mean that gender is numerical. It is not. It is nominal (ie categorical)

Numerical vs Categorical Variables

- *Categorical variables* record the quality of something. Predetermined non-overlapping categories were set up ahead of time by the researcher.
- Examples include gender, blood type, ethnicity, outcome of a plate appearance in baseball, etc.
- Variables non-numerical in nature (can be coded numerically)
- Each observation must fit in a single class or category (e.g., male or female)
- Nominal and ordinal variables are categorical.

Discrete vs Continuous Variables

- Numerical variables can be classified as discrete or continuous
- *Discrete variables* take on a finite number of values and are usually countable objects.
  - For example, you can count chairs, people, oranges
- *Continuous variables* deal with characteristics that cannot be counted directly, like age, weight, height, volume, heat, speed, and area. A continuous variable is one whose values are obtained as a result of measuring some characteristic of an object
- A continuous variable can take on an infinite number of values:
  - Reporting age to the second, or weight to the decimal



Numerical vs Categorical Variables

Example: Determine if variables are *categorical or numerical*. If categorical, decide whether *nominal or ordinal*. If numerical, decide whether it is *discrete or continuous*.

a) Marital status (single, divorced, married, separated, widowed)

Categorical, nominal

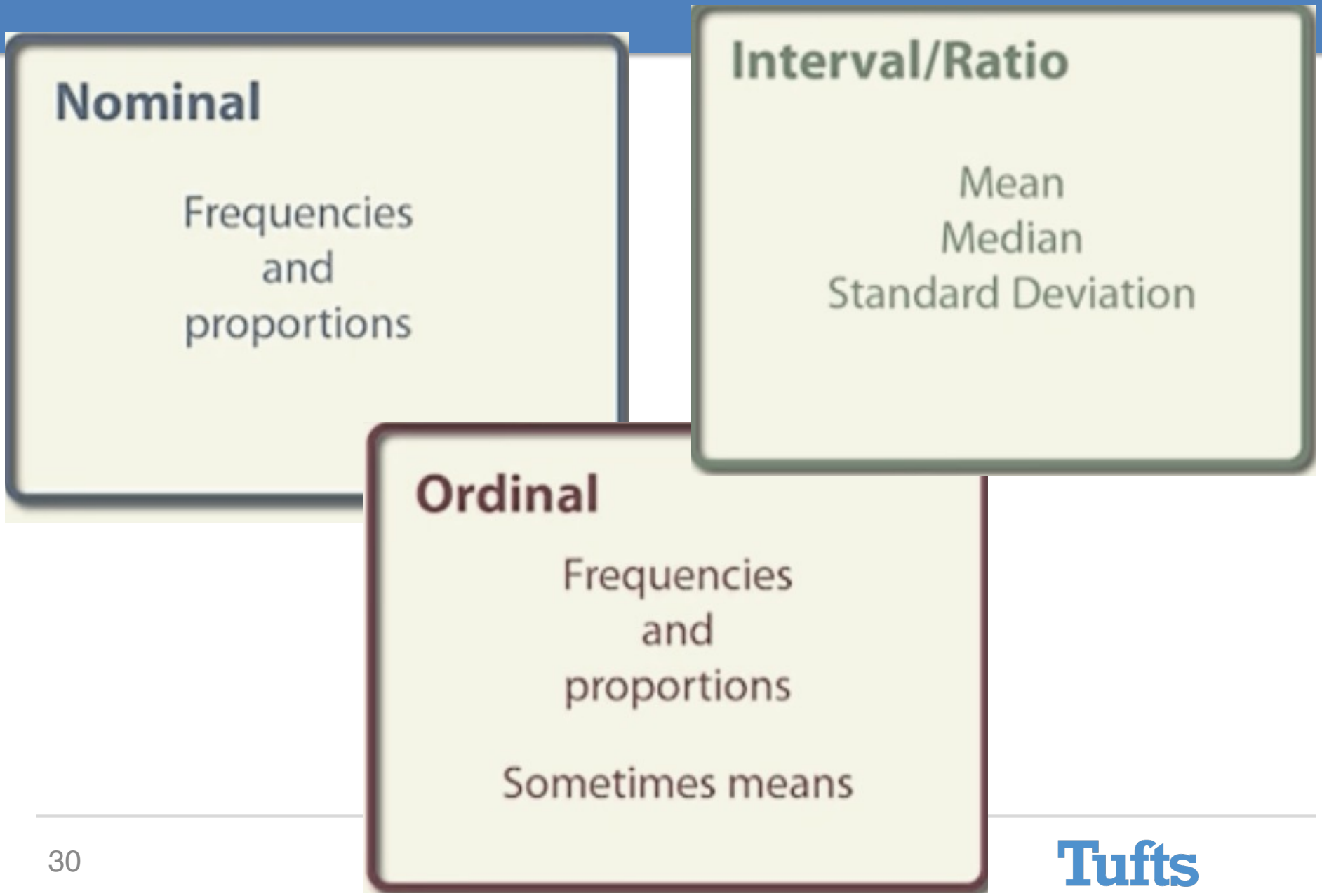
b) Number of courses you are taking this semester (0, 1, 2, 3, 4, 5, or 6)

Numerical, discrete

c) Age (below 18, 18-19, 20-21, over 21)

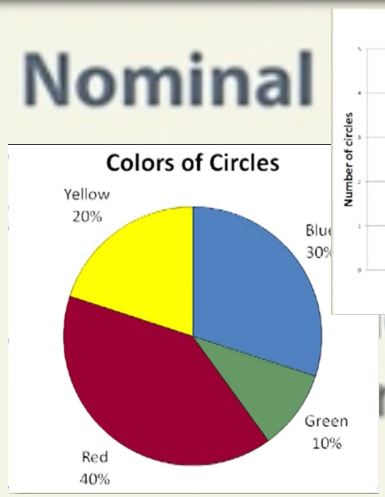
Categorical (being reported as categories), ordinal

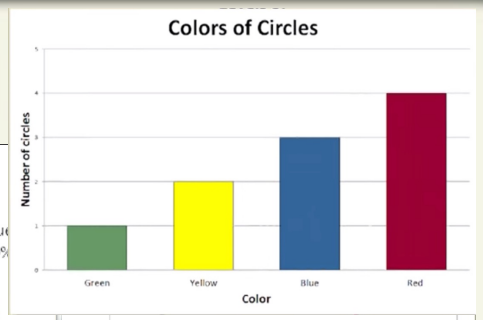
Graphical Representation of Different Data Types

