EDA 2 of 2

- Take home: **EDA is where you will spend 60%-80% of your time!!!** I cannot emphasize this enough
- The next part of EDA is understanding/exploring relationships in your variables

Last Lecture Recap

- **Data Quality**
  - Missingness: MCAR, MNAR, MAR
  - Outliers: Errors, contextual, natural extremes, multivariate, sampling
  - Duplicates: exact, Key duplicates, near duplicates, time-based
- **Data Structure**
  - IDs, data ordered by time, grouping/hierarchy, random order
- **Data Insight**
  - Different distributions

Why Look Beyond One Variable?

- Single-variable summaries only give part of the picture
- Most questions involve relationships between variables
- Examples:
  - “Do different groups have different averages?”
  - “Does missingness cluster in one category?”
  - “Are two features moving together?”
- Key: ask comparative questions

Visualizing Distributions (Advanced)

- **Histograms**: good for shape, but sensitive to bin size
- **Density plots**: smooth version, good for comparisons
- **Boxplots**: compare medians/spread, highlight outliers
- **Violin plots**: combine density + boxplot
- **Example prompt**: Compare Age distribution between Groups A and B

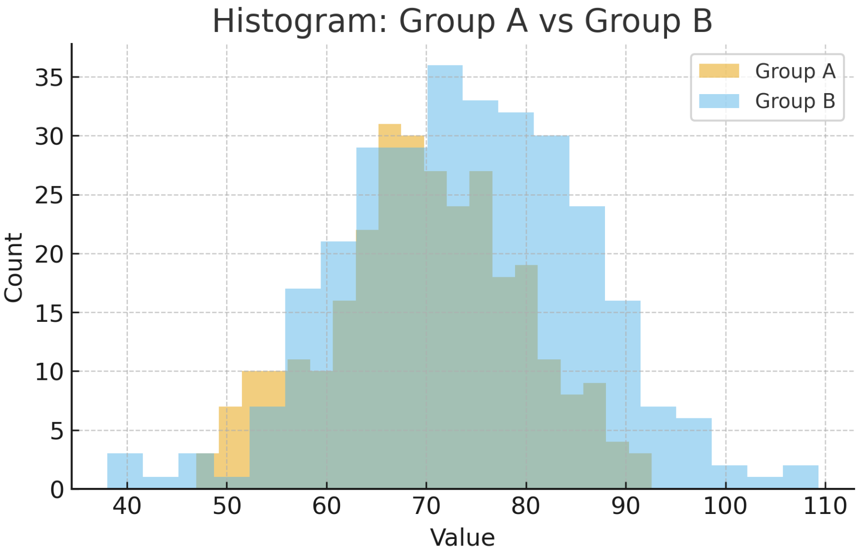
Histogram

- Generally your first stop when visualizing data (can even apply to time series data)
- 1 main parameter: number of bins: more bins → higher resolution

