Exploratory Data Analysis

Faculty of DS & AI
Spring semester, 2026

Trong-Nghia Nguyen

Business AI Lab

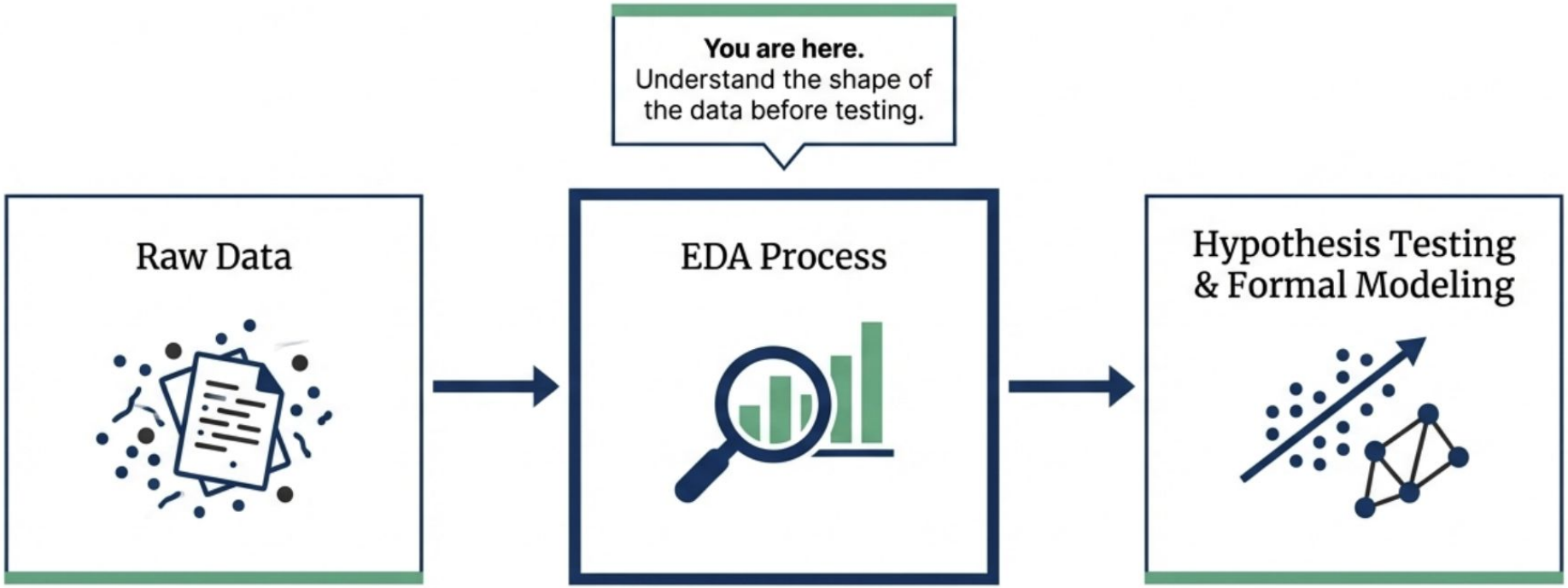
Content

- **Exploratory Data Analysis (EDA)**
- Descriptive Statistics
- Summary Tables & Data Segmentation
- PivotTables

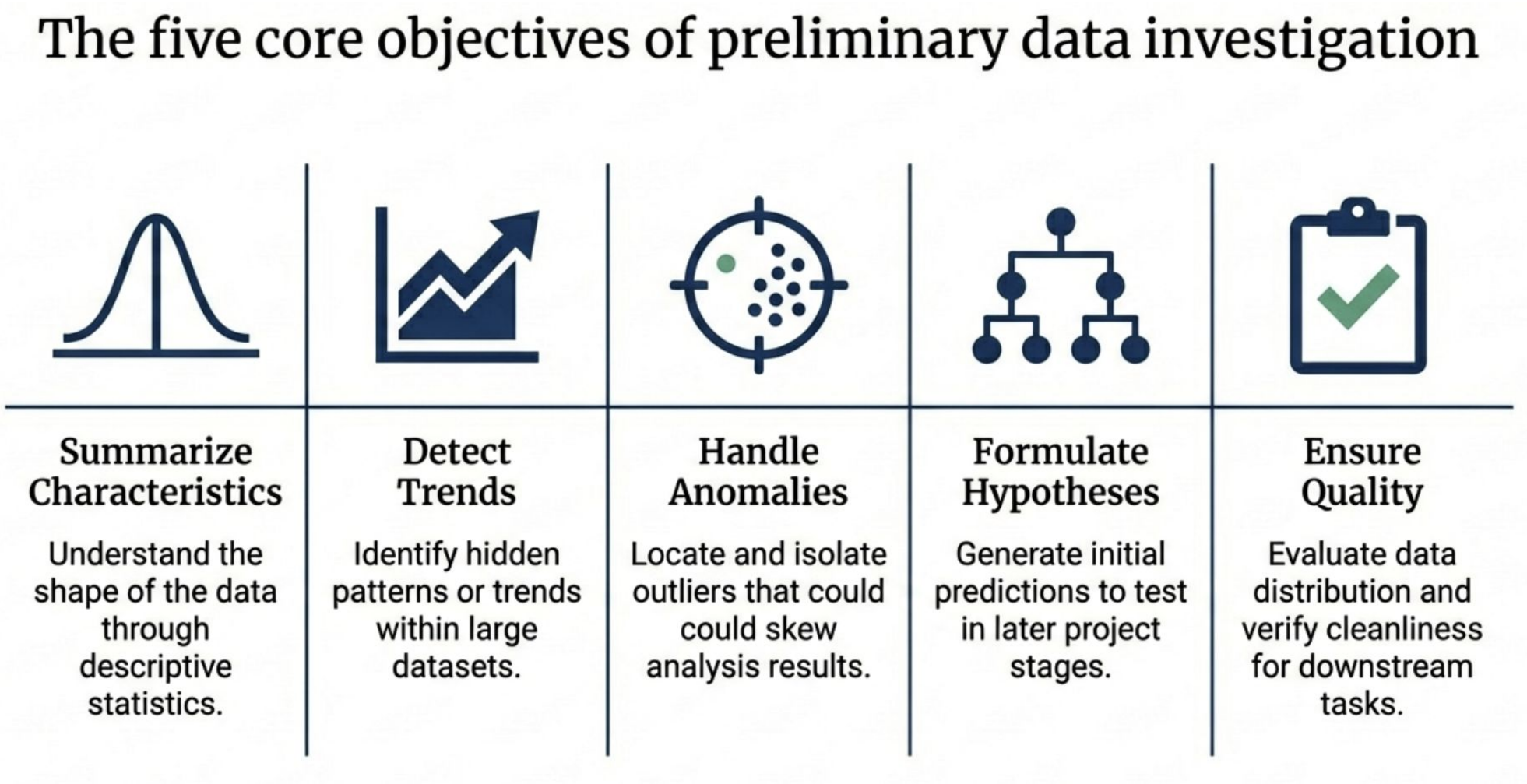
Exploratory Data Analysis (EDA)

EDA is the mandatory stepping stone to formal modeling

Exploratory Data Analysis (EDA) is the preliminary investigation of data to summarize its main characteristics, typically utilizing visualization and summary statistics.



Exploratory Data Analysis (EDA)



Exploratory Data Analysis (EDA)

Example of EDA in a scientific paper

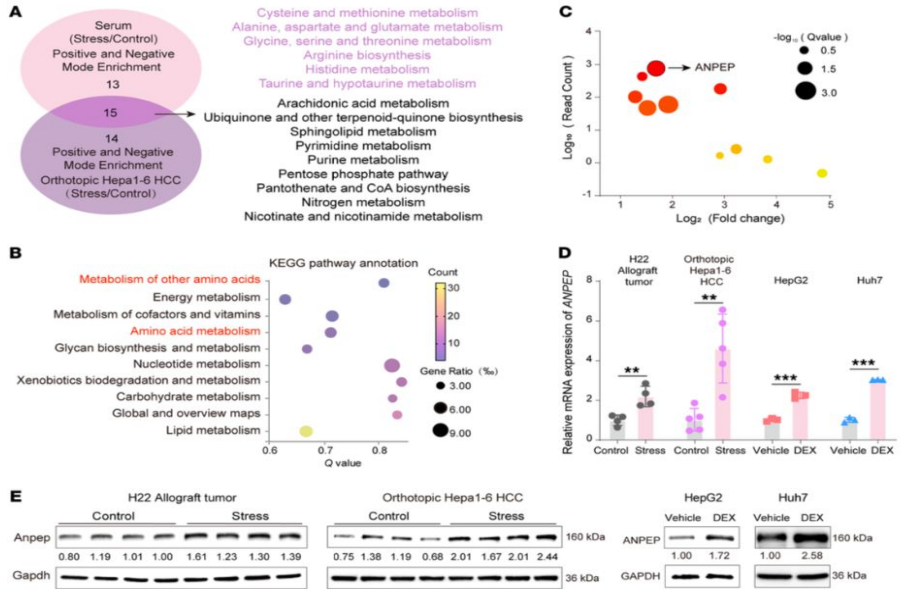
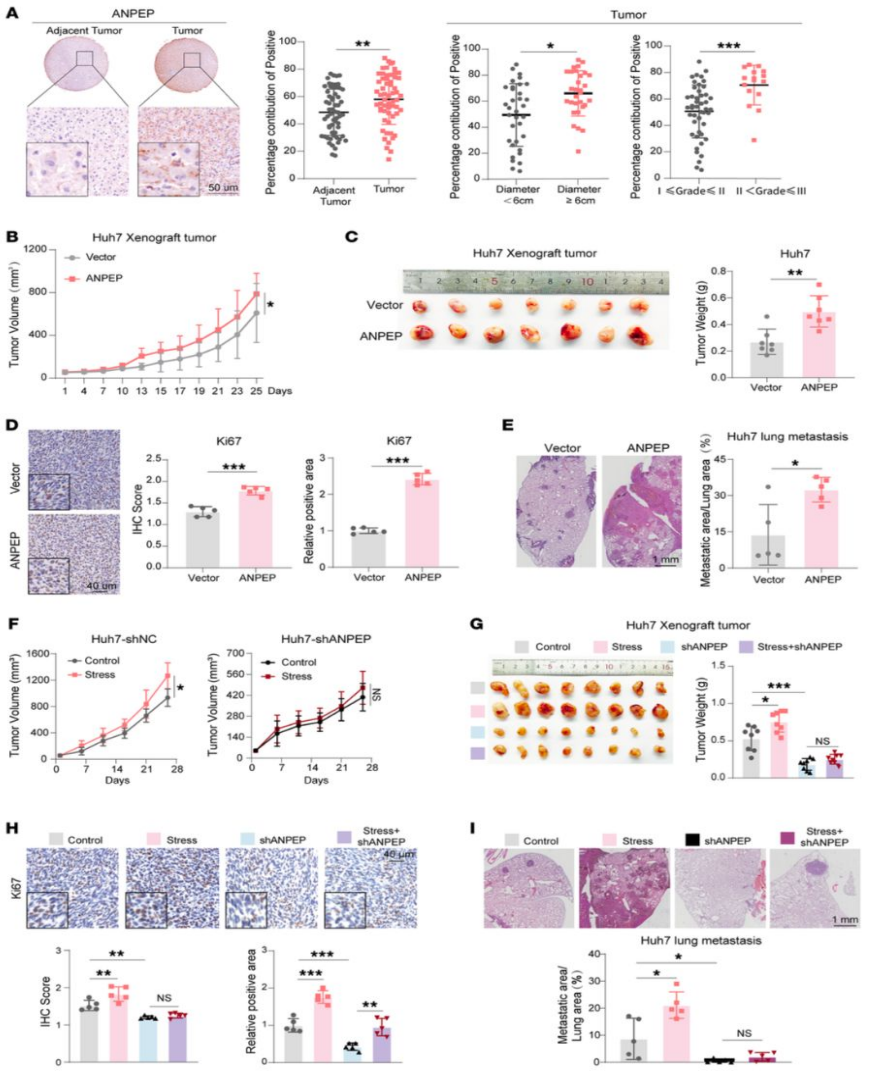


Figure 2. Chronic stress reprograms amino acid metabolism and upregulates ANPEP expression in liver cancer. (**A**) Venn diagram displayed intersecting pathway enrichment analysis of differential metabolites in serum of mice with H22 allograft tumor (n = 3) and in liver of mice with orthotopic Hepa1-6 tumor (n = 5). (**B**) The bubble chart of stress-upregulated genes in metabolic signaling pathways, as identified by KEGG pathway analysis. (**C**) The bubble plot of stress-upregulated genes in amino acid–associated pathways. (**D** and **E**) ANPEP mRNA and protein levels in primary H22 (n = 4) and Hepa1-6 tumors from control and stressed mice (n = 5) and in Huh7/HepG2 cells after dexamethasone (DEX; 1 μM) for 48 hours (n = 3). Data are presented as mean ± SD. Significance was assessed by 2-tailed unpaired Student's t test (**D**). $^{**}P < 0.01$; $^{***}P < 0.001$.

were reduced in stressed mice (Supplemental Figure 1A; supplemental material available online with this article; https://doi.org/10.1172/JCI195685DS1). Serum corticosterone levels, TG levels, and free fatty acid levels were elevated, whereas glucose levels remained unchanged in mice subjected to chronic restraint (Figure 1C and Supplemental Figure 1, B and C). As expected, chronic restraint promoted liver cancer progression, displaying as higher tumor growth rate and lung metastatic ability in mice with chronic restraint than those in control mice (Figure 1, D and E). Similarly, chronic restraint promoted orthotopic liver tumor growth (Figure 1, F and G). To investigate the systemic metabolic changes, the sera of allograft tumor models and orthotopic liver tumor models were collected for metabolomics (Supplemental Tables 1–4). The principal component analysis (PCA) showed that the metabolites from both models were well separated in chronic restraint and control groups (Supplemental Figure 1, D and E). The metabolite set enrichment analysis (MSEA) revealed that many amino acid metabolic pathways, such as cysteine and methionine metabolism; alanine, aspartate, and glutamate metabolism; and arginine biosynthesis, were commonly enriched in both models with chronic restraint (Figure 2A and Supplemental Table 5). These data suggest that chronic restraint promotes liver cancer progression associated with systemic amino acid metabolic reprogramming.

