

Practical 1

AIM:

Data Wrangling, I

Perform the following operations using Python on any open source dataset (e.g., data.csv)

1. Import all the required Python Libraries.
2. Locate an open source data from the web (e.g. <https://www.kaggle.com>). Provide a clear description of the data and its source (i.e., URL of the web site).
3. Load the Dataset into pandas data frame.
4. Data Preprocessing: check for missing values in the data using `pandas.isnull()`, `describe()` function to get some initial statistics. Provide variable descriptions. Types of variables etc. Check the dimensions of the data frame.
5. Data Formatting and Data Normalization: Summarize the types of variables by checking the data types (i.e., character, numeric, integer, factor, and logical) of the variables in the data set. If variables are not in the correct data type, apply proper type conversions.
6. Turn categorical variables into quantitative variables in Python.

REQUIREMENT:

AnacondaInstaller

Windows10OS

JupyterNotebook

THEORY:

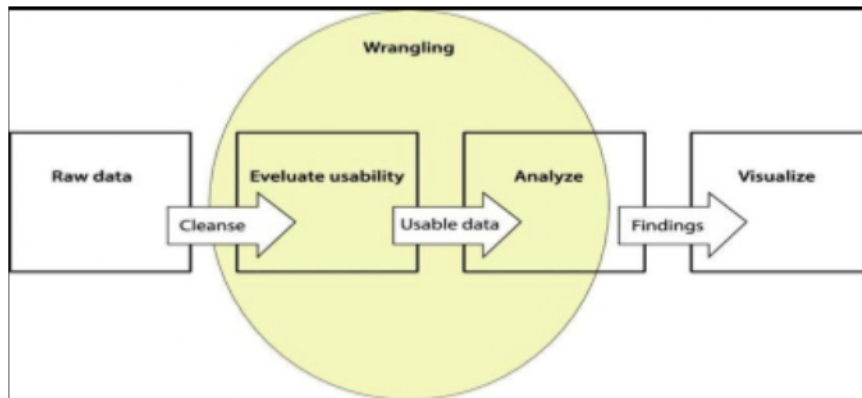
Data Wrangling:

- Data wrangling is the process of **cleaning and organizing messy data**
- It converts raw data into a **usable format for analysis**
- Involves **transforming and mapping data** from one form to another
- Helps in handling **large and complex datasets**

Goals of Data Wrangling:

- Collect data from **multiple sources** for better insights
- Provide **accurate and useful data** for analysis
- Reduce time spent on **data cleaning and preparation**
- Allow analysts to focus more on **data analysis**
- Support **better decision-making** in organizations
-

Key steps to DataWrangling:



LibrariesUsed:

Pandas

Numpy

Input:

[NOTE: Students need to paste their own Code here which they have performed]

Output:

[NOTE: Students need to paste the Screenshots of the Output of their code own Code here which they have performed, Individual Students Name should be there in the Screenshot]

CONCLUSION:

This experiment demonstrated basic data wrangling in Python, including loading a dataset, checking missing values, analyzing data types, and converting categorical variables into numerical form for further analysis.