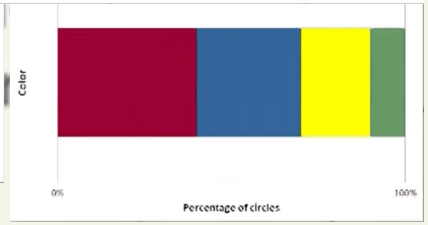


Graphical Representation of Different Data Types

Nominal







Interval/Ratio

Mean
Median
Standard Deviation

Ordinal

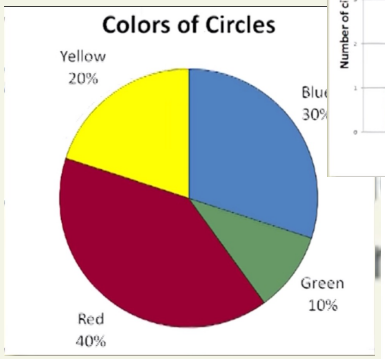
Frequencies
and
proportions

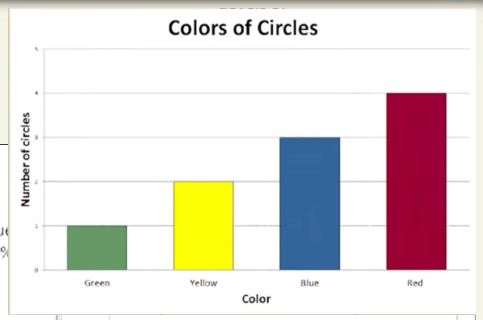
Sometimes means

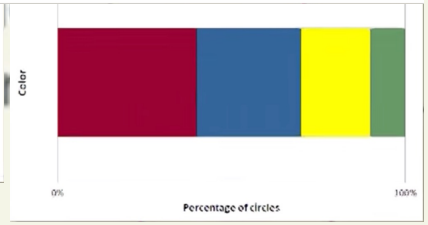
Tufts

Graphical Representation of Different Data Types

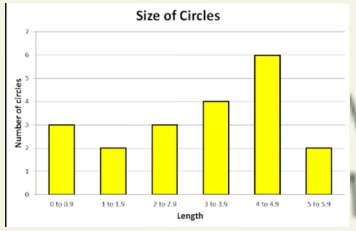
Nominal

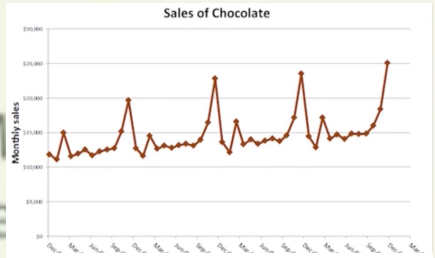


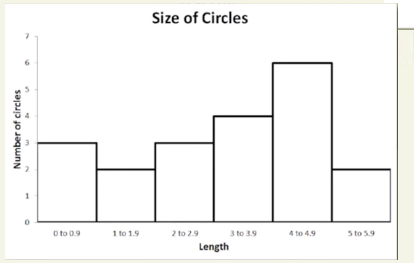


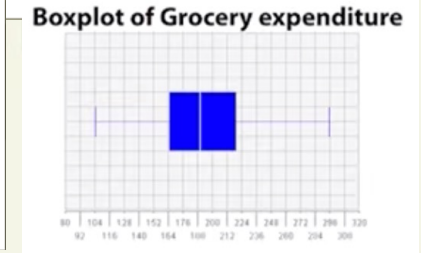


Interval/Ratio









Ordinal

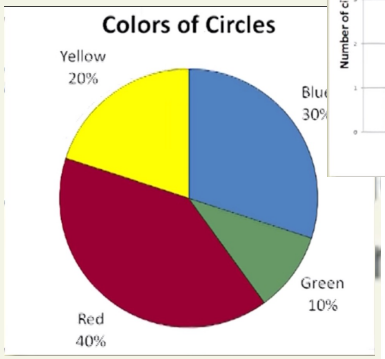
Frequencies
and
proportions

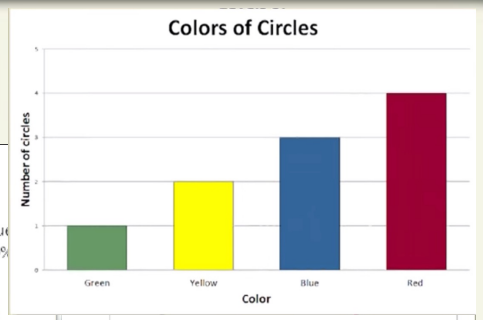
Sometimes means

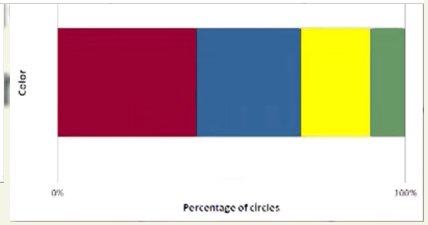
Tufts

Graphical Representation of Different Data Types

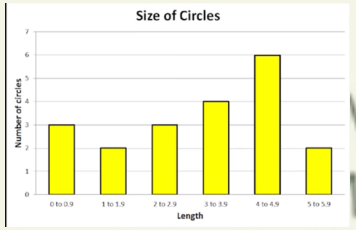
Nominal

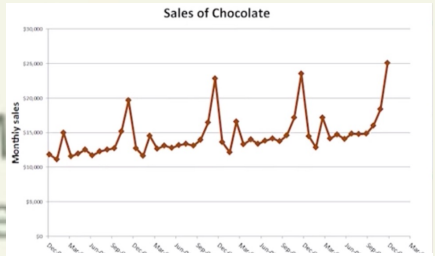


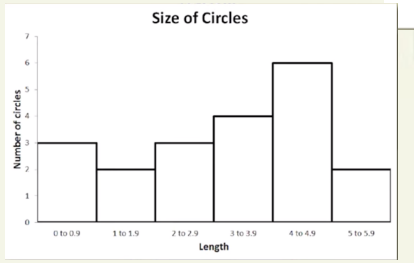


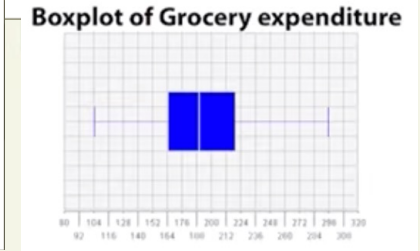


Interval/Ratio



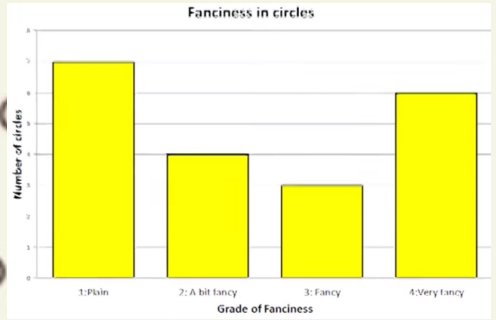






Ordinal

Fre



times means

Tufts

Data Quality: A Problem

It's Not Just Us

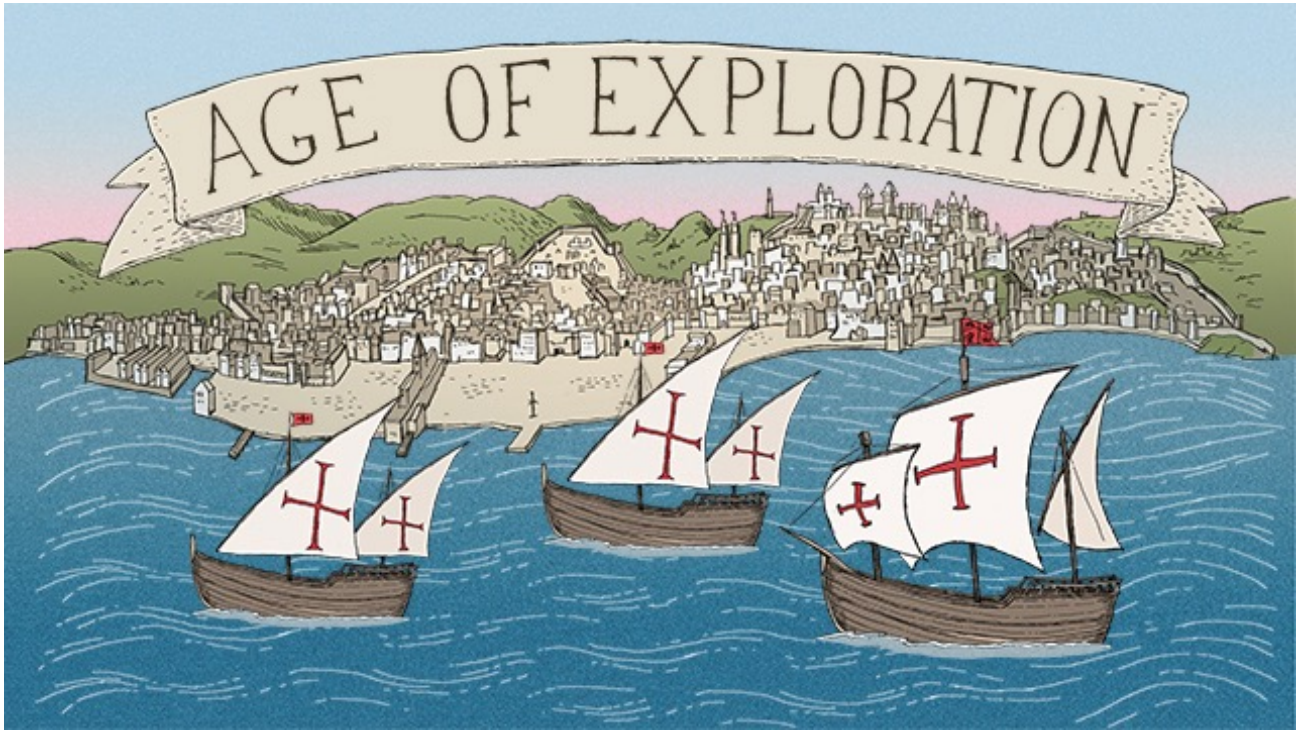
In just about any organization, the state of information quality is at the same low level

- Olson, Data Quality

Some Consequences of Poor Data Quality

- Affects quality (precision) of result
- Can't do modeling project because of data problems
- If errors not found– modeling blunder

Data Exploration in Predictive Modeling



Exploratory Data Analysis (EDA)

- Typically the 1st step in analyzing data
- Makes heavy use of graphical techniques
- Also makes use of simple descriptive statistics
- Purpose
  - Find outliers (and errors)
  - Explore structure of the data

Exploratory Data Analysis (EDA)

EDA is part of a statistical practice concerned with reviewing, communicating and using data when there's a low level of knowledge about its cause system.

