Feature Analysis and Visualization

MIE223

Winter 2025

1 Feature Analysis and Visualization

1.1 Exploratory Data Analysis (EDA)

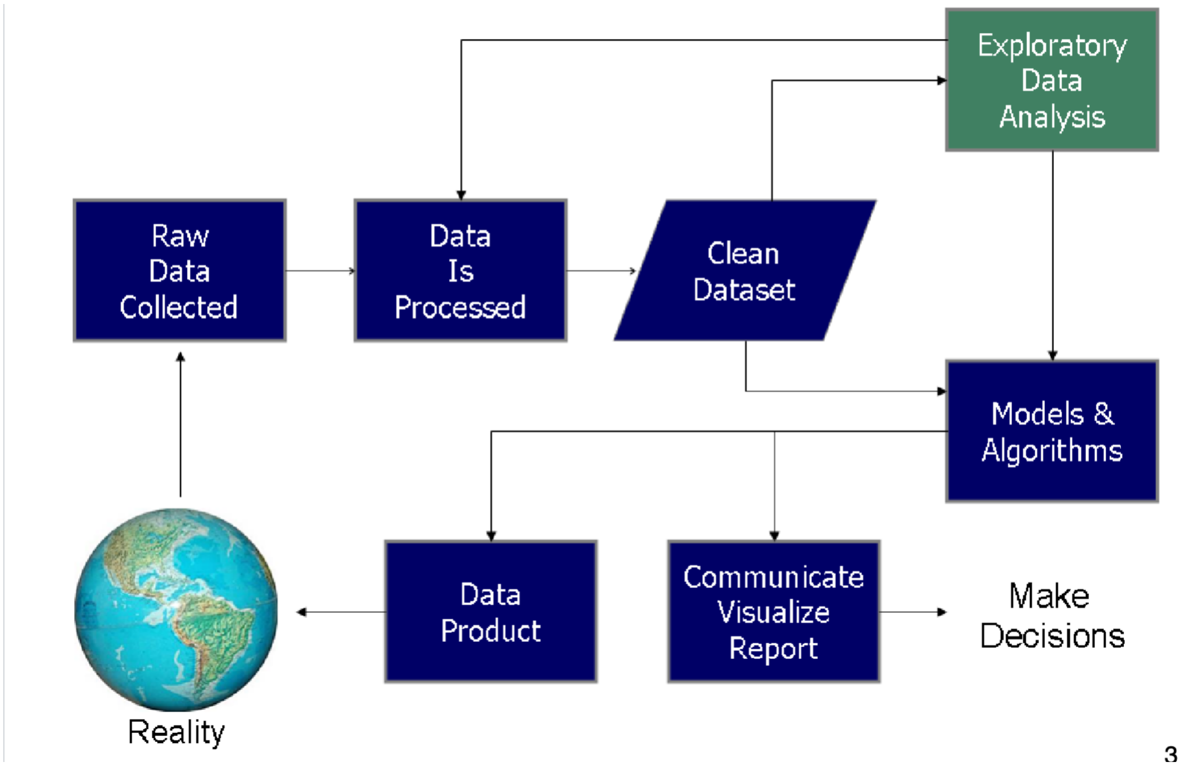
John Tukey (Bell Labs, 1970's) (Inventor of the box-plot)

- The first time corporations had a lot of digital data
- Data on semiconductor processes, networks

Observed that statistics was preoccupied with hypothesis testing

- But there were no established methodologies for **generating (data-driven) hypotheses**
- Espoused a visual methodology and a new language S (R became the free version)

1.2 Data Science Process



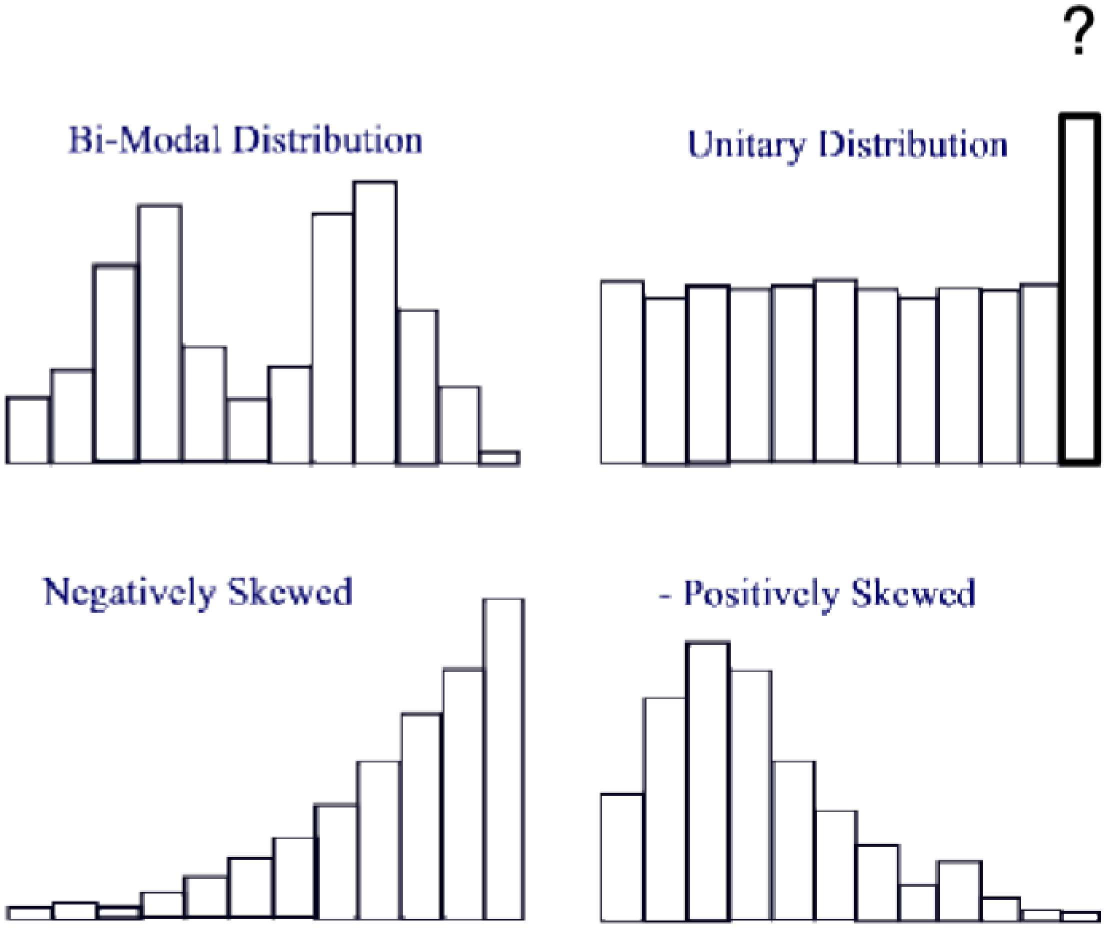
1.3 Data Cleaning: Brief Recap

80% of your Data Science time

Some pointers:

