CSE 4/587
Data Intensive Computing

Dr. Eric Mikida
epmikida@buffalo.edu
208 Capen Hall

Day 03
Data Strategy and EDA

Announcements and Feedback

- Start reading Chapter 2 in Doing Data Science
  - Work through EDA example on your own
- Start reading Chapter 10 in Data Science from Scratch
  - If you are not comfortable with Python, also read Chapter 1

Recap from Last Class

- Intelligent Applications
  - Good intelligence requires a lot of information
  - Good information requires a lot of data
  - Intelligent applications are inherently data-intensive
- Data Strategy
  - How will you gather, process, visualize your data?
  - What considerations do you need to take into account?

Steps to Consider

1. **Frame the problem:** Understand the use case
2. **Understand the data:** Exploratory data analysis
3. **Extract features:** What parts of the data are important to you…
4. **Model the data and analyze:** How do we plan to get meaning from the data
5. **Design, code and experiment:** Use tools to clean, extract, plot, view
6. **Present and test results:** Two types of clients - humans and systems
7. **Iterate:** Go back to any of the steps based on the insights!

1. Frame the Problem

Frame the Problem

- Have a standard use case format (What, why, how, stakeholders, data in, info out, challenges, limitations, scope etc.)
- Refer to your software engineering course
- Statement of work (SOW): clearly state what you will accomplish

2. Understand the Data

Understand the Data

- Data represents the traces of real-world processes
  - What traces we collect depends on the sampling methods
  - You build models to understand the data and extract meaning and information from the data: statistical inference
- Two sources of randomness and uncertainty
  - The process that generates data is random
  - The sampling process itself is random
- Your mind-set should be “statistical thinking in the age of big-data”
  - Combine statistical approach with big-data

Questions to ask

- How big is the data?
- Any outliers?
- Missing data? How to address it? (Clean our data…)
- Sparse or dense?
- Collision of identifiers in different sets of data

New Kinds of Data

- **Traditional:** numerical, categorical, or binary
- **Text:** emails, tweets, NY times articles
- **Records:** user-level data, time-stamped event data, json formatted log files
- **Geo-based location data**
- **Network data** (How do you sample and preserve network structure?)
- **Sensor data**
- **Images**

Uncertainty and Randomness

- A mathematical model for uncertainty and randomness is offered by probability theory.
- A world/process is defined by one or more variables. The model of the world is defined by a function:
  - **Model** = f(w) or f(x,y,z) (A multivariate function)
  - The function is unknown, model is unclear, at least initially. Typically our task is to come up with the model, given the data.
- **Uncertainty**: is due to lack of knowledge - consider predicting the weather
- **Randomness**: is due lack of predictability - consider a die roll
- Both can be expressed by probability theory

Statistical Inference

- From **the world** -> collect data
- From the data -> capture the meaning through models or functions
- From the meaning -> devise statistical estimators for predicting things about **the world**

Statistical inference: development of procedures, methods, and theorems that allow us to extract meaning and information from data that has been generated by *stochastic (random) processes*

Population and Sample

- Population is complete set of traces/data points
  - US population: 314 Million, world population: 7 billion for example
  - All voters, all things
- Sample is a subset of the complete set (or population): **how we select the sample introduces biases into the data**
- See an example in http://www.sca.isr.umich.edu/
  - Here out of the 314 Million US population, 250,000 households form the sample
- Population -> mathematical model -> sample

Population and Sample

- Example: Emails sent by people in the CSE dept. in a year.
  - Method 1: 1/10 of all emails over the year randomly chosen
  - Method 2: 1/10 of people randomly chosen; all their email over the year
- Both are reasonable sample selection method for analysis.
- However estimations (probability distribution functions) of the emails sent by a person for the two samples will be different.

Big Data vs Statistical Inference

- Sample size N
  - For statistical inference N < All
  - For big data N == All
  - For some atypical big data analysis N == 1
    - World model through the eyes of a prolific twitter user
    - Followers of Ashton Kutcher: If you analyze the twitter data you may get a world view from his point of view

