Exploratory Data Analysis (EDA)

Before building any sophisticated model, we need to do EDA first.

EDA is the first step in your data analysis. You take a broad look at patterns, trends, outliers, unexpected results and so on in your existing data, using visual and quantitative methods to get a sense of the story this tells. You're looking for clues that suggest your logical next steps, questions or areas of research.

- Dataset summary
- Missing data
- Basic Statistics
- Basic graphs
- Basic relationship

https://www.sisense.com/blog/exploratory-data-analysis/

Some of the tasks in EDA

- Spotting mistakes and missing data
- Mapping out the underlying structure of the data
- Identifying the most important variables
- Listing anomalies and outliers
- Test a hypotheses / check assumptions related to a specific model
- Establish a parsimonious model (one that can be used to explain the data with minimal predictor variables)
- Estimate parameters and figuring out the associated confidence intervals or margins of error.

Data Cleansing (Garbage in Garbage out, 80/20 rules)

Most data scientists spend only 20 percent of their time on actual data analysis and 80 percent of their time finding, cleaning, and reorganizing huge amounts of data

- Duplicate data removed
- Missing values need to be filled (or handled)
- Data elements should be comparable (similar units)
- Continuous values may need to be binned
- Outlier data need to be removed
- Ensure dataset has no systematic biases for the phenomena under analysis
- Be sure dataset has enough information density

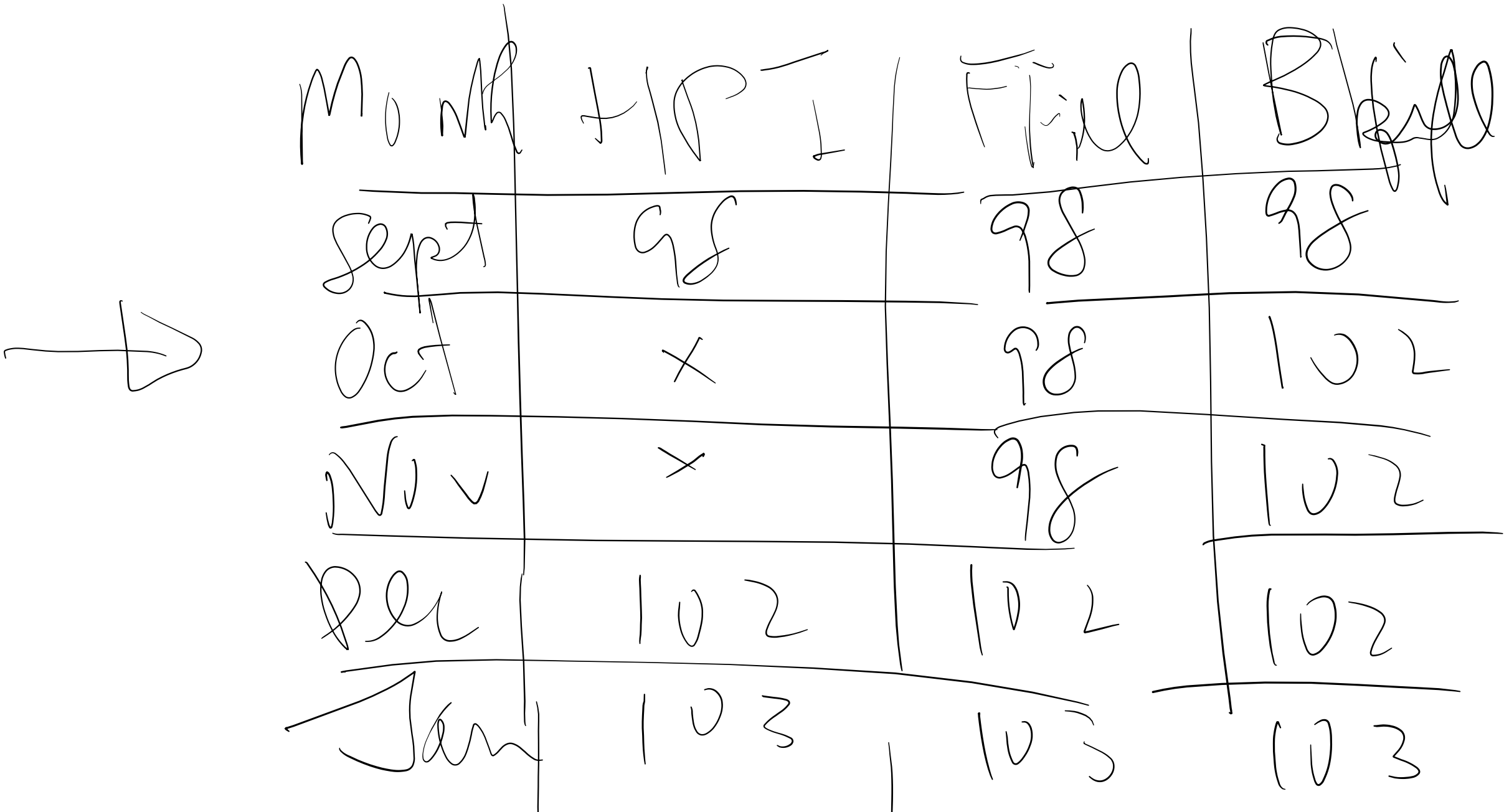
How to handle missing values —> Two ways to handle: Delete or Modify

- Deletion
  - Pro: most easy way and no ambiguity
  - Con: can apply only if we have enough data, may introduce systematic bias

- Imputation
  - Use Mean, Median or Mode
    - Pro: Easy to understand, ok most of the time
    - Con: may introduce systematic bias

  - For Time Series data,
    - Use last observed data (forward fill) (df.fillna(method='ffill'))
    - Use latest available data (backward fill) (df.fillna(method='bfill'))

  - More advanced method such as use nearest neighbor


Month
TIP
Fin
Bill
Sept
95
98
98
Oct
×
98
102
Nov
×
98
102
Dec
102
102
102
Jan
103
103
103

How to handle missing values

There is no one way to handle missing value

DEPENDS ON THE SITUATION

There is no silver bullet

That’s why a critical mind is important

Other aspects of Exploratory Data Analysis

Ask the right questions

Exploring Data Analysis

The goal of EDA is to explore and develop a high-level intuition and understanding of the data before we dive into any more sophisticated models

Exploratory Data Analysis

Learning by doing