Name: Ujwal Sahu

Branch: B.Voc(AIDS)

Div: B

Experiment No.4

Title: Demonstrate preparing data for Modelling: Preparing Rows and Columns in Machine Learning**.**

Tools: Anaconda (Jupyter Notebook)

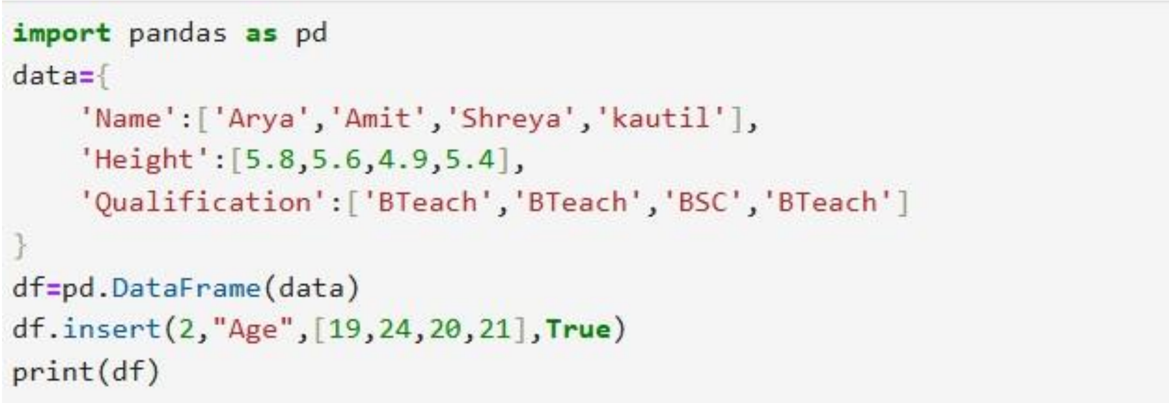
Theory: A Data frame is a two-dimensional data structure, i.e., data is aligned in a tabular fashion in rows and columns. We can perform basic operations on rows/columns like selecting, deleting, adding, and renaming.

- Rows (Instances/Samples/Observations): Each row represents a single data point or example. Think of it as one individual, one transaction, one image, etc. In a dataset of houses, each row would represent a different house.
- Columns (Features/Attributes/Variables): Each column represents a specific characteristic or property of the data. In the house dataset, columns might include square footage, number of bedrooms, location, price, etc.

1] Adding new column to existing DataFrame.

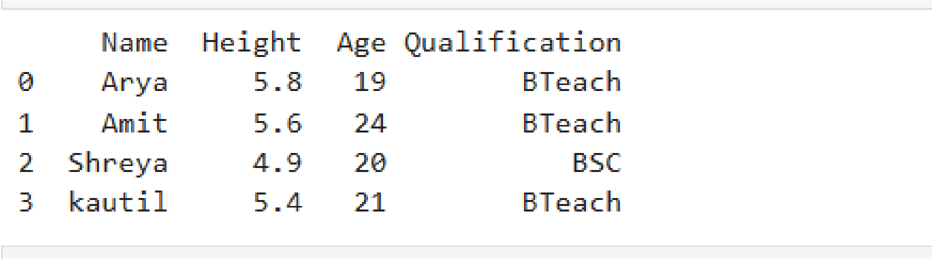
Code:

```
import pandas as pd
data={
   'Name':['Arya','Amit','Shreya','kautil'],
   'Height':[5.8,5.6,4.9,5.4],
   'Qualification':['BTeach','BTeach','BSC','BTeach']
}
df=pd.DataFrame(data)
df.insert(2,"Age",[19,24,20,21],True)
print(df)
```



```
import pandas as pd
data={
    'Name':['Arya','Amit','Shreya','kautil'],
    'Height':[5.8,5.6,4.9,5.4],
    'Qualification':['BTeach','BTeach','BSC','BTeach']
}
df=pd.DataFrame(data)
df.insert(2,"Age",[19,24,20,21],True)
print(df)
```