- Missing values – do not replace with 0 or -999

- Treat as missing
- Look for frequencies of values or outliers (-999)
  - **Histograms** invaluable
  - Examine **outliers**



2 Correlation

We'll be look at feature distributions independently (per column), but also relationships between features (columns). Correlation is one of the first tools you think of for looking at the relationship between two features (but has limitations)

2.1 Correlated variables

Two variables are correlated if changes in one variable, correspond to changes in the other
Correlation in general refers to the statistical association between two variables

- This statistical association allows us to make estimates for the one variable based on the value of the other

- While we typically think of linear relationships between variables, nonlinear might exist too

2.2 Linear correlation

- Two variables x and y are said to be linearly correlated if their relationship can be described by the following equation:

$$y = \alpha + \beta x, \beta \neq 0 \tag{1}$$

- This relationship will not be exact
- There will be an error associated with it, and hence, in reality:
- ϵ is the error term for the relationship

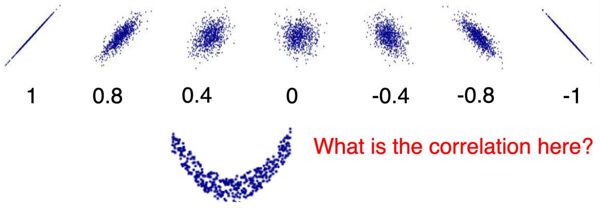
$$y = \alpha + \beta x + \epsilon, \beta \neq 0 \tag{2}$$

2.3 Pearson correlation coefficient

- In order to test the linear association between two variables x and y we can use the Pearson correlation coefficient r$_{xy}$

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Pearson correlation is in the range [-1,1]
  - 1: perfect/strong and positive linear correlation
  - -1: perfect/strong and negative linear correlation
  - 0: no linear correlation
- It is very crucial to understand that this correlation coefficient can only examine linear associations between two variables



What is the correlation here?

Note 1.

- Pearson correlation coefficient is sensitive to outliers
- It is not robust to outliers
- It is not a good measure of correlation for non-linear relationships

3 Feature Analysis: Probability Recap

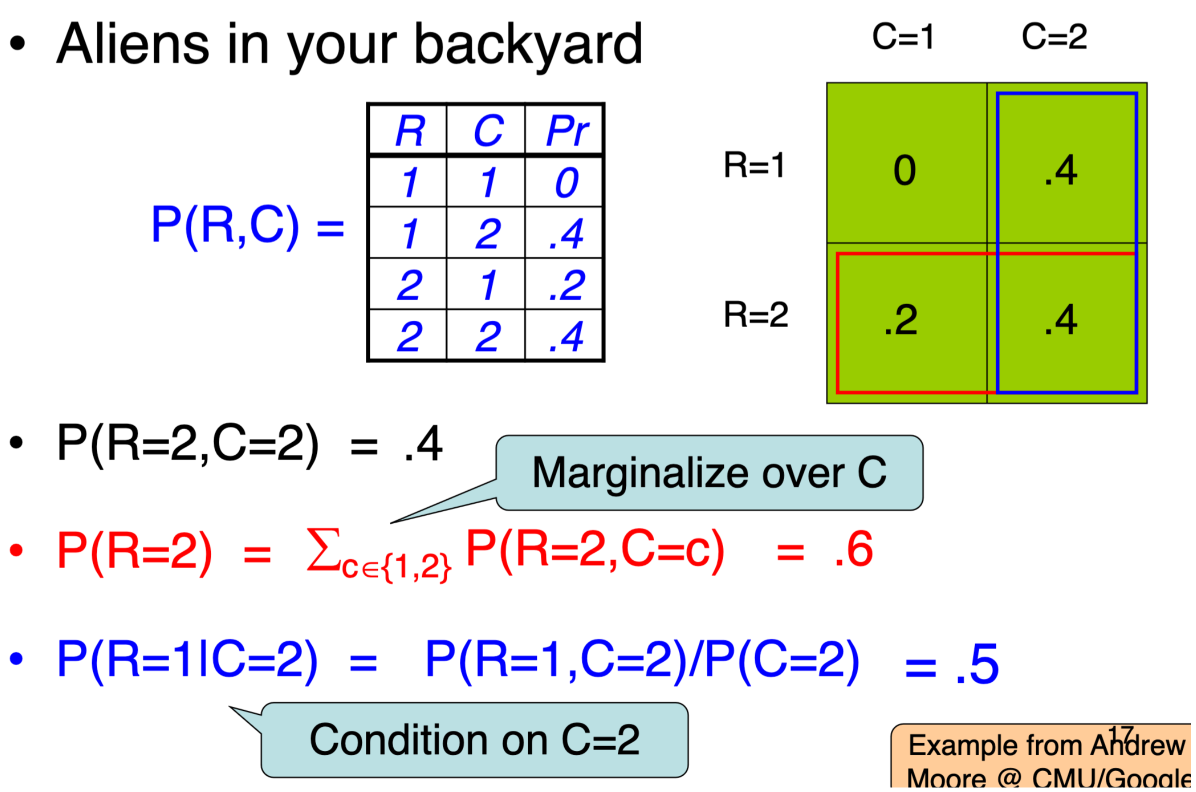
This is mathematical so we first we need to make sure everyone has the same understanding and intuition for notation