$$\text{bin\_size} = \frac{\max(x) - \min(x)}{\text{number of bins}}$$

$$\text{edges} = \min(x), \min(x) + \text{bin\_size}, \ldots, \max(x)$$

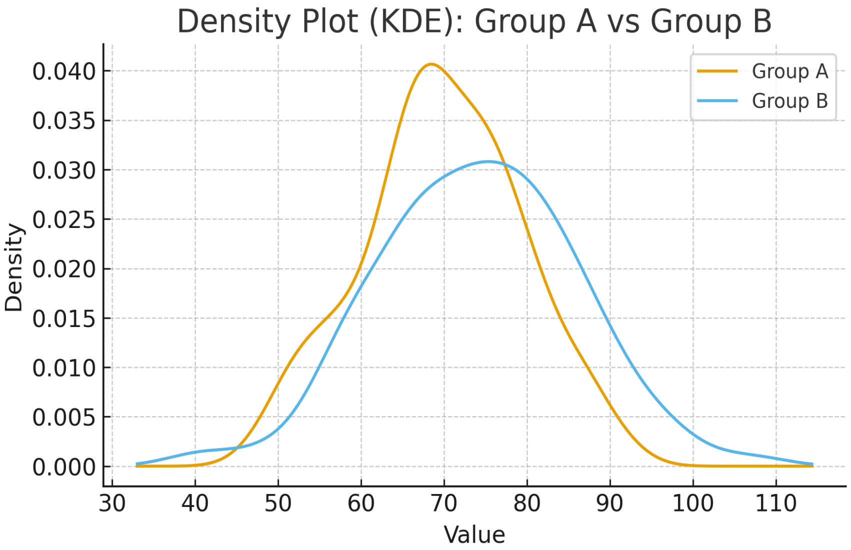
$$\text{center}_i = \frac{\text{edge}_i + \text{edge}_{i+1}}{2}$$



Density

- Example is of a kernel density estimator
- Smoother plots (probability curves) used for comparison (better than historgram)
- Smoothing parameter h controls fitting (smaller h overfit/wiggly)

$$\hat{f}_h(x) = \frac{1}{nh\sqrt{2\pi}} \sum_{i=1}^{n} \exp\left(-\frac{1}{2}\left(\frac{x - x_i}{h}\right)^2\right)$$



Boxplot

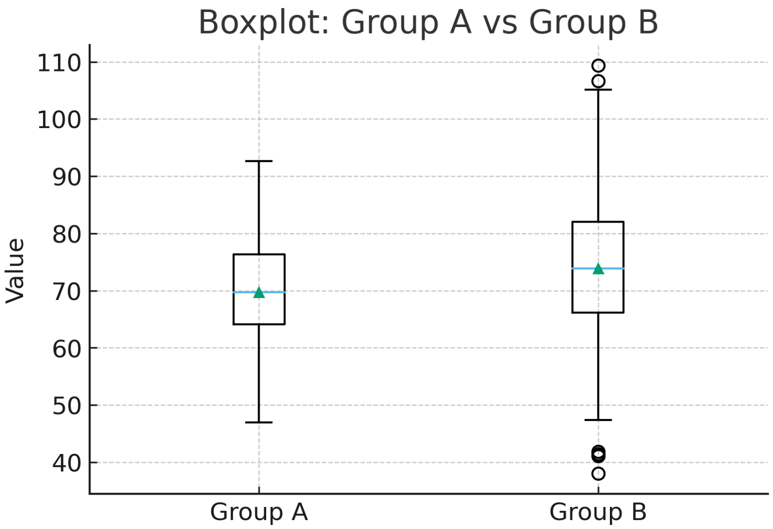
- Summarizes a distribution quickly — shows median, spread, and outliers at a glance
- Compares groups easily — side-by-side boxplots highlight differences in central tendency & variability
- Robust to skew & outliers — focuses on medians and quartiles, not just the mean
- Outlier detection — values beyond whiskers can be flagged for further inspection

