

Practical No. 01

Data Wrangling 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 `isnull()`, `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.

```
In [ ]: import pandas as pd
```

```
In [ ]: df = pd.read_csv('/home/kartik/Documents/Python Notebooks/StudentsPerformance.csv')
```

```
In [ ]: df
```

```
Out[ ]:
```

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group B	bachelor's degree	standard	none	72	72	74
1	female	group C	some college	standard	completed	69	90	88
2	female	group B	master's degree	standard	none	90	95	93
3	male	group A	associate's degree	free/reduced	none	47	57	44
4	male	group C	some college	standard	none	76	78	75
...
995	female	group E	master's degree	standard	completed	88	99	95
996	male	group C	high school	free/reduced	none	62	55	55
997	female	group C	high school	free/reduced	completed	59	71	65
998	female	group D	some college	standard	completed	68	78	77
999	female	group D	some college	free/reduced	none	77	86	86

1000 rows × 8 columns

```
In [ ]: df.isnull()
```

Out[]:

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False
...
995	False	False	False	False	False	False	False	False
996	False	False	False	False	False	False	False	False
997	False	False	False	False	False	False	False	False
998	False	False	False	False	False	False	False	False
999	False	False	False	False	False	False	False	False

1000 rows × 8 columns

In []: df.describe()

Out[]: math score reading score writing score

count	1000.000000	1000.000000	1000.000000
mean	66.08900	69.169000	68.054000
std	15.16308	14.600192	15.195657
min	0.00000	17.000000	10.000000
25%	57.00000	59.000000	57.750000
50%	66.00000	70.000000	69.000000
75%	77.00000	79.000000	79.000000
max	100.000000	100.000000	100.000000

In []: df.isnull().sum()

Out[]:

gender	0
race/ethnicity	0
parental level of education	0
lunch	0
test preparation course	0
math score	0
reading score	0
writing score	0
dtype: int64	

In []: df.notnull().sum()

```
Out[ ]: gender           1000  
race/ethnicity      1000  
parental level of education 1000  
lunch              1000  
test preparation course 1000  
math score          1000  
reading score       1000  
writing score       1000  
dtype: int64
```

```
In [ ]: df.notnull()
```

```
Out[ ]:
```

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	True	True	True	True	True	True	True	True
1	True	True	True	True	True	True	True	True
2	True	True	True	True	True	True	True	True
3	True	True	True	True	True	True	True	True
4	True	True	True	True	True	True	True	True
...
995	True	True	True	True	True	True	True	True
996	True	True	True	True	True	True	True	True
997	True	True	True	True	True	True	True	True
998	True	True	True	True	True	True	True	True
999	True	True	True	True	True	True	True	True

1000 rows × 8 columns

```
In [ ]: df.size
```

```
Out[ ]: 8000
```

```
In [ ]: df.ndim
```

```
Out[ ]: 2
```

```
In [ ]: df.shape
```

```
Out[ ]: (1000, 8)
```

```
In [ ]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 8 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   gender            1000 non-null    object  
 1   race/ethnicity     1000 non-null    object  
 2   parental level of education 1000 non-null    object  
 3   lunch              1000 non-null    object  
 4   test preparation course 1000 non-null    object  
 5   math score          1000 non-null    int64  
 6   reading score       1000 non-null    int64  
 7   writing score        1000 non-null    int64  
dtypes: int64(3), object(5)
memory usage: 62.6+ KB
```

```
In [ ]: df['writing score'].astype(int)
```

```
Out[ ]: 0      74
1      88
2      93
3      44
4      75
..
995    95
996    55
997    65
998    77
999    86
Name: writing score, Length: 1000, dtype: int64
```

```
In [ ]: df.dropna()
```

```
Out[ ]:
```

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group B	bachelor's degree	standard	none	72	72	74
1	female	group C	some college	standard	completed	69	90	88
2	female	group B	master's degree	standard	none	90	95	93
3	male	group A	associate's degree	free/reduced	none	47	57	44
4	male	group C	some college	standard	none	76	78	75
...
995	female	group E	master's degree	standard	completed	88	99	95
996	male	group C	high school	free/reduced	none	62	55	55
997	female	group C	high school	free/reduced	completed	59	71	65
998	female	group D	some college	standard	completed	68	78	77
999	female	group D	some college	free/reduced	none	77	86	86

1000 rows × 8 columns

```
In [ ]: df['writing score'] = df['writing score'].astype(int)
```

```
In [ ]: df
```

```
Out[ ]:
```

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group B	bachelor's degree	standard	none	72	72	74
1	female	group C	some college	standard	completed	69	90	88
2	female	group B	master's degree	standard	none	90	95	93
3	male	group A	associate's degree	free/reduced	none	47	57	44
4	male	group C	some college	standard	none	76	78	75
...
995	female	group E	master's degree	standard	completed	88	99	95
996	male	group C	high school	free/reduced	none	62	55	55
997	female	group C	high school	free/reduced	completed	59	71	65
998	female	group D	some college	standard	completed	68	78	77
999	female	group D	some college	free/reduced	none	77	86	86

1000 rows × 8 columns

```
In [ ]: df["gender"] = df["gender"].replace({"female":0,"male":1})
```

```
In [ ]: df
```

Out[]:

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	0	group B	bachelor's degree	standard	none	72	72	74
1	0	group C	some college	standard	completed	69	90	88
2	0	group B	master's degree	standard	none	90	95	93
3	1	group A	associate's degree	free/reduced	none	47	57	44
4	1	group C	some college	standard	none	76	78	75
...
995	0	group E	master's degree	standard	completed	88	99	95
996	1	group C	high school	free/reduced	none	62	55	55
997	0	group C	high school	free/reduced	completed	59	71	65
998	0	group D	some college	standard	completed	68	78	77
999	0	group D	some college	free/reduced	none	77	86	86

1000 rows × 8 columns