Exploratory Data Analysis (EDA)

- Take home: **EDA is where you will spend 60%-80% of your time!!!** I cannot emphasize this enough
- Today we will cover Data Quality checks and Distributions
- Bottom line – as a data scientist it is your first and primary job to validate the data!

EDA Intro

- Definition: The process of investigating datasets to summarize their main characteristics, often using visual methods, before formal modeling.
- Origin: Term popularized by John Tukey (1970s) as an alternative to purely confirmatory analysis.
- Goal: Build an intuitive understanding of the data before committing to a model.

Example

- Mars Climate Orbiter lost in 1999 because one engineering team used imperial units (pounds of force) while another used metric (newtons).
- If someone had done thorough EDA on the input pipeline, they might have caught the mismatch before launch.
- Cost: $125 million.

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

• Helps generate and refine hypotheses
• Guides selection of modeling approaches

Why EDA Matters

- Avoids false conclusions
- Reveals data quality issues
- Helps generate and refine hypotheses
- Guides selection of modeling approaches
- Practically, saves a lot of time/money and saves you from looking like an idiot!

The Three Core Goals of EDA

- **Data Quality** – What's missing, wrong, or suspicious?
- **Data Structure** – How is the data organized? What's the distribution of variables?
- **Data Insight** – What trends or patterns jump out immediately?

Data Quality Checks

- Missing values
  - % missing per column
  - Patterns of missingness (spot visually)
- Outliers
  - Context vs. error (deer antler/gender example)
    - Will female deer have entlers?
- Duplicates
  - Exact vs near-duplicate records

Missingness

- 1. **MCAR** – Missing Completely At Random Definition: The probability of a value being missing is unrelated to the data (observed or unobserved).
  - Example: A lab tech accidentally drops a test tube and loses the blood sample → the missingness is random and unrelated to patient characteristics.
  - Consequence: Safe to analyze the remaining data — no systematic bias, though you lose power.
- 2. **MAR** – Missing At Random Definition: Missingness depends on observed data but not the missing value itself.
  - Example: Older participants are less likely to respond to a digital survey → missingness depends on age (observed), but not directly on the unreported values.
  - Consequence: Can be handled if you condition on the related observed variables.
- 3. **MNAR** – Missing Not At Random Definition: Missingness depends on the missing value itself.
  - Example: People with higher incomes are less likely to report their income → the probability of missingness depends on the true (unobserved) value.
  - Consequence: Very tricky — requires domain assumptions or specialized models.

Outliers

- Errors (measurement/data entry)– Typos, sensor glitches, unit mismatches
  - e.g., Height = 300 cm
- Contextual– Unusual only in certain situations
  - e.g., 30°C in winter
- Natural Extremes– Rare but valid tail values
  - e.g., very tall athlete
- Multivariate– Odd combinations of features
  - e.g., Math = 100, English = 5
- Sampling/Processing Artifacts– Wrong population or merge error
  - e.g., dog weights in human dataset

Duplicates

- Exact duplicates
  - Every column identical across rows
  - Usually from merging or re-importing data
- Key duplicates– Same ID appears more than once
  - May be errors or multiple records per entity (needs checking)
- Near-duplicates– Almost identical but small differences
  - e.g., “Jon Smith” vs. “John Smith”
- Time-based duplicates
  - Multiple rows with same timestamp/value
  - May indicate resampling or logging error

Data Structure

- How are the data organized?
- What are the data distributions?

Data Structure

- Is there an identifier column?
  - Example: Customer ID, patient number, experiment ID
  - IDs should not be used as numeric features — they're labels, not measurements.
- Is the data ordered by time?
  - Time-series or longitudinal data requires preserving order.
  - Example: Stock prices, sensor readings, patient vitals over time
  - Pitfall: Shuffling time-series breaks temporal dependencies.
- Is there a grouping or hierarchy?
  - Example: Schools → Classes → Students or Company → Department → Employee
  - Ignoring hierarchy can inflate significance (pseudo-replication).
  - Is it sorted by magnitude or size?
  - Example: Top 100 products by sales — can bias descriptive statistics.
- Is the order random?
  - Random ordering is fine for many analyses, but you need to confirm.

