Stat C131A: Statistical Methods for Data Science

Fall 2025

Lecture 2

a) Class resources (online)
b) Syllabus
c) Exploratory Data Analysis (EDA): a start

Sep 2 2025

Key links

- **bCourses:** https://bcourses.berkeley.edu/courses/1549075
- **Course Website:** https://datasciencey.github.io/statc131a/
  - The lecture slides will be maintained here
    - Typically posted *after* the particular lecture
  - Other references and readings as we go along
- Also coming up
  - Ed:
  - Gradescope:

R

- R for Data Science (2e)
  - Garrett Grolemund and Hadley Wickham
  - Free online book that covers the "tidyverse" set of R packages
- The lab assignments will be centric to doing work in R

LLM Policy

- Only a **controlled use** of LLMs is permitted
- Gemini Pro
  - Instructions will be provided in the coming days

Assignments and Exams

- Assignments and Lab
- One midterm
- Final project&

Grading

- Breakdown
  - Assignments / Labs
  - Midterm
  - Project & Final Exam I Final project only
    - TBD: discussing with the department
- Grades
  - We follow the standard university grading system
    - 97% = A+, 93% = A, 90% = A-, 87% = B+, etc.
  - This class is **not** curved. However, we may curve *upwards* if the exams/projects prove harder than intended.

Topics we will cover

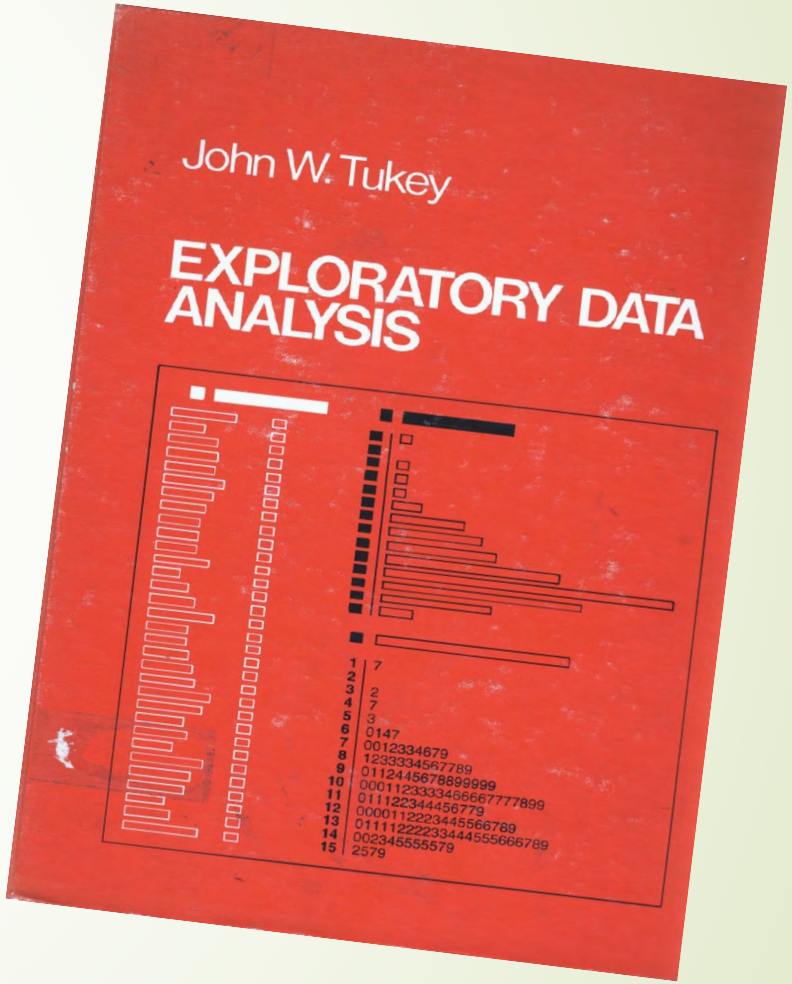
- Distributional Summary of Data. 2.1
- Visualization (Boxplots, Histograms, etc.). 2.1
- Discrete and continuous distributions. Normal Distribution, Central Limit Theorem 2.1, 2.3, 2,4.3
- Probability. Bayes' theorem. Naive Bayes algorithm. 2.1
- Sampling distributions. Bootstrapping. 2.4
- Confidence intervals. Parametric hypothesis testing. 3
- Hypothesis testing. Type I and II errors. 3
- Linear regression. 4.1-4.3, 6