3.1 Random Variables

- Random variable (RV) denoted by uppercase letter (e.g. X)
- RVs take value assignments X = x where X is a
  - Discrete RV if x is in a countable set (binary: $x \in 0,1$; or dice)
  - Continuous RV if x is in an uncountable set (real: $x \in$ Reals)
- Write $x \in X$ for possible value assignments of X
- For all x, $P(X = x) \in [0,1]$
- P(X = x) is a proper distribution
  - Discrete RV:
  - $\sum_{x \in X} P(X = x) = 1$
  - Continuous RV:
  - $\int_{x \in X} P(X = x)\, dx = 1$
- Write P(x) for P(X=x), write P(X) for full distribution
- Representing probability distributions
  - Discrete (finite) RV: tabular
  - Continuous (infinite) RV: function

3.2 Joint Distributions on RVs

Aliens in your backyard



- Aliens in your backyard

P(R,C) =

R	C	Pr
1	1	0
1	2	.4
2	1	.2
2	2	.4

	C=1	C=2
R=1	0	.4
R=2	.2	.4

- P(R=2,C=2) = .4

Marginalize over C

- P(R=2) = $\Sigma_{c\in\{1,2\}}$ P(R=2,C=c) = .6
- P(R=1|C=2) = P(R=1,C=2)/P(C=2) = .5

Condition on C=2

Example from Andrew Moore @ CMU/Google

17

3.3 Rules of Probability

- Joint and conditional distributions:

$$P(X,Y) = P(X|Y)P(Y) = P(Y|X)P(X) \tag{3}$$

- Marginalization:

$$P(X) = \sum_y P(X,Y) \tag{4}$$

- Conditional probability & Bayes rule:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \tag{5}$$

3.4 Manipulating Distributions

- Sometimes we don't just want P(R=1,C=2) = .4
- We want to work with full distributions P(R,C)
- How to apply previous rules to full distributions?
- easy, just do once for each case and store in table
- Marginalization:

$$\sum_b P(A,b) \quad = \quad P(A)$$

$\sum_b$

A	B	Pr
0	0	.1
0	1	.3
1	0	.2
1	1	.4

=

A	Pr
0	.4
1	.6

- Binary Multiplication

$$P(A) \quad \cdot \quad P(B|A) \quad = \quad P(A,B)$$

A	Pr
0	.7
1	.3

·

A	B	Pr
0	0	.1
0	1	.9
1	0	.2
1	1	.8

=

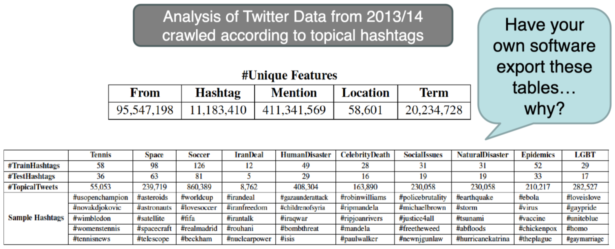
A	B	Pr
0	0	.07
0	1	.63
1	0	.06
1	1	.24

4 Feature Analysis

Frequency, Correlation, (Pointwise) Mutual Information, Conditional Entropy

4.1 Feature Analysis: Frequency

- Always produce summary frequency statistics
- Always look at samples of data



Analysis of Twitter Data from 2013/14 crawled according to topical hashtags

Have your own software export these tables… why?

#Unique Features

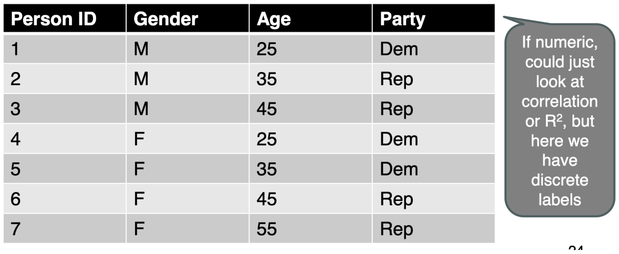
From	Hashtag	Mention	Location	Term
95,547,198	11,183,410	411,341,569	58,601	20,234,728

	Tennis	Space	Soccer	IranDeal	HumanDisaster	CelebrityDeath	SocialIssues	NaturalDisaster	Epidemics	LGBT
#TrainHashtags	58	98	126	12	49	28	31	31	52	29
#TestHashtags	36	63	81	5	29	16	19	19	33	17
#TopicalTweets	55,053	239,719	860,389	8,762	408,304	163,890	230,058	230,058	210,217	282,527
Sample Hashtags	#usopenchampion	#asteroids	#worldcup	#irandeal	#gazaunderattack	#robinwilliams	#policebrutality	#earthquake	#ebola	#loveislove
	#novakdjokovic	#astronauts	#lovesoccer	#iranfreedom	#childrenofsyria	#ripmandela	#michaelbrown	#storm	#virus	#gaypride
	#wimbledon	#satellite	#fifa	#irantalk	#iraqwar	#ripjoanrivers	#justice4all	#tsunami	#vaccine	#uniteblue
	#womenstennis	#spacecraft	#realmadrid	#rouhani	#bombthreat	#mandela	#freetheweed	#abfloods	#chickenpox	#homo
	#tennisnews	#telescope	#beckham	#nuclearpower	#isis	#paulwalker	#newnjgunlaw	#hurricanekatrina	#theplague	#gaymarriage

4.2 Feature (In)dependence

Often we want to know whether one feature column predicts another (bivariate feature analysis)

- What do you see here?
- How would quantify / visualize / automatically detect this?

