COGS 9 – Discussion Section A01 and A02

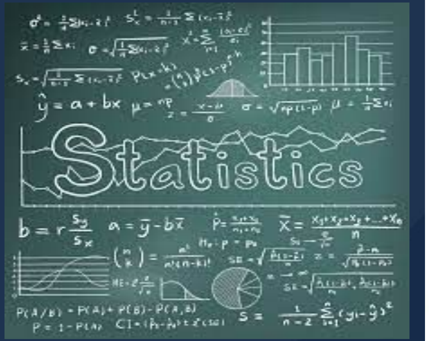
Kunal Rustagi (TA): Mon 9AM (Zoom)
Connor McManigal (IA)
Yupei Sun(IA)

50 Years of Data Science Part. 1

- Connects the discipline of DS to its 50+ years of history (John Tukey in 1960s)
- DS as an **extension** of statistics?
- Common Task Framework (e.g., Netflix Challenge)

UC San Diego

Tukey’s introductory paragraphs



DS as an **extension** of statistics ?

(*) Multidisciplinary investigations (25%)

(*) Models and Methods for Data (20%)

(*) Computing with Data (15%)

(*) Pedagogy (15%)

(*) Tool Evaluation (5%)

(*) Theory (20%)

DS as an **extension** of statistics ?

Inference model: To [infer] how nature is associating the response variables to the input variables.

Prediction model: To be able to predict what the responses are going to be to future input variables;

DS as an **extension** of statistics ?

Inference model: To [infer] how nature is associating the response variables to the input variables.

Prediction model: To be able to predict what the responses are going to be to future input variables;

Professor Breiman’s paper is an important one for statisticians to read. He and Statistical Science should be applauded ... His conclusions are consistent with how statistics is often practiced in business. -Bruce Hoadley

Common Task Framework and the secret sauce

- A publicly available training dataset
- A set of enrolled competitors
- A scoring referee

Common Task Framework and the secret sauce

Common Task Framework and the secret sauce

1. Error rates decline by a fixed percentage each year, to an asymptote depending on task and data quality.

2. Progress usually comes from many small improvements; a change of 1% can be a reason to break out the champagne.

3. Shared data plays a crucial role—and is reused in unexpected ways.

Donoho’s six divisions

- Data Gathering, Preparation, and Exploration
- Data Representation and Transformation
- Computing with Data
- Data Modeling
- Data Visualization and Presentation
- Science about Data Science

Data Gathering, Preparation, and Exploration

For example, a data team can ***gather*** data

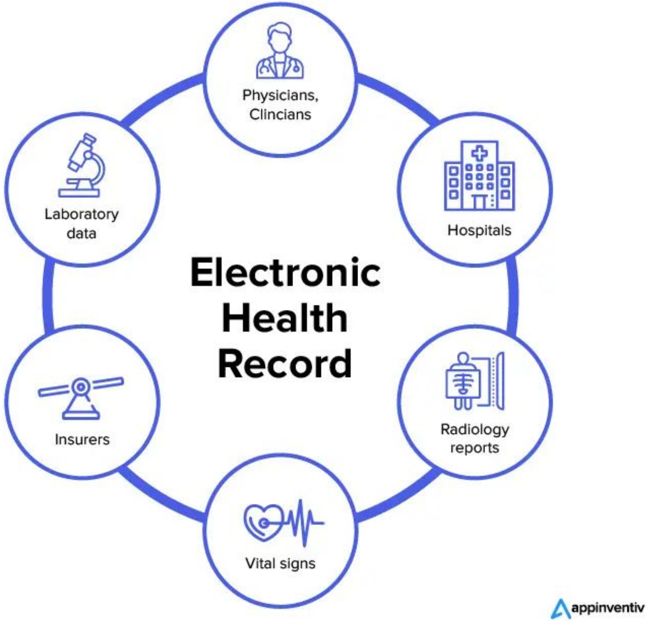
- about patient demographics, medical history and drug efficacy,
- from clinical trials, electronic health record, and public datasets,