To identify a key regulator involved in amino acid metabolic reprogramming in liver cancer with chronic stress, the allograft tumor samples were collected for RNA sequencing (RNA-Seq). RNA-Seq identified a total of 889 differentially expressed genes (DEGs) ($|\log_2 FC| \geq 1.00$, adjusted $P \leq 0.05$) (Supplemental Figure 1F and Supplemental Table 6). Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis of upregulated DEGs showed that amino acid metabolism–related pathways were among the most substantially altered metabolic pathways in tumors from mice exposed to chronic stress (Figure 2B). Further analysis demonstrated that *ANPEP*, a gene related to amino acid metabolism, was the most strongly upregulated gene in tumors from mice subjected to chronic restraint (Figure 2C). The elevated ANPEP mRNA and protein levels were confirmed in mouse tumors with chronic restraint (Figure 2, D and E). Dexametha-



Exploratory Data Analysis (EDA)

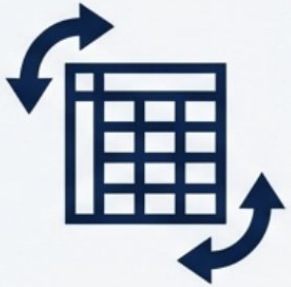
The Excel toolkit relies on three primary capabilities

Basic & Advanced Functions

Statistical functions to calculate central tendency and dispersion.

Summary Tables & Segmentation

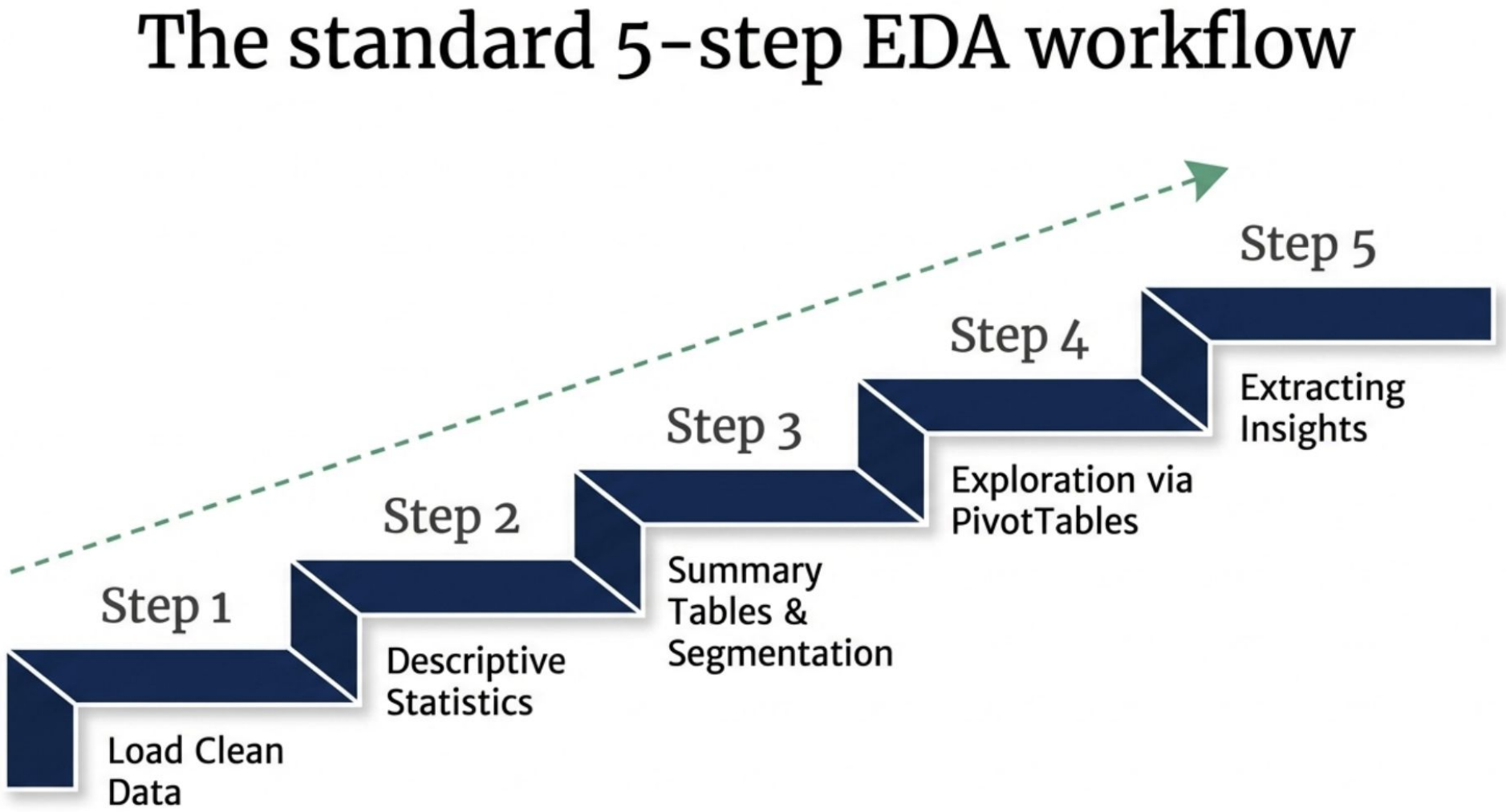
Grouping data by category to isolate specific performance metrics.



PivotTables

The primary engine for rotating and exploring multi-dimensional data relationships.

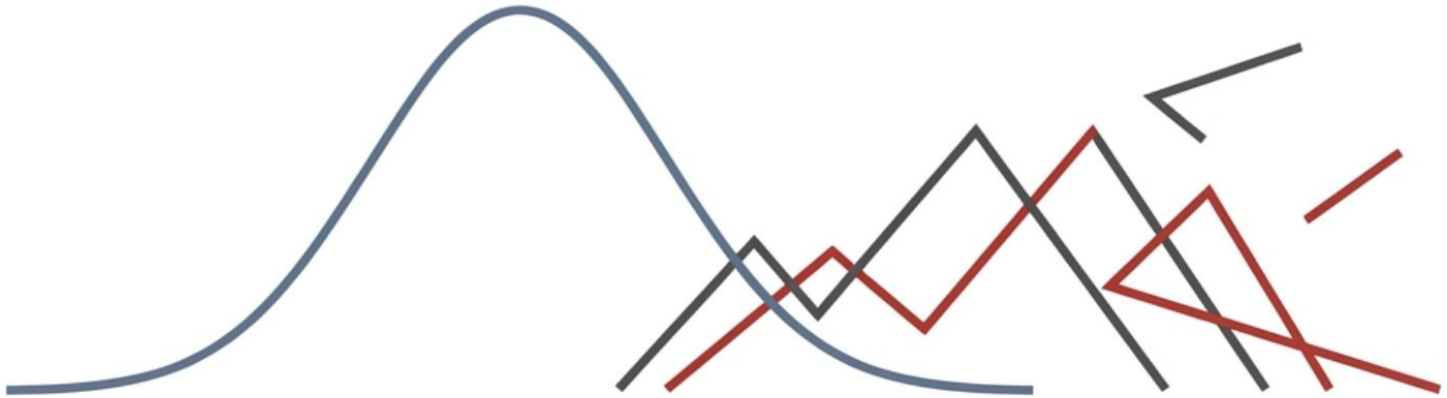
Exploratory Data Analysis (EDA)



Exploratory Data Analysis (EDA)

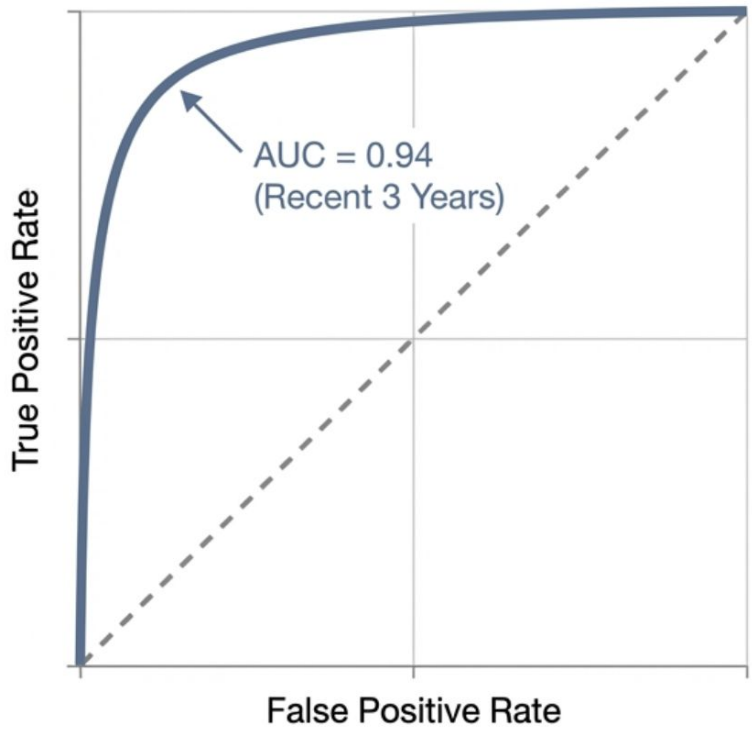
The Fatal Drift

Why foundational Exploratory Data Analysis outlives deep learning complexity in clinical prediction models.



Exploratory Data Analysis (EDA)

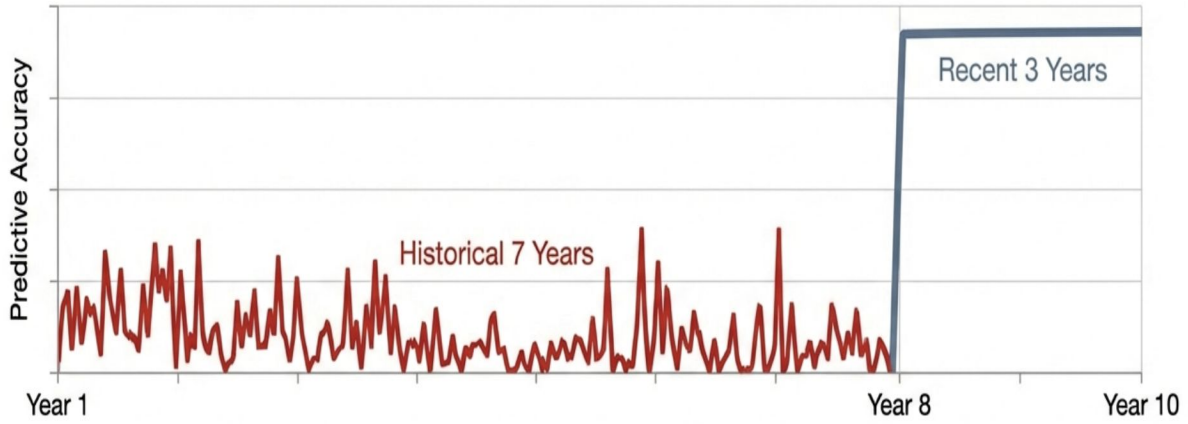
Initial deep learning models showed remarkable predictive success.



- Objective: Build a deep learning model for early prediction of CPR or intubation requirements.
- Baseline Data: Trained on the most recent 3 years of clinical data.
- Result: Exceptional AUC; the model demonstrated high sensitivity to critical vital signs.
- Next Step: Scale the model across a 10-year dataset to prove enterprise AI capabilities.

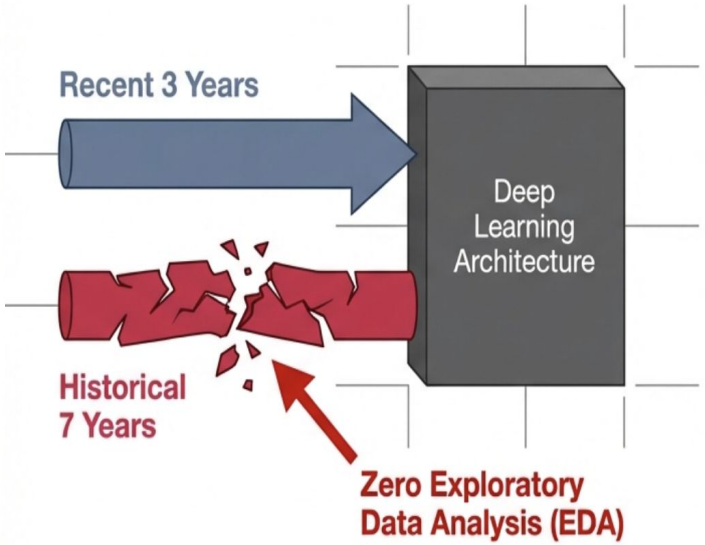
Exploratory Data Analysis (EDA)

Scaling the data volume triggered a catastrophic collapse in accuracy.



- The model was deployed across a full 10-year historical dataset.
- Overall predictive accuracy degraded severely.
- The model produced erratic predictions and missed highly critical patient cases.
- The state-of-the-art architecture completely broke down under the weight of more data.

More data broke a data-hungry algorithm algorithm.



- Deep learning traditionally requires massive datasets to optimize.
- The failure was not rooted in coding logic or algorithmic architecture.
- The root cause lay in 7 years of unverified historical data.
- Skipping foundational EDA blinded the model to underlying temporal realities.

Exploratory Data Analysis (EDA)

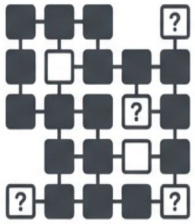
Three hidden mechanisms of data drift bypassed the initial evaluation.

Instrument Shift

Hardware upgrades altering baseline readings.