- 

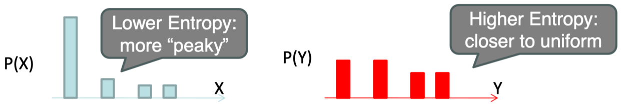
Person ID	Gender	Age	Party
1	M	25	Dem
2	M	35	Rep
3	M	45	Rep
4	F	25	Dem
5	F	35	Dem
6	F	45	Rep
7	F	55	Rep

If numeric, could just look at correlation or R^2, but here we have discrete labels

4.3 Feature Independence

- X is independent of Y means that knowing Y does not change our belief about X
- p(X—Y=y) = p(X)
- In words: knowing Y=y has no impact on our belief in the distribution of X
- Occurs if: p(X=x, Y=y) = p(X=x)*p(Y=y)

4.4 Information Theory

- P(X) encodes our uncertainty about X
- Some variables are more uncertain that others
- 

- How can we quantify this intuition?
- Entropy: average number of bits required to encode X
- $$H_P(X) = E\left[\log \frac{1}{p(x)}\right] = \sum_x P(x) \log \frac{1}{P(x)} = -\sum_x P(x) \log P(x)$$

  Higher entropy is more uncertain, lower entropy is more certain!
- In short: low bits, low entropy → low uncertainty
- We can define conditional entropy (CE) similarly
- $$H_P(X \mid Y) = E\left[\log \frac{1}{p(x \mid y)}\right] = H_P(X,Y) - H_P(Y)$$
- i.e. once Y is known, we only need H(X,Y) – H(Y) bits
- Higher CE is more uncertain (y is less predictive), lower CE is more certain (y is more predictive)

4.5 (Pointwise) Mutual Information

Note 2. Mutual information is just a test of independence. It only applies to discrete variables. Correlation does not apply to binary variables. If X and Y are non-linear, MI will capture that unlike correlation coefficient.

- Mutual Information
- $$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p(x)\,p(y)}\right)$$
- For discrete random variables: can also be generalized to continuous case
- PMI: target specific values x and y (pointwise)
- $$\operatorname{pmi}(x;y) \equiv \log\frac{p(x,y)}{p(x)p(y)} = \log\frac{p(x|y)}{p(x)} = \log\frac{p(y|x)}{p(y)}$$
- Lower MI and PMI → More independent, Higher MI and PMI → More dependent

4.6 Discrete Feature Analysis Summary

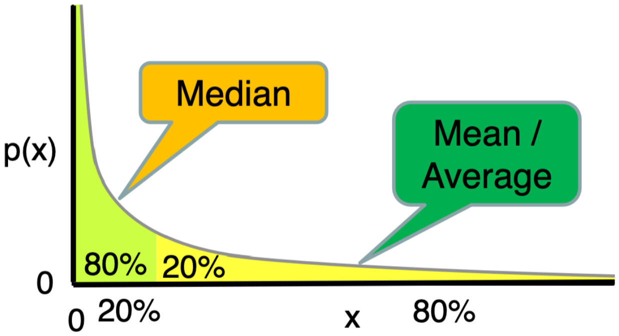
If we look at two feature columns X and Y, can analyze:

- MI(X,Y): test independence (fix neither variable)
  - MI=0 → independent,
  - The higher the value, the more dependence
  - Chi2(X,Y) / Chi-squared(X,Y): independence test similar to MI
- CE(X—Y=b): test how well Y=b predicts X (fix Y)
  - Low conditional entropy when X is well-predicted, why?
  - CE=0 → perfect prediction! CE=1 → complete noise!
- PMI(X=a,Y=b): test if X=a, Y=b dependent (fix X and Y)
  - Hard to interpret absolute numerical meaning (0 is still independent)
  - Primarily useful when ranking features... why?

4.7 Predictive Feature Analysis: Mutual Information (MI) or Pointwise MI

We can predict topics based on popular terms. The top 5 terms from a selected topic are the terms that have high mutual information. E.g., top 5 term features for Twitter topics – Use mutual information to know what is predictive / important

4.8 Mean, Median, and Power Laws

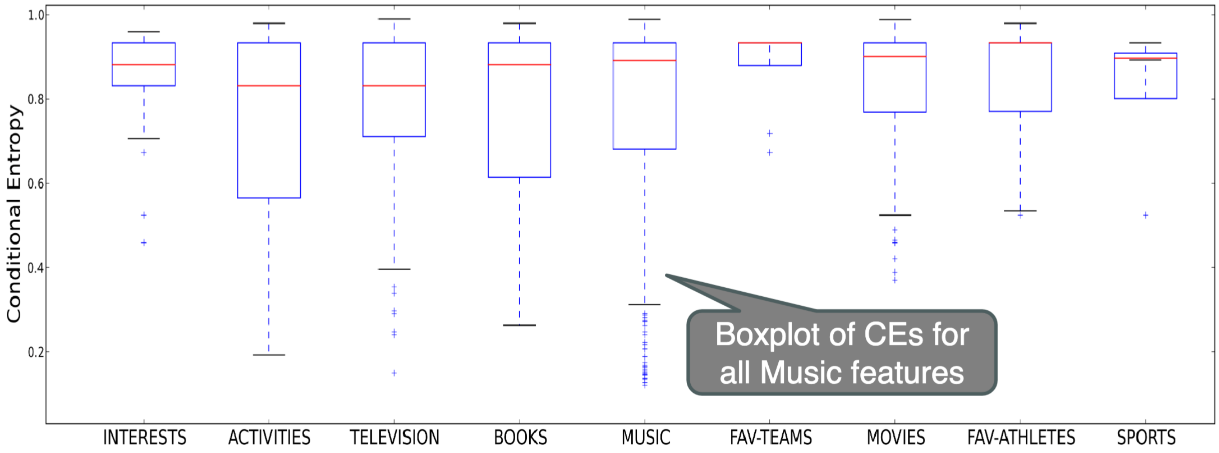
- A lot of human data is power law (income, #friends)
- Discrete power law (Zipf) for freq f:
- $p(f) = \alpha f^{-1-1/s}$
- Continuous power law:
- $P(X > x) \sim L(x)x^{-\alpha+1}$
- E.g., Probability mass of population with different income levels, or #friends on Facebook
- 

- Not symmetric like Gaussian, mean $\neq$ median
- I.e., the average case is much closer to the max case!
- 80/20 rule: top 20% of values account for 80% of mass

4.9 Box Plots

Most data is not Gaussian, often power law or log-Normal. Lower is more predictive.

How well can I predict a user's liking behavior conditioned on observing a feature, e.g., you like "U2" or "Breaking Bad"



An example where most informative = mean informative (but not median)... why?