- Explanation: Feature generation. Transformations. 6
- Evaluation: Cross-validation. Bias-variance tradeoff. Logistic regression. Classification error metrics. 7 7.5
- Non-parametric methods. Kernel density estimation (KDE). 2.5. 4.4-4.5
- Principal components analysis (PCA). 5
- Clustering. 5.4
- Decision trees. Random Forests. 8
- Causal Inference
- Project prep

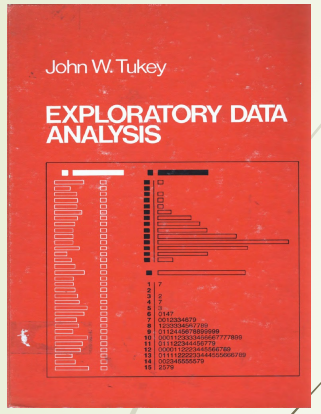
Exploratory Data Analysis (EDA)

John Tukey

- First “data scientist” ? ☺
- Credit with
  - FFT: the Fast Fourier Transform
  - coined the term “bit”
  - EDA !!



Evolution



- **1. Pen-and-Paper Era (pre-computing, ~1600s–1950s)**
  - Foundational probability theory and **statistical inference**
  - Reliance on **mathematical derivations and manual calculation** (log tables, slide rules).
- **2. Computing Era (1960s–1970s)**
  - Widespread access to computers for statistical work.
  - Tukey's **Exploratory Data Analysis (EDA)**
  - Software packages (e.g., SAS, SPSS, later R)
- **3. Big Data Era (early 2010s)**
  - Explosion of **digital data sources** (social media, sensors, transactions).
  - Techniques adapted for **large-scale data storage & processing** (Hadoop, Spark).
  - **Machine learning** (random forests, SVMs, neural nets) embraced as statistical tools.
- **4. Today: Statistics in the Age of AI (2020s–)**
  - Re-emergence of **statistics as the "safety and rigor layer"** ensuring reliability of AI-driven decisions in science, medicine, economics, and defense.

EDA

- Distributional summarization
- Visualizing these distributional summaries
- The **DESCRIPTIVE**

measurements of central tendency

Mean Median Mode

Mean	Median	Mode
Defined as the arithmetic average of all observations in the data set.	Defined as the middle value in the data set arranged in ascending or descending order.	Defined as the most frequently occurring value in the distribution; it has the largest frequency.

Mean Median Mode

Mean	Median	Mode
Defined as the arithmetic average of all observations in the data set.	Defined as the middle value in the data set arranged in ascending or descending order.	Defined as the most frequently occurring value in the distribution; it has the largest frequency.
Requires measurement on all observations.	Does **not require** measurement on all observations.	Does **not require** measurement on all observations.

Mean Median Mode

Mean	Median	Mode
Defined as the arithmetic average of all observations in the data set.	Defined as the middle value in the data set arranged in ascending or descending order.	Defined as the most frequently occurring value in the distribution; it has the largest frequency.
Requires measurement on all observations.	Does **not require** measurement on all observations.	Does **not require** measurement on all observations.
Uniquely and comprehensively defined.	Cannot be determined under all conditions.	Not uniquely defined for multi-modal situations.

Mean Median Mode

Mean	Median	Mode

Mean Median Mode

Mean	Median	Mode
Affected by outliers	Resists the outliers pretty well	Resists the outliers pretty well

Mean Median Mode

Mean	Median	Mode
Affected by outliers	Resists the outliers pretty well	Resists the outliers pretty well
Can be treated algebraically (means of several groups can be combined).	Cannot be treated algebraically (medians of several groups cannot be combined).	Cannot be treated algebraically (modes of several groups cannot be combined).

sneakers

Startup : new sneaker

- Surveyed **500 people**.
- **Purchase intent:** *Absolutely buy / Maybe / Not a chance*
- **Free-text:** "Why/how would you use these sneakers?"
- **Numeric to summarize: Willingness to Pay (WTP)** in USD (max they'd pay if they had to buy)

Survey results

Absolutely buy (n≈180)

Mode: $99 (most common)
Median: $95 (middle buyer)
Mean: $104 (average)
A few high outliers: $200, $220

Maybe (n≈210)

Mode: $79
Median: $85
Mean: $92
Some high outliers: $200, $240

Not a chance (n≈110)

Mode: $0 (most say "no price")
Median: $30
Mean: $58
Occasional oddball: $250

Median: the "typical buyer" anchor

- **What it tells us:** 50% would pay **at least** this much; robust to outliers.
  - **Here:** $95 (Absolutely), $85 (Maybe), $30 (Not a chance).