$$Q2 = \text{50th percentile of the data}$$

$$Q1 = \text{25th percentile}, \quad Q3 = \text{75th percentile}$$

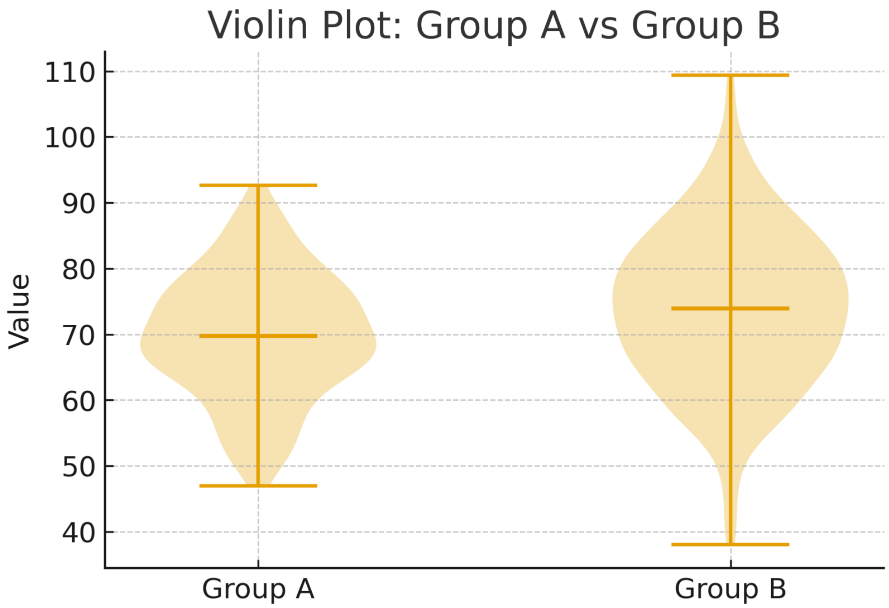
$$IQR = Q3 - Q1$$

$$Q1 - 1.5 \times IQR \qquad\qquad Q3 + 1.5 \times IQR$$



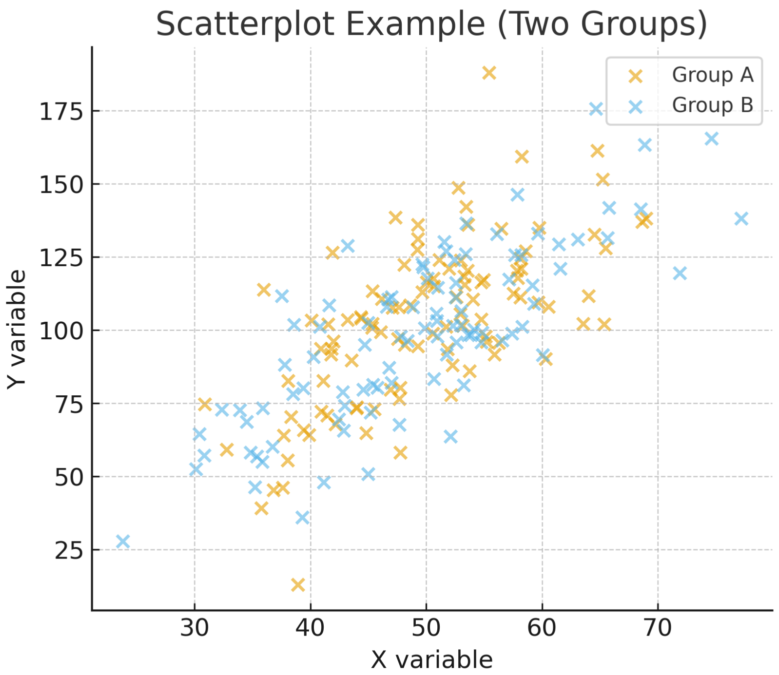
Violin Plots

- Combines **boxplot + density plot**
- Width = density (from KDE) at that value
- **Center line** = median
- **Box** inside = interquartile range (Q1–Q3)
- **Whiskers/outliers** optional
- Always **symmetric left–right**, but vertical shape shows skew/multimodality



