Introduction to Data Science

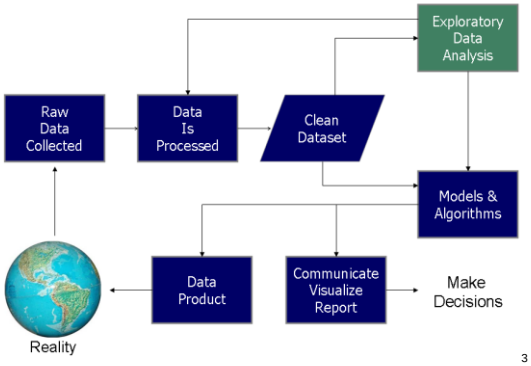
Scott Sanner

Exploratory Data Analysis

- John Tukey (Bell Labs, 1970's)
  - 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)

2

Data Science Process



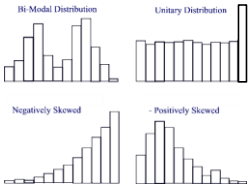
3

https://en.wikipedia.org/wiki/Exploratory_data_analysis

Data Cleaning

- 50% of your Data Science time
  - It gets two slides in this lecture… good luck!
- 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**

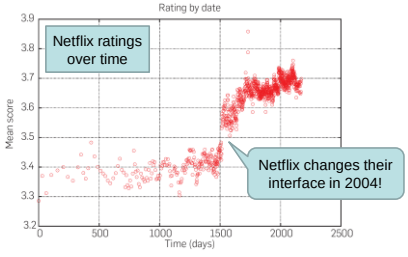
?



4

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

5

Data Analysis: Some Key Principles

- **Most human-generated data is not Gaussian**
  - Be cautious with summary statistics like mean, standard deviation
  - Median, quartiles, quantiles much better (e.g., boxplot, histogram)
- Explore and visualize whenever possible
  - Tabular summaries: frequency and mutual information
  - Histogram, Box plot, Violin plot for univariate data
  - Scatter plot, Heatmap / Density plot for bivariate data
  - Bubble plot for three dimensional data
  - Choropleth for geographical data
- Beware of bimodalities
  - Indicate important latent variables

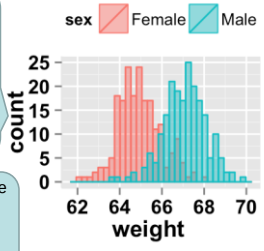
7

Overlapping Histograms

- Is there a data dependence on a discrete variable?

How would you do this for purchase behavior vs. price? Plot #purchases per person vs. price = 0 or not.

If you believe there is an unknown latent variable, neural networks can help.



8

Data Description

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

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	#tripmandela	#michaelbrown	#storm	#virus	#gaypride
		#wimbledon	#satellite	#fifa	#irantalk	#iranpeace	#tripjeanrivers	#justiceforall	#tsunami	#vaccine	#nuteltblue
		#womenstennis	#spacecraft	#realmadrid	#rouhani	#bombthreat	#mandela	#freetheweed	#allbloods	#chickenpox	#homo
		#tennisnews	#telescope	#beckham	#nuclearpower	#isis	#paulwalker	#newyiguanlaw	#hurricanekatrina	#theplague	#gaymarriage

9

(Pointwise) Mutual Information

- Mutual Information

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right)$$

If y constant, just rank by numerator!

- PMI: target specific values x and y

$$\text{pmi}(x;y) \equiv \log \frac{p(x,y)}{p(x)p(y)} = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)}$$

- Good descriptions of MI and PMI here:
  https://en.wikipedia.org/wiki/Mutual_information
  https://en.wikipedia.org/wiki/Pointwise_mutual_information

10

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

- E.g., top 5 term features for Twitter topics
  - Use mutual information to know what is predictive / important

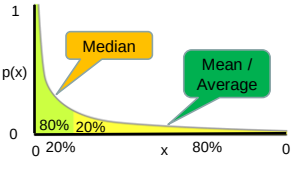
Topics/Top10	NaturalDisaster	Epidemics	IranDeal	SocialIssues	LBGT	HumanDisaster	CelebrityDeath	Space	Tennis	Soccer
Term	philippines	health	iran	police	obama	israel	robin	online	murray	madrid
Term	donate	ebola	regime	protesters	gun	gaza	williams	movistar	tennis	goal
Term	typhoon	aces	nuclear	officer	rights	israeli	nelson	encro	federer	cup
Term	affected	medical	iranian	protest	america	killed	mandela	imperdible	djokovic	manchester
Term	relief	virus	resistance	cops	gop	children	cory	greet	nadal	match

11

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., Rank Google queries from most to least frequent and look at their probability:

