Basic Sanity checks of Data

GROKKERS
AI FOR EVERYONE

Pre-reqs

- Python
- NumPy and PANDAS, SciPy, Visualizations
- Elementary stats and maths
- Some preprocessing steps – may need ML as well, for advanced topics

Background

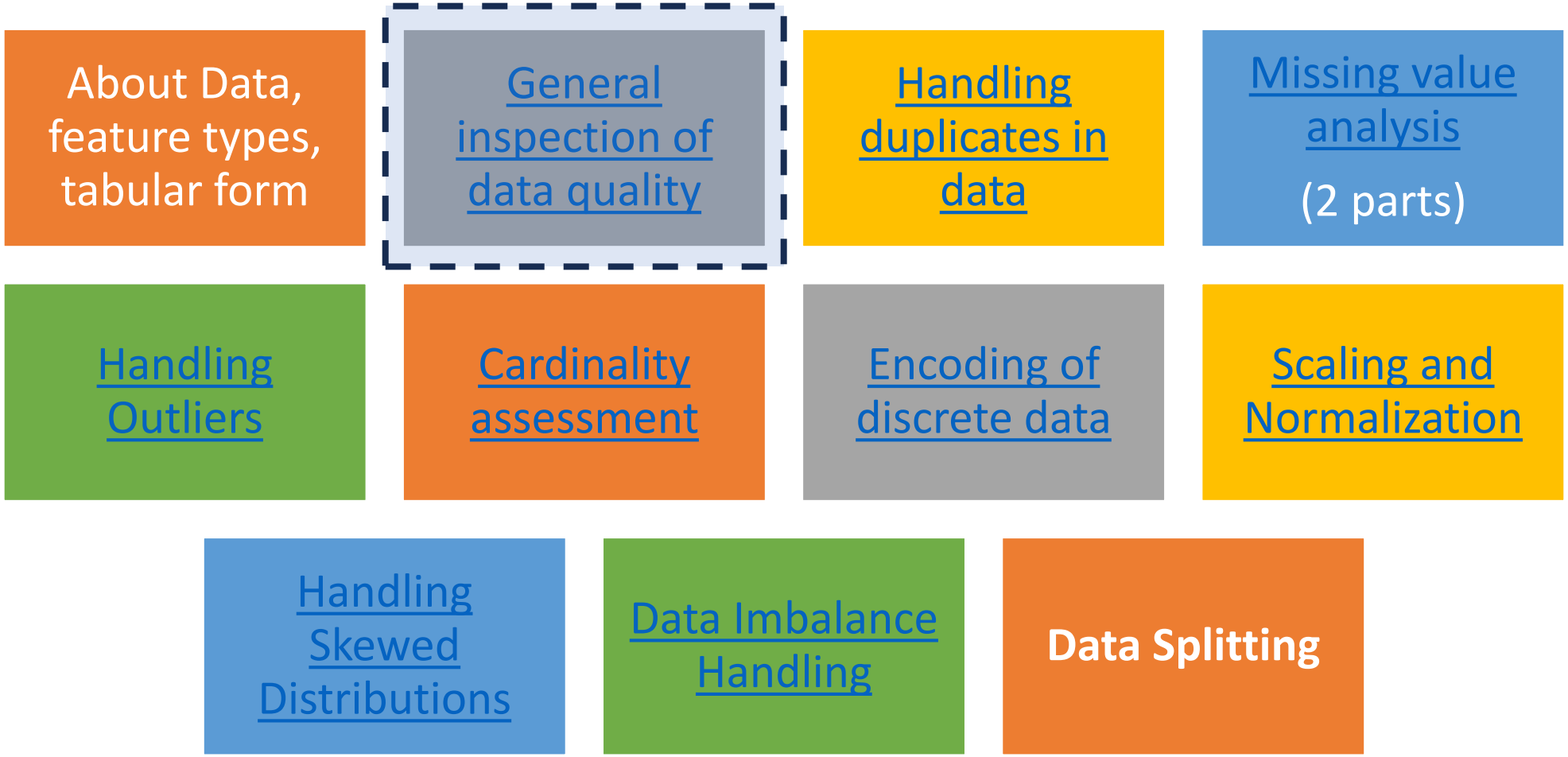
Numeric Data: Preprocessing involves handling missing values, scaling to a similar range, and possibly normalizing the distribution.