Protocol Drift

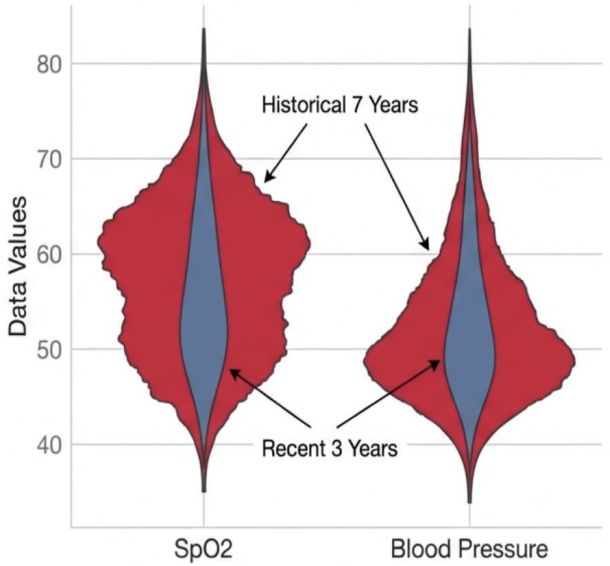
Changes in clinical workflows altering the meaning of metrics.



Missing Data Patterns

Storage architecture changes altering data density over time.

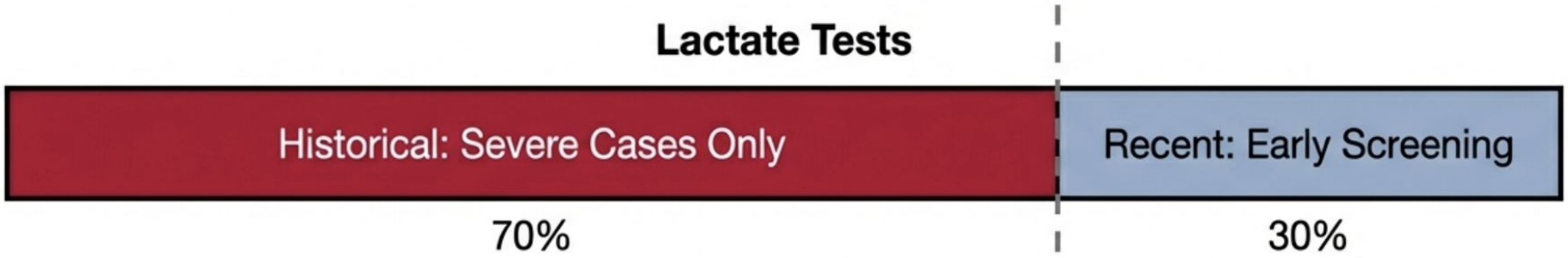
Hardware upgrades introduced silent distribution shifts



- The Reality: The hospital replaced vital sign monitors twice over the 10-year span.
- The Impact: Legacy machines produced inherently noisier data with distinct baselines.
- The Result: The model misinterpreted hardware noise in historical SpO2 and Blood Pressure readings as physiological anomalies.

Exploratory Data Analysis (EDA)

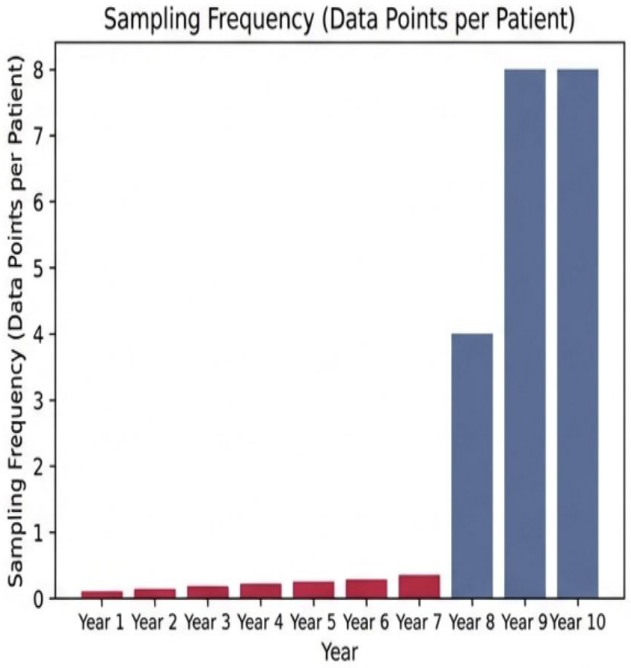
The clinical definition of a metric changed over the decade.



- Historical Protocol: 7 years ago, Lactate tests were reserved strictly for critically ill patients.
- Modern Protocol: In the last 3 years, Lactate became an early, routine screening tool.
- The Result: The exact same numerical lab value carried drastically different clinical severity depending on the year it was recorded.

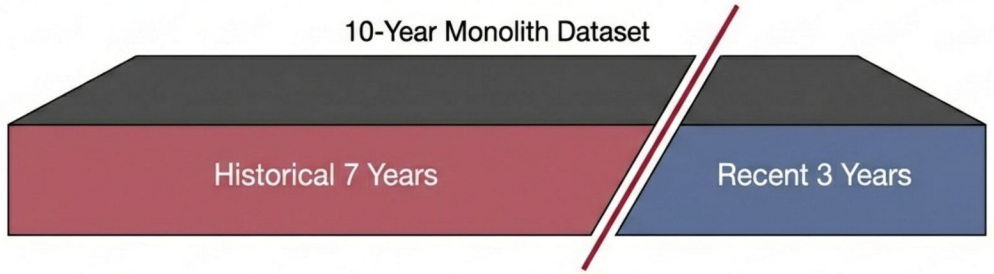
Exploratory Data Analysis (EDA)

Legacy storage systems starved the model of data density.



- Superficial data checks only looked at historical averages.
- They missed that the actual sampling frequency doubled over the decade.
- Older records contained significantly fewer data points per patient.
- The deep learning model, optimized for high-density modern data, was starved of context in the historical segments.

Temporal Segmented Analysis exposes hidden fractures in longitudinal data.



- Treating a decade of data as a static monolith guarantees failure.
- Methodology: Break longitudinal data into distinct chronological bins.
- Analyze descriptive statistics and distributions across time, rather than in aggregate.

Exploratory Data Analysis (EDA)

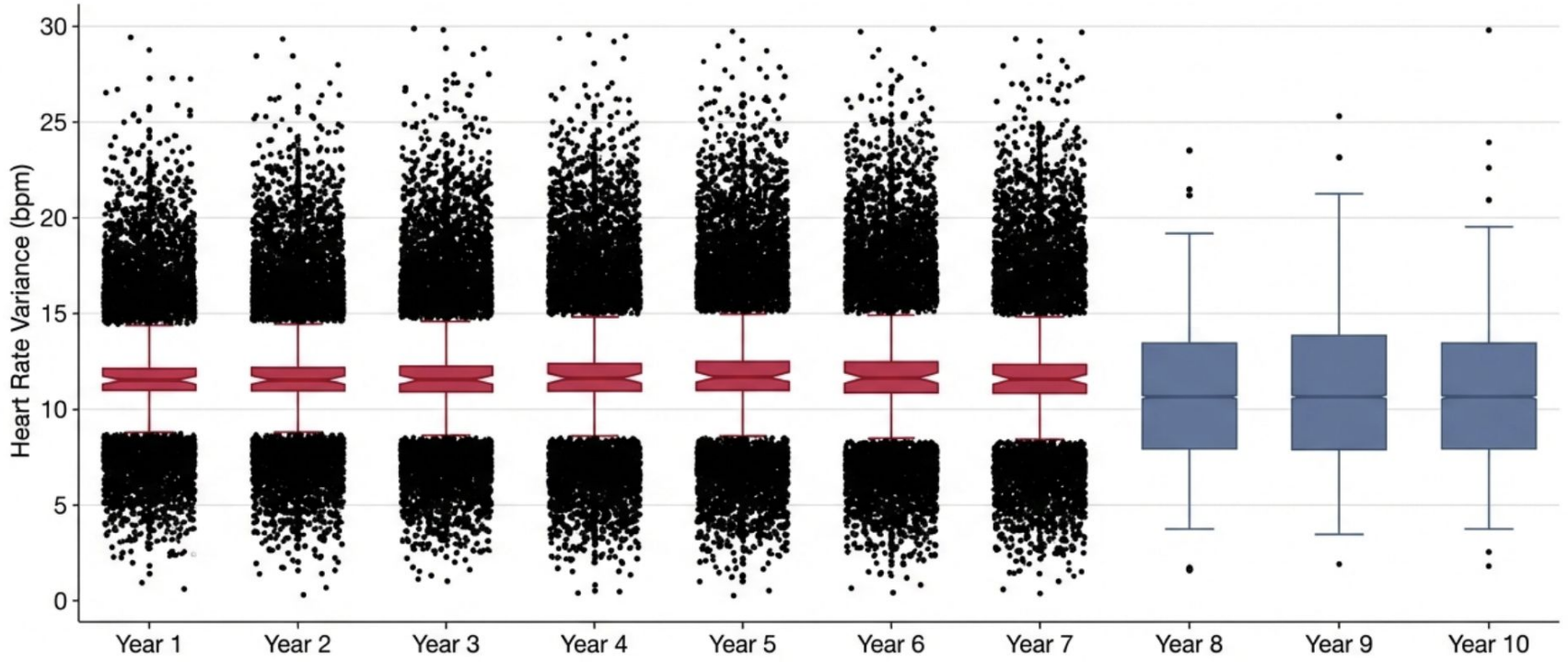
Binned descriptive statistics isolate era-based variances.

Metric	Historical 7 Years	Recent 3 Years
Time Between Lab Tests	Mean: 14.2h | Median: 12h | SD: 8.5	Mean: 6.1h | Median: 4h | SD: 2.1
Creatinine Levels	Mean: 1.1 | Median: 1.0 | SD: 0.9	Mean: 1.2 | Median: 1.1 | SD: 0.2

- Compare fundamental statistics across temporal segments.
- Changes in Standard Deviation reveal shifting clinical environments
- Shifts in intervals between lab tests expose hidden operational changes.

Exploratory Data Analysis (EDA)

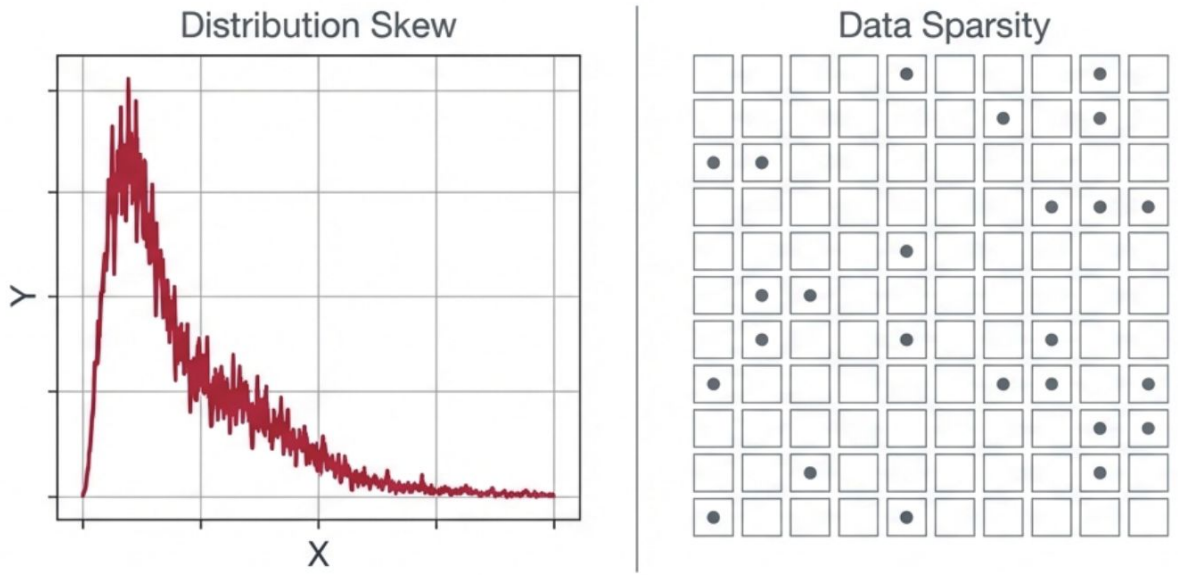
Visualizing temporal drift maps the exact location of hardware anomalies.



• Box plots map physiological metrics across individual years. • Visualizing the data exposes 'fake' outliers generated by legacy equipment errors.
• Proves that aggregating data obscures critical operational context.

Exploratory Data Analysis (EDA)

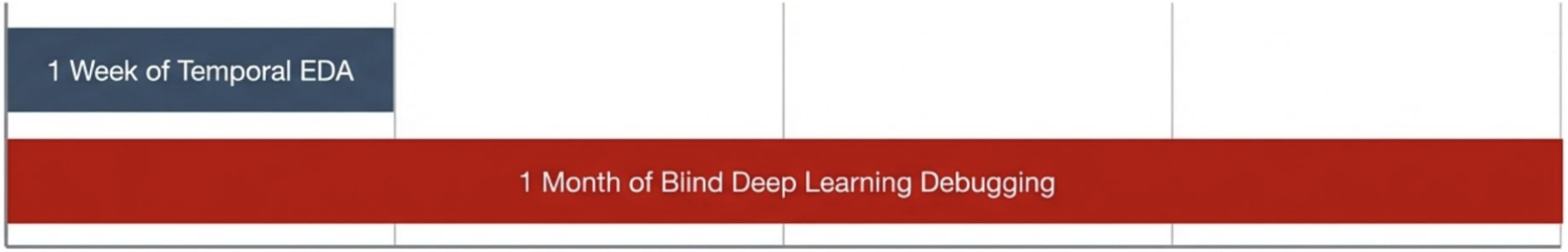
Exploratory analysis acts as a diagnostic tool for data skew and sparsity.



- **Distribution Skew**: Older vital signs operated on entirely different mathematical baselines.
- **Data Sparsity**: The sheer lack of rows per patient in early years structurally disadvantaged the neural network.
- Deep learning cannot independently correct for undocumented historical context.

Exploratory Data Analysis (EDA)

Rigorous temporal analysis prevents massive debugging waste.



- Attempting to fix data drift through code changes is futile.
- Proper temporal EDA forces teams to confront reality before training begins.
- One week of strict data verification saves months of misdirected algorithmic tuning.