EDA techniques have been adopted into data mining and are being taught to young students as a way to introduce them to statistical thinking.[1]

1. www.wikipedia.org

Example Data

- Private passenger auto

- Some variables are:
  - Age
  - Gender
  - Marital Status
  - Zip code
  - Earned premium
  - Number of claims
  - Incurred losses
  - Paid losses

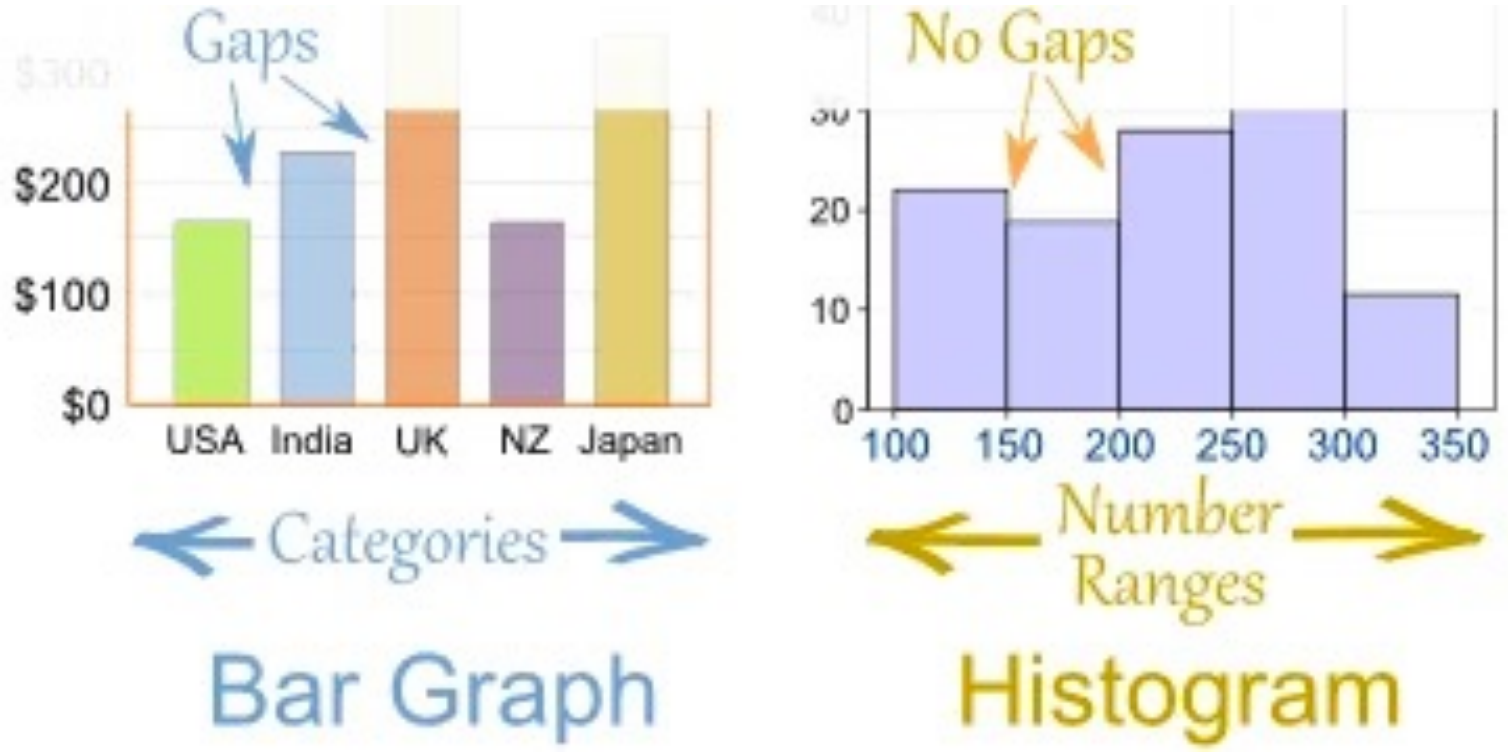
Some Methods for Numeric Data

- Visual
  - Histograms
  - Box and Whisker Plots
  - Stem and Leaf Plots

- Statistical
  - Descriptive Stats
  - Data spheres

Statistic	Policyholder Age
Mean	36.9
Standard Error	0.1
Median	35.0
Mode	32.0
Standard Deviation	13.2
Sample Variance	174.4
Kurtosis	0.5
Skewness	0.7
Range	84
Minimum	16
Maximum	100
Sum	1114357
Count	30226
Largest(2)	100
Smallest(2)	16

Histograms vs Bar Graphs

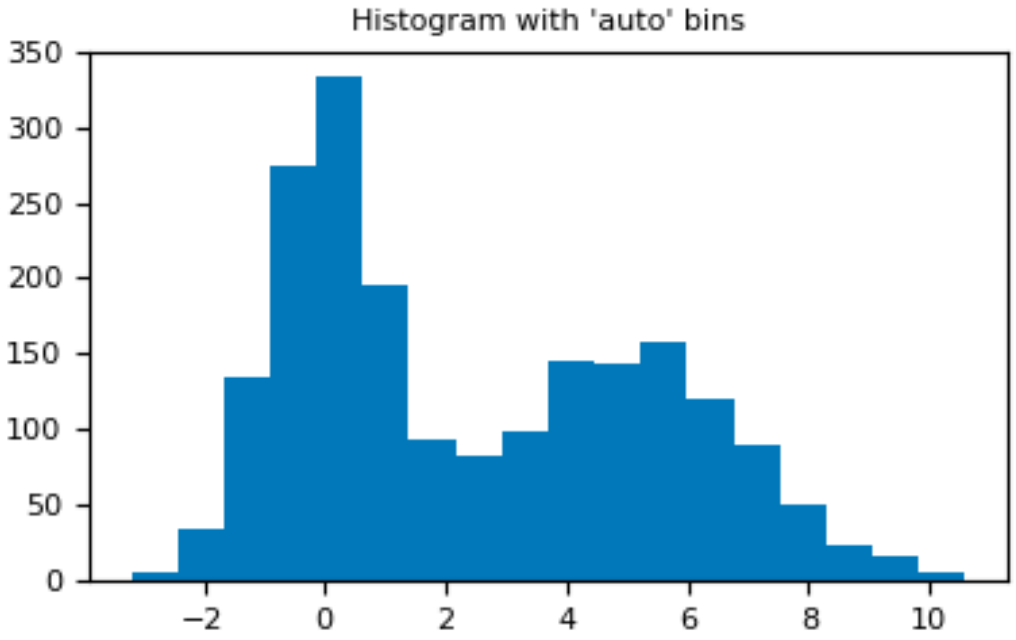


Tufts

Histograms vs Bar Graphs

numpy

```
import numpy as np
import matplotlib.pyplot as plt
rng = np.random.RandomState(10)  # deterministic random data
a = np.hstack((rng.normal(size=1000), rng.normal(loc=5, scale=2, size=1000)))
plt.hist(a, bins='auto')  # arguments are passed to np.histogram
plt.title("Histogram with 'auto' bins")
plt.show()
```



