# □ Suraj Sawant TEB-38

## DSBDA Practical No A-1: Data Wrangling I

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 (i.e., URL of the web site).
3. Load the Dataset 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 descriptions. Types of variables etc. Check the dimensions 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 variables 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. In addition to the codes and outputs, explain every operation that you do in the above steps and explain everything that you do to import/read/scrape the data set.

```
import pandas as pd
```

```
pip install openpyxl
```

```
Collecting openpyxl
  Downloading openpyxl-3.1.5-py2.py3-none-any.whl.metadata (2.5 kB)
Collecting et-xmlfile (from openpyxl)
  Downloading et_xmlfile-2.0.0-py3-none-any.whl.metadata (2.7 kB)
Downloading openpyxl-3.1.5-py2.py3-none-any.whl (250 kB)
Downloading et_xmlfile-2.0.0-py3-none-any.whl (18 kB)
Installing collected packages: et-xmlfile, openpyxl
Successfully installed et-xmlfile-2.0.0 openpyxl-3.1.5
Note: you may need to restart the kernel to use updated packages.
```

```
df = pd.read_excel(r"C:\Users\Admin\Desktop\research_student.xlsx")
```

```
df.shape
```

```
(223, 24)
```

```
df.head(5)
```

| | Branch | Marks[10th] | Marks[12th] | Gender | Board[10th] | Board[12th] | Category | GPA 1 | Rank | Normalized Rank | ... | GPA 3 | GPA 4 | GPA 5 | GPA 6 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | NaN NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN | NaN | |
| 1 | CIVIL<br>5.41 | 77.57<br>6.25 | 64.6<br>6.13 | Male | BSEB Patna | | BSEB Patna | | OBC | 6.29 | 44718.0 | 15.970714 ... | | 5.94 | |
| 2 | CSE 86.40 | 71.8 | Male | CBSE | CBSE | GEN | 6.47 | 24222.0 | 8.650714 | ... | 5.88 | 5.53 | 6.44 | 6.19 | |
| 3 | CSE 88.14 | 78.0 | Male | ICSE | ICSE | GEN | 7.35 | 24723.0 | 8.829643 | ... | 6.54 | 6.41 | 6.50 | 6.69 | |
| 4 | CSE 65.40 | 59.8 | Female | CBSE | CBSE | ST | 6.41 | 232157.0 | 82.913214 | ... | 5.71 | 5.24 | 5.88 | 6.25 | |

```
df.tail(5)
```

| | Branch | Marks[10th] | Marks[12th] | Gender | Board[10th] | Board[12th] | Category | GPA 1 | Rank | Normalized Rank | ... | GPA 3 | GPA 4 | G |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 218 | PROD | 91.2 | 80.6 | Male | ICSE | CBSE | GEN | 74.70 | 39792.0 | 14.211429 ... | | 67.20 | 72.90 | 81.9 |
| 219 | PROD | 79.4 | 63.2 | Male | OBC | CENTRAL<br>BOARD OF<br>6.71<br>SECONDARY<br>EDUCATION | CENTRAL<br>BOARD OF<br>114306.0<br>SECONDARY<br>EDUCATION | 40.823571 ... | | 6.41 | 6.88 | 7.4 | | |
| 220 | PROD | 87.4 | 83.2 | Male | CBSE | CBSE | GEN | 7.18 | 40000.0 | 14.285714 ... | | 7.06 | 7.88 | 8.5 |
| 221 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | Na |