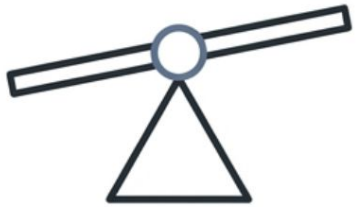
Content

- Exploratory Data Analysis (EDA)
- **Descriptive Statistics**
- Summary Tables & Data Segmentation
- PivotTables

Descriptive Statistics

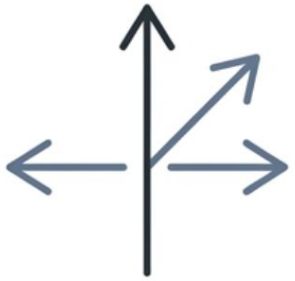
Descriptive Statistics Form the Bedrock of EDA

- Summarize dataset characteristics through quantitative metrics.
- Enable comprehension of data before executing advanced analysis.



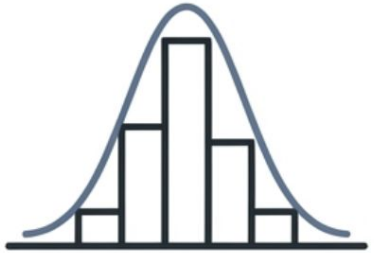
Center

Where is the middle?



Spread

How varied is the data?



Shape

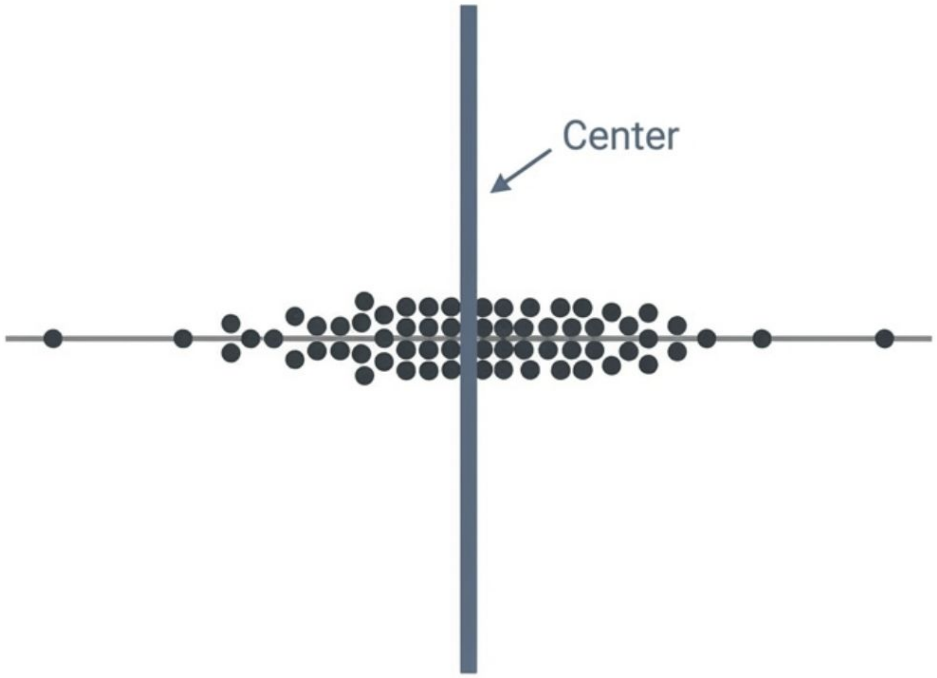
What is the distribution profile?

Descriptive Statistics

Central Tendency Locates the Representative Middle

Goal: Identify the exact value that best represents the entire dataset.

Answers the critical question: Where does the data cluster?

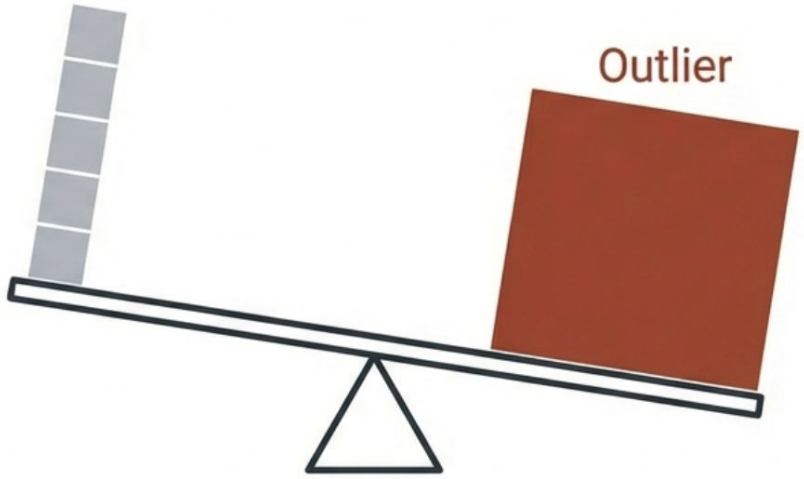


Descriptive Statistics

The Mean is Precise but Highly Sensitive to Outliers

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- $\bar{x}$: Sample mean
- $\sum$: Sigma (sum of all values)
- x_i : Value of the i-th observation
- n : Total number of observations (sample size)



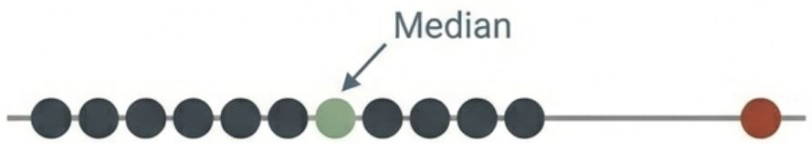
Vulnerability: A single extreme value drastically skews the result.

Descriptive Statistics

The Median and Mode Provide Robust Alternatives

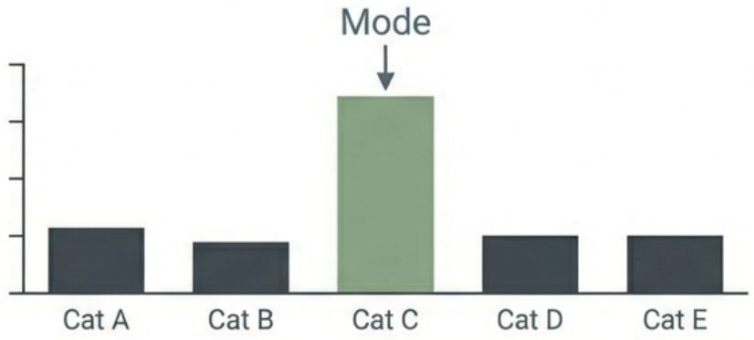
The Median

- Definition: The exact middle value in a sorted dataset.
- Characteristic: Robust; completely unaffected by extreme outliers. Ideal for skewed data.



The Mode

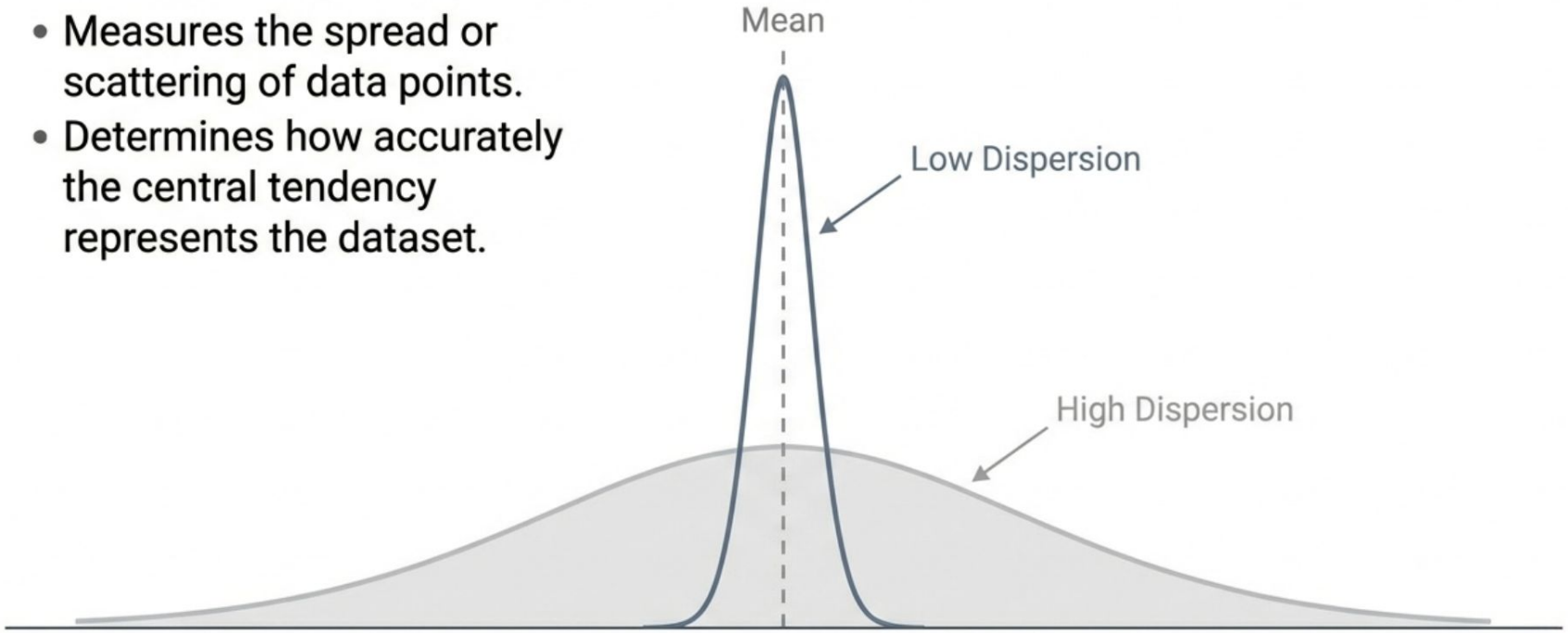
- Definition: The most frequently occurring value.
- Characteristic: Optimal for categorical data or discrete sets with narrow ranges.



Descriptive Statistics

Dispersion Quantifies Volatility Around the Center

- Measures the spread or scattering of data points.
- Determines how accurately the central tendency represents the dataset.

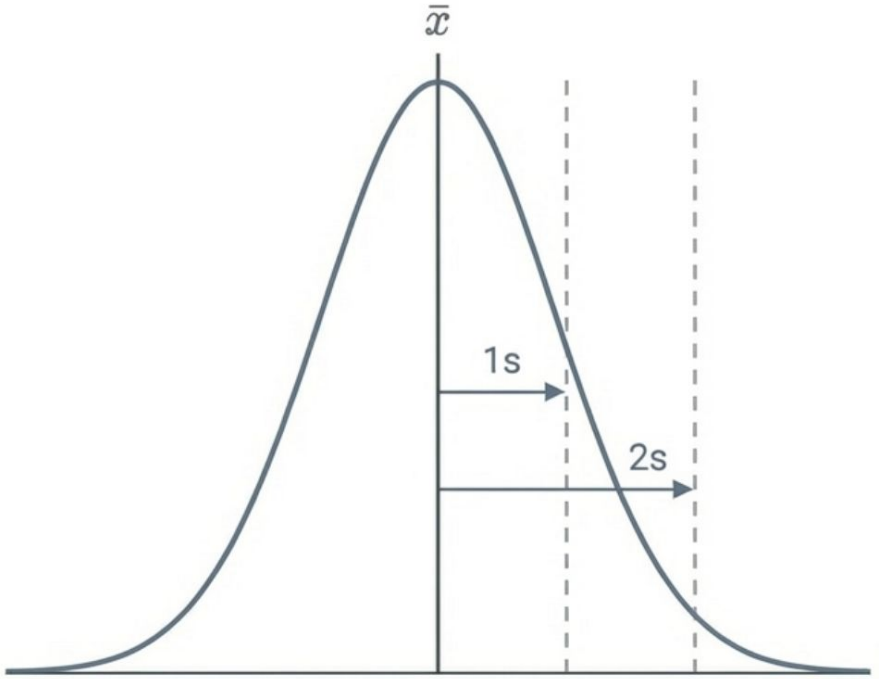


Descriptive Statistics

Standard Deviation Measures Average Variation

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

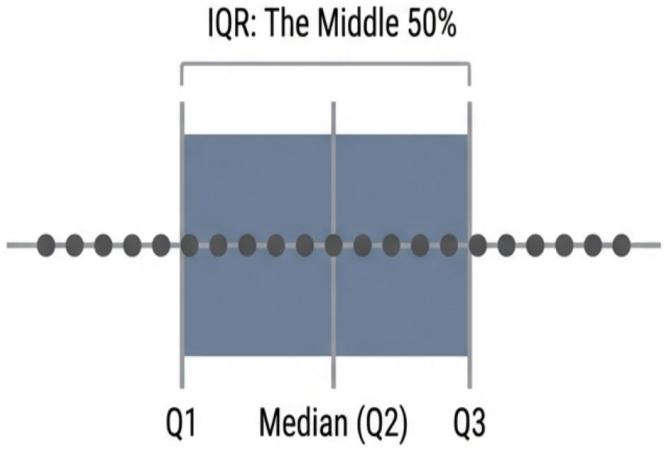
- s : Sample Standard Deviation.
- $(x_i - \bar{x})^2$: Squared distance from the mean (eliminates negative signs, penalizes large errors).
- n−1 : Degrees of freedom (Bessel's correction for unbiased population estimates).



Descriptive Statistics

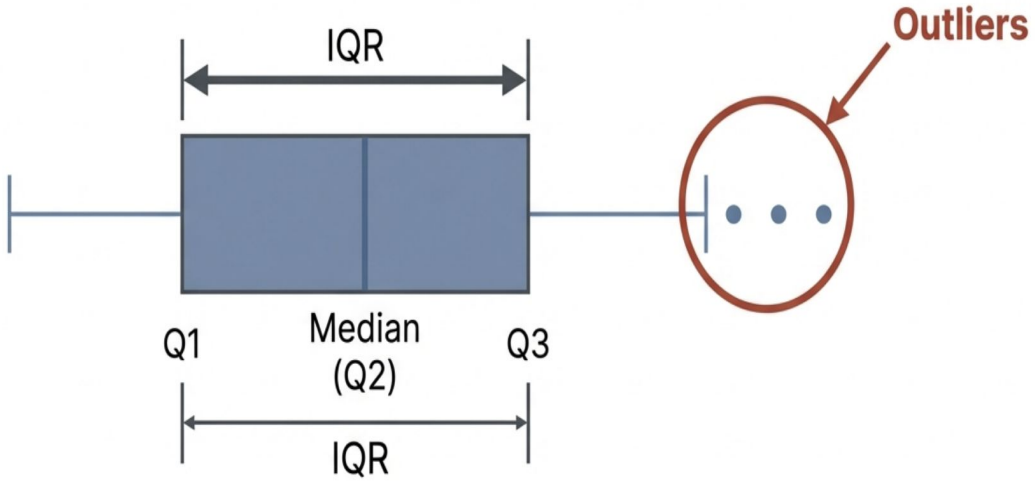
Quartiles and the IQR Isolate the Middle 50%

- Segments sorted data into four equal parts.
- Q1 (25th Percentile): 25% of data is ≤ this value.
- Q2 (50th Percentile): The Median.
- Q3 (75th Percentile): 75% of data is ≤ this value.
- IQR: Calculated as Q3 - Q1.



