Name: Patil Sanket
Roll no: COTB031

Group A
Assignment No: 1

Sinhgad Institutes | Date

objective of the Assignment:
student should be able to perform the data wrangling operation using python on any open source dataset.

Prerequisite:-
1. Basic of Python Programming
2. concept of data preprocessing, Data formatting Data Normalization and Data cleaning.

Concepts for theory:-
1. Introduction to dataset
A dataset is a collection of records, similar to rotational database table. Records are similar to table rows, but the columns can contain not only strings or numbers, but also nested data structures such as lists, maps and other records.

2. Instance- A single row of data is called an instance. It is a observation from the domain

3. Features- A single column of data is called a feature. It is component of an observation and is also called an attribute of data instance.

4. Data type: Features have a data type. They may be real or integer-valued or may have a categorial or ordinal value, you can have strings, dates,

times and more complex types, but typically they are reduced to real or categorial values when working with traditional machine learning methods.

5. Datasets:- A collection of instances is a dataset and when working with machine learning methods we typically need a few dataset for different purposes.

6) Training Dataset: A dataset that we feed into our machine learning algorithm to train our model. It may be called the validation dataset.

7) Data Represented in a table:

Data should be arranged in a two dimensional space made up of rows and columns. This type of data structure makes it easy to understand the data stored pinpoint any problems. An example of some raw data stored as a CSV (comma seperated values)

8) Pandas Data Types:

A data type is essentially an internal construct that a programming language use to understand how to store and manipulate data. A possible confusing point about pandas data types there is same overlap between pandas, python and numpy. This table summaries the key points.

2. python Libraries for Data science:-

a) Pandas

Pandas is an open-source python Package that provides high performance easy-to-use data structures and data analysis tools for the labeled data in Python programming language.

Sinhgad Institutes | Date

What can you do with pandas?

1. Indexing, manipulating, renaming, sorting, merging data frame.
2. Update Add, Delete, columns from a data frame.
3. Impute missing files, handle missing data or NANs.
4. Plot data with histogram or box plot.

b. Numpy

One of the most fundamental package in python, Numpy is a general purpose array-processing package. It provides high performance multidimensional array objects and tools to work with array. Numpy is an efficient container of generic multidimensional data.

In Numpy, dimensions are called axes and the number of axes is called rank. Numpy's array class is called ndarray aka array.

What can you do with Numpy?

1. Basic array operations: add, multiply, slice, flatten, reshape, index, arrays.
2. Advanced array operations: stack arrays, split into section, broadcast arrays.
3. Work with Data, Time or Linear Algebra.
4. Basic Slicing and advanced Indexing in Numpy Python.

C. Matplotlib

This is undoubtedly my favourite and quintessential python library. You can create stories with data, visualized with Matplotlib. Another library from the Scipy stack. Matplotlib plots 2D figures.

Sinhgad Institutes | Date

What can you do with Matplotlib?

Histogram, bar plots, Scatters plot area Plot to pie plot, Matplotlib can depict a wide range of visualizations. with a bit of effort and hint of visualization, capabilities, with matplotlib

You can create just any visualizations Line plots:

- Scatterm plots
- Area plots
- Bar charts and Histograms
- pie charts
- stem plots
- Countour plots
- Quiver plots
- spectrograms
-

d. Seaborn

So when you read the official documentation on seaborn it is defined as the data visualization library based on Matplotlib that provides a high level interference for drawing attractive and informative stastical graphics. putting it simply, seaborn is an extension of matplotlib with advanced features.

What can you do with seaborn?

1. Determine relationship between mutiple variables (correlation)
2. observe categorial variables for aggregate statics.
3. Analyze univariate or bi variate distribution and compare them between different datasubsets.
4. plot linear regression models for dependent variable

Date

e.s. Scikit Learn:

Introduced to the word as a google summer of code project. scikit learn is robust machine learning library for Python. It features ML algorithm like SVMs random forests. k means clustering. Spectral clustering, mean shift, cross-validation and more...

what can you do with scikit learn?

1. classification: spam detection, image Recongnition.
2. clustering: Drug response, stock price.
3. Regression: customer segmentation, grouping experiment outcomes
4. Dimensionally reduction, visualization, increased efficiency.

3. Description of Dataset:

The Iris dataset was used in R.A fisher's classic 1936 Paper, The use of multiple measured in taxonomic problems and can also be found on UCI machine Learning Repository. It includes three iri's species with 50 sample each as well as some properties about each flower.

Total sample - 150

The columns in this dataset are:

1. Id
2. SepalLength cm
3. sepalwidth cm
4. Petalwidth cm
5. Petallongth cm
6. Species.

Sinhgad Institutes | Date

4. Panda Dataframe function

Sr NO.	Data frame function	Description
1	dataset head (n=3)	Return the first n rows
2	dataset tail (n=5)	Return the last n rows
3	dataset index	The index (row labels) of dataset
4	dataset . columns	The column labels of dataset
5	dataset shape	Return a tuple representing the dimensionality of the dataset.
6	dataset. dtypes	Return the dtypes in the dataset.
7	dataset. columns ~~values~~ values	Return the columns values in the Dataset in array format
8	dataset. describe (include = "all")	Generate descriptive statistics
9	dataset ['column name']	Read the data column wise
10	dataset. sort_idex (axis =1. ascending= false)	sort object by labels (along an axis)
11	dataset. sort-values (by = 'column name")	sort values by column name
12	dataset.iloc [s]	purely integer location based indexing for selection by position
13	dataset [0:3]	selecting via [], which slices
14	dataset. lac [:1 "col	~~rows~~ s. selecting by label
15.	dataset . iloc [:n. i)	a subset of first n rows of original data.
16	dataset. iloc [:, :n]	a subset of first n columns of original data.
17	dataset. iloc[:m, :n]	a sub set of first m rows & the first n columns

6) Panda functions for data formatting and normalization

a] Data Formatting: Ensuring all data formats are correct (e.g object, text floating number, integer, etc) is another part of this initial cleaning process

Functions used for data formatting

Sr No	Data frame Function	Description	Output
1.	dt.types	To check the data types	df.dtypes sepal length (cm) float 64 sepal width (cm) float 64 peral length (cm) float 64
2	df['Petal length (cm) = df ['petal length (cm)']] astype	To change the data type of petal length (cm) changed to int	df.dtypes sepal. length (cm) float sepal. width(cm) float 64 petal length (cm) int 64 petal width(cm) float 64 dtype: object

Conclusion: In this way we have explored the functions of the python library for data preprocessing. Data wrangling techniques and How to Handle missing values on Iris Dataset.