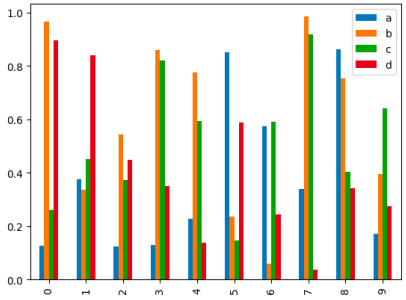
Tufts

Histograms vs Bar Graphs

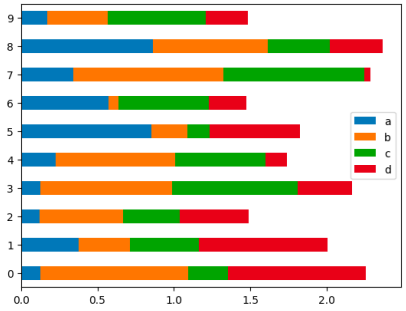
pandas

```
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
```

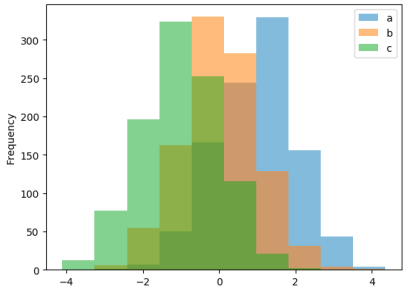
```
df1=pd.DataFrame(np.random.rand(10,4),columns=['a','b','c','d'])
df1.plot.bar();
```




```
df2 = pd.DataFrame(np.random.rand(10, 4), columns=['a', 'b', 'c', 'd'])
df2.plot.bar(stacked=True)
df2.plot.bar()
df2.plot.barh(stacked=True)
```




```
df3 = pd.DataFrame({'a': np.random.randn(1000) + 1, 'b': np.random.randn(1000),
                    'c': np.random.randn(1000) - 1}, columns=['a', 'b', 'c'])
plt.figure()
df3.plot.hist(alpha=0.5)
```

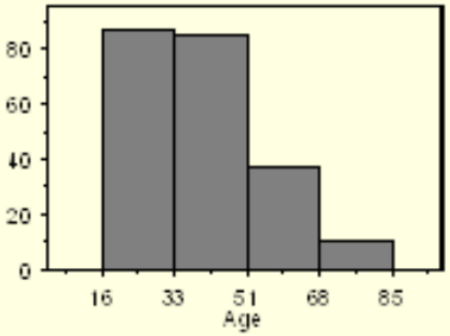


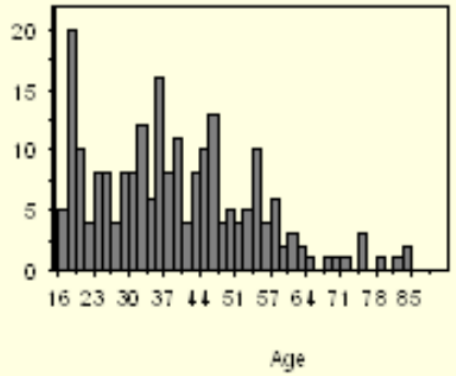
Tufts

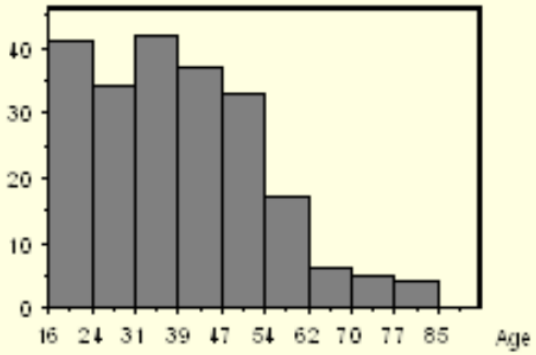
Window Width

Varying window size

Bin	Frequency
20	2853
25	3709
30	4372
35	4366
40	4097
45	3588
50	2707
55	1831
60	1140
65	615
70	397
75	271
80	148
85	83
90	32
95	12
More	5



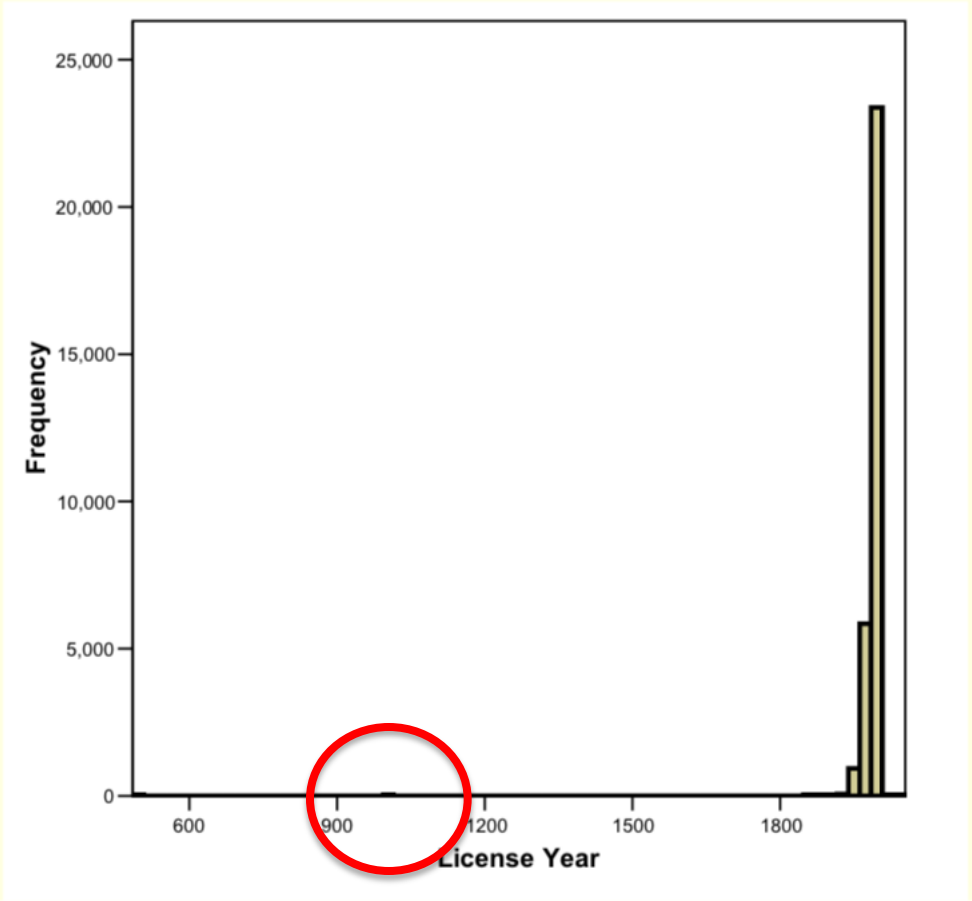




$$h = \frac{3.5\sigma}{\sqrt[1/3]{N}}$$

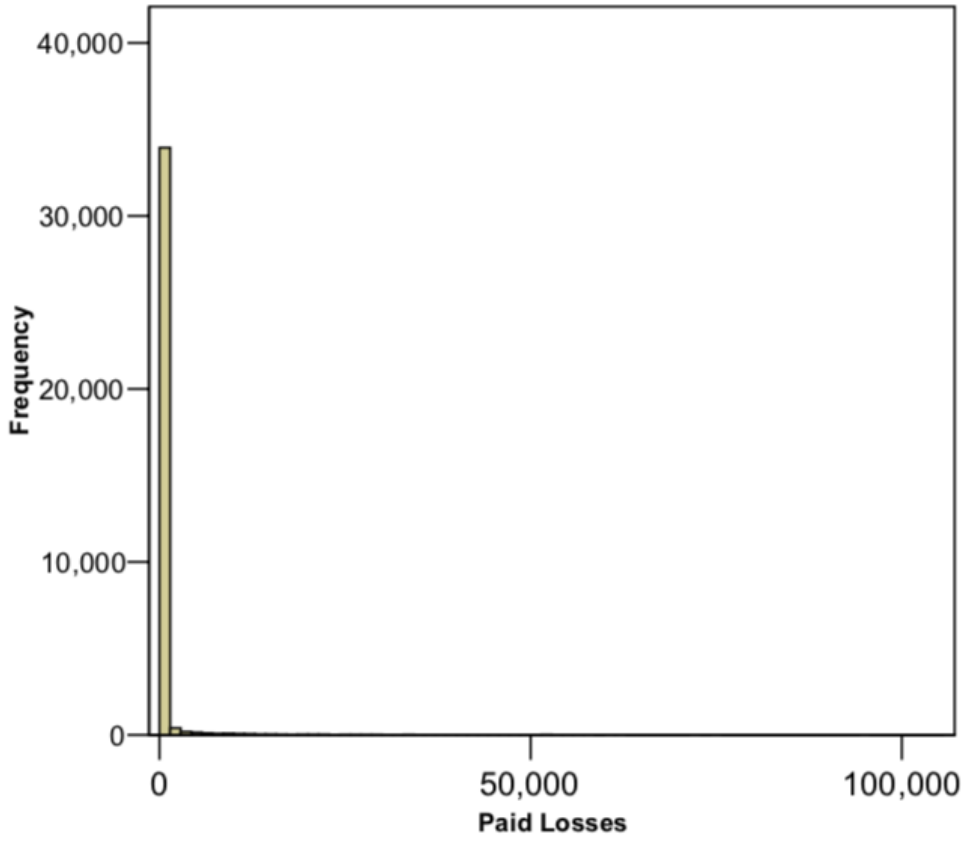
σ = standard deviation
N = Sample Size
h = Window width

