Problem Statement:-

1. Perform the following operations using R/Python on suitable data sets:

a) read data from different formats (like csv, xls)

b) indexing and selecting data, sort data,

c) describe attributes of data, checking data types of each column,

d) counting unique values of data, format of each column, converting variable data type (e.g. from long to short, vice versa),

e) identifying missing values and fill in the missing values.

Library:

Pandas, Matplotlib.

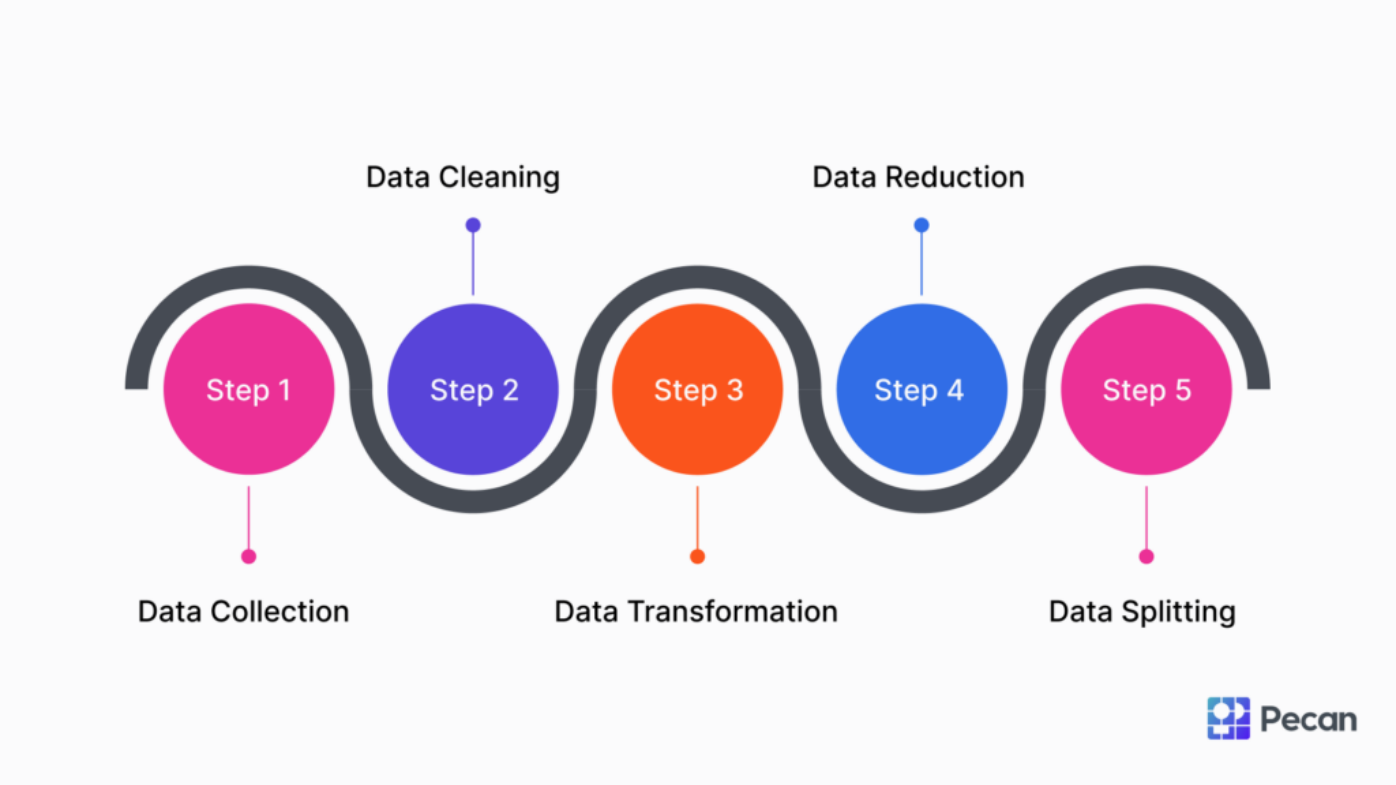
Theory:

Data Preparation in Machine Learning

Data Preparation is the process of cleaning and transforming raw data to make predictions accurately through using ML algorithms. Although data preparation is considered the most complicated stage in ML, it reduces process complexity later in real-time projects. Various issues have been reported during the data preparation step in machine learning as follows:

- **Missing data:** Missing data or incomplete records is a prevalent issue found in most datasets. Instead of appropriate data, sometimes records contain empty cells, values (e.g., NULL or N/A), or a specific character, such as a question mark, etc.
- **Outliers or Anomalies:** ML algorithms are sensitive to the range and distribution of values when data comes from unknown sources. These values can spoil the entire machine learning training system and the performance of the model. Hence, it is essential to detect these outliers or anomalies through techniques such as visualization technique.
- **Unstructured data format:** Data comes from various sources and needs to be extracted into a different format. Hence, before deploying an ML project, always consult with domain experts or import data from known sources.
- **Limited Features:** Whenever data comes from a single source, it contains limited features, so it is necessary to import data from various sources for feature enrichment or build multiple features in datasets.
- **Understanding feature engineering:** Features engineering helps develop additional content in the ML models, increasing model performance and accuracy in predictions.

Diagram:



Conclusion:

Data preparation is one of the key players in developing high-quality machine learning models. Data preparation allows us to explore, clean, combine, and format data for sampling and deploying ML models. It is essential as most ML algorithms need data to be in numbers to reduce statistical noise and errors in the data, etc. In this topic, we have learned about data preparation, the importance of data preparation in building predictive modeling machine learning projects, etc.