

Name : Sakshi Dalvi

Batch : B3

Roll no. : 60

Panel : 2

DEC

Lab Assignment : 1

Aim : Data handling with Pandas

Objective :

- Learn to load use different data formats
- Prepare Data for Analysis

Theory :

- Various panda methods for loading and reading data from various file formats.
- (1) Load the CSV into a dataframe: import pandas as pd.
`df = pd.read_csv('data.csv')`
- (2) Print the DataFrame without the `to_string()` method: import pandas as pd.
- (3) Check the number of maximum returned rows: import pandas as pd.
- (4) Increase the maximum numbers of rows to display the entire dataframe: import pandas as pd.

Functions like the pandas `read_csv()` method enable you to work with file effectively.

Pandas functions for reading the contents of file are named using the pattern `read_<file_type>()`, where `<file_type>` indicates the type of the file to read.

- 1) read_json()
- 2) read_html()
- 3) read_sql()
- 4) read_pickle()

Various methods for central tendency in pandas

→ There are three main measures of central tendency: the mode, the median and mean. Each of these measures describes a different indication of the typical or central value in the distribution.

Input : Datasets

Output : ipynb file with all operations

Platform : windows & linux

FAQs :

Q1 What are different types data File Format ?

⇒ There are three types of data File Format :

- **File based data Format :** This type of data Format includes either one ~~or~~ file or more than one file. These files are then stored in any of the arbitrary folders. In most of the cases it uses the single file only for ex DGN. But then there are cases, which even includes atleast three files. The filename extension of all these three files is different from each other.
- **Directory Based Data Format :** In this type of data Format, whether there is one file or there is more than one file they are all stored in the parent folder in a particular manner. There are some cases where the requirement of an additional folder is there in the file tree in some other location so that it can be accessed easily. It is a possibility that data source is the directory itself. There are many files present in the directory, which are represented at the available data's layers.
- **Database connection :** In one respect, the database connections are quite similar to the above mentioned data formats that are file and directory-based data format. For interpreting, for map server, they give the geographic coordinate

data. One needs to access the coordinates inside the map server, that's creating vector datasets.

Q2 Compare various central tendency measures

⇒

Mode Median Mean

- most frequent data point
- value that divides ranked data pts. into halves: 50% of data is larger than it, 50% is smaller
- mode exists as a data point, median may not exist as a data point, mean may not exist as a data point
- unaffected by extreme values
- influenced by position of items, but not extreme values
- affected by extreme values
- useful for qualitative data, may have more than 1 value

C/G