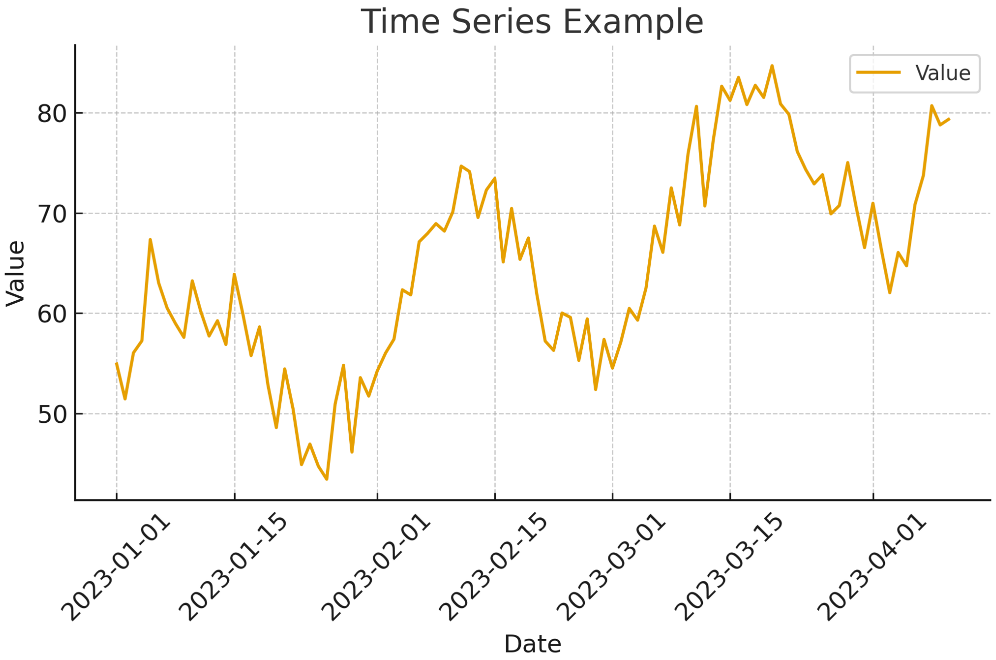
Scatter Plot

- Show the relationship between two numerical variables
- Each point = one observation (x = predictor, y = response)
- Useful for spotting: Trends (positive, negative, nonlinear)
- Clusters or subgroups
- Outliers / leverage points
- Can use color/shape to add a categorical grouping
- First step toward correlation & regression analysis



Time Series

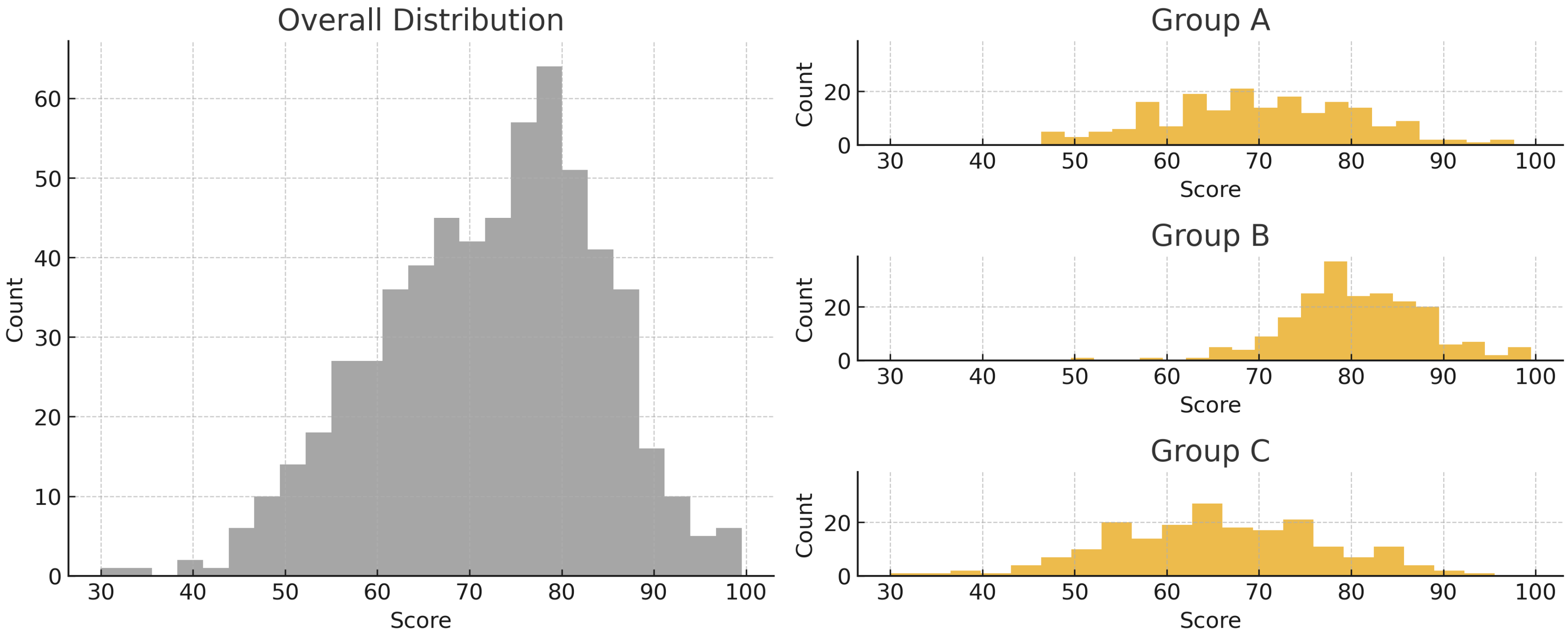
- Useful for detecting:
  - Trends (long-term increase/decrease)
  - Seasonality (regular repeating patterns)
  - Cycles (longer-term fluctuations)
  - Anomalies (spikes, sudden drops)
- Line plots are most common; scatter or area plots can also be used
- Often need resampling (daily → weekly, monthly, etc.) to see structure clearly
- First step before forecasting or decomposition