prepare the data,

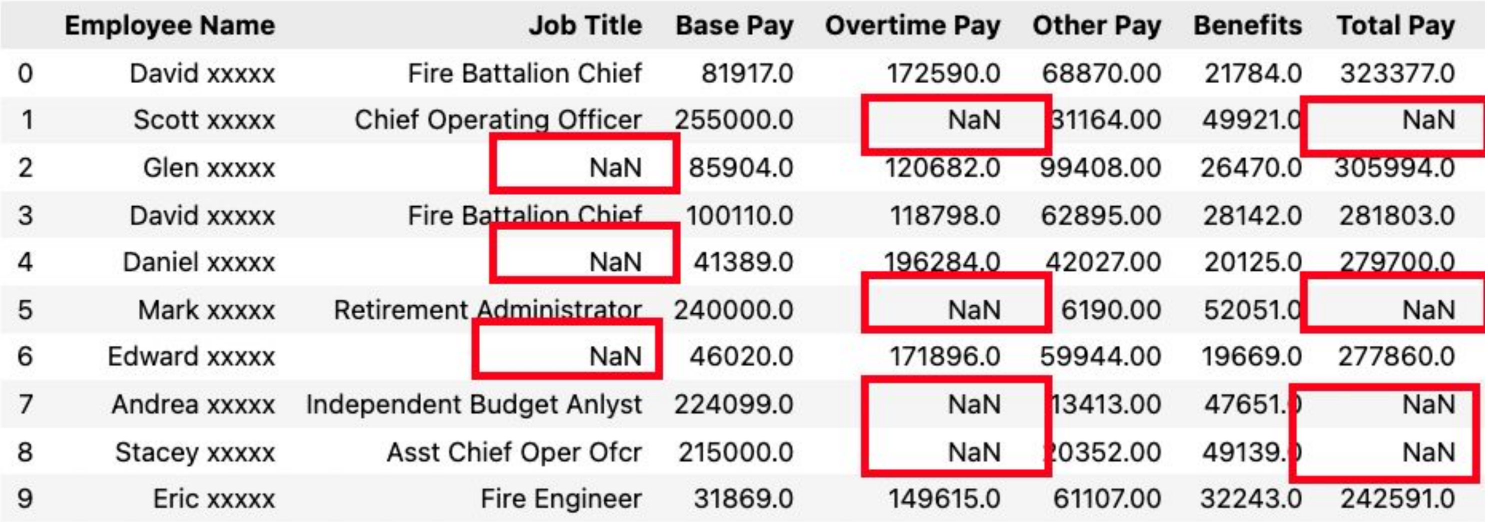
- by data cleaning,
- such as removing any missing or inconsistent values,

and ***explore*** the data,

- by creating visualizations,
- such as histograms and scatter plots,
- to understand the distribution and identify patterns from the data.



Opps, there are missing values in the data frame, which need to be handled properly.



	Employee Name	Job Title	Base Pay	Overtime Pay	Other Pay	Benefits	Total Pay
0	David xxxxx	Fire Battalion Chief	81917.0	172590.0	68870.00	21784.0	323377.0
1	Scott xxxxx	Chief Operating Officer	255000.0	NaN	31164.00	49921.0	NaN
2	Glen xxxxx	NaN	85904.0	120682.0	99408.00	26470.0	305994.0
3	David xxxxx	Fire Battalion Chief	100110.0	118798.0	62895.00	28142.0	281803.0
4	Daniel xxxxx	NaN	41389.0	196284.0	42027.00	20125.0	279700.0
5	Mark xxxxx	Retirement Administrator	240000.0	NaN	6190.00	52051.0	NaN
6	Edward xxxxx	NaN	46020.0	171896.0	59944.00	19669.0	277860.0
7	Andrea xxxxx	Independent Budget Anlyst	224099.0	NaN	13413.00	47651.0	NaN
8	Stacey xxxxx	Asst Chief Oper Ofcr	215000.0	NaN	20352.00	49139.0	NaN
9	Eric xxxxx	Fire Engineer	31869.0	149615.0	61107.00	32243.0	242591.0

Data Representation and Transformation

- After exploring the data, the team would **represent** and **transform** the data in a way that is suitable for analysis and modeling.
- This could include **feature engineering**, **normalization**, and **dimensionality reduction**.

Computing with Data

- Involves using computational techniques to analyze the data, such as **statistical inference, machine learning, and data mining**.

- These can include popular languages such as **R** and **Python**, and many more.

Data Modeling

- In this stage, the data science team would build mathematical models to explain the underlying patterns in the data.