Na

```
df.describe()
```

| | Marks[10th] | Marks[12th] | GPA 1 | Rank | Normalized Rank | CGPA | Current Back | Ever Back | GPA 2 | GPA 3 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **count** | 220.000000 | 220.000000 | 220.000000 | 221.000000 | 220.000000 | 220.000000 | 220.000000 | 219.000000 | 220.000000 | 220.000000 | 220 |
| **mean** | 84.307455 | 77.783227 | 7.534409 | 42312.122172 | 14.725877 | 7.199468 | 0.077273 | 1.442922 | 7.266500 | 7.173500 | 7 |
| **std** | 8.519507 | 9.044172 | 4.602384 | 38503.691510 | 12.528090 | 0.700186 | 0.267633 | 2.657892 | 4.426861 | 4.134494 | 4 |
| **min** | 53.700000 | 56.800000 | 5.760000 | 11814.000000 | 4.219286 | 5.890000 | 0.000000 | 0.000000 | 5.760000 | 4.880000 | 4 |
| **25%** | 79.000000 | 71.550000 | 6.710000 | 23949.000000 | 8.544196 | 6.637500 | 0.000000 | 0.000000 | 6.417500 | 6.317500 | 6 |
| **50%** | 86.550000 | 79.100000 | 7.180000 | 30080.000000 | 10.694107 | 7.125000 | 0.000000 | 0.000000 | 6.880000 | 6.820000 | 6 |
| **75%** | 91.000000 | 85.250000 | 7.760000 | 41527.000000 | 14.718036 | 7.702500 | 0.000000 | 2.000000 | 7.470000 | 7.410000 | 7 |
| **max** | 96.600000 | 96.500000 | 74.700000 | 279839.000000 | 99.942500 | 9.010000 | 1.000000 | 13.000000 | 71.800000 | 67.200000 | 72 |

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
 RangeIndex: 223 entries, 0 to 222
 Data columns (total 24 columns):
  #   Column                 Non-Null Count  Dtype
 ---  ------                 --------------  -----
 0   Branch                 220 non-null    object
 1   Marks[10th]            220 non-null    float64
 2   Marks[12th]            220 non-null    float64
 3   Gender                 220 non-null    object
 4   Board[10th]            220 non-null    object
 5   Board[12th]            220 non-null    object
 6   Category               220 non-null    object
 7   GPA 1                  220 non-null    float64
 8   Rank                   221 non-null    float64
 9   Normalized Rank        220 non-null    float64
 10  CGPA                   220 non-null    float64
 11  Current Back           220 non-null    float64
 12  Ever Back              219 non-null    float64
 13  GPA 2                  220 non-null    float64
 14  GPA 3                  220 non-null    float64
 15  GPA 4                  220 non-null    float64
 16  GPA 5                  220 non-null    float64
 17  GPA 6                  220 non-null    float64
 18  Olympiads Qualified    220 non-null    float64
 19  Technical Projects     220 non-null    float64
 20  Tech Quiz              220 non-null    float64
 21  Engg. Coaching         220 non-null    float64
 22  NTSE Scholarships      220 non-null    float64 23  Miscellany Tech Events  220
     non-null    float64 dtypes: float64(19), object(5) memory usage: 41.9+ KB
```

```
df.isnull()
```

| | Branch | Marks[10th] | Marks[12th] | Gender | Board[10th] | Board[12th] | Category | GPA 1 | Rank | Normalized Rank | ... | GPA 3 | GPA 4 | GPA 5 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | True | True | True | True | True | True | True | True | True | True | ... | True | True | True |
| **1** | False<br>F | False | False | False | False | False | False | False | False | False | ... | False | False | False |
| **2** | False<br>F | False | False | False | False | False | False | False | False | False | ... | False | False | False |
| **3** | False<br>F | False | False | False | False | False | False | False | False | False | ... | False | False | False |
| **4** | False<br>F | False | False | False | False | False | False | False | False | False | ... | False | False | False |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **218** | False<br>F | False | False | False | False | False | False | False | False | False | ... | False | False | False |
| **219** | False<br>F | False | False | False | False | False | False | False | False | False | ... | False | False | False |
| **220** | False<br>F | False | False | False | False | False | False | False | False | False | ... | False | False | False |
| **221** | True | True | True | True | True | True | True | True | True | True | ... | True | True | True |
| **222** | True | True | True | True | True | True | True | True | False | True | ... | True | True | True |