Text Data**:** Common preprocessing steps include text cleaning (removing stop words, punctuation, etc.), tokenization, and vectorization (converting text into numerical form, such as TF-IDF or word embeddings).

Image Data: Techniques like resizing, normalization of pixel values, and data augmentation (creating variations of existing images) are often used.

Time Series Data: Dealing with temporal aspects, handling missing values over time, and creating lag features are important steps in preprocessing time series data.

Topics

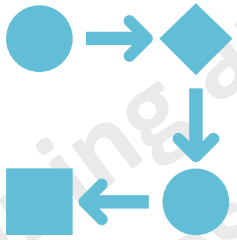


Preprocessing

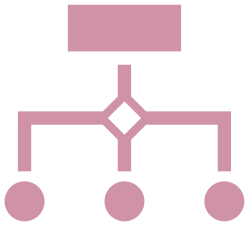
- General inspection
- Handling duplicates
- Missing data – imputation
- Handling outliers/noise/novelty
- Cardinality assessment
- Encoding - dummy variables
- Scaling , normalization
- Attribute transformations
- Multi-collinearity assessment
- Feature creation

General inspection of data

conducting a preliminary assessment of your raw data to identify any initial issues, inconsistencies, or patterns that might affect your analysis.



gain an understanding of the data's quality, distribution, and potential challenges before you proceed with more detailed preprocessing steps.



General inspection serves as a starting point to determine the extent of data cleaning, transformation, and feature engineering required.

General inspection

Data Overview

Data Size: Determine the number of rows and columns in your dataset.

Data Types: Identify the types of data in each column (numeric, categorical, text, etc.).

Basic Statistics: Calculate basic summary statistics like mean, median, and standard deviation for numerical columns.

Missing Data

Missing Values: Identify columns with missing data and assess the percentage of missing values for each column.

Data Distributions

Histograms: Create histograms to visualize the distribution of numerical variables.

Bar Plots: Plot bar charts to visualize the distribution of categorical variables.

General inspection

Data Quality

Data Consistency: Look for inconsistent data formats or values that don't make sense.

Unique Values: Identify columns with a high number of unique values, which might indicate data quality issues.

Domain Knowledge

Contextual Understanding: Apply domain knowledge to interpret any unusual patterns or inconsistencies in the data.

General inspection

Diligence	Comments
nominal or ordinal scale (where there are a fixed number of possible values)	inspect all possible values to uncover mistakes, duplications and inconsistencies. variable Company may include a number of different spellings for the same company such as “General Electric Company,” “General Elec. Co.,” “GE,” “Gen. Electric Company,” “General electric company,” and “G.E. Company.”
Numeric variables with inclusion of nonnumeric terms.	a variable generally consisting of numbers may include a value such as “above 50” or “out of range.”

General inspection

Diligence	Comments
Timeliness of data	• how up-to-date the observations are and whether the quality is the same across different sources of data.
Data been collected over time	• changes related to the passing of time may no longer be relevant to the analysis. • *Cost of production field* - collected over many years, the rise in costs attributable to inflation may need to be considered for the analysis.

Duplicate data

- **Duplicate** Records:
  - Duplicates can skew statistical analyses and lead to incorrect conclusions.
  - They can also result in **overestimations** or **underestimations** of certain patterns.

Handling duplicates

Customer Database	Clinical Trial Data	Social Media Engagement data	Sensor Readings
• contains information about individuals who have made purchases. Due to errors or system glitches, some customer records are duplicated.	• Data from a clinical trial includes duplicate entries for some patients, possibly due to data entry errors.	• contains duplicated entries for user engagement metrics, possibly due to tracking issues.	• includes duplicate readings, possibly due to communication errors.

Best practices on containing duplicates

Identify and Understand the Source of Duplicates

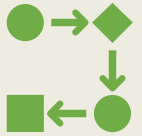
identify why duplicates exist. .. data entry errors, system glitches, or other reasons,

Deduplication During Data Collection

Implement deduplication mechanisms during the data collection phase.

Utilize Advanced Matching Techniques

… such as fuzzy matching or record linkage, to identify duplicates that may have variations in spelling or formatting.



Iterate and Refine:

Deduplication is often an iterative process.

Demo using python/sklearn

Simple example on duplicates with Pandas

Thanks !!