Example: Suspicious Value



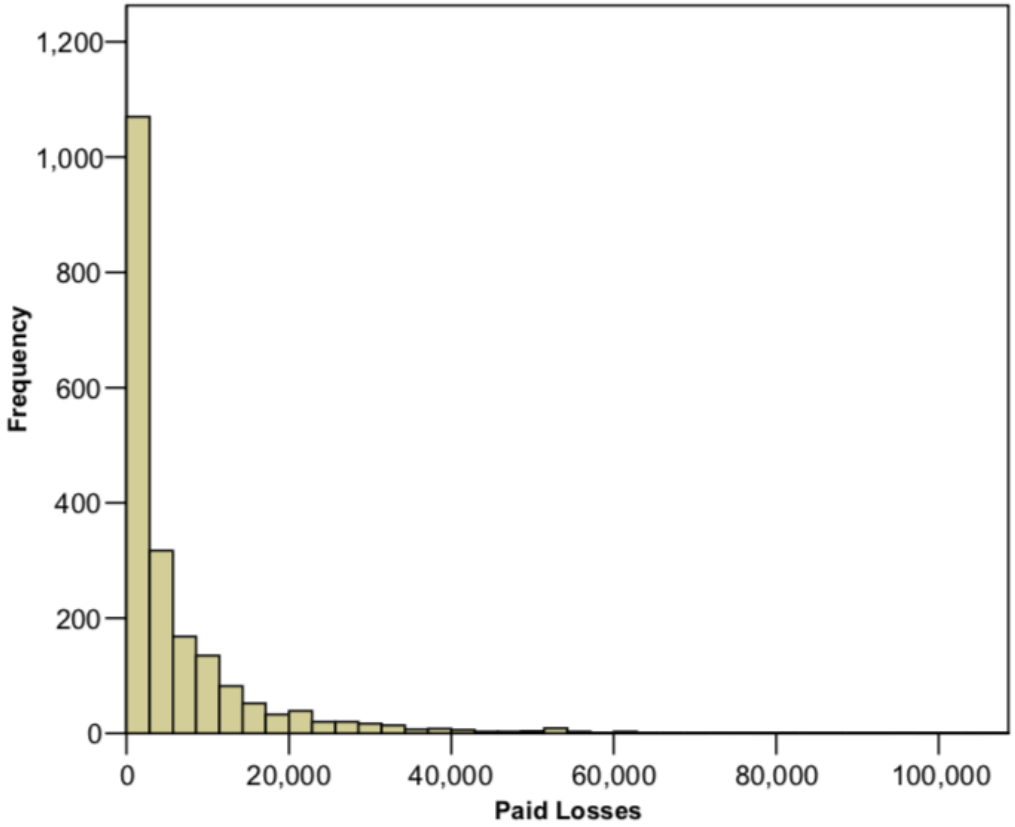
Tufts

Example: Discrete-Numeric Data



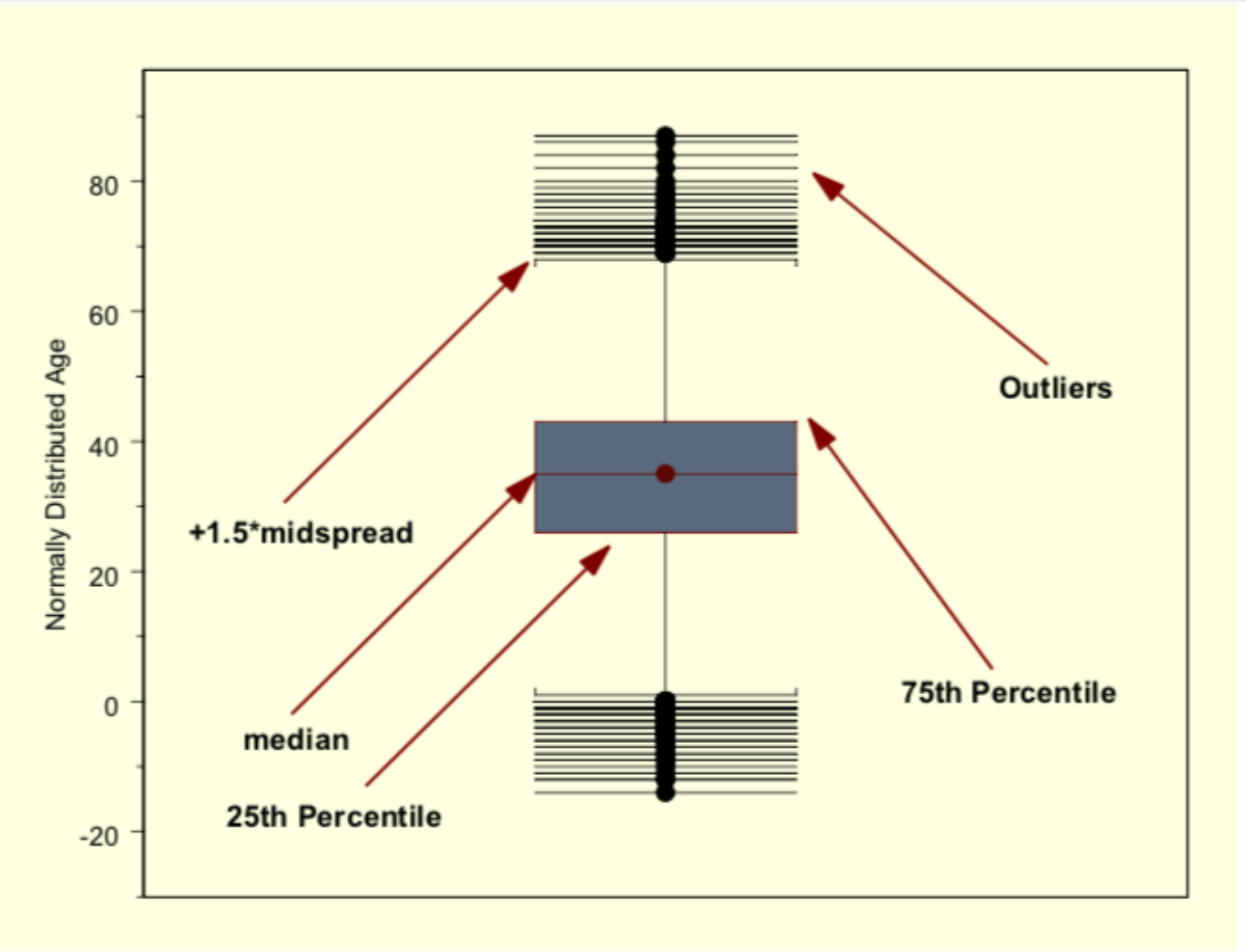
Tufts

Example: Filtered Data (filter out unwanted records)