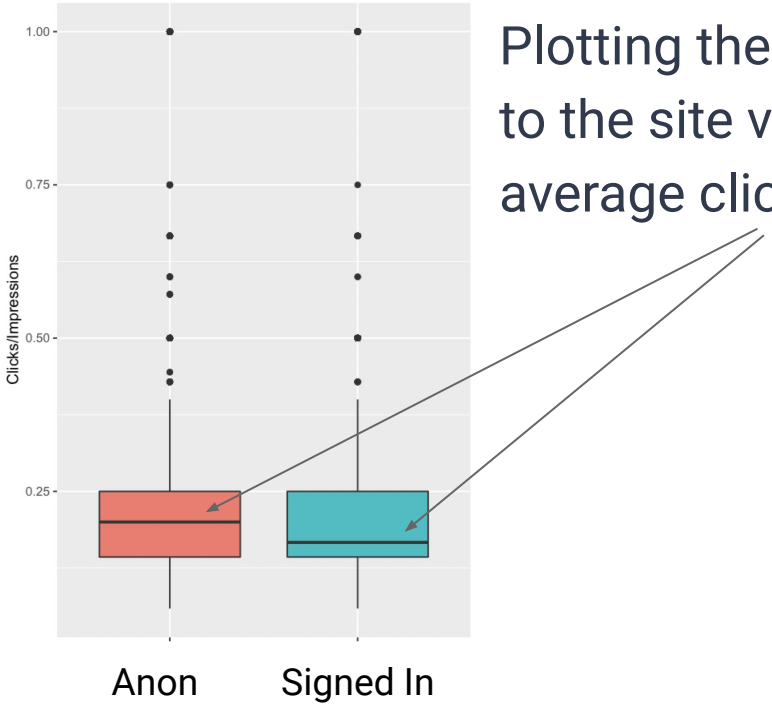
Big Data Context

- Sampling is still a valid solution, it depends on your needs
  - For quick analysis, or inference purposes you don't need all the data
  - At Google (at the originator big data algs.) people sample all the time.
- However, if you want to serve and render information in a UI, you cannot sample.
- Some DNA-based search you cannot sample.
- **Just because you have an entire population does not mean there is no bias**
  - Even taking, for example, the entirety of the data on Twitter…conclusions made cannot be extended beyond the population which uses Twitter
  - Another example is of the tweets pre- and post- hurricane Sandy..
  - Think about Yelp

Exploratory Data Analysis (EDA)

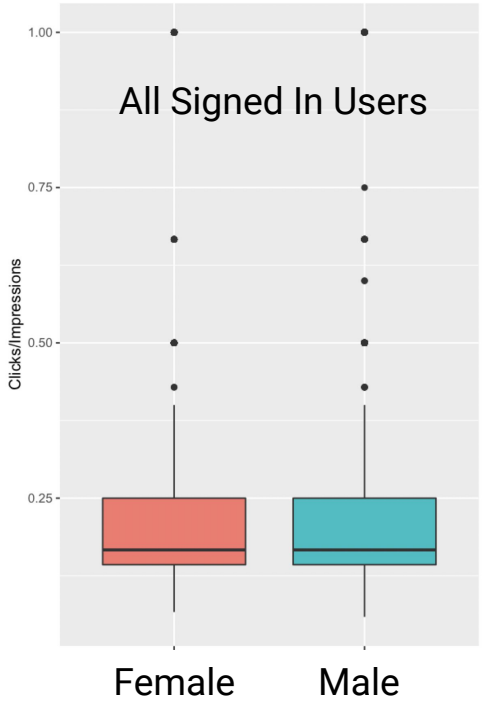
- By doing EDA, you achieve two things to get you started:
  - You get an intuitive feel for the data
  - You can start to make a list of hypotheses
- EDA is the prototype phase of ML and other sophisticated approaches
- Basic tools of EDA are plots, graphs, and summary stats (a lot of histograms)...
- It is a method for “systematically” going through data, plotting distributions, plotting time series, looking at pairwise relationships using scatter plots, generating summary stats.eg. mean, min, max, upper, lower quartiles, identifying outliers.
- **EDA is done to understand big data before using expensive big data methodology.**

Example from Doing Data Science (Ch. 2)



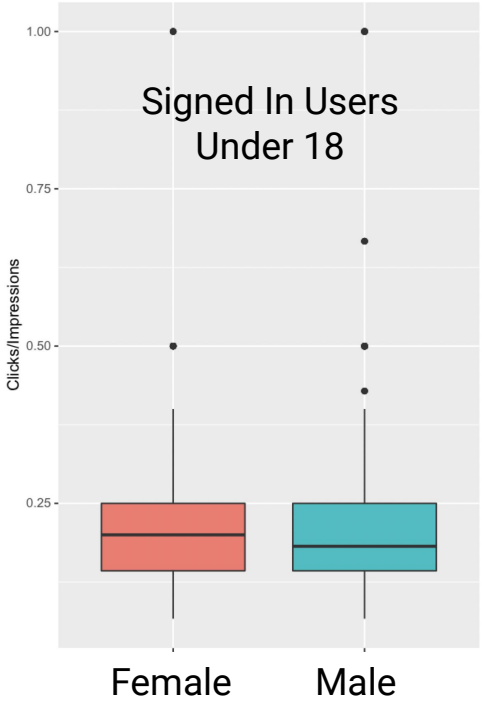
Plotting the click-through rate for anonymous visitors to the site vs those that are signed in shows a higher average click-through rate for anonymous users

Example from Doing Data Science (Ch. 2)



Looking at all users, there is no significant difference between male and female click-through rate…

…But restricting the population to those under 18 changes these assumptions



Example from Doing Data Science (Ch. 2)

- Previous plots were just from a single day
  - How does the data look across the whole month?
  - What if we look at the change across days?
- What are the sample sizes of each grouping?
- What are the outliers?