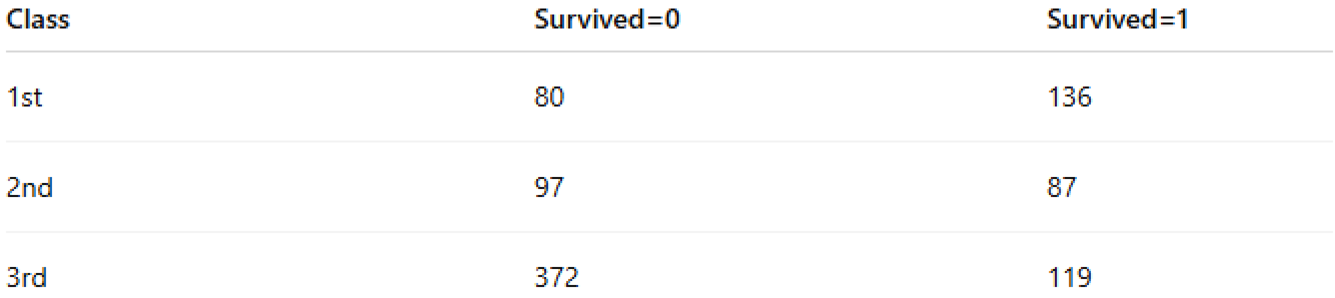
Start Investigating Relationships

- Idea: break one plot into panels by category
- Lets you see structure within groups
- Example:
  - Income distribution by Gender
  - Score distribution by Group A vs Group B
- EDA = slicing the data to ask better questions