Median Informative

Median Informative Favourites by Category

Movies	Music	Television
Forrest Gump	John Lennon	Futurama
Pretty Woman	U2	Star Trek
Napoleon Dynamite	AC/DC	The Trap Door
Harry Potter	The Smashing Pumpkins	Drawn Together
Toy Story 3	Gotye	Sherlock(Official)
The Godfather	The Rolling Stones	Hitchhiker's Guide to the Galaxy
Mulan	All Axess	Buffy The Vampire Slayer
How to Train Your Dragon	Steve Aoki	South Park

Most Informative

Most Informative Favourites by Category

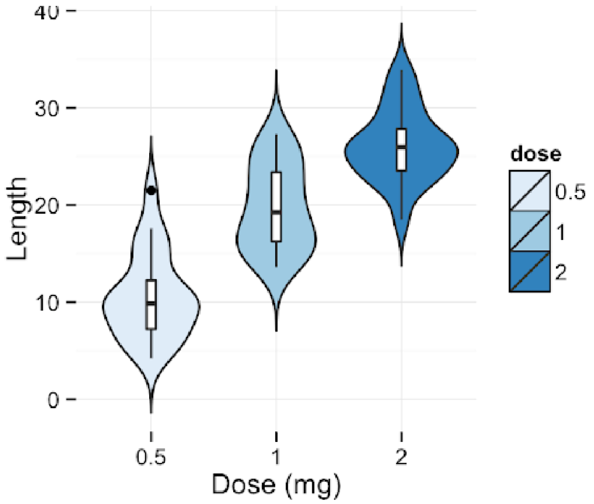
Movies	Music	Television
Billy Madison	Avascular Necrosis	Metalocalypse
Team America: World Police	Tortured	Beast Wars
Pan's Labyrinth	Elysian	Hey Arnold!
Pirates of the Caribbean	Anno Domini	Sherlock
Aladdin	Darker Half	Hey Hey It's Saturday
Starship Troopers	Hellbringer	Neil Buchanan and Art Attack!
Happy Gilmore	Johnny Roadkill	Breaking Bad
Timon and Pumbaa	Aeon of Horus	Red vs. Blue

- Median favorites were largely generic. The less popular it is the more informative it is.
- Most informative (max ≈ avg) were largely specialized

4.10 Violin Plots

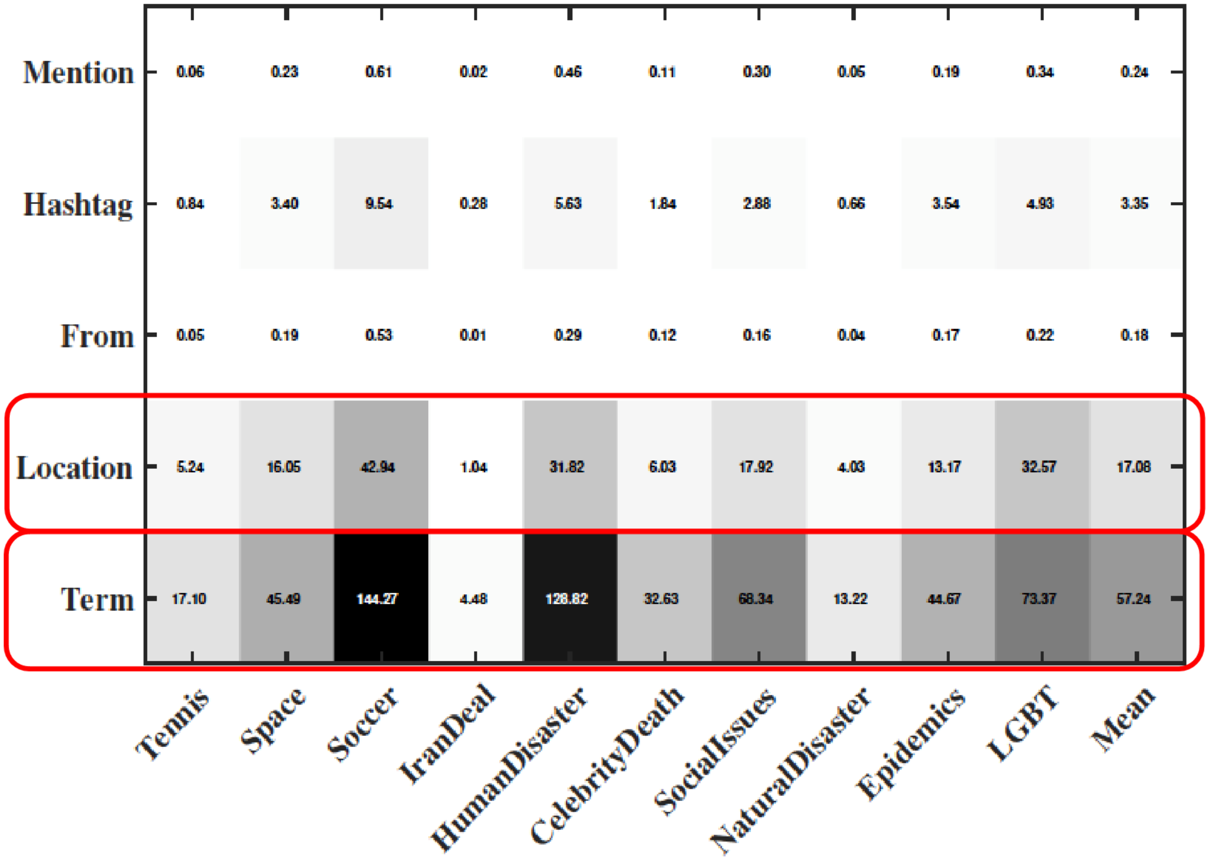
Compactly show full distributions side-by-side. Like compact histogram, can show bimodality unlike Boxplot