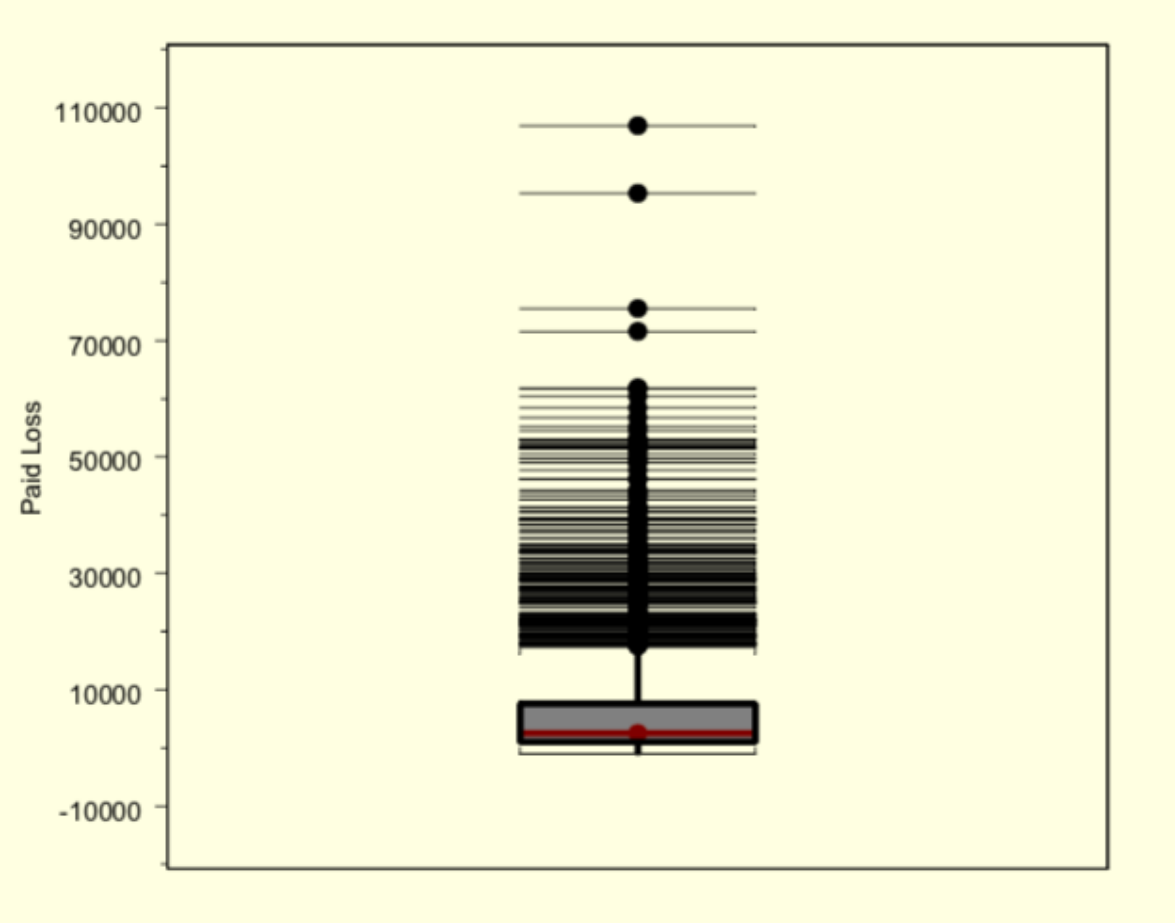
Box Plot Basics: Five– Point Summary
Box and Whisker Plot

