Problem statement -

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.
3. Load the data sets into pandas data frame.
4. Data preprocessing : check for missing values in the data using pandas ,insult(), describe() function to get some initial statistics. Provide variable description. Types of variables etc. check the dimension 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 variable 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.

Objective -

eliminate redundant data (storing the same data in more than one table) and ensure data dependensies make sense (only storing related data in a table).

Theory -

Data wrangling -

It is also called data cleaning, data remediation or data munging - refers to a variety of processes designed to transform raw data into more readily used formats.

The exact methods differ from project to project depending on the data you are leveraging and the goal you are trying to archieve.

Importing all the required python Libraries

Here we using numpy and pandas libraries.

We execute that libraries by

- import numpy as np
- import pandas as pd.

Importing Data and reading into a Pandas DataFrame

- data = pd.read_csv('csv file')

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

Describe() function return the statistical summary of the data frame.

- data.describe()

Data formating -

- The last step in the data cleaning and making sure that all data is in the correct format (int, float, text or other)
- In pandas we use
  - .dtype() - to check the data type
  - .astype() - to change the data type.

MinMaxScaler -

Transform features by scaling each feature to a given range.

- sklearn.preprocessing.MinMaxScaler

Standard scaler -

standardize features by removing the mean and scaling to unit variance

The standard score of a sample x is calculated as

$z = (x-u)/s$

By using

- class sklearn.preprocessing.StandardScaler()

Result.

In this way we import libraries and perform the operations of data wrangling on the opensource dataset file.