Median: the "typical buyer" anchor

- **What it tells us:** 50% would pay **at least** this much; robust to outliers.
  - **Here:** $95 (Absolutely), $85 (Maybe), $30 (Not a chance).
- **How do we USE it ?**

Median: the "typical buyer" anchor

- **What it tells us:** 50% would pay **at least** this much; robust to outliers.
  - **Here:** $95 (Absolutely), $85 (Maybe), $30 (Not a chance).
- **How do we USE it ?** Set **floor expectations** and design **student/launch promos !**

Median: the "typical buyer" anchor

- **What it tells us:** 50% would pay **at least** this much; robust to outliers.
  - **Here:** $95 (Absolutely), $85 (Maybe), $30 (Not a chance).
- **How do we USE it ?** Set **floor expectations** and design **student/launch promos !**
  - Pricing near **$95** captures half of the "Absolutely" segment.

Median: the "typical buyer" anchor

- **What it tells us:** 50% would pay **at least** this much; robust to outliers.
  - **Here:** $95 (Absolutely), $85 (Maybe), $30 (Not a chance).
- **How do we USE it ?** Set **floor expectations** and design **student/launch promos !**
  - Pricing near **$95** captures half of the "Absolutely" segment.
  - A limited promo at **$85** can flip fence-sitters in "Maybe".

Mode : the “price magnet”

- **What it tells us:** the **most popular single price** people typed.
  - **Here:** $99 (Absolutely), $79 (Maybe), $0 (Not a chance).

Mode : the “price magnet”

- **What it tells us:** the **most popular single price** people typed.
  - **Here:** $99 (Absolutely), $79 (Maybe), $0 (Not a chance).
- **How do we use it ?**

Mode : the “price magnet”

- **What it tells us:** the **most popular single price** people typed.
  - **Here:** $99 (Absolutely), $79 (Maybe), $0 (Not a chance).
- **How do we use it ?** Pick **psychological price points** and **entry tiers**.

Mode : the “price magnet”

- **What it tells us:** the **most popular single price** people typed.
  - **Here:** $99 (Absolutely), $79 (Maybe), $0 (Not a chance).
- **How do we use it ?** Pick **psychological price points** and **entry tiers**.
  - Launch flagship at **$99** (aligns with your hottest segment’s instinct).

Mode : the “price magnet”

- **What it tells us:** the **most popular single price** people typed.
  - **Here:** $99 (Absolutely), $79 (Maybe), $0 (Not a chance).
- **How do we use it ?** Pick **psychological price points** and **entry tiers**.
  - Launch flagship at **$99** (aligns with your hottest segment’s instinct).
  - Offer a **budget trim at $79** to convert part of “Maybe”.

Mode : the “price magnet”

- **What it tells us:** the **most popular single price** people typed.
  - **Here:** $99 (Absolutely), $79 (Maybe), $0 (Not a chance).
- **How do we use it ?** Pick **psychological price points** and **entry tiers**.
  - Launch flagship at **$99** (aligns with your hottest segment’s instinct).
  - Offer a **budget trim at $79** to convert part of “Maybe”.
  - If many “Not a chance” say **$0**, they’re true **non-buyer**s: don’t chase them with price alone.

Mean: planning, but outlier-sensitive

- **What it tells us:** the **average** across everyone, **pulled by extremes**.
  - **Here:** $104 (Absolutely) > $95 median because a few sneakerheads at $200+ tug the mean up !

Mean: planning, but outlier-sensitive

- **What it tells us:** the **average** across everyone, **pulled by extremes**.
  - **Here:** $104 (Absolutely) > $95 median because a few sneakerheads at $200+ tug the mean up !
- **How do we USE it ?**

Mean: planning, but outlier-sensitive

- **What it tells us:** the **average** across everyone, **pulled by extremes**.
  - **Here:** $104 (Absolutely) > $95 median because a few sneakerheads at $200+ tug the mean up !
- **How do we USE it ?** Forecast **average order value (AOV)** and revenue

Mean: planning, but outlier-sensitive

- **What it tells us:** the **average** across everyone, **pulled by extremes**.
  - **Here:** $104 (Absolutely) > $95 median because a few sneakerheads at $200+ tug the mean up !
- **How do we USE it ?** Forecast **average order value (AOV)** and revenue
  - Try a "collector's edition" to **monetize the high tail**

Mean: planning, but outlier-sensitive

- **What it tells us:** the **average** across everyone, **pulled by extremes**.
  - **Here:** $104 (Absolutely) > $95 median because a few sneakerheads at $200+ tug the mean up !
- **How do we USE it ?** Forecast **average order value (AOV)** and revenue
  - Try a "collector's edition" to **monetize the high tail**
  - But **don't** set base price by mean

Conclusion

- In distributional summarization we covered
  - measurement of **central tendency**
- Next (lecture) we will go onto
  - Measurement of **dispersion**
  - More distributional summarization elements