```
df.isnull().sum()
```

```
Branch                   3
Marks[10th]              3
Marks[12th]              3
Gender                   3
Board[10th]              3
Board[12th]              3
Category                 3
GPA 1                    3
Rank                     2
Normalized Rank          3
CGPA                     3
Current Back             3
Ever Back                4
GPA 2                    3
GPA 3                    3
GPA 4                    3
GPA 5                    3
GPA 6                    3
Olympiads Qualified      3
Technical Projects       3
Tech Quiz                3
Engg. Coaching           3
NTSE Scholarships        3
Miscellany Tech Events   3


dtype: int64
```

```
df=df.drop([0,221,222])
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 220 entries, 1 to 220
Data columns (total 24 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
0   Branch                  220 non-null    object
1   Marks[10th]             220 non-null    float64
2   Marks[12th]             220 non-null    float64
3   Gender                  220 non-null    object
4   Board[10th]             220 non-null    object
5   Board[12th]             220 non-null    object
6   Category                220 non-null    object
7   GPA 1                   220 non-null    float64
8   Rank                    220 non-null    float64
9   Normalized Rank         220 non-null    float64
10  CGPA                    220 non-null    float64
11  Current Back            220 non-null    float64
12  Ever Back               219 non-null    float64
13  GPA 2                   220 non-null    float64
14  GPA 3                   220 non-null    float64
15  GPA 4                   220 non-null    float64
16  GPA 5                   220 non-null    float64
17  GPA 6                   220 non-null    float64
18  Olympiads Qualified     220 non-null    float64
19  Technical Projects      220 non-null    float64
20  Tech Quiz               220 non-null    float64
21  Engg. Coaching          220 non-null    float64
22  NTSE Scholarships       220 non-null    float64 23  Miscellany Tech Events  220
    non-null    float64 dtypes: float64(19), object(5) memory usage: 43.0+ KB
```

```
df.shape
```

```
(220, 24)
```

```
df.dtypes
```

```
Branch                    object
Marks[10th]              float64
Marks[12th]              float64
Gender                    object
Board[10th]               object
Board[12th]               object
Category                  object
GPA 1                    float64
Rank                     float64
Normalized Rank          float64
CGPA                     float64
```

```
Current Back              float64
Ever Back                 float64
GPA 2                     float64
GPA 3                     float64
GPA 4                     float64
GPA 5                     float64
GPA 6                     float64
Olympiads Qualified       float64
Technical Projects        float64
Tech Quiz                 float64
Engg. Coaching            float64
NTSE Scholarships         float64


Miscellany Tech Events    float64


dtype: object
```

```
df=df.fillna(0)
```

```
df.isnull().sum()
```

```
Branch                   0
Marks[10th]              0
Marks[12th]              0
Gender                   0
Board[10th]              0
Board[12th]              0
Category                 0
GPA 1                    0
Rank                     0
Normalized Rank          0
CGPA                     0
Current Back             0
Ever Back                0
GPA 2                    0
GPA 3                    0
GPA 4                    0
GPA 5                    0
GPA 6                    0
Olympiads Qualified      0
Technical Projects       0
Tech Quiz                0
Engg. Coaching           0
NTSE Scholarships        0
Miscellany Tech Events   0
dtype: int64
```

```
df.columns
```

```
Index(['Branch', 'Marks[10th]', 'Marks[12th]', 'Gender', 'Board[10th]',
       'Board[12th]', 'Category', 'GPA 1', 'Rank', 'Normalized Rank', 'CGPA',
       'Current Back', 'Ever Back', 'GPA 2', 'GPA 3', 'GPA 4', 'GPA 5',
       'GPA 6', 'Olympiads Qualified', 'Technical Projects', 'Tech Quiz',
       'Engg. Coaching', 'NTSE Scholarships', 'Miscellany Tech Events'],
      dtype='object')
```

```
imp_columns=['Marks[10th]', 'Marks[12th]', 'GPA 1', 'Rank', 'Normalized Rank', 'CGPA',
'Current Back', 'Ever Back', 'GPA 2', 'GPA 3', 'GPA 4', 'GPA 5',
'GPA 6', 'Olympiads Qualified', 'Technical Projects', 'Tech Quiz',
'Engg. Coaching', 'NTSE Scholarships', 'Miscellany Tech Events']
```