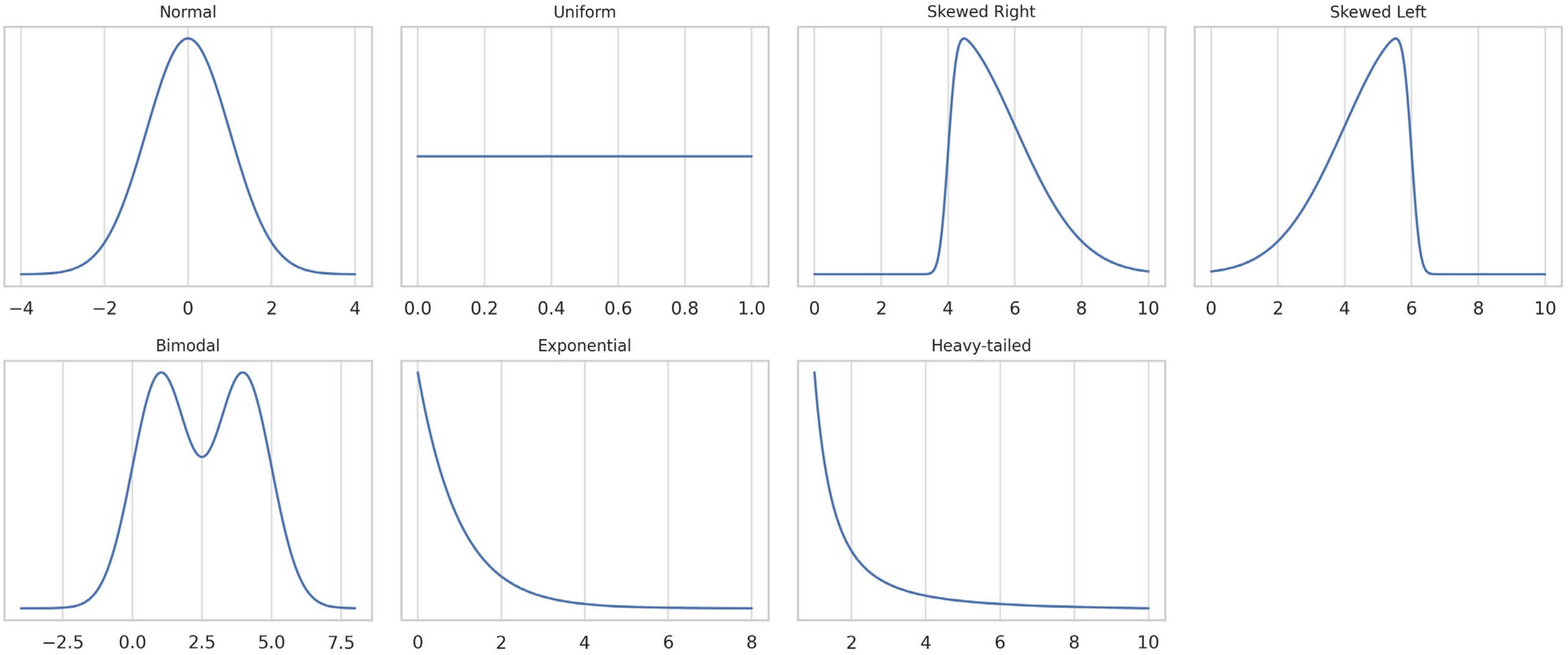
Data Structure Examples

- **Mini Examples:**
  - **Bad:** Running a train/test split on time-ordered sales data without shuffling → the “future” leaks into training.
  - **Good:** Recognizing that patient data is grouped by hospital and stratifying splits to preserve group balance.

Data Distribution Importance

- The shape of a variable’s distribution affects the summaries, statistical tests, and models you can use.
- Always visualize distributions — numbers alone can hide skew, multimodality, or outliers.
- Common shapes: normal, uniform, skewed, bimodal, exponential, heavy-tailed.
- Skewed data may need transformations (log, square root) before modeling.
- Multimodal patterns often indicate distinct subgroups in your data.

Data Structure - Distributions



Python!

- Early visualization and summary statistics, bin size