DATA WRANGLING

Data Wrangling is a crucial topic for Data Science and Data Analysis. Pandas Framework of python is used for Data Wrangling.

Data wrangling involves processing the data in various formats like - merging, grouping, concatenating etc. for the purpose of analyzing or getting them ready to be used with another set of data.

Python has built-in features to apply these wrangling methods to various data set to achieve the analytical goal.

Data wrangling in Python deals with the below functionalities:

1. Data exploration: In this process, the data is studied, analyzed, and understood by visualizing representations of data.

2. Dealing with missing values: Most of the datasets 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 frequent value of the column, or simply by dropping the row having a NaN value.

3. Reshaping 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: Some times datasets are composed of unwanted rows or columns which are required to be removed or filtered.

Data exploration in Python

```
import pandas as pd
data={'Name':['Priya','Riya','Shivangi','Amit','Neha']
      'Age':[20,21,22,23,24],
      'Gender':['F','F','F','M','F'],
      'Marks':[90,76,'NaN',84,'NaN']}

df = pd.Dataframe(data)

df
```

Output

```
     Name      Age   Gender   Marks
0    Priya     20    F        90
1    Riya      21    F        76
2    Shivangi  22    F        NaN
3    Amit      23    M        84
4    Neha      24    F        NaN
```

Dealing with missing data:

```
people.dropna(subset="weight_kg")
people.fillna({"weight_kg":100})

df1 = df.copy()
df1.dropna(subset='Gender')
```

~~Output~~

```
     Name      Age   Gender   Marks
0    Jai       20    F        90.0
2    Riya      21    F        NaN
3    Priti     22    F        76.0
4    Shivangi  23    F        NaN
5    Neha      24    F        87.0
```

```
df1.fillna({'Gender':100})
```

```
     Name     | Age  | Gender | Marks
0    Jai        20     F        90.0
1    Priya      20     100      94.0
2    Riya       21     F        NaN
3    Priti      22     F        76.0
4    Shivangi   23     F        NaN
5    Neha       24     F        87.0
```

Filtering data

```
df= df[df['Marks']>80].copy()
df.drop('Age', axis=1, inplace =True)
df
```

Output

```
     Name       Gender   Marks
0    Priya        F       90
~~2  Shivangi     F       80.2~~
3    Amit         M       84
~~4  Neha         F       80.2~~
```