```
df[imp_columns]
```

| | Marks[10th] | Marks[12th] | GPA 1 | Rank | Normalized Rank | CGPA | Current Back | Ever Back | GPA 2 | GPA 3 | GPA 4 | GPA 5 | GPA 6 | Olympiads Qualified | Technical Projects |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 77.57<br>1.0 | 64.6<br>4.0 | 6.29 | 44718.0 | 15.970714 | 6.02 | 1.0 | 4.0 | | 6.12 | 5.94 | | 5.41 | 6.25 | 6.13 |
| 2 | 86.40<br>2.0 | 71.8<br>2.0 | 6.47 | 24222.0 | 8.650714 | 6.10 | 1.0 | 7.0 | | 6.12 | 5.88 | | 5.53 | 6.44 | 6.19 |
| 3 | 88.14<br>1.0 | 78.0<br>1.0 | 7.35 | 24723.0 | 8.829643 | 6.65 | 1.0 | 1.0 | | 6.35 | 6.54 | | 6.41 | 6.50 | 6.69 |
| 4 | 65.40<br>2.0 | 59.8<br>0.0 | 6.41 | 232157.0 | 82.913214 | 6.09 | 1.0 | 11.0 | | 6.00 | 5.71 | | 5.24 | 5.88 | 6.25 |
| 5 | 81.00<br>2.0 | 74.0<br>0.0 | 6.80 | 23252.0 | 8.304286 | 6.13 | 1.0 | 0.0 | | 6.06 | 5.88 | | 6.00 | 5.93 | 5.44 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **216** | 78.80<br>3.0 | 66.0<br>3.0 | 6.35 | 100000.0 | 35.714286 | 6.44 | 0.0 | 3.0 | 6.35 | 6.06 | 6.00 | 6.94 | 7.00 |
| **217** | 91.00<br>0.0 | 81.0<br>4.0 | 7.00 | 36706.0 | 13.109286 | 7.07 | 0.0 | 0.0 | 6.65 | 6.47 | 6.71 | 7.94 | 7.75 |
| **218** | 91.20<br>4.0 | 80.6<br>1.0 | 74.70 | 39792.0 | 14.211429 | 7.36 | 0.0 | 2.0 | 71.80 | 67.20 | 72.90 | 81.90 | 7.69 |
| **219** | 79.40<br>1.0 | 63.2<br>1.0 | 6.71 | 114306.0 | 40.823571 | 6.89 | 0.0 | 0.0 | 6.12 | 6.41 | 6.88 | 7.44 | 7.69 |
| **220** | 87.40<br>2.0 | 83.2<br>1.0 | 7.18 | 40000.0 | 14.285714 | 7.69 | 0.0 | 0.0 | 6.88 | 7.06 | 7.88 | 8.56 | 8.69 |

