Pune District Education Association's
College Of Engineering
Manjari (Bk.), Hadapsar, Pune-412307.
Accredited by NAAC

# ASSIGNMENT NO. 1

Title : Data Wrangling, I

Objectives : Students should be able to perform the data wrangling operation using python on only any open source dataset.

Aim :

Data Wrangling : I perform the following operations using python on any open source dataset (eg. data.csv)

1. Import all the required python libraries.
2. Locate an open source data from the web provide a clear discription of the data & its source.
3. Load the dataset into pandas data frame.
4. Data Preprocessing : Check for missing values in it using pandas isnull() describe function to get some initial statistics. Provide variable descriptions, types of variables, etc. check the dimension of the data frame.
5. Data formatting & Data Normalisation : Summarize the type of variables by checking the data types of data frame.

→ P.T.O.

Pune District Education Association's
College Of Engineering
Manjari (Bk.), Hadapsar, Pune-412307.
Accredited by NAAC

6. Turn categorical variables into quantitative variables in python. In addition to the codes & o/p, explain every operation that you do in the above steps & explain everything that you do to import/ read / scrape the data set.

Requirement : 1. Basic of python progeamming.
2. Concept of data preprocessing, Data formatting, Data Normalization & Data cleaning.

Theory:

· Data wrangling in python

Data wrangling is the process of gathering, collecting & transforming raw data into another format for better understanding, decision-making, accessing & analysis in less time. Data wrangling is also known as Data mugging.

· Importance of Data Wrangling -

Data wrangling is a very important step. The below example will explain its importance as: Books selling website want to show top-selling books of different domains, according to user preference. For example a new user search for motivational books which sell the most or having a

→ P.T.O.

Pune District Education Association's
College Of Engineering
Manjari (Bk.), Hadapsar, Pune-412307.
Accredited by NAAC

high rating, etc. But on their website, there are plenty of raw data from different users. Here the concept of data mugging or data wrangling is used. As we know Data is not wrangled by system. This process is done by Data scientists.

Data wrangling in python is a crucial topic for Data Science & Data Analysis. Pandas framework of python is used for Data wrangling. Pandas is an open source library specifically developed for data analysis & data science. The process like data sorting, Data filteration, Data grouping, etc.

Data Wrangling in python deals with the below funtionalities.

1. Data exploration : In this process, the data is studied, analyzed & understood by visualizing representation of data.
2. Dealing with missing values:

Most of the dataset having a vast amount of data contain missing values of NaN, they are needed to be taken care of by replacing them with mean, mode, the most fequent value of the column or simply by dropping the row having a NaN value.

P.T.O.

Pune District Education Association's
College Of Engineering
Manjari (Bk.), Hadapsar, Pune-412307.
Accredited by NAAC

3. Rephasing Data: In this process, data is manipulated according to the requirements, where new data can be added or pre-existing data can be modified.
4. Filtering Data: Sometimes dataset are comprised of unwanted rows or columns which are required to be removed or filtered.
5. Other: After dealing with the raw dataset as per our requirements & then it can be used for required purpose like data analyzing, machine learning, data visualization, model training, etc.

Below is an example which implements the above funtionalities on a raw dataset:

- Data exploration, here we assign the data & then we visualize the data in a tabular format.



csv file / Dataset - titanic_train.csv

- Required python libraries

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

→ P.T.O.

Pune District Education Association's
College Of Engineering
Manjari (Bk.), Hadapsar, Pune-412307.
Accredited by NAAC

· Required syntax

```
#Load the Dataset into pandas data feame
df = pd.read_csv('titanic_train.csv')
df

# For showing top result
df.head()

# For showing bottom result
df.tail()

# Calculating Null Values.
df.isnull().sum()

# Calculating Null Values in 'Age' & 'Cabin' column
df['Age'].isnull().sum()
df['Cabin'].isnull().sum()

# Get some initial statistics
df.describe()

# Getting some information about dataset
df.info()
```



P.T.O.

Pune District Education Association's
College Of Engineering
Manjari (Bk.), Hadapsar, Pune-412307.
Accredited by NAAC

```
# Finding Data types
df.dtypes

# Finding Dimensions of data frame
df.shape

# Making impute function for filling Null Values
def impute_age(cols):
    Age = cols[0]
    Pclass = cols[1]

    if pd.isnull(Age):
        if Pclass == 1:
            return 37
        elif Pclass == 2:
            return 29
        else:
            return 24
    else:
        return Age

# Displaying two-dimensional data in grid format where the color intensity represents value.
sns.heatmap()
sns.heatmap(df.isnull(), yticklabels=False, cbar=False,
    cmap='viridis')
```

→ P T O.