Plot of tooth length vs. vitamin C dose



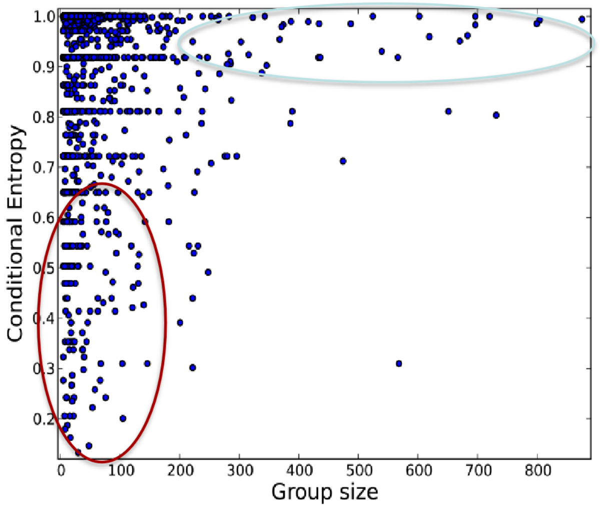
4.11 Discrete Heatmaps

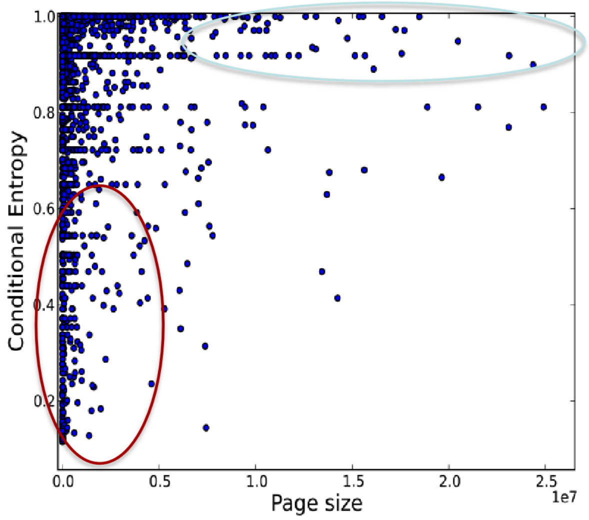
Better than a table... visualize numbers!



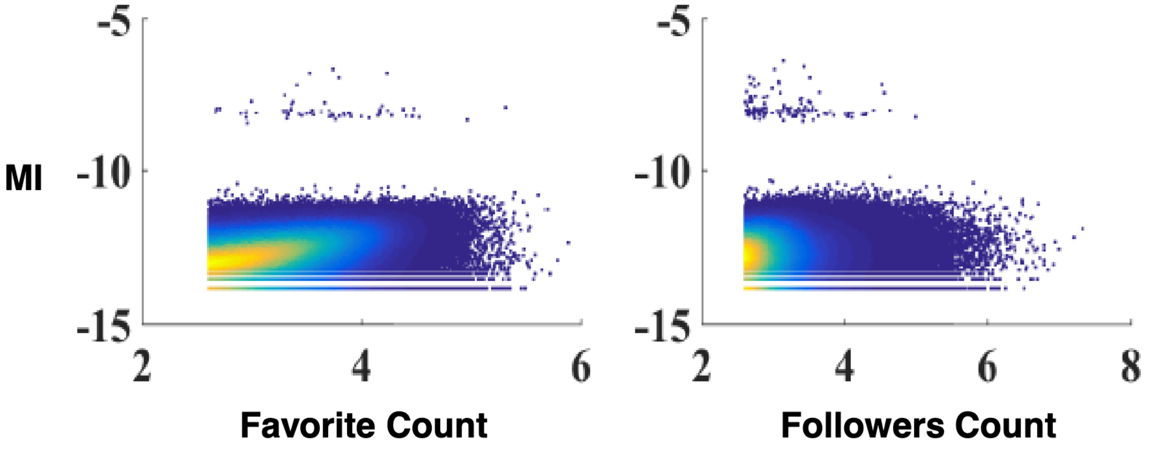
4.12 Scatterplots: don't need linear relationship to be informative!

- Large groups tend not to be predictive
- Most informative groups were small
- The larger the group/page size, the less predictive it is





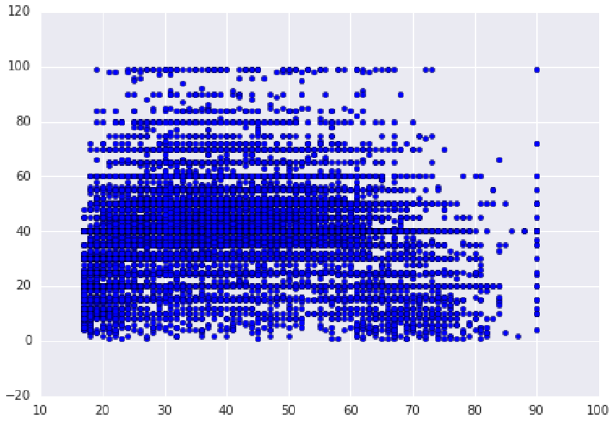
4.13 Heatmaps / Density Plots



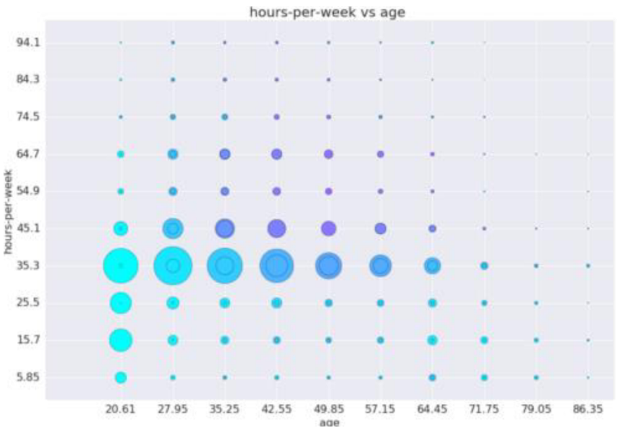
Important to use when points in a scatterplot are dense and do not reflect density. These show density.