Example from Data Science from Scratch (Ch. 5)

- Consider dividing up data scientists you know into bins based on which coast they are from and finding the average number of friends in for each bin

Coast	# of members	Avg. # of friends
West Coast	101	8.2
East Coast	103	6.5

It would appear that the West Coast scientists are, by this metric, “friendlier”

…however

Example from Data Science from Scratch (Ch. 5)

- Consider dividing up data scientists you know into bins based on which coast they are from and finding the average number of friends in for each bin

If we also bin by degree, we see a different story…

East coast has a higher percentage of PhD members, but each bin is “friendlier”

Coast	Degree	# of members	Avg. # of friends
West Coast	PhD	35	3.1
East Coast	PhD	70	**3.2**
West Coast	No PhD	66	10.9
East Coast	No PhD	33	**13.4**

Example from Data Science from Scratch (Ch. 5)

- Consider dividing up data scientists you know into bins based on which coast they are from and finding the average number of friends in for each bin

If we also bin by degree, we see a different story…

East coast has a higher percentage of PhD members, but each bin is “friendlier”

Coast	Degree	# of members	Avg. # of friends
W			3.1
Ea			**3.2**
W			10.9
East Coast	No PhD	33	**13.4**

From this we can conclude that the degree may also be a relevant factor in “friendliness” and form new hypotheses

3. Extracting Features

Extract Features

- Data is clean, we've done EDA, now we need to extract what is useful
- Filter out only the important fields or features, say from a json file
- Often defined by the problem analysis (EDA) and use case defined
  - Example: location and temperature are the only important data in a tweet for a particular analysis
  - Consider the example from Doing Data Science (Ch. 2)
    - Depending on what you are trying do get, do you need information about age? gender? neither?

4. Modeling

Modeling

- Abstraction of a real world process
- Let's say we have a data set with two columns x and y and y is dependent on x, we could write (for example):
  - $y = \beta 1 + \beta 2x$ (linear relationship)
  - We do not know $\beta 1$, or $\beta 2$…we must find them
- How to build a model?
  - Unfortunately…that's not always straightforward
- Probability distribution functions (pdf) are building blocks of statistical models

Probability Distributions

- Normal, uniform, Cauchy, t-, F-, Chi-square, exponential, Weibull, lognormal, ...
- They are known as continuous density functions
- Any random variable, x, can be assumed to have probability distribution p(x), which maps x to a positive real number
- To be a probability density function, integrating p(x) to find the area must give 1
    - This allows us to determine the probability of a certain outcome by integrating area under the curve

Probability Distributions

- We can also combine functions to serve as joint distributions or conditional distributions:
  - p(x,y) is a multivariate function that determines the probability of both x and y occurring (area under the plane must also be 1)
  - p(x|y) is a conditional distribution: it is the probability of x occurring, given a particular value of y

Fitting a Model

- Estimating the parameters of the model: what distribution to use, what are the values of min, max, mean, stddev, coefficients for the distribution, etc.
- This functionality is readily provided in tools like R and python libraries
- It involves algorithms such as maximum likelihood estimation (MLE) and optimization methods…
  - Example: $y = \beta1+\beta2x \rightarrow y = 7.2 + 4.5*x$

5. Design, Code, Deploy

Design, code, deploy

- Design first before you code: an important principle
- Code using best practices and “Software engineering” principles
- Choose the right language and development environment
- Document within the code and outside
- Clear state the steps in deploying the code
- Provide troubleshooting tips

6. Present the Results

Present the Results

- Good annotated graphs and visuals are important explaining the results
- Annotate using text, markup and markdown
- Extras: provide ability to interact with plots and assess what-if conditions
- Explore
  - d3.js : https://d3js.org/
  - Tableau: https://www.tableau.com/academic
  - Python viz libraries (see Data Science from Scratch)
- And a lot of creativity! Do not underestimate this…you are the best person to figure out how to present your results effectively.

7. Iterate

ITERATE!

- Iterate through any of steps as warranted by the feedback and the results
- Data science process is an iterative process