Pune District Education Association's
College Of Engineering
Manjari (Bk.), Hadapsar, Pune-412307.
Accredited by NAAC

# Applying impute function & droping column cabin respectively

```
df['Age'] = df[['Age','Pclass']].apply(impute_age, axis=1)
df.drop('Cabin', axis = 1, inplace = True)
```

# df.drop_duplicates()

# Data type conversion

```
df['Age'] = df['Age'].astype('int')
df['Age'] = df['Age'].round(0).astype('int')
```

# Converting categorial variables to Quantitative variables

```
cat = pd.get_dummies(df, columns = ['sex'])
```

# Female = 0 & male = 1

```
cat['Sex_female']
cat['sex_male']
```

# df.columns

Conclusion: Hence we have throughly studied how to perform the following operations on python on any open source dataset.(eg. data.csv)

1. Import the all required libraries.

→ P.T.O.

Pune District Education Association's
College Of Engineering
Manjari (Bk.), Hadapsar, Pune-412307.
Accredited by NAAC

2. Locate an open source data from the web, provide clear description of data & its source.
3. Load the dataset into pandas dataframe
4. Data preprocessing : check for missing values in description. Type of Variables, etc., check dimension of data frame.
5. Data formatting & data normalization : Summarize the type of variables by checking the data type of variables in the data set. If variables aren't in the correct data type, apply proper type conversion
6. Turn categorial variables into Quantitative variables in python. In addition to codes & o/p, explain every operation that you do in above steps & explain everything that you do to import / read / scrape the dataset.

Q1 → Explain Dataframes with suitable examples.

Data frames are data displayed in a format as a table. Data frames can have different types of data inside it.

Example :- Lets have - Data of training pulse duration

| Training | Pulse | duration |
|---|---|---|
| 1. Strength | 100 | 60 |
| 2. Stamina | 150 | 30 |
| 3 other | 120 | 45 |

→ P.T.O.

Pune District Education Association's
College Of Engineering
Manjari (Bk.), Hadapsar, Pune-412307.
Accredited by NAAC

Q2. What is the limitations of label encoding method?

→ Limitation of label encoding method :-

Label encoding converts the data in machine readable form, but it assigns a unique number to each class of priority issues in training of data sets. A label with a high value may be considered to have high priority than the label having lower value.

Q3. What is need of data normalization ?

→ The main objective of database normalization is to eliminate redundant data, minimize data modification errors & simplify the query process.

Q4. What are different techniques for handling the missing data?

→ Common Techniques :-

1. Mean or Median Imputation
2. Multivariate Imputation by chained equations
3. Random forest

Q5. What is meant by Data preprocessing ?

→ 1. Checking of missing values using pandas isnull() function
2. By using describe() function to get some initial statistics.

→ P.T.O.

Pune District Education Association's
College Of Engineering
Manjari (Bk.), Hadapsar, Pune-412307.
Accredited by NAAC

3. Provide variable description.
4. Type of variables.
5. Checking the dimension of data frame.

Q6. What is meant by Data Wrangling?
→ 1. Data exploration: The data is studied, analyzed & understood by visualizing representation of data.
2. Dealing with missing values.
3. Data reshaping: Data is manipulated according to the requirement, where new data can be added or pre-existing data can be modified.
4. Filtering data: Some data set exists with unwanted rows & columns which are required to be removed or filtered.

Q7. Importance of Data Wrangling.
→ · Improve data usability
· Converts data into compatible format.

Q8. What is meant by data wrangling process?
→ Cleaning, organising & enriching raw data so that it can be used for decision making process.

Q9. Uses of Pandas library.
→ Panda is open source library.

→ P.T.O

Pune District Education Association's
College Of Engineering
Manjari (Bk.), Hadapsar, Pune-412307.
Accredited by NAAC

Uses:
1. Data cleaning
2. Data fill
3. Data normalization
4. merges & joins
5. Data visualization
6. Statistical analysis
7. Data inspection
8. Loading & saving data

Q10. Uses of numpy library.
→ Numpy is a numerical python library.
Uses:
1. Working with arrays.
2. Working in domain of linear algebra.
3. Fourier transformation
4. Working with matrices. It is also open source library.
5. Working in vector - vector multiplication.

Q11. Uses of matplotlib library.
→ Matplotlib is used for data visualization & graphical plotting library (histogram, scatter plots, bar chart set)

Q12. Uses of seaborn library.
→ Seaborn library is used for making statistical graphics in python.

→P.T.O.

Pune District Education Association's
College Of Engineering
Manjari (Bk.), Hadapsar, Pune-412307.
Accredited by NAAC

Q13. Difference between matplotlib & seaborn.

→ • Matplotlib :

It is a python library used to plot various graphs with the help of additional libraries like numpy & Pandas.

Matplotlib creates simple graphs, including histograms, bargraphs, pie charts, scatter plots, lines & other visual representation of data.

Mainly used to plot 2D graphs of arrays.

It uses syntax that is relatively complicated & extensive.

eg: matplotlib.pyplot.bar(x-axis, y-axis) is syntax for bar graph.

• Seaborn :

It is also a python library that utilizes matplotlib, pandas & numpy to plot graphs.

It is a superset of matplotlib library.

It has relatively simple syntax.

eg: seaborn.barplot(x-axis, y-axis) syntax for bar graph.

Seaborn is more comfortable with panda data frames.

It prevents overlapping with the help of default themes.

Q14. How to install any library in python program.

→ pip install package_name.

eg: pip install seaborn.