- Minimum
- 1st quartile
- Median
- 2nd quartile
- Maximum

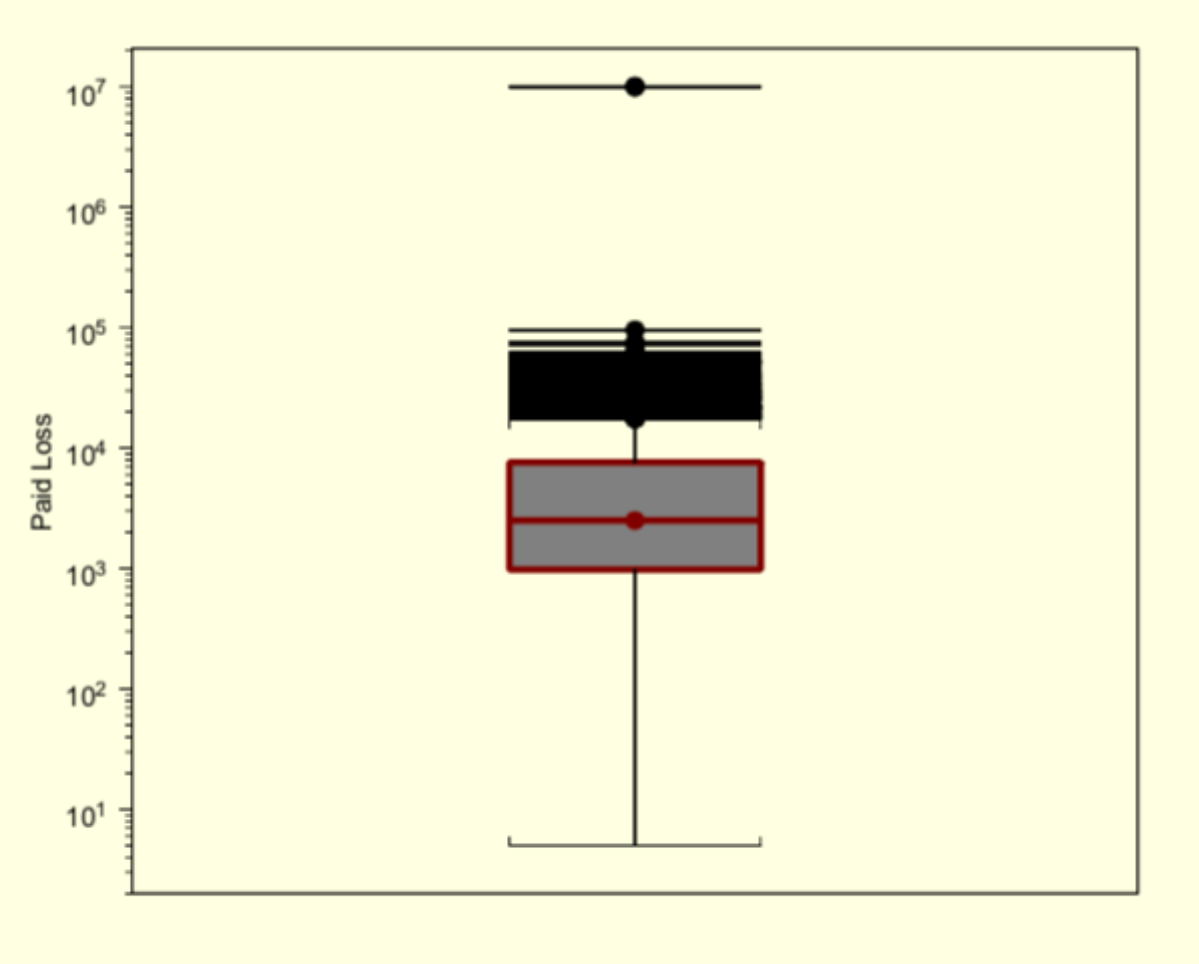


Tufts

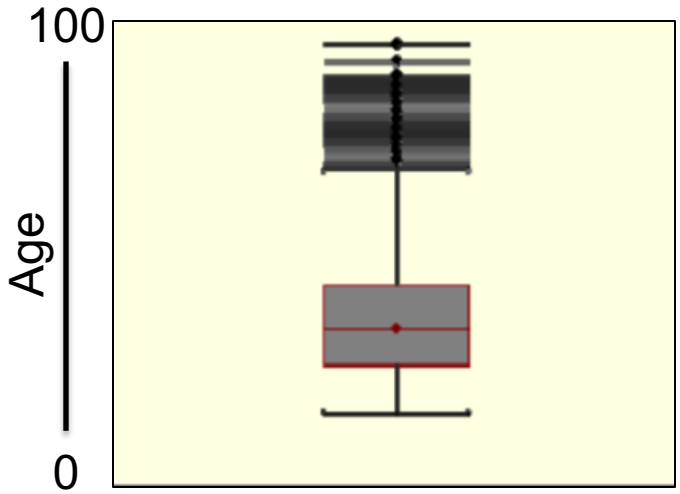
Plot of Heavy Tailed Data
Paid Losses

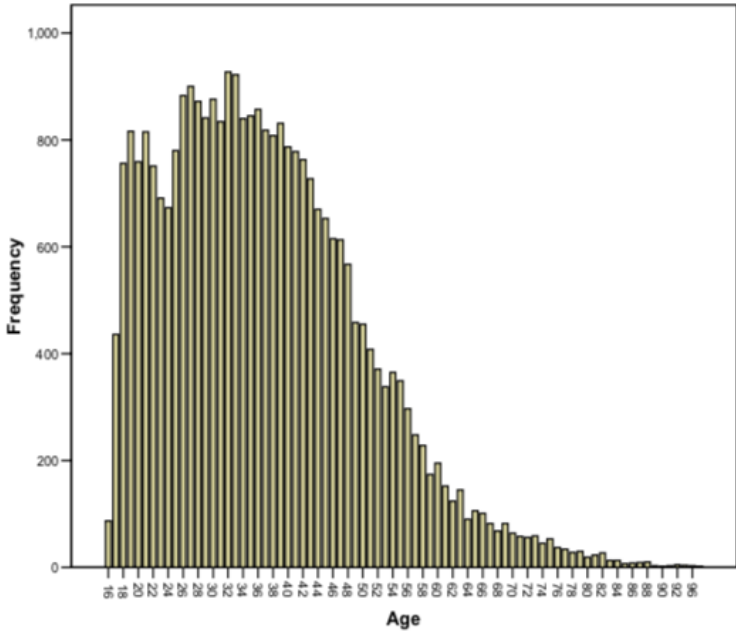


Heavy Tailed Data: Log Scale



Box and Whisker Example



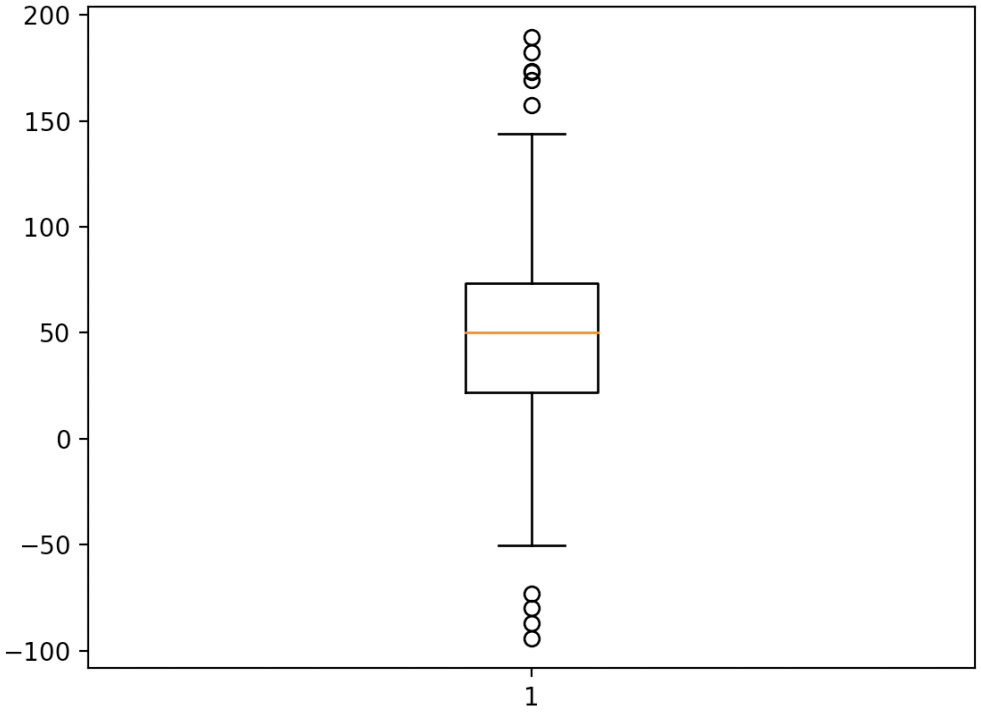


Tufts

Box and Whisker

numpy

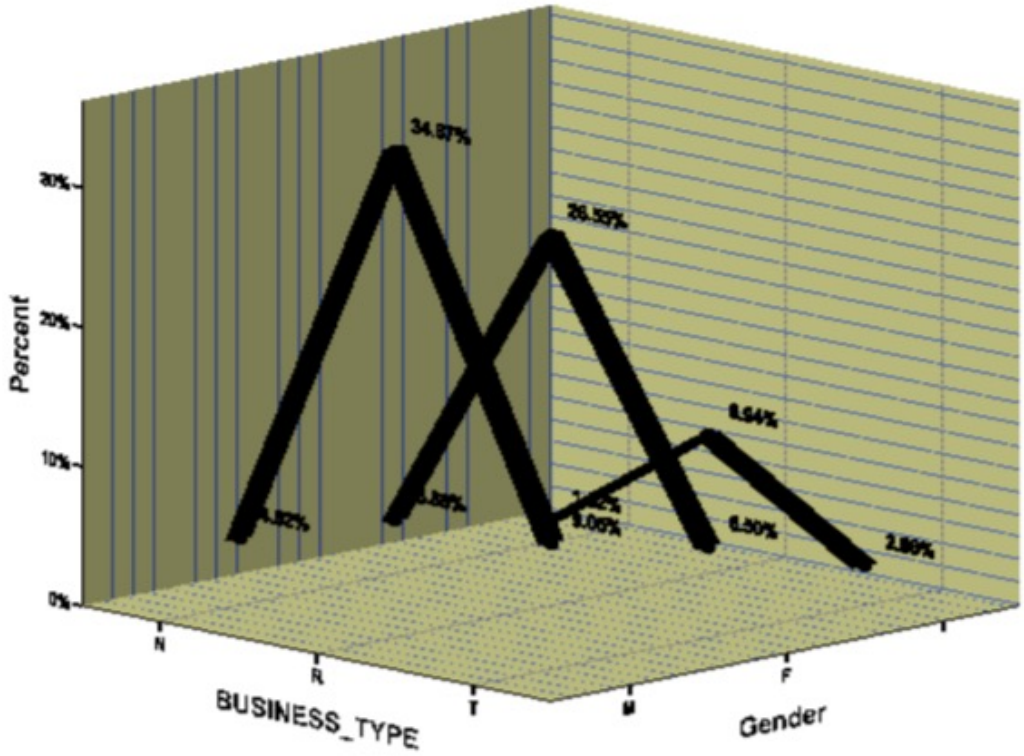
```
import matplotlib.pyplot as plt
import numpy as np

# fake up some data
spread = np.random.rand(50) * 100
center = np.ones(25) * 50
flier_high = np.random.rand(10) * 100 + 100
flier_low = np.random.rand(10) * -100
data = np.concatenate((spread, center, flier_high, flier_low), 0)

# basic plot
plt.boxplot(data)
```



Categorical Data: Data Cubes

- Data Cubes
  - Usually frequency tables
  - Search for missing values coded as blanks