Box Plots Visually Translate the IQR and Outliers

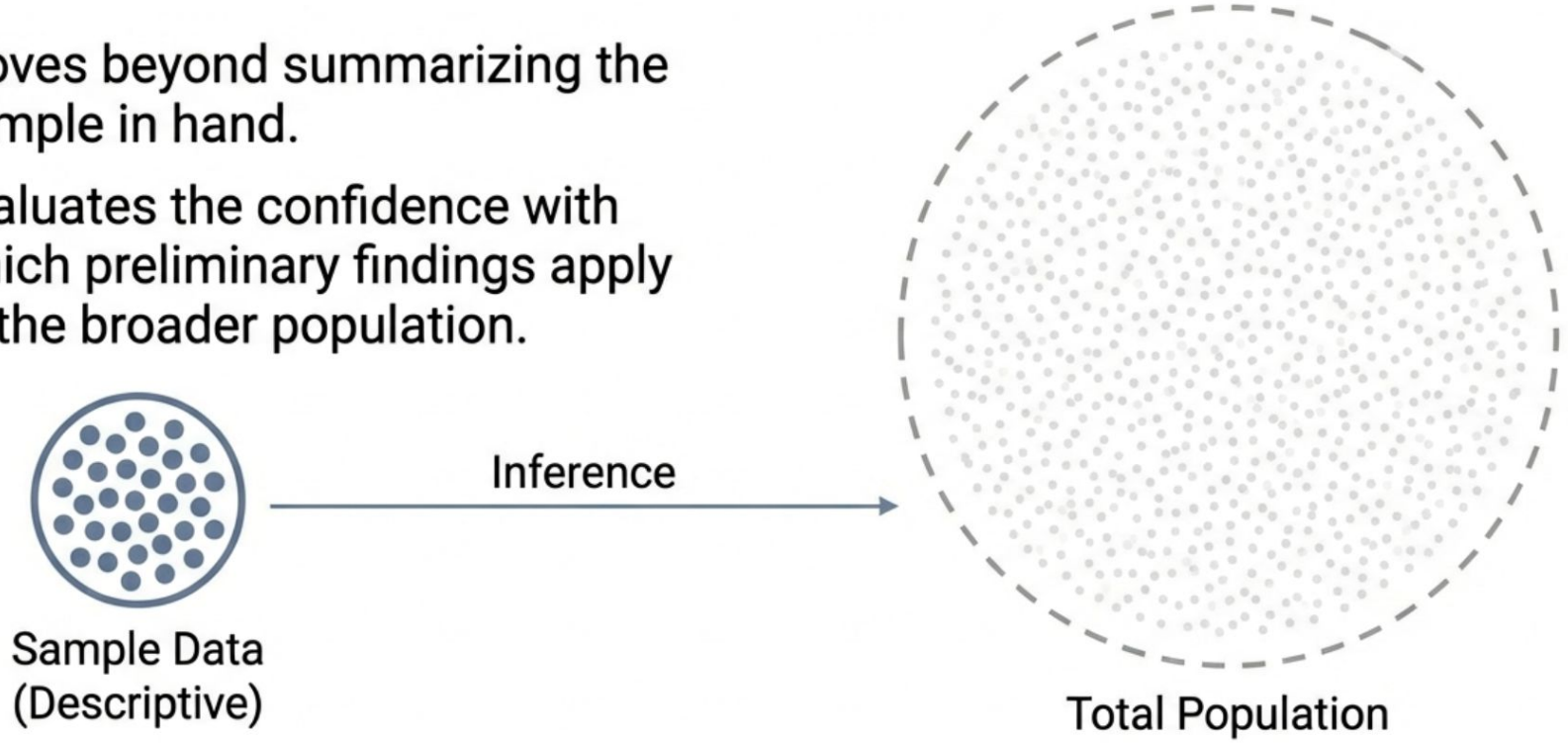
- The IQR provides a mathematical boundary to systematically identify outliers.



Descriptive Statistics

Inferential Statistics Assess Finding Reliability

- Moves beyond summarizing the sample in hand.
- Evaluates the confidence with which preliminary findings apply to the broader population.

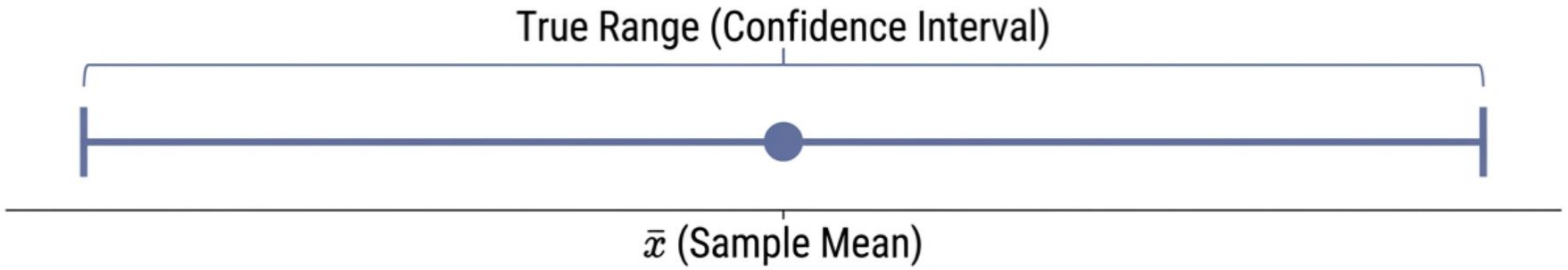


Descriptive Statistics

Confidence Intervals Estimate the True Population Mean

$$CI = \bar{x} \pm \left(t^* \cdot \frac{s}{\sqrt{n}} \right)$$

- $\bar{x}$: Sample mean.
- t^* : Critical value from t-distribution (based on confidence level, e.g., 95%, and n-1).
- $\frac{s}{\sqrt{n}}$: Standard Error (measures the volatility of sample means across repeated sampling).



Descriptive Statistics

Example of CI in a scientific paper

Table 3
Experiment results on CNUH dataset. Custom window size (D) = [8; 16; 24]; sliding step size k = 1. Predict the next 1 hour's clinical status.

Window time size (D)	Model	Metric (95% CI)		
		AUROC	AUPRC	Late alarm
8	BiLSTM + Attention [21]	0.799 (0.770 – 0.827)	0.572 (0.517 – 0.627)	80.8 (66.8 – 94.7)
	FCN	0.898 (0.885 – 0.911)	0.693 (0.658 – 0.728)	40.6 (35.5 – 45.6)
	RNN [20]	0.871 (0.854 – 0.888)	0.714 (0.687 – 0.741)	51.6 (42.3 – 60.8)
	DCNN [17]	0.902 (0.879 – 0.926)	0.734 (0.691 – 0.886)	37.6 (28.9 – 46.2)
	FCNN [18]	0.921 (0.905 – 0.937)	0.800 (0.777 – 0.823)	30.4 (24.0 – 36.7)
	XGBM [16]	0.955 (0.946 – 0.963)	0.811 (0.801 – 0.829)	18.0 (14.5 – 21.4)
	TVAE	**0.970 (0.958 – 0.982)**	**0.829 (0.752 – 0.906)**	**11.6 (6.8 – 16.3)**
16	BiLSTM + Attention [21]	0.843 (0.801 – 0.886)	0.647 (0.529 – 0.765)	59.2 (45.1 – 73.2)
	FCN	0.917 (0.895 – 0.939)	0.763 (0.712 – 0.813)	31.4 (22.6 – 40.2)
	RNN [20]	0.950 (0.928 – 0.981)	0.846 (0.794 – 0.899)	16.6 (7.7 – 25.5)
	DCNN [17]	0.944 (0.917 – 0.972)	0.811 (0.763 – 0.859)	20.6 (12.1 – 29.0)
	FCNN [18]	0.953 (0.934 – 0.973)	0.863 (0.827 –0.899)	17.8 (9.6 – 26.0)
	XGBM [16]	0.960 (0.945 – 0.974)	0.911 (0.870 – 0.953)	15.2 (10.1 – 20.2)
	TVAE	**0.970 (0.940 – 0.990)**	**0.858 (0.809 – 0.906)**	**10.8 (2.4 – 19.2)**
24	BiLSTM + Attention [21]	0.759 (0.575 – 0.944)	0.496 (0.167 – 0.826)	85.8 (26.1 – 145.4)
	FCN	0.910 (0.894 – 0.825)	0.758 (0.736 – 0.781)	32.6 (25.7 – 39.4)
	RNN [20]	0.962 (0.936 – 0.987)	0.857 (0.801 – 0.913)	13.8 (4.0 – 23.6)
	DCNN [17]	0.949 (0.929 – 0.969)	0.844 (0.797 – 0.892)	18.2 (11.1 – 25.3)
	FCNN [18]	0.957 (0.933 – 0.981)	0.866 (0.825 – 0.908)	15.4 (6.6 – 24.1)
	XGBM [16]	0.960 (0.941 – 0.979)	0.814 (0.805 – 0.923)	14.6 (7.2 – 21.9)
	TVAE	**0.973 (0.955 – 0.989)**	**0.887 (0.846 – 0.927)**	**9.6 (3.5 – 15.6)**

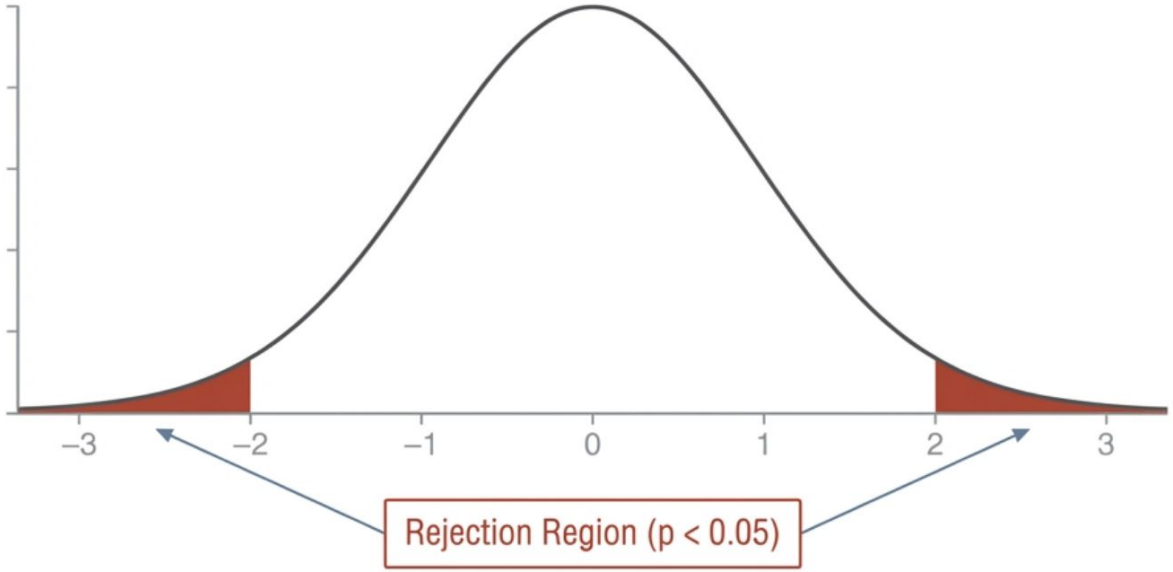
Table 4
Experiment results on UV dataset. Custom window size (D) = [8; 16; 24]; sliding step size k = 1. Predict the next 1 hour's clinical status.