4.14 Bubble Plots

Sometimes scatterplots not dense enough for a heatmap



- And may have a third data dimension
- So don't want density=color

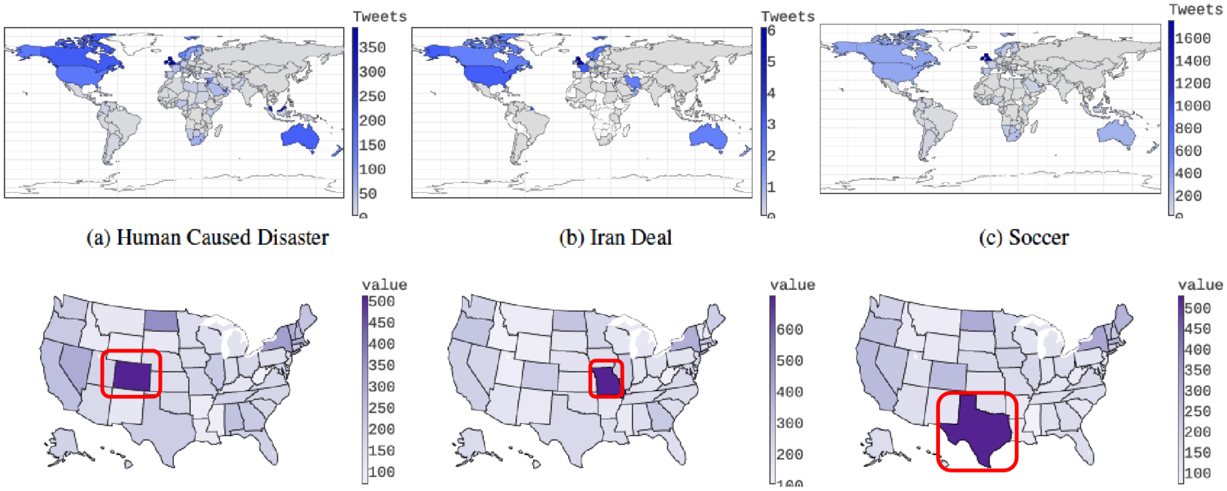


Enter the bubble plot

- X-axis: age
- Y-axis: hours worked per week
- Bubble size is relative mass (density in a heatmap)
- Color is income (purple=higher)

4.15 Choropleths

Example: per capita tweet frequency across different international and U.S. locations for different topics. These are all English tweets, so the data is skewed to English-speaking countries. French Guyana is identified as a territory of France so it's colored as France.
d) There were bubonic plague outbreaks in Colorado in 2014
e) Start of BLM, murder of Michael Brown
f) NASA space centres are in Texas and Florida
Most tweets about an event are located where the event occurred.



(a) Human Caused Disaster

(b) Iran Deal

(c) Soccer

(d) Health Epidemics

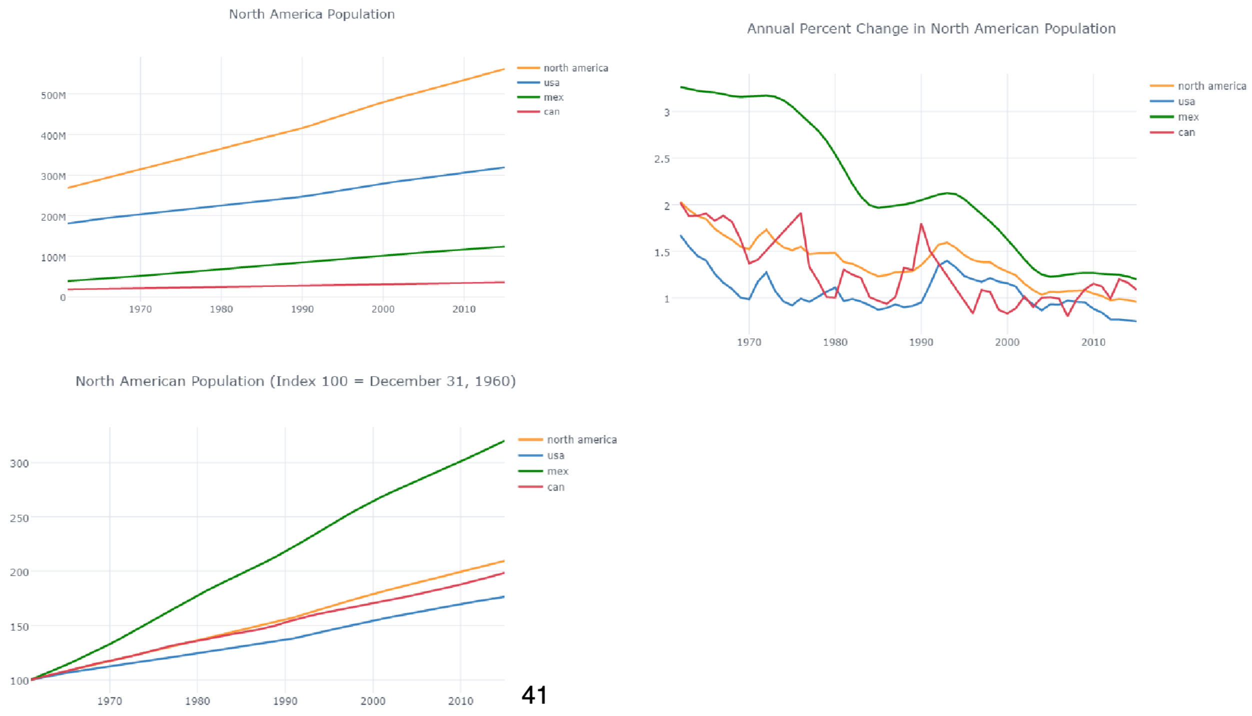
(e) Social Issues

(f) Space

4.16 More than one way to plot time series!

E.g. population vs year
E.g. percent change in population
E.g. proportion 1960 population to be 100 (indexing)

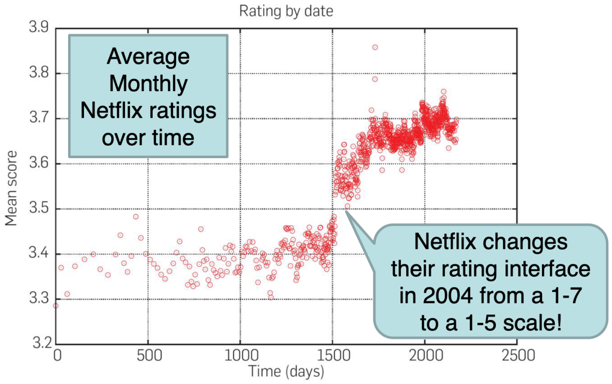
- Sometimes % change or relative growth vs. index is better than the obvious plot of values vs. time
- Note the different insights from these three examples of the same data!