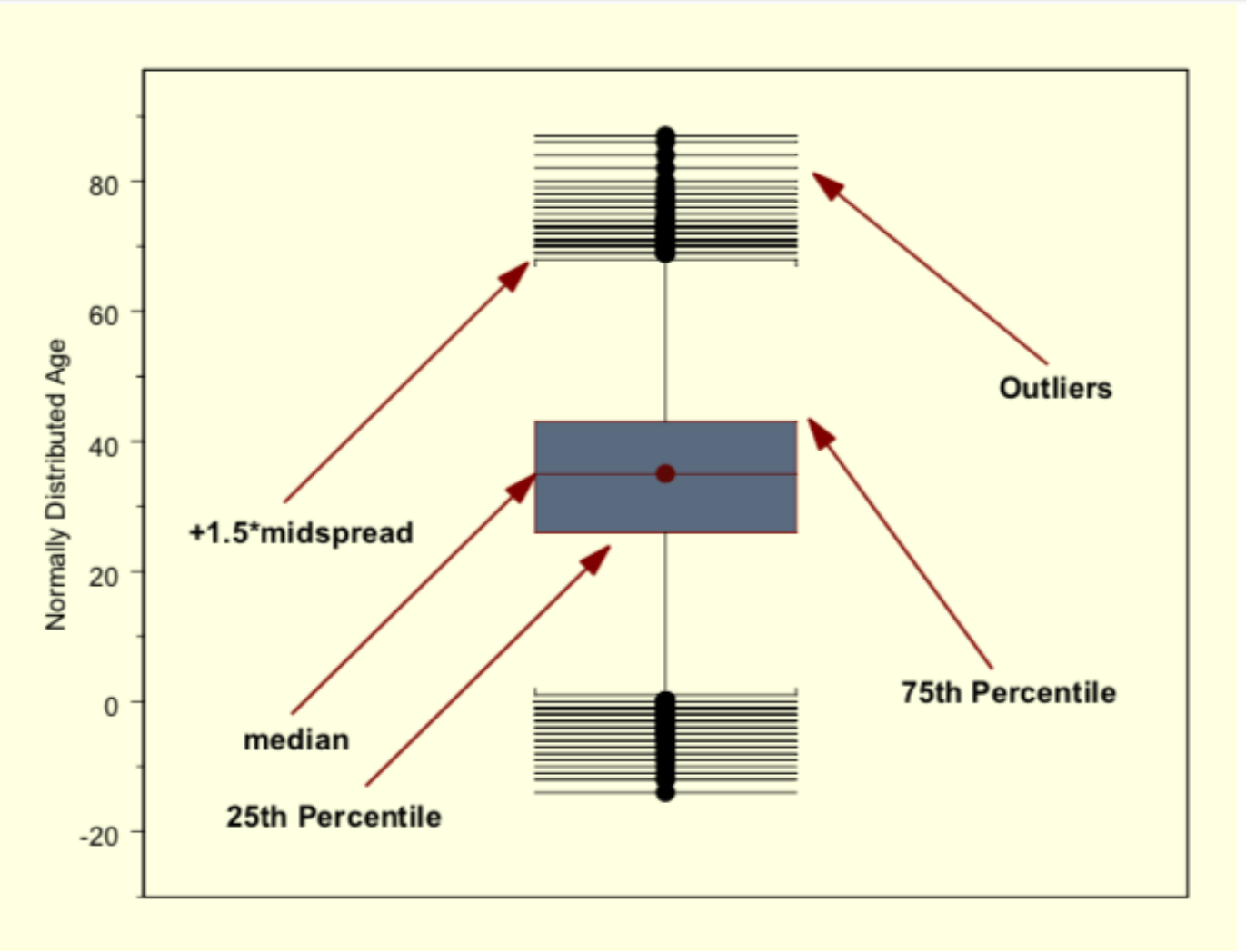
Gender		
	Frequency	Percent
Blank	5,054	14.4
F	13,032	36.9
M	17,198	48.7
Total	32,284	100



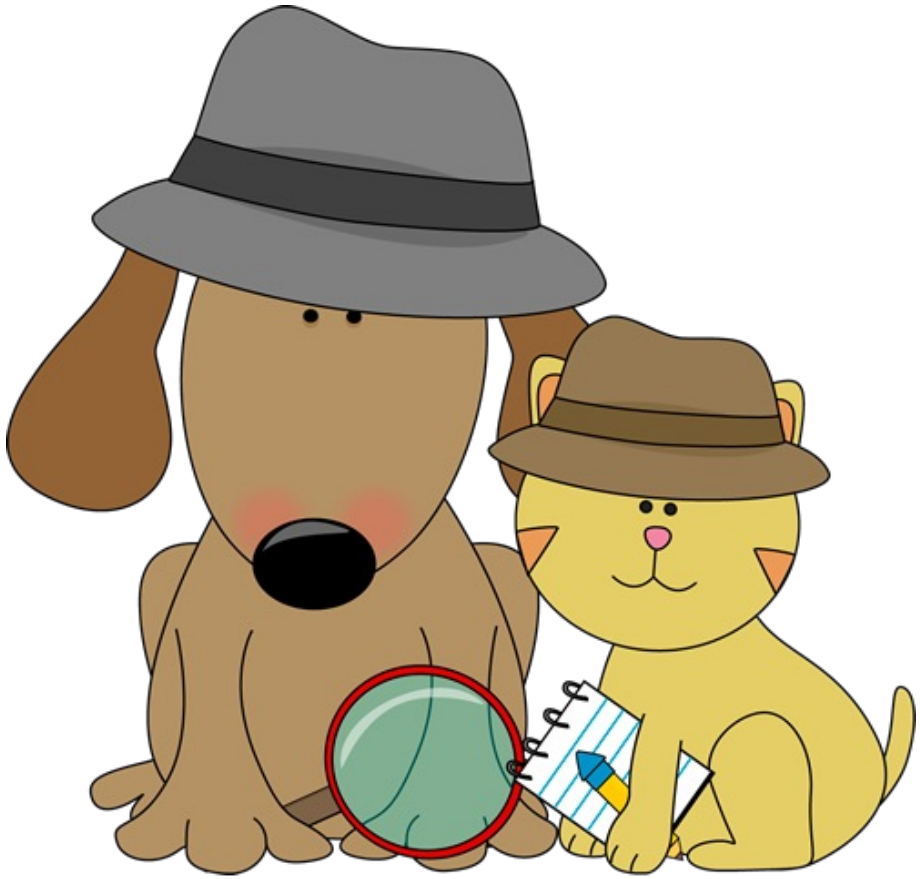
Tufts

Plot of Heavy Tailed Data
Paid Losses

- Minimum
- 1st quartile
- Median
- 2nd quartile
- Maximum



Missing Data



Screening for Missing Data

		BUSINESS TYPE	Gender	Age	License Year
N	Valid	35,284	35,284	30,242	30,250
	Missing	0	0	5,042	5,034
Percentiles	25			27.00	1,986.00
	50			35.00	1,996.00
	75			45.00	2,000.00

Blanks as Missing

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid		5,054	14.3	14.3	14.3
	F	13,032	36.9	36.9	51.3
	M	17,198	48.7	48.7	100.0
	Total	35,284	100.0	100.0	

Types of Missing Values

- Missing completely at random
- Missing at random
- Informative missing

Methods for Missing Values

- Drop record if any variable used in model is missing
- Drop variable
- Data Imputation
- Other
  - CART, MARS use surrogate variables
  - Expectation Maximization

CART: Classification and Regression Trees
MARS: Multivariate Adaptive Regression Splines

Imputation

- A method to “fill in” missing value

- Use other variables (which have values) to predict value on missing variable
- Involves building a model for variable with missing value
  - $Y = f(x_1, x_2, \ldots, x_n)$

Example: Age Variable

- About 14% of records missing values
- Imputation will be illustrated with simple regression model
  - Age = $a + b_1X_1 + b_2X_2 + \ldots + b_nX_n$

Model for Age

Tests of Between-Subjects Effects

Dependent Variable: Age

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
	Corrected Model	3,218,216	24	134,092	1,971.2	0.000
	Intercept	9,255	1	9,255	136.0	0.000
	ClassCode	3,198,903	18	177,717	2,612.4	0.000
	CoverageType	876	3	292	4.3	0.005
	ModelYear	7,245	1	7,245	106.5	0.000
	No of Vehicles	2,365	1	2,365	34.8	0.000
	No of drivers	3,261	1	3,261	47.9	0.000
	Error	2,055,243	30,212	68		
	Total	46,377,824	30,237			
	Corrected Total	5,273,459	30,236			

Missing Values

- A problem for many traditional statistical models
  - Elimination of records missing on anything from analysis
- Many data mining procedures have techniques built in for handling missing values
- If too many records missing on a given variable, probably need to discard variable

Metadata

Metadata

- Data about data
  - A reference that can be used in future modeling projects

- Detailed description of the variables in the file, their meaning and permissible values

Marital Status Value	Description
1	Married, data from source 1
2	Single, data from source 1
4	Divorced, data from source 1
D	Divorced, data from source 2
M	Married, data from source 2
S	Single, data from source 2
Blank	Marital status is missing