

FROM RAW DATA TO INSIGHTS: THE 5 ESSENTIAL EDA STEPS



AKASH CHANDRASEKAR

STEP 1: IDENTIFY AND SOURCE RELEVANT DATA

The first step in EDA is to identify and collect a dataset that matches your problem statement or business goal. Data can be sourced from public platforms like Kaggle, UCI, or company databases. Make sure the data is relevant, accurate, and has enough observations. Also check the data format, granularity, and update frequency. A strong foundation in the form of good data saves a lot of time during cleaning and analysis.



STEP 2: UNDERSTAND THE DATA THROUGH DESCRIPTIVE STATISTICS

Once the data is loaded, understand its structure using basic statistical summaries. Use Pandas functions like `.info()`, `.describe()`, `.shape`, and `.value_counts()` to know data types, null counts, ranges, mean, median, and distribution. This step gives a bird's-eye view of your dataset and helps identify potential data quality issues. You can also calculate standard deviation, skewness, and kurtosis to assess spread and distribution shape before choosing the right transformation or visualization.



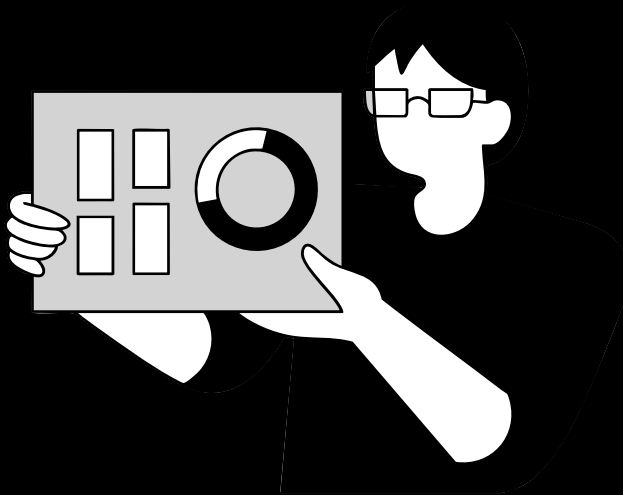
STEP 3: HANDLE MISSING VALUES STRATEGICALLY

Missing values are inevitable in most real-world datasets. They can arise from manual errors, system glitches, or data unavailability. Use `.isnull().sum()` to detect them. Depending on the nature of the column, you can impute missing values using the mean, median, or mode. Alternatively, if a column has too many missing values, consider dropping it. Use caution—dropping rows or columns can lead to data loss, so weigh it based on the impact on analysis.



STEP 4: DETECT AND TREAT OUTLIERS

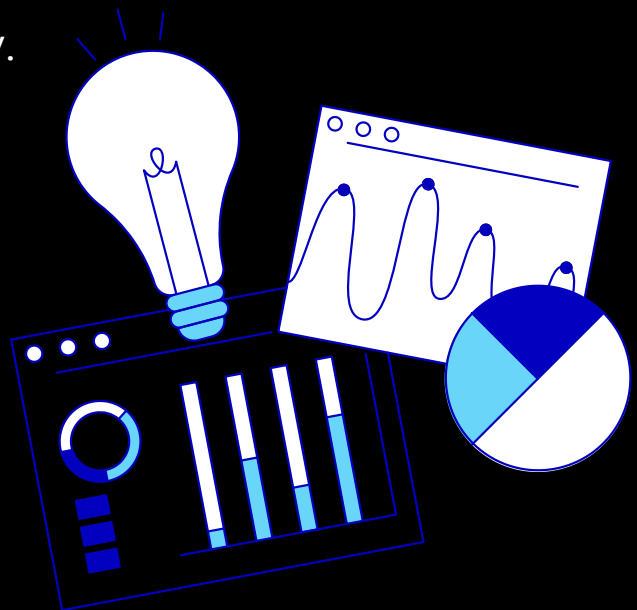
Outliers can distort model performance and mislead conclusions. Use boxplots, Z-score, or IQR methods to detect them. For numerical columns, values outside $1.5 * \text{IQR}$ range are often considered outliers. Once detected, you can either cap them (winsorization), transform them (e.g., log scale), or remove them entirely. However, not all outliers are bad—some may indicate important anomalies, especially in fraud detection or defect analysis. Always interpret outliers in the context of your domain.



AKASH CHANDRASEKAR

STEP 5: VISUALIZE DATA TO EXPLORE DISTRIBUTIONS AND RELATIONSHIPS

Visualization helps convert complex numbers into meaningful insights. Use Seaborn or Matplotlib to create histograms, scatter plots, boxplots, heatmaps, and pairplots. Histograms show distribution, scatter plots reveal trends and clusters, and heatmaps show feature correlation. Visuals help validate statistical summaries and communicate findings effectively to non-technical audiences. Good visualization highlights key insights at a glance and guides your modeling decisions. Always label axes and include titles to make graphs self-explanatory.



AKASH CHANDRASEKAR

CONCLUSION

- Data Discovery lays the groundwork by identifying relevant, quality datasets.
- Statistical Summarization helps in understanding central tendencies, spread, and structure.
- Handling Missing Values ensures completeness and reduces potential biases in analysis.
- Outlier Detection & Removal improves model performance and accuracy.
- Data Visualization brings clarity by revealing hidden patterns and distributional trends.