Window time size (D)	Model	Metric (95% CI)		
		AUROC	AUPRC	Late alarm
8	BiLSTM + Attention [21]	0.823 (0.593 – 0.998)	0.744 (0.552 – 0.935)	1604.6 (546.3 – 3755.5)
	FCN	0.898 (0.885 – 0.911)	0.830 (0.808 – 0.851)	225.0 (180.7 – 269.3)
	RNN [20]	0.969 (0.962 – 0.975)	0.890 (0.881 – 0.899)	263.6 (212.3 – 314.8)
	DCNN [17]	0.930 (0.920 – 0.940)	0.808 (0.793 – 0.824)	621.0 (523.9 – 718.0)
	FCNN [18]	0.974 (0.970 – 0.979)	0.919 (0.912 – 0.927)	221.0 (172.9 – 269.0)
	XGBM [16]	0.907 (0.906 – 0.908)	0.803 (0.800 – 0.807)	842.4 (831.5 – 853.2)
	TVAE	**0.983 (0.976 – 0.990)**	**0.909 (0.906 – 0.913)**	**136.4 (68.9 – 203.8)**
16	BiLSTM + Attention [21]	0.947 (0.935 – 0.960)	0.863 (0.854 – 0.872)	437.0 (327.1 – 546.8)
	FCN	0.986 (0.982 – 0.991)	0.923 (0.919 – 0.926)	103.0 (70.5 – 135.4)
	RNN [20]	0.969 (0.962 – 0.975)	0.991 (0.990 – 0.992)	63.0 (58.4 – 67.5)
	DCNN [17]	0.959 (0.944 – 0.974)	0.880 (0.859 – 0.900)	343.4 (214.3- 472.4)
	FCNN [18]	0.980 (0.974 – 0.986)	0.912 (0.903 – 0.954)	165.8 (110.8 – 220.7)
	XGBM [16]	0.922 (0.916 – 0.927)	0.835 (0.822 – 0.848)	680.6 (641.8 – 719.3)
	TVAE	**0.989 (0.987 – 0.991)**	**0.915 (0.904 – 0.925)**	**82.4 (55.4 – 109.3)**
24	BiLSTM + Attention [21]	0.946 (0.922 – 0.971)	0.867 (0.842 – 0.891)	431.6 (225.2 – 637.9)
	FCN	0.991 (0.988 – 0.993)	0.927 (0.922 – 0.931)	65.8 (45.0 – 86.5)
	RNN [20]	0.991 (0.99 – 0.993)	0.920 (0.914 – 0.925)	55.2 (44.0 – 66.3)
	DCNN [17]	0.980 (0.977 – 0.983)	0.915 (0.913 – 0.917)	155.6 (129.4 – 181.7)
	FCNN [18]	0.986 (0.979 – 0.992)	0.922 (0.92 – 0.924)	109.4 (50.9 – 167.8)
	XGBM [16]	0.938 (0.934 – 0.941)	0.868 (0.861 – 0.875)	523.4 (497.1 – 549.7)
	TVAE	**0.992 (0.988 – 0.996)**	**0.926 (0.922 – 0.93)**	**51.6 (13.1 – 90.0)**

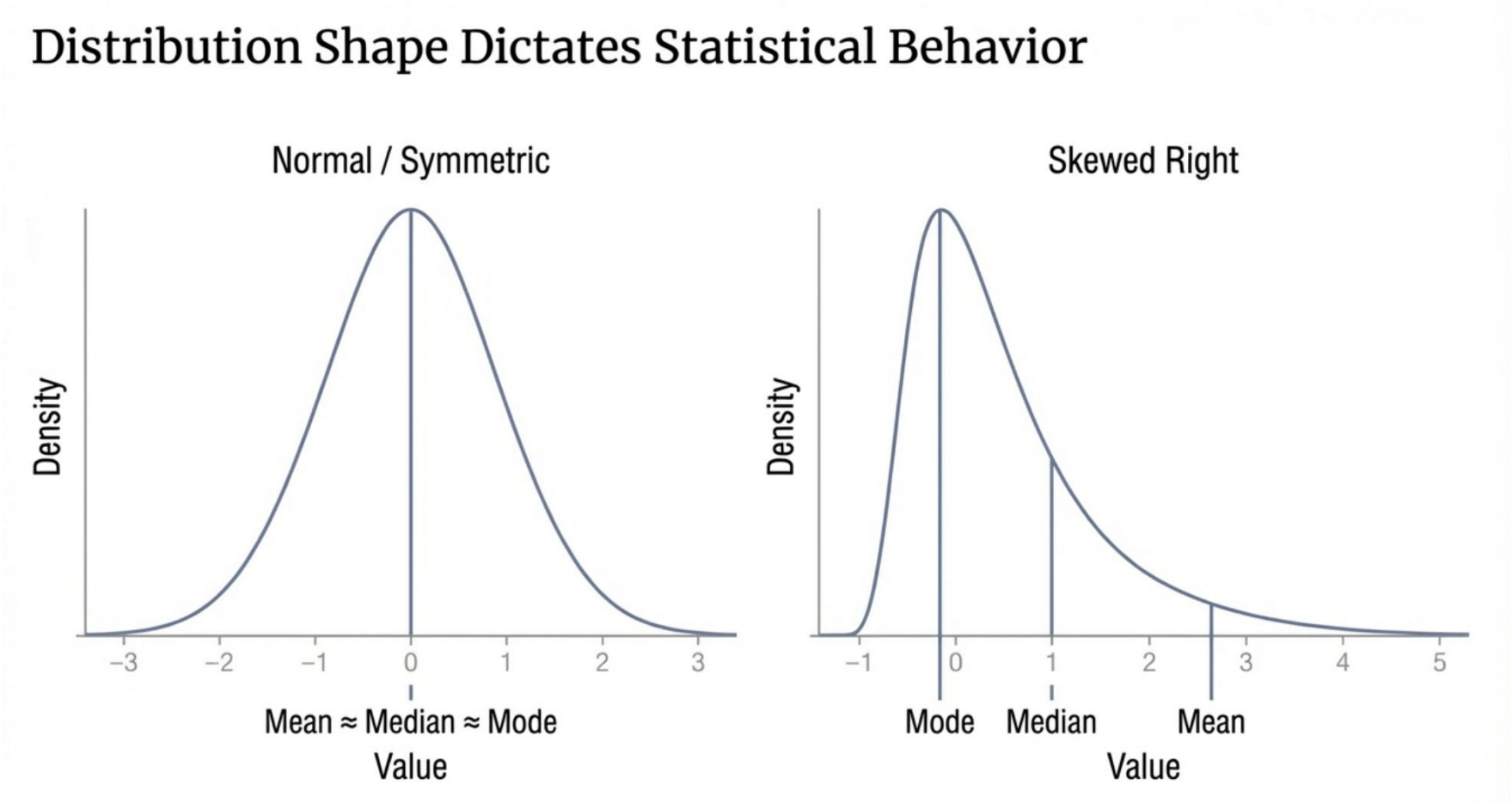
Descriptive Statistics

The p-value Determines Statistical Significance

- Measures the probability of observing the current result if the null hypothesis (H_0 - no difference) is true.
- Academic Threshold: If $p < 0.05$, the difference is statistically significant, not due to random chance.

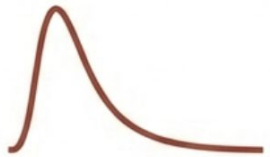


Descriptive Statistics



Descriptive Statistics

Metric Selection Matrix: Matching Tools to Distributions

Distribution Shape	Optimal Metrics
 Normal (Symmetric) Characteristic: Data is balanced; Mean ≈ Median ≈ Mode.	Use Mean for Center Use Standard Deviation (SD) for Spread
 Skewed (Asymmetric) Characteristic: Outliers stretch the tail, pulling the Mean away from the true center.	Use Median for Center Use Interquartile Range (IQR) for Spread

Content

- Exploratory Data Analysis (EDA)
- Descriptive Statistics
- **Summary Tables & Data Segmentation**
- PivotTables

Summary Tables

Step Zero is Converting Raw Data to Excel Tables

Creating an official Excel Table (Ctrl + T) is a mandatory preparation step to ensure consistency and efficiency before building any summary tables.

Before

Date	Region	Product	Sales
2023-01-01	North	Widget A	$150
2023-01-02	South	Widget B	$200
2023-01-03	East	Widget A	$175

Ctrl + T

After

Date	Region	Product	Sales
2023-01-01	North	Widget A	$150.00
2023-01-02	South	Widget B	$200.00
2023-01-03	East	Widget A	$175.00

> Auto-Expansion: When new data is added to the bottom of an Excel Table, all related formulas and summary tables automatically update their calculation ranges.

Summary Tables

Structured References Read Like Plain English

Naming conventions within Excel tables make formulas highly legible and significantly minimize calculation errors.

The Old Way

```
=SUM($A$2:$A$100)
```

Relies on cell coordinates.
Hard to interpret.

The Structured Way

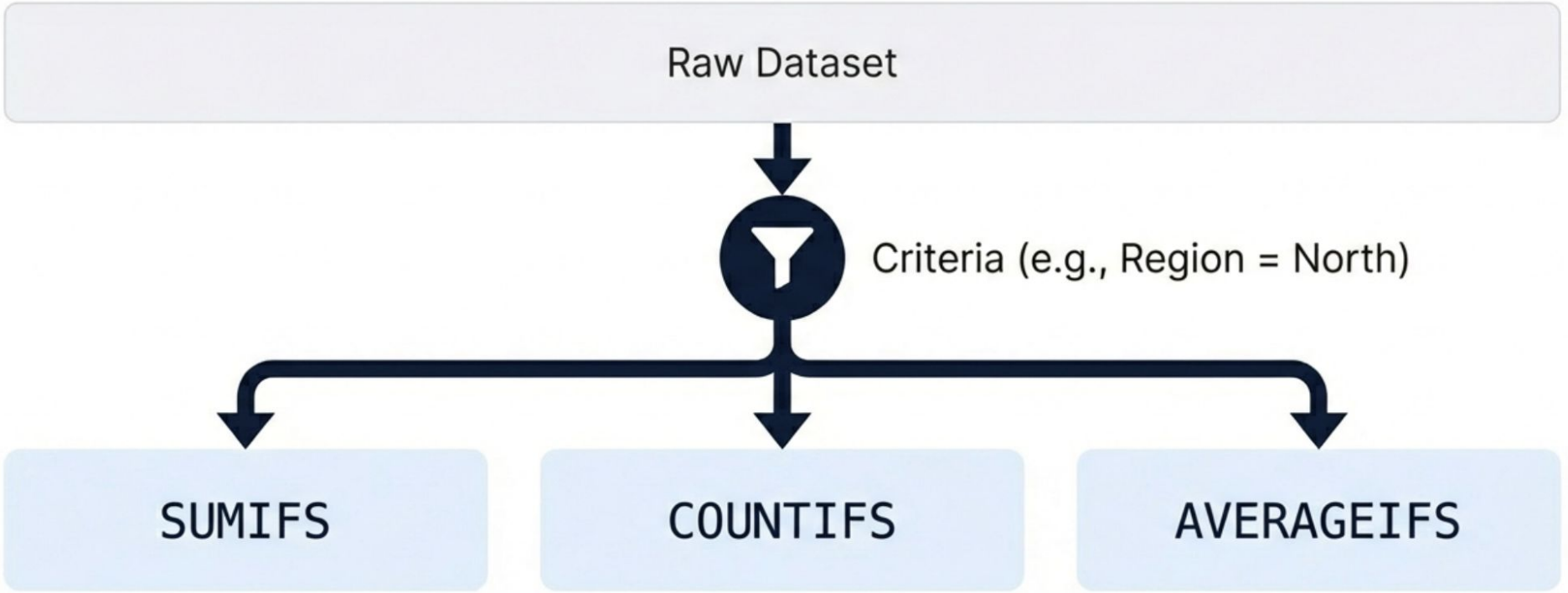
```
=SUM(Table1[Sales])
```

Uses table and column names.
Instantly clear.

Summary Tables

Building Summary Tables with Conditional Aggregation

Summary tables rely heavily on the *IFS family of functions. These formulas calculate metrics by evaluating one or multiple criteria against the dataset.



Summary Tables

SUMIFS Isolates and Adds Specific Data Points

```
=SUMIFS(sum_range, criteria_range1, criteria1, ...)
```

sum_range: The column containing the numeric values to be added together (e.g., Sales column).

criteria_range1: The column containing the categorical values to check against the condition (e.g., Region column).

criteria1: The specific condition required for the data to be included (e.g., 'North').

Applying SUMIFS to Calculate Regional Metrics

By utilizing multiple criteria ranges, SUMIFS can aggregate complex, multi-layered business questions into a single cell.

Date	Region	Sales
Jan 1	North	$1,200
Jan 5	South	$800
Jan 10	East	$950
Jan 12	North	$1,500
Jan 15	West	$1,100

Region	Total Sales
North	$1,200
South	$1,400
East	$950
West	$1,100

Summary Tables

COUNTIFS Tallies Records Based on Strict Criteria

```
=COUNTIFS(criteria_range1, criteria1, ...)
```

criteria_range1: The data range that will be evaluated.

criteria1: The specific standard a cell must meet to be counted as a valid record.

Segment	Number of Transactions
North	45

AVERAGEIFS Finds the Mean of Targeted Data Subsets

```
=AVERAGEIFS(average_range, criteria_range1, criteria1, ...)
```

average_range: The specific range containing the numeric values to average.

criteria_range1: The categorical data used to filter the dataset prior to the calculation.

Region	Average Deal Size
West	$4,500

Data Segmentation

Data Segmentation Answers Specific Business Questions

Segmentation is the technique of dividing a large dataset into smaller subgroups based on categorical variables. This allows analysts to compare performance and extract highly specific insights.

Which region generates the highest profit?

Which customer segment is currently growing?

Region	Customer Segment	Total Profit	YTD Growth
North	Enterprise	$500,000	+15%
South	Enterprise	$450,000	+8%
East	Enterprise	$520,000	+12%
West	Enterprise	$480,000	+10%
North	SMB	$150,000	+5%
South	SMB	$180,000	+20%
East	SMB	$160,000	+7%
West	SMB	$170,000	+9%
North	Consumer	$80,000	+2%
South	Consumer	$90,000	+4%
East	Consumer	$85,000	+3%
West	Consumer	$95,000	+6%

Data Segmentation

Three Core Techniques for Segmenting Datasets

AutoFilter

Applying quick filters directly on the data table to hide or show rows matching exact criteria.

Grouping

Combining granular values into higher-level categories (e.g., grouping individual monthly data into Quarters) to observe macro trends.