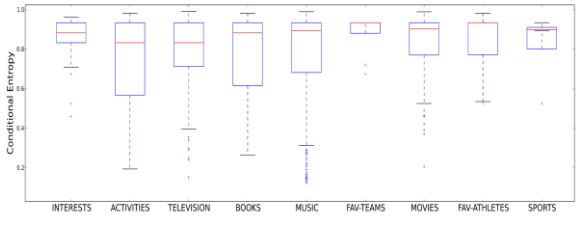


- Not symmetric like Gaussian, mean ≠ median
  - I.e., the average case is much closer to the max case!
  - 80/20 rule: top 20% of values account for 80% of mass

12

Box Plots

- Most data is not Gaussian, often power law or log-Normal



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

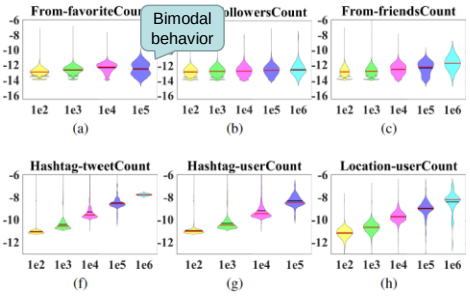
Most Informative Favourites by Category		
Movies	**Music**	**Television**
Billy Madison	Avascular Necrosis	Metalocalypse
Team America: World Police	Tortured	Beast Wars
Pan's Labyrinth	Elysian	Hey Arnold!
Pirates of the Caribbean	Anno Domini	Sherlock
Aladdin	Darker Half	Hey Hey It's Saturday
Starship 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
- Most informative (max ≈ avg) were largely specialized

14

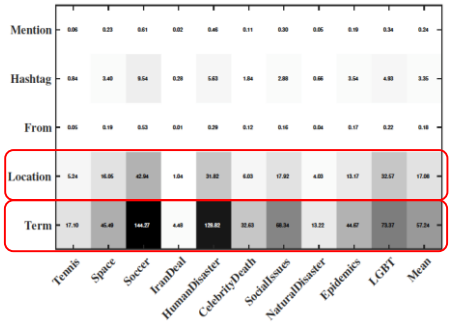
Violin Plot

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



Discrete Heatmaps

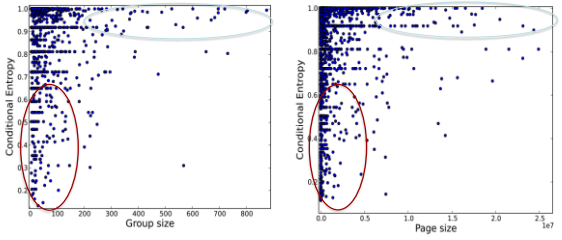
- Better than a table… visualize numbers!



16

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

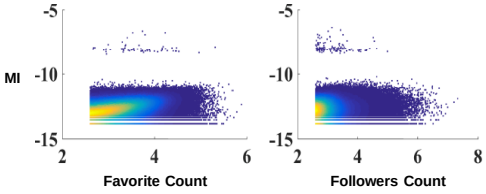
Large groups tend not to be predictive



Most informative groups were small

17

Heatmaps / Density Plots

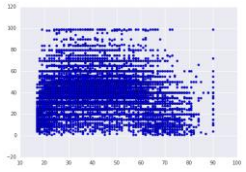


- Important to use when points in a scatterplot are dense and do not reflect density

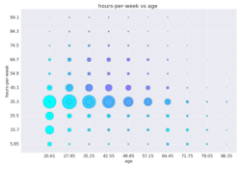
18

Bubble Plots

- Sometimes scatterplots not dense enough for a heatmap
  - And may have a **third data dimension**
  - So don't want density=color

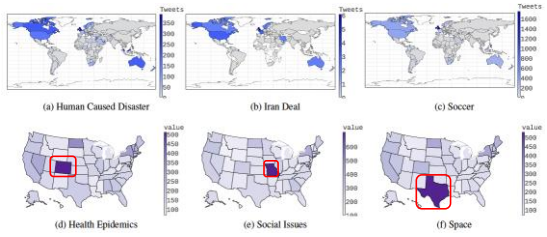


- Enter the bubble plot
  - X-axis: age
  - Y-axis: hours worked per week
  - Bubble size is relative mass (density in a heatmap)
  - Color is income (purple=higher)



Choropleths

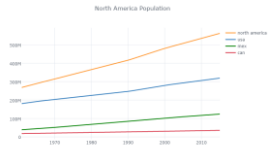
- Example: per capita tweet frequency across different international and U.S. locations for different topics

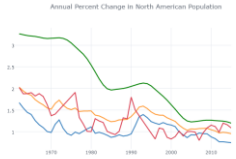


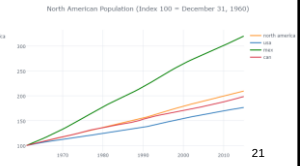
20

More than one way to plot time series!

- 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!







21