Data Preprocessing - Introduction

What is Data Preprocessing?

Data preprocessing is a collective term used to describe a collection of approaches that help get the data ready for analysis

It is the first step in the process of data analysis and goes by many different names like data cleaning/cleansing, data preparation, data scrubbing etc.

This is generally very contextual and it requires good domain understanding to do this well

Why data preprocessing?

Real world data is messy

- It is often incomplete, inconsistent and has many other such fallacies
- This makes it inapt for any statistical analysis
  - It might lead to wrong insights
  - Decisions taken from this data can be counter productive for the organization
- So you can't directly go from data to insights

In real world, the journey looks more like this...

Raw Data → Structure Data → Exploratory Data Analysis ⇅ Data Preprocessing → Insights, Visualizations
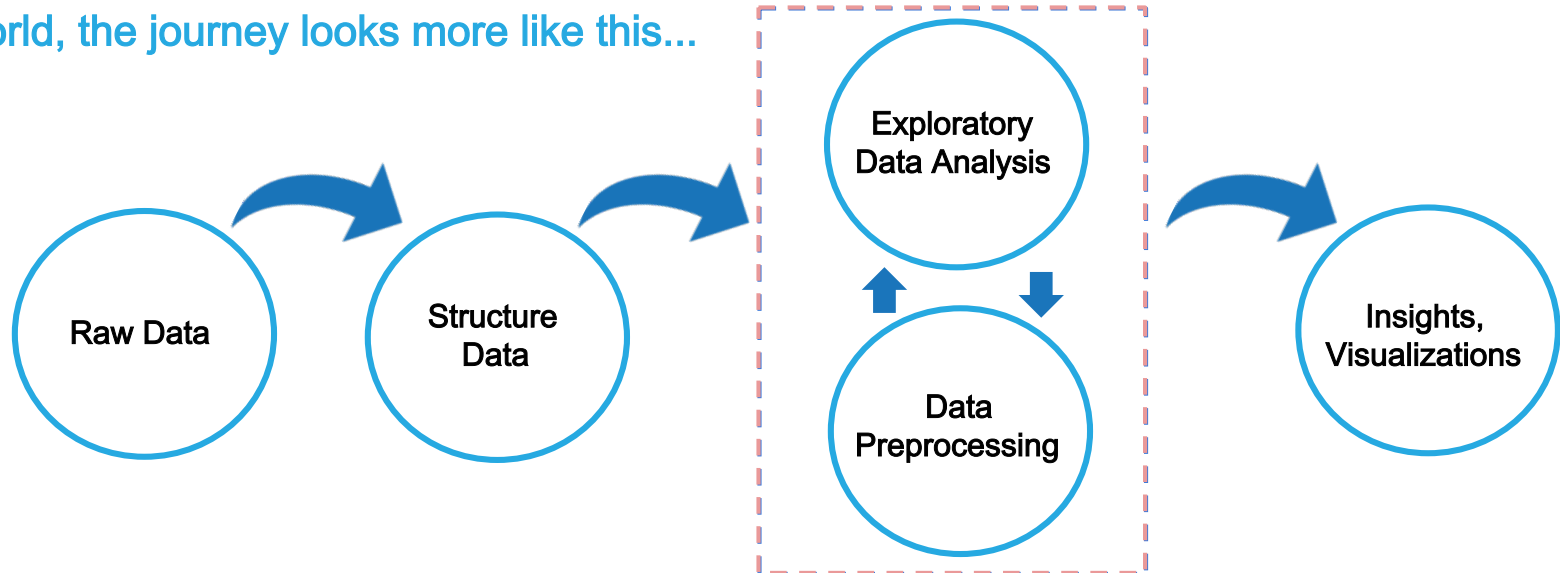
Broadly, the process can be thought of in these three steps

1

Data format checks

- data dimension
- data types

2

Data Consistency

- Missing values
- Extreme value / Outliers
- Distributions / Skewness

3

Feature Engineering

- Variable transformations
- Feature extraction

greatlearning
Power Ahead

Happy Learning !

Key Steps in Data Preprocessing

There are following key steps involved while doing data preprocessing:

a.   Data inconsistency - check data dimension, data types
b.   Check numerical and categorical data separately
c.   Looking at the data distribution - generally the first thing you do in any analysis
    i.   Missing values
    ii.   Outliers / Extreme values
    iii. Skewness
    iv.   Business sense check
d.   Variable Transformations
    i.   Encoding for Categorical variables - because some algorithms work only on numerical data
    ii.   Scaling
    iii. Other transformations
e.   Feature engineering - Creating new variables using domain knowledge