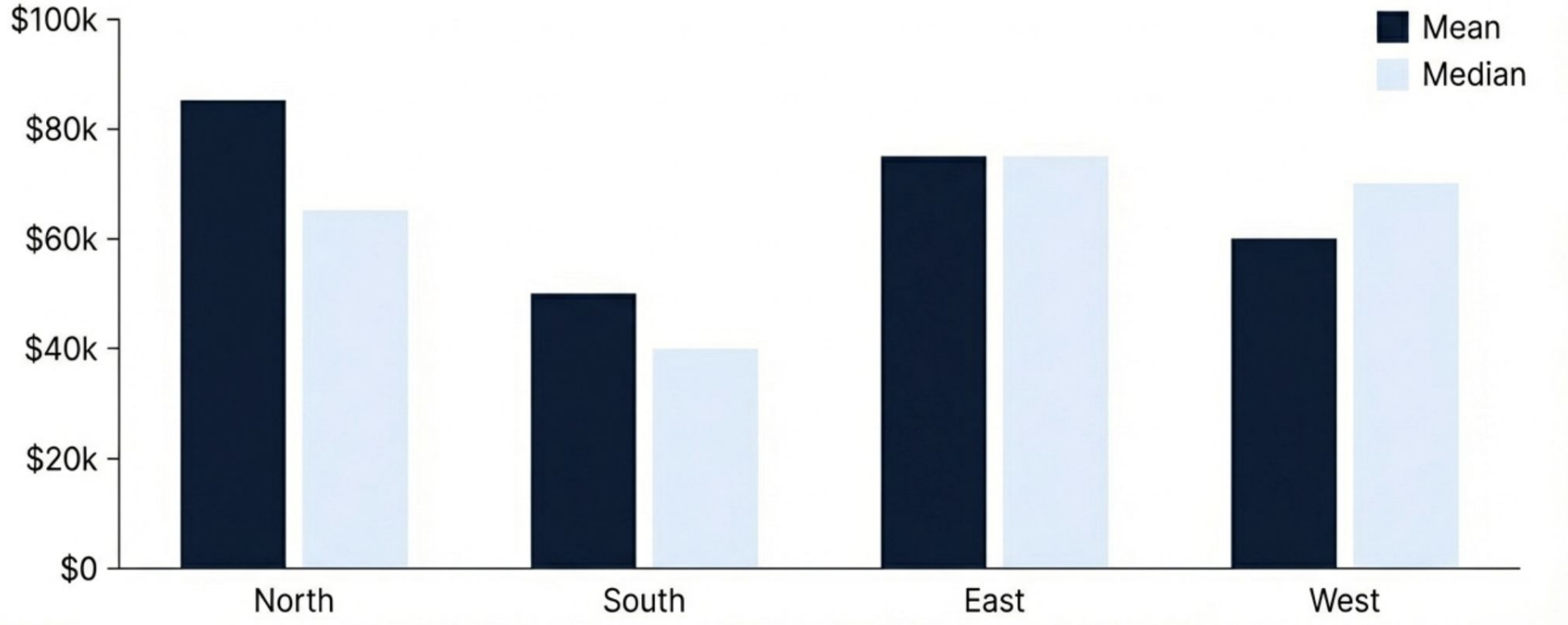
Comparison

Utilizing metrics like Mean and Median to evaluate the fundamental differences between the newly created segments.

Data Segmentation

Visualizing Segment Comparisons Reveals Intrinsic Differences

Once data is segmented and aggregated, visual comparison makes the intrinsic differences between categorical groups immediately obvious.



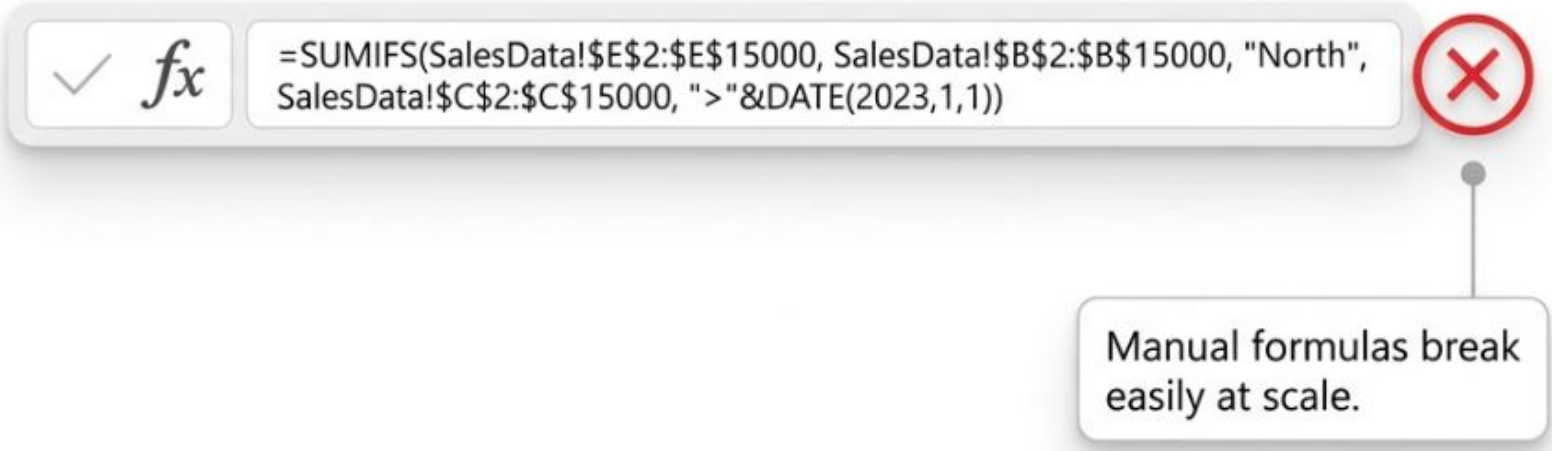
Content

- Exploratory Data Analysis (EDA)
- Descriptive Statistics
- Summary Tables & Data Segmentation
- **PivotTables**

PivotTables

Move beyond complex manual formulas

Analyzing thousands of rows manually requires writing and maintaining complex functions like SUMIFS or COUNTIFS. This manual approach is slow, highly prone to errors, and difficult to adapt when business questions change.



PivotTables

The ultimate interactive summary tool

A PivotTable allows you to rearrange (or "pivot") rows and columns to view your data from entirely new angles in seconds.

Speed:
Create instant summaries without writing a single formula.

Flexibility:
Instantly swap rows and columns to find hidden relationships.

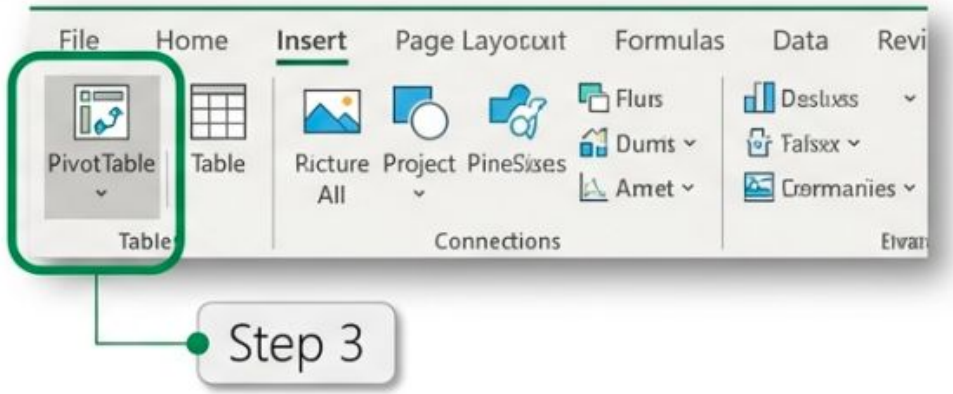
Scale:
Process massive datasets effortlessly.

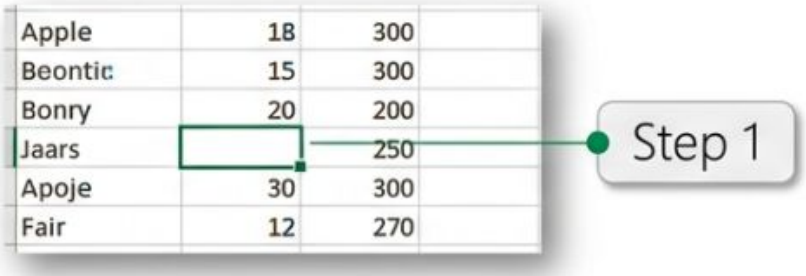
Agility:
Easily refresh your report when source data changes.

PivotTables

Launching your PivotTable starts in the Insert ribbon

1. Select any single cell inside your source data (formatting your your data as an Excel Table first first is highly recommended).
2. Navigate to the Insert tab.
3. Click PivotTable.

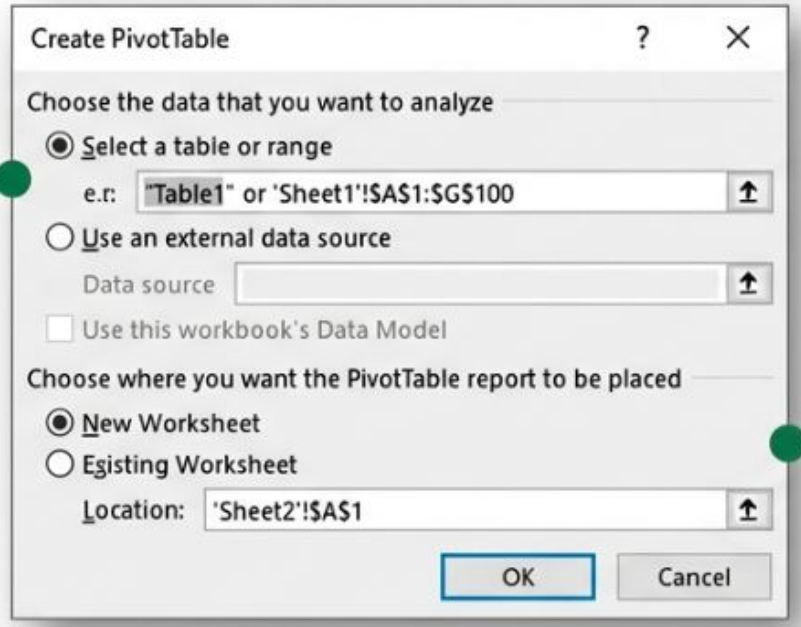




PivotTables

Configure your data source and destination

Choose the data:
Select either a local Table/Range or pull from the Data Model.

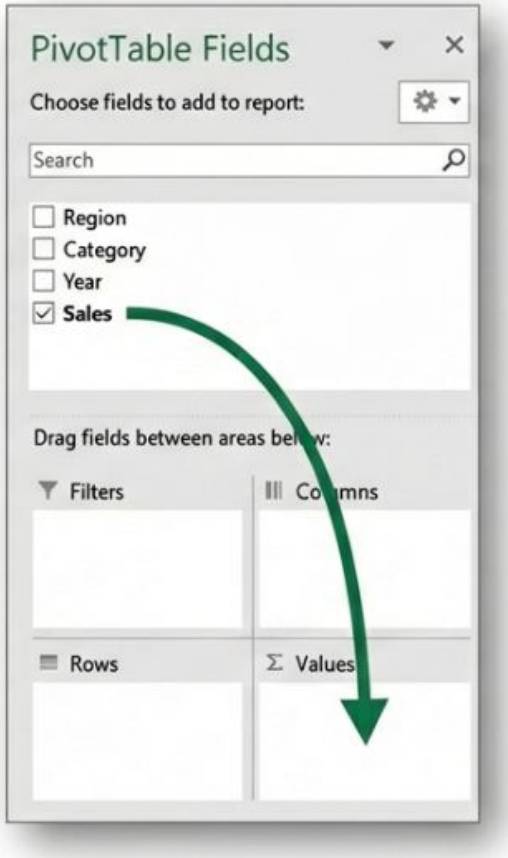


Choose the location:
Place your new report on a New Worksheet to keep things clean, or an Existing Worksheet to view it alongside other data.

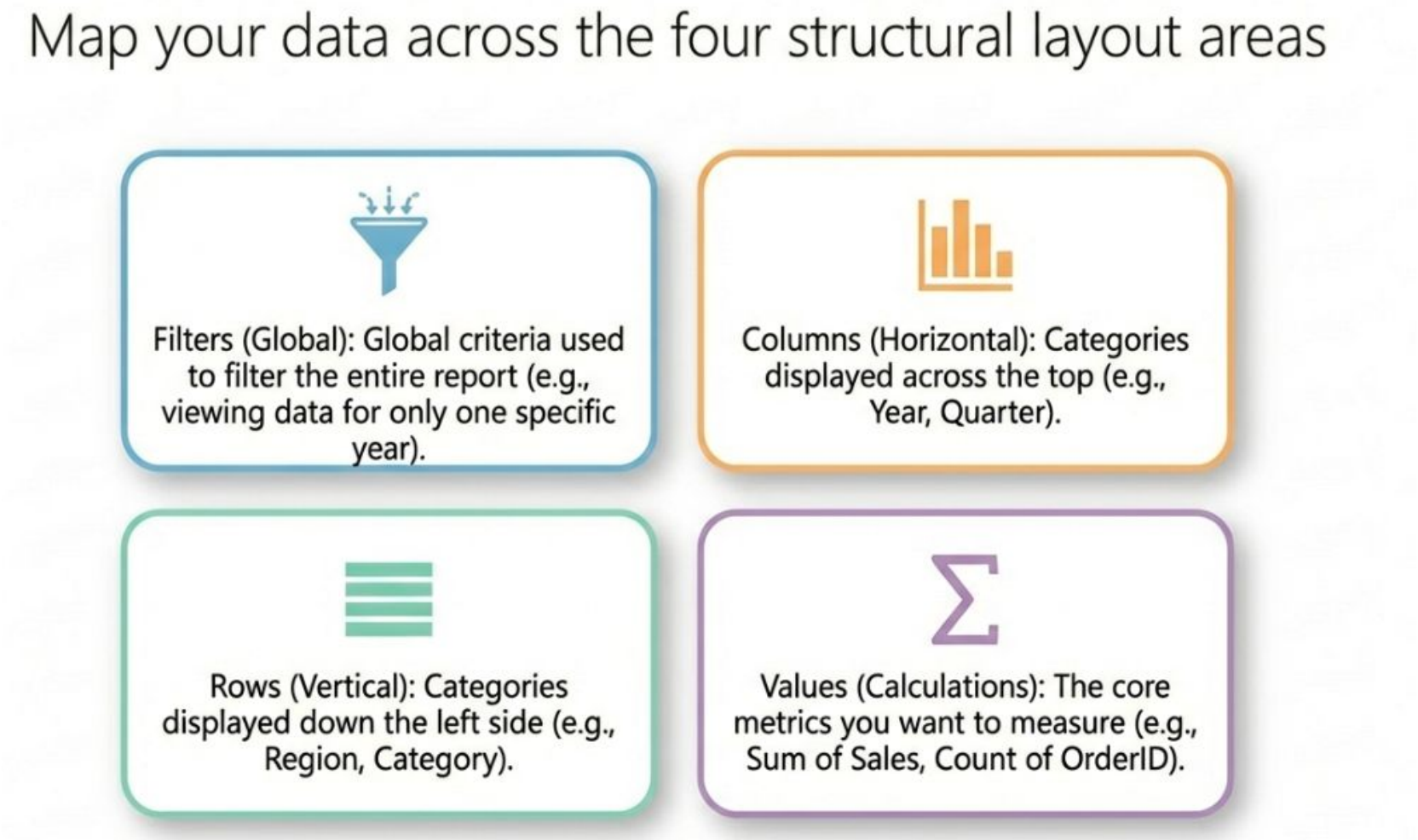
PivotTables

The PivotTable Fields pane is your command center

Once your table is created, this pane appears. Building your report is simply a matter of checking boxes or dragging data fields from the top list into one of the four functional areas at the bottom.