Output:

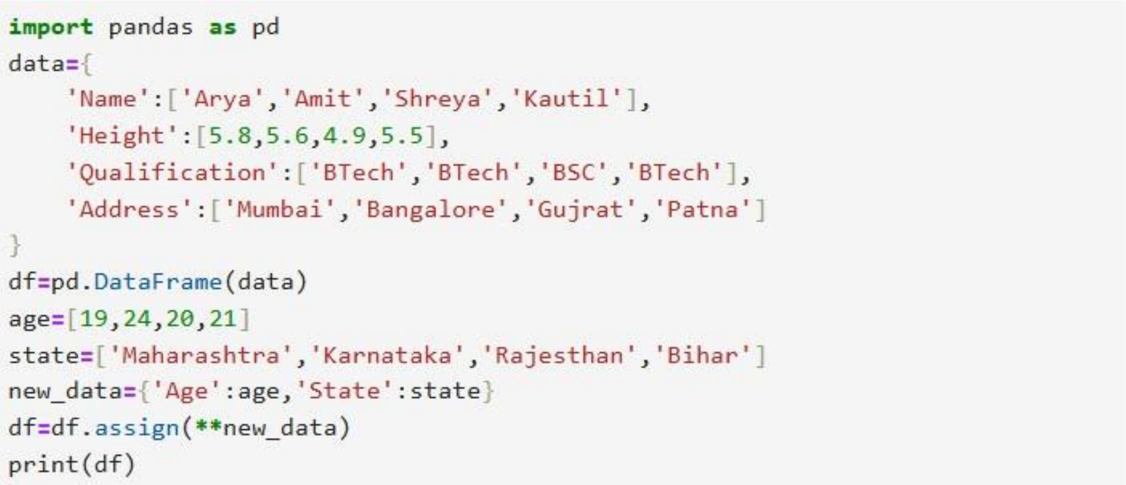


```
     Name  Height  Age Qualification
0    Arya     5.8   19        BTeach
1    Amit     5.6   24        BTeach
2  Shreya     4.9   20           BSC
3  kautil     5.4   21        BTeach
```

2] Adding more than one column in exiting dataframe .

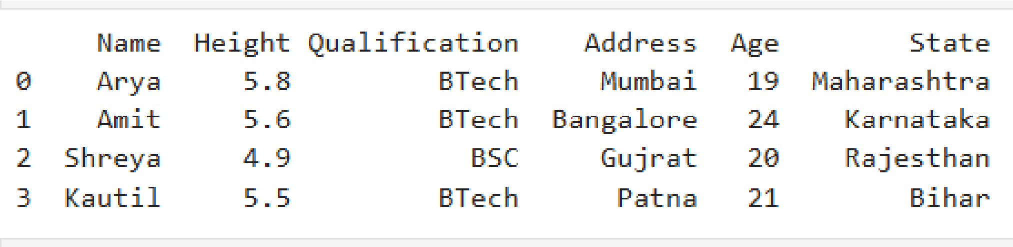
Code:

```
import pandas as pd
data={
   'Name':['Arya','Amit','Shreya','Kautil'],
   'Height':[5.8,5.6,4.9,5.5],
   'Qualification':['BTech','BTech','BSC','BTech'],
   'Address':['Mumbai','Bangalore','Gujrat','Patna']
}
df=pd.DataFrame(data)
age=[19,24,20,21]
state=['Maharashtra','Karnataka','Rajesthan','Bihar']
new_data={'Age':age,'State':state}
df=df.assign(**new_data)
print(df)
```



```
import pandas as pd
data={
    'Name':['Arya','Amit','Shreya','Kautil'],
    'Height':[5.8,5.6,4.9,5.5],
    'Qualification':['BTech','BTech','BSC','BTech'],
    'Address':['Mumbai','Bangalore','Gujrat','Patna']
}
df=pd.DataFrame(data)
age=[19,24,20,21]