5 Multimodality and Confounders

Your data is a mix of generative processes: try to understand what those generative processes are.

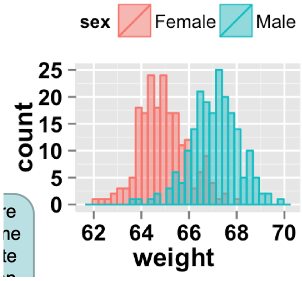
5.1 Time Series and Data Cleaning



- Always evaluate sufficient statistics of data over time
- Don’t ignore a shift... evidence that something changed
    - Can try to correct if understand source of shift (this was caused by UI change, changed to 1-5 scale)

5.2 Overlapping Histograms

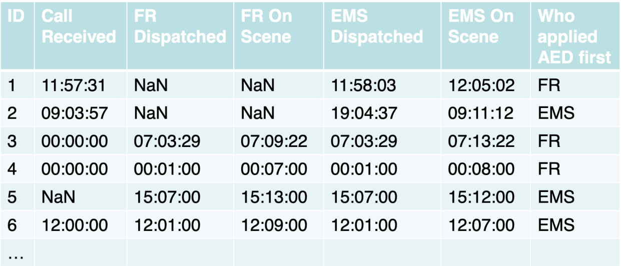
Is there a data dependence on a discrete variable? Multimodalities are always due to some underlying discrete confounder(s): can be explicit or latent!



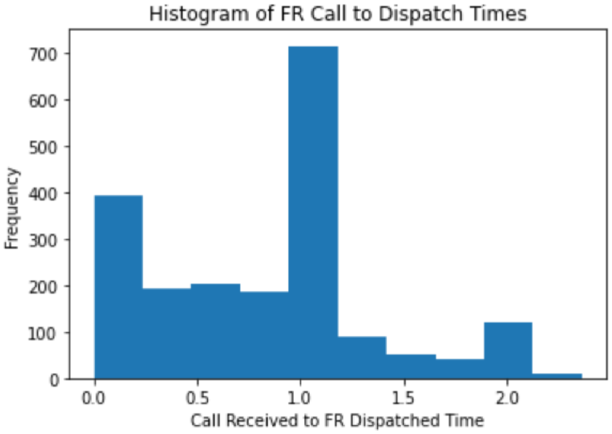
6 EDA Case Study with Confounders

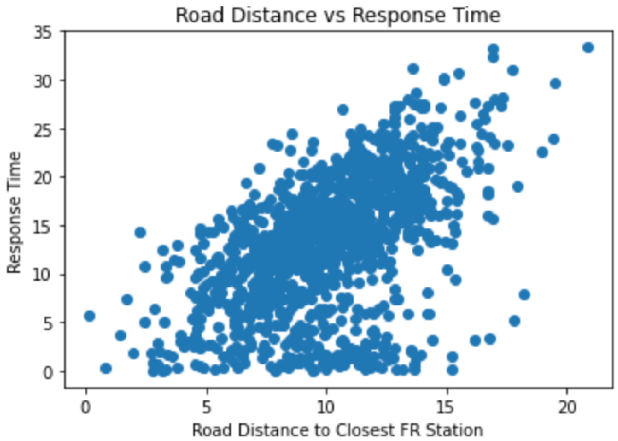
Analysis of Ambulance Response Times

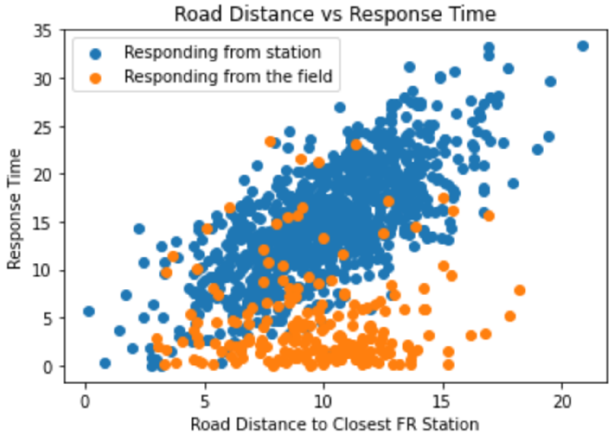
6.1 Cardiac Arrest - Case Study



ID	Call Received	FR Dispatched	FR On Scene	EMS Dispatched	EMS On Scene	Who applied AED first
1	11:57:31	NaN	NaN	11:58:03	12:05:02	FR
2	09:03:57	NaN	NaN	19:04:37	09:11:12	EMS
3	00:00:00	07:03:29	07:09:22	07:03:29	07:13:22	FR
4	00:00:00	00:01:00	00:07:00	00:01:00	00:08:00	FR
5	NaN	15:07:00	15:13:00	15:07:00	15:12:00	EMS
6	12:00:00	12:01:00	12:09:00	12:01:00	12:07:00	EMS
…						







What abnormalities would you investigate further?
Odd times, don't know what format they're in. 00:00:00 is missing data
What would you plot?

Cause of modes: artifacts from data being recorded to the minute.
The farther the distance the longer the response time. But, ambulances positioned on field are quicker to respond regardless of distance

7 Feature Analysis and Viz: Recap

- Explore and visualize whenever possible
- Tabular summaries: frequency, summary statistics
  - Much (human-generated) data is not Gaussian
  - Be cautious with summary statistics like mean, standard deviation
  - Median, quartiles, quantiles, full distribution much better (boxplot, histogram)
- Dependency analysis: (pointwise) mutual information, CE, Chi2
- Histogram, Box plot, Violin plot for univariate / bivariate data
- Scatter plot, Heatmap / Density plot for bivariate data
- Bubble plot for 3 or 4 dimensional data
- Choropleth for geographical data
- Beware of bimodalities
  - Indicate important latent variables

7.1 Extra

Univariate has discrete or numeric/continuous columns.

Discrete: bar plot of frequency

Numeric/Continuous: histogram, box plot, violin plot

Discrete heatmap of MI, PMI can produce correlation between D and d

Scatter plot, heatmap is good for C and C

Box, violin, overlapping histogram plots for D and C