PivotTables

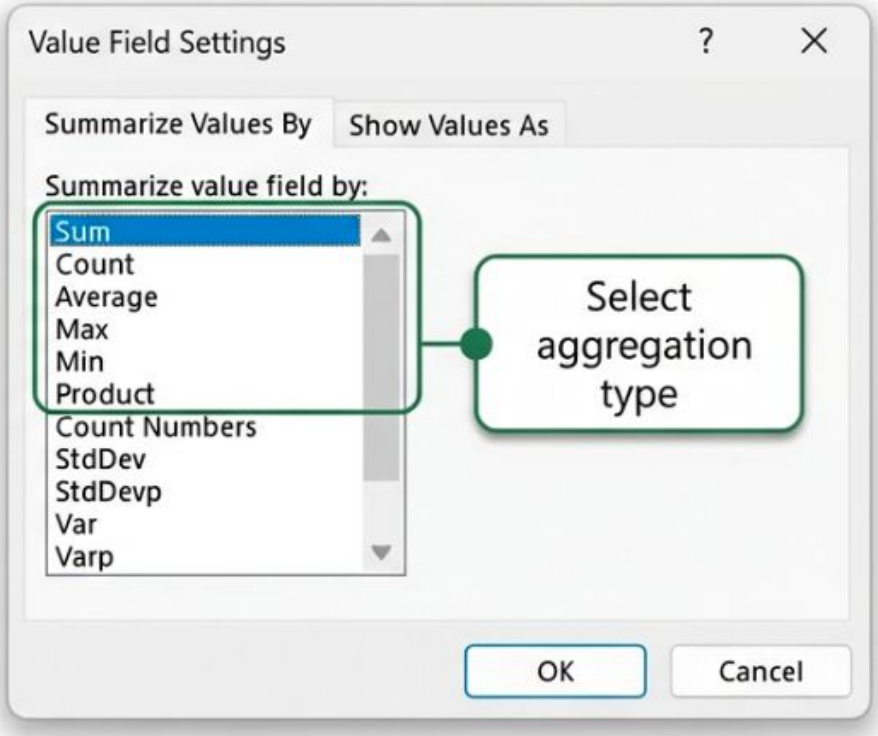


PivotTables

Control your calculations with Value Field Settings

You are not limited to basic addition. Right-click any value and open **Value Field Settings** to change how data aggregates:

- **Sum**: Total revenue or profit.
- **Count**: Tally the number of records (Order IDs, customer counts).
- **Average / Min / Max**: Find means or extremes.
- **Pro-Tip**: You can drag the exact same field into the Values area multiple times to view both the Sum and the Average simultaneously.

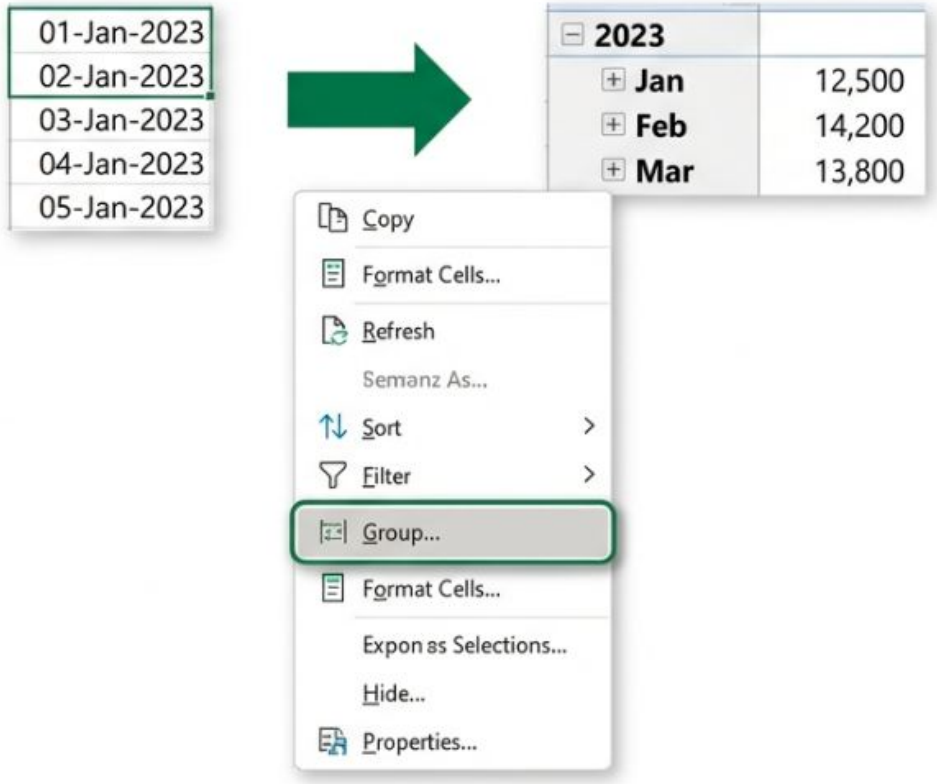


PivotTables

Group micro-data to reveal macro trends

Grouping is vital for Exploratory Data Analysis (EDA). Right-click any row or column value and select Group:

- Date Grouping: Instantly roll up individual daily dates into Months, Quarters, or Years.
- Numeric Grouping: Create distinct "bins" to categorize continuous numbers (e.g., bucketing orders into price ranges of 0-50, 50-100).



PivotTables

Filter visually and interactively using Slicers

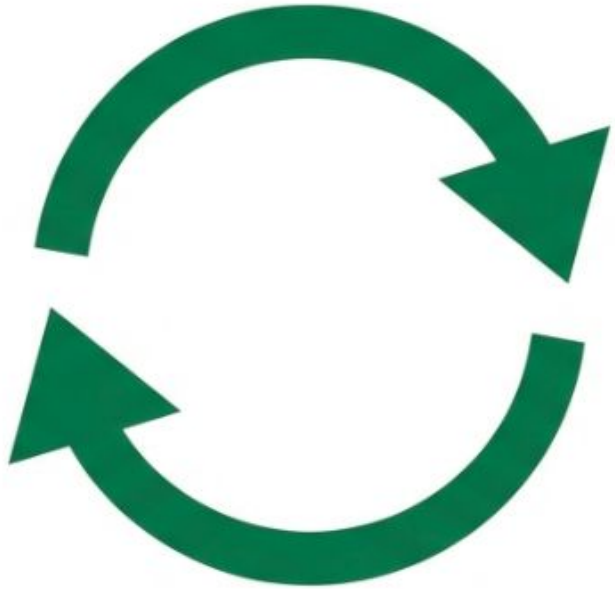
Standard drop-down filters can be hidden and clunky. Slicers replace them with visual, clickable buttons placed directly on your canvas.

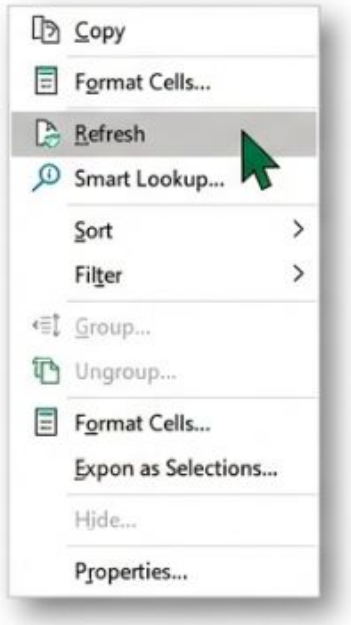
They make your reports highly interactive, allowing anyone looking at the dashboard to instantly slice the data with a single click.

PivotTables

Always remember to refresh your data manually

Unlike standard Excel formulas, a PivotTable does not automatically update when your source data changes or expands. Whenever you add new records or modify the original numbers, you must right-click anywhere inside the PivotTable and select Refresh to pull in the latest metrics.



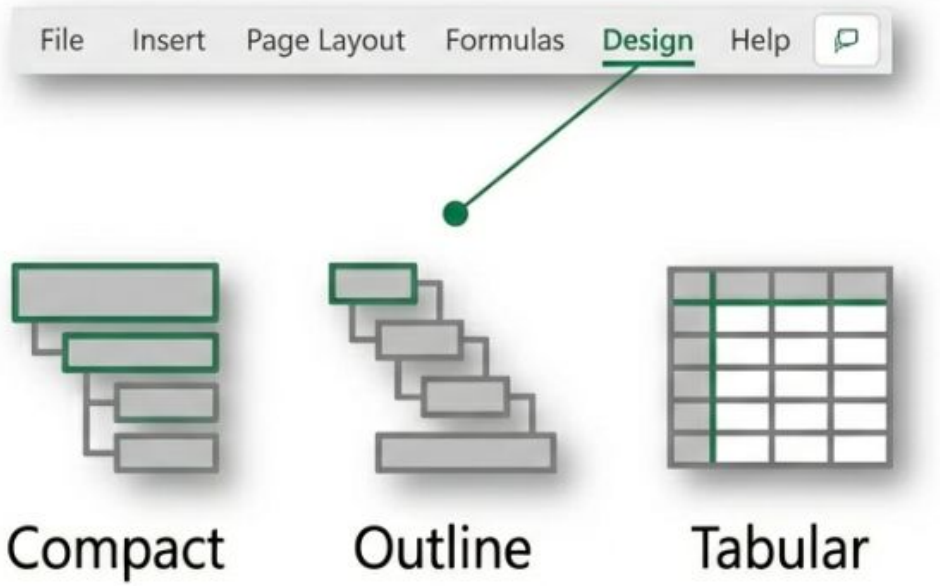


PivotTables

Design and format the final output structure

Navigate to the Design tab to finalize the presentation of your report:

- Toggle Subtotals and Grand Totals on or off to reduce visual clutter.
- Change the overall layout structure to fit your needs: choose between Compact, Outline, or Tabular forms.

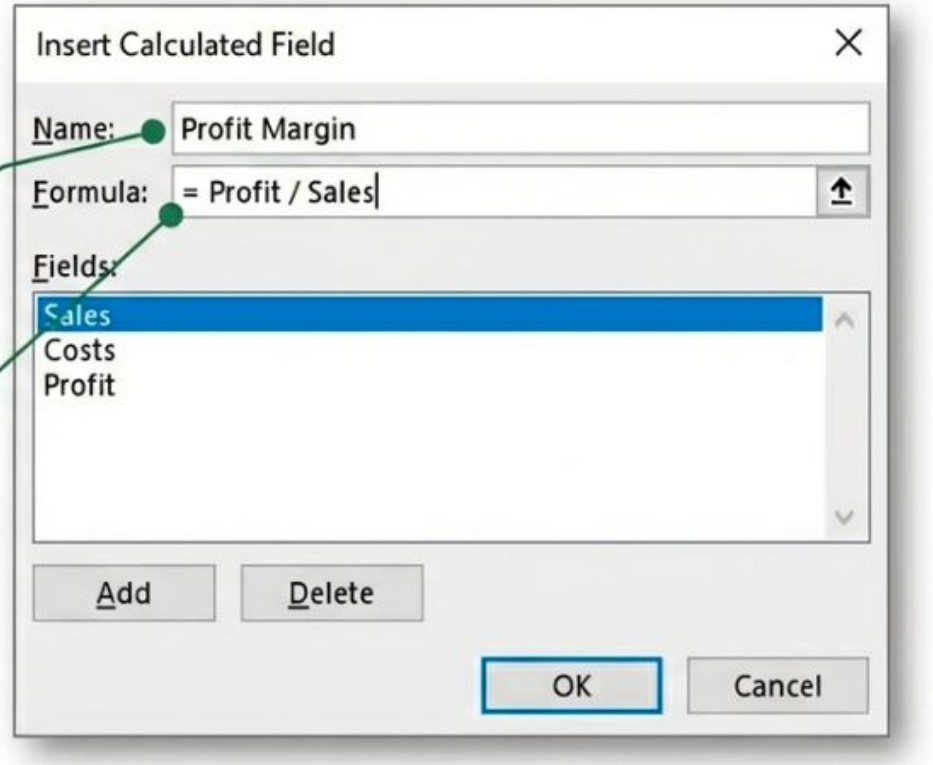


PivotTables

Generate new metrics internally with Calculated Fields

You don't need to alter your raw data source to calculate new ratios. Calculated Fields allow you to build completely new columns inside the PivotTable environment using formulas based on existing fields.

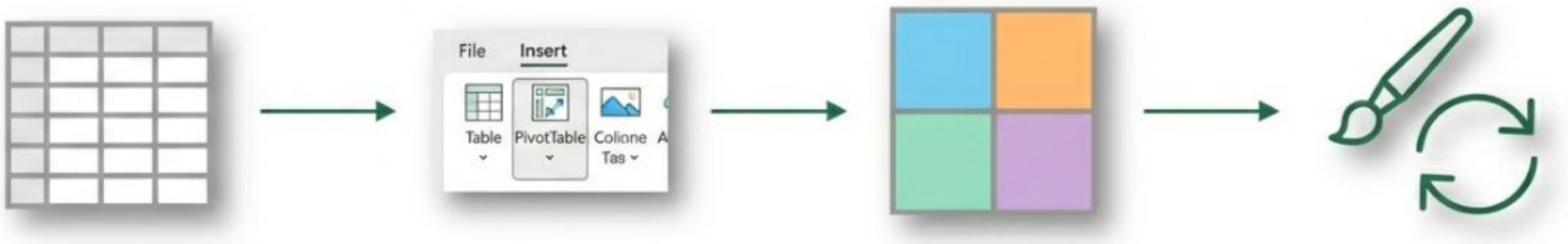
(e.g., Creating a new "Profit Margin" metric by dividing the existing Profit field by the Sales field).



PivotTables

The complete data transformation workflow

By mastering this sequential flow, you possess the ability to instantly organize chaos, isolate key metrics, and extract the precise insights your business needs.



Thank you!