Facets/Small Multiples



Tables for Visualization



Class	Survived=0	Survived=1
1st	80	136
2nd	97	87
3rd	372	119

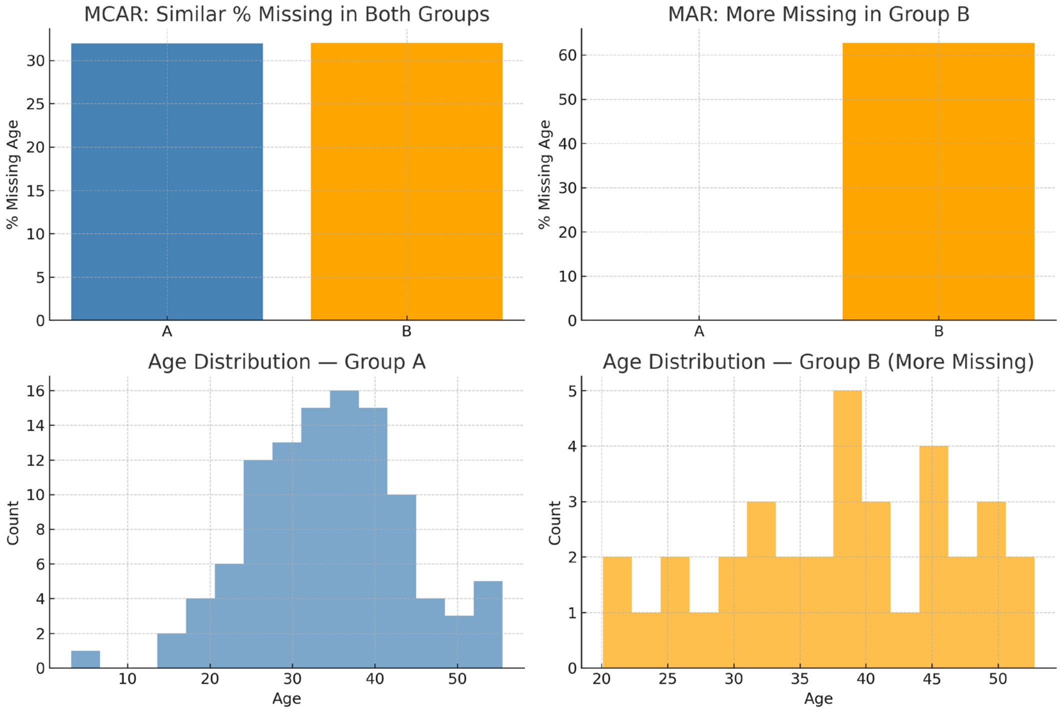
- You can and will create summary tables for different variables as you explore. For this example, we are looking at survivors of the titanic based on ticket class. You could also look at this by gender, overall, age etc.

Visualizing Missingness

- Recap: MCAR, MAR, MNAR
- In practice, visualize missingness patterns
  - Heatmaps or missingness matrices
  - Count missing values per row/column
- Question to ask: Is missingness random or patterned?
  - Example: Age missing more often in Group B

Missingness Visualization

- MCAR: Both groups lose about the same fraction of data → random, no bias, just less power
- MAR: Group B has systematically more missing values → patterned, introduces bias if ignored
- Group A distribution: Full and balanced, enough observations
- Group B distribution: Thinner, shows impact of higher missingness



Handling Missingness

- **Drop rows/columns (listwise deletion)**
  - Easy but can waste data
  - Risk of bias if missingness is not MCAR
  - *Example: Drop all rows missing Age → smaller dataset*
- **Simple imputation (mean, median, mode)**
  - Fills gaps with a single summary statistic
  - Can shrink variance, distort distributions
  - *Example: Replace missing Age with median Age*
- **Forward/backward fill (time series)**
  - Carries forward last known value or fills with next value
  - Assumes stability between measurements
  - *Example: Missing stock price on Tuesday filled with Monday's*
- **Model-based imputation**
  - Use regression, k-NN, or ML model to predict missing values
  - More powerful but requires assumptions and computation
  - *Example: Predict missing Age using Income and Education*

Imputation vs Removal

- ALWAYS DOCUMENT!
- When to Drop Missing Data (Listwise Deletion)
  - Very few rows are missing (e.g., <5%-10% of the dataset)
  - The missingness is MCAR (random, not patterned)
  - Losing those rows won’t bias results or reduce statistical power much
- When to Impute
  - **Substantial proportion** of data is missing (e.g., 10–30%)
  - Missingness is **MAR** (depends on observed variable)
  - The variable is **important for analysis** (you can’t just ignore it)

One More Wrinkle (Features)