```
from sklearn.preprocessing import StandardScaler
scaler=StandardScaler()
```

```
df[imp_columns]=scaler.fit_transform(df[imp_columns])
```

```
df[imp_columns]
```

| | Marks[10th] | Marks[12th] | GPA 1 | Rank | Normalized Rank | CGPA | Current Back | Ever Back | GPA 2 | GPA 3 | GPA 4 | GPA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | -0.792630 | -1.460973 | -0.271000 | 0.099590 | 0.099590 | -1.688349 | 3.455601 | 0.968301 | -0.259578 | -0.299024 | -0.417347 | -0.3219 |
| **2** | 0.246178 | -0.663065 | -0.231801 | -0.486029 | -0.486029 | -1.573833 | 3.455601 | 2.101418 | -0.259578 | -0.313569 | -0.390673 | -0.2845 |
| **3** | 0.450881 | 0.024023 | -0.040160 | -0.471714 | -0.471714 | -0.786536 | 3.455601 | -0.164817 | -0.207504 | -0.153573 | -0.195065 | -0.2728 |
| **4** | -2.224374 | -1.992912 | -0.244867 | 5.455168 | 5.455168 | -1.588147 | 3.455601 | 3.612242 | -0.286747 | -0.354780 | -0.455134 | -0.3945 |
| **5** | -0.389107 | -0.419259 | -0.159935 | -0.513744 | -0.513744 | -1.530889 | 3.455601 | -0.542523 | -0.273162 | -0.313569 | -0.286200 | -0.3847 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **216** | -0.647926 | -1.305824 | -0.257934 | 1.679129 | 1.679129 | -1.087140 | -0.289385 | 0.590595 | -0.207504 | -0.269934 | -0.286200 | -0.1864 |
| **217** | 0.787347 | 0.356485 | -0.116381 | -0.129332 | -0.129332 | -0.185327 | -0.289385 | -0.542523 | -0.139581 | -0.170542 | -0.128380 | 0.0099 |
| **218** | 0.810876 | 0.312156 | 14.626933 | -0.041157 | -0.041157 | 0.229793 | -0.289385 | 0.212889 | 14.610954 | 14.551571 | 14.584464 | 14.5331 |
| **219** | -0.577339 | -1.616122 | -0.179535 | 2.087885 | 2.087885 | -0.442988 | -0.289385 | -0.542523 | -0.259578 | -0.185087 | -0.090592 | -0.0882 |
| **220** | 0.363824 | 0.600290 | -0.077181 | -0.035214 | -0.035214 | 0.702171 | -0.289385 | -0.542523 | -0.087507 | -0.027515 | 0.131690 | 0.1316 |

```
from sklearn.preprocessing import LabelEncoder encoding_list =
['Branch','Gender','Board[10th]','Board[12th]','Category']
df[encoding_list] = df[encoding_list].apply(LabelEncoder().fit_transform)
```

```
df[encoding_list]
```

| | Branch | Gender | Board[10th] | Board[12th] | Category |
|---|---|---|---|---|---|
| **1** | 0 | 1 | 11 | 11 | 1 |
| **2** | 1 | 1 | 18 | 19 | 0 |
| **3** | 1 | 1 | 25 | 29 | 0 |
| **4** | 1 | 0 | 18 | 19 | 3 |
| **5** | 1 | 1 | 18 | 19 | 0 |
| **...** | ... | ... | ... | ... | ... |
| **216** | 6 | 1 | 18 | 19 | 3 |
| **217** | 6 | 1 | 18 | 19 | 0 |
| **218** | 6 | 1 | 25 | 19 | 0 |
| **219** | 6 | 1 | 20 | 22 | 1 |
| **220** | 6 | 1 | 18 | 19 | 0 |

220 rows × 5 columns

```
df. info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 220 entries, 1 to 220
Data columns (total 24 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   Branch                  220 non-null    int64  
 1   Marks[10th]             220 non-null    float64
 2   Marks[12th]             220 non-null    float64
 3   Gender                  220 non-null    int64  
 4   Board[10th]             220 non-null    int64  
 5   Board[12th]             220 non-null    int64  
 6   Category                220 non-null    int64  
 7   GPA 1                   220 non-null    float64
 8   Rank                    220 non-null    float64
 9   Normalized Rank         220 non-null    float64
 10  CGPA                    220 non-null    float64
 11  Current Back            220 non-null    float64
 12  Ever Back               220 non-null    float64
 13  GPA 2                   220 non-null    float64
 14  GPA 3                   220 non-null    float64
 15  GPA 4                   220 non-null    float64
 16  GPA 5                   220 non-null    float64
 17  GPA 6                   220 non-null    float64
 18  Olympiads Qualified     220 non-null    float64
 19  Technical Projects      220 non-null    float64
 20  Tech Quiz               220 non-null    float64
 21  Engg. Coaching          220 non-null    float64
 22  NTSE Scholarships       220 non-null    float64
 23  Miscellany Tech Events  220 non-null    float64
dtypes: float64(19), int64(5)
memory usage: 43.0 KB
```