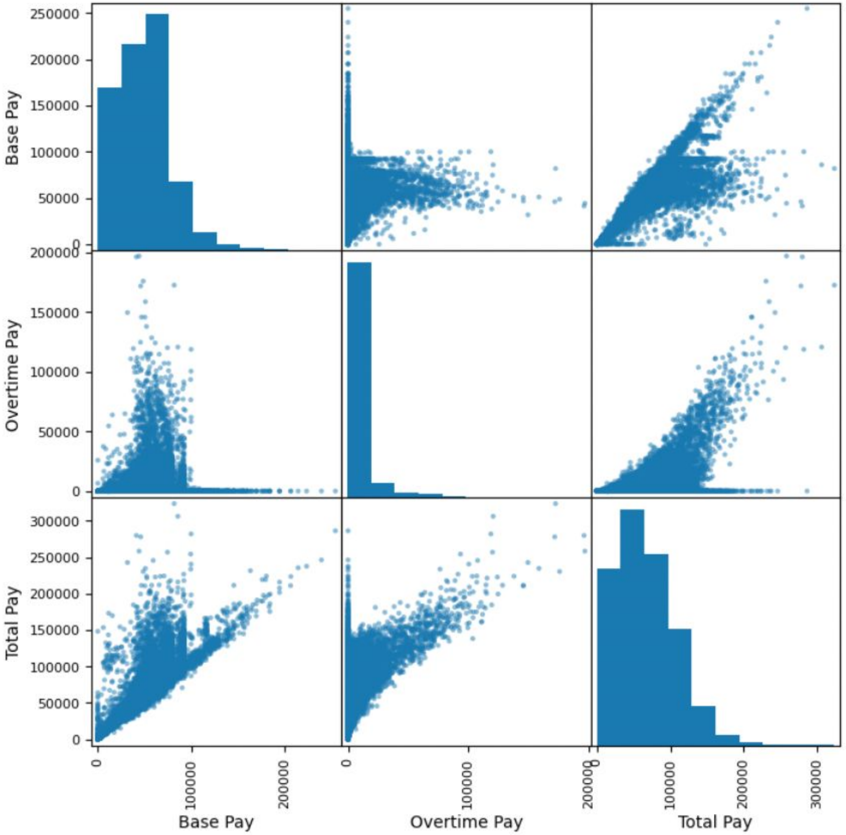
- For example, the data team predict the likelihood of drug efficacy based on patient characteristics using machine learning algorithms, such as logistic regression and decision trees.

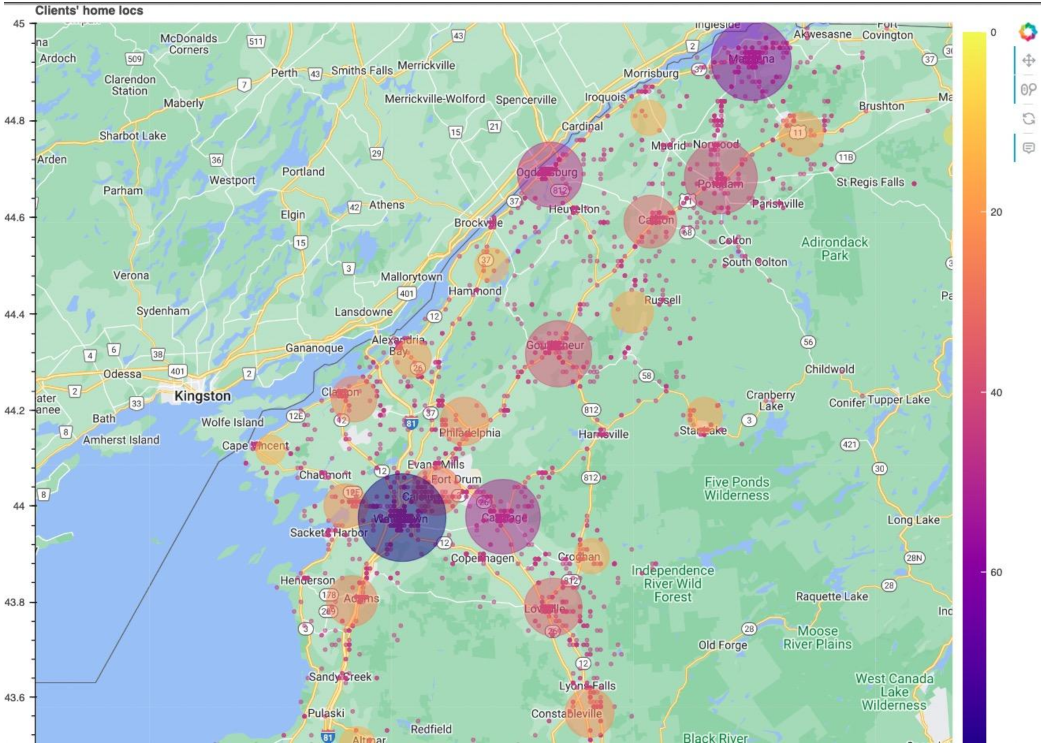
Data Visualization and Presentation

- The data team would create visual representations of the data, such as heatmaps and bar charts, to make it easier to understand and interpret the data.

- For example, they could create interactive dashboards that allow the medical team to explore the data and gain insights, and also prepare the results of the project in a way that is easy to understand and present to stakeholders.

Examples of Data Visualization





Science about Data Science

- “Tukey proposed that ‘a science of data analysis’ exists and should be recognized as among the most complicated of all sciences.”
- It involves monitoring the performance of the model, validating the findings, and understanding the ethical and legal implications of the results.
- Additionally, it involves staying current with the latest developments and trends in data science and being able to reflect on the processes and methods used throughout the project.