- Drop observations (rows)
- Use when only a few records are missing values
  - Safer if missingness is random (MCAR)
  - Risk: lose individual cases, but dataset stays wide (all features preserved)
- Drop features (columns)
  - Use when a variable has too many missing values (e.g., >40–50%)
  - Keeps most cases, but sacrifices that variable
  - Risk: lose potentially important predictor/insight

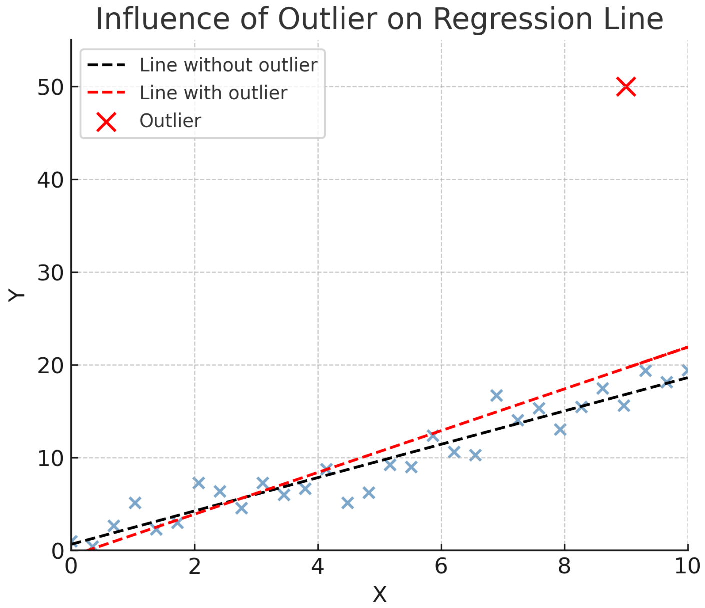
Missing Data Summary

- **Summary — Handling Missing Data**
- There is **no single "right" answer**
- Best approach depends on:
  - What data are missing
  - How they are missing (MCAR, MAR, MNAR)
- Philosophical balance between two extremes:
  - **Imputation** → making up what *might* have been there
  - **Deletion** → discarding data that *was* collected
- It's highly context-dependent, and we'll revisit this throughout the semester as we analyze real datasets

Outliers Revisited

- Last time:
  - Errors (typos, sensor glitches)
  - Natural extremes (valid but rare)
  - Multivariate oddities (weird combos of features)
- Today: Focus on Influence
- A single extreme can distort:
  - Mean → pulls average away from the bulk of the data
  - Correlation → one outlier can inflate or flip the relationship
  - Regression line → slope and intercept shift dramatically

Outlier Effects



- Mean is sensitive: one extreme value can shift it dramatically
- Median is robust: resistant to outliers
- Example:Data:
  - [5, 6, 7, 8, 9] → Mean = 7, Median = 7
  - Add an outlier: [5, 6, 7, 8, 9, 100] → Mean ≈ 22.5, Median = 7.5
- Lesson: Always compare mean vs. median when checking for outliers

Overall Process (Iterative)

- **Practical Steps for EDA**
- **Start with structure**
  - Identify IDs, categorical vs numerical variables
  - Check ordering (time, grouping)
- **Check data quality**
  - Look for duplicates (exact, key, near-duplicates)
  - Summarize missingness (% overall, by subgroup)
  - Identify outliers (errors vs real extremes)
- **Explore distributions**
  - Plot histograms, boxplots, violin plots
  - Compare distributions across groups (facets / small multiples)
  - Watch for skewness, multimodality
- **Investigate relationships**
  - Cross-tabulations, grouped summaries
  - Scatterplots for pairs of numeric variables
  - Split patterns by subgroup (e.g., Group A vs Group B)
- **Iterate & document**
  - Clean obvious errors (e.g., impossible ages)
  - Re-check after cleaning — new issues may emerge
  - Keep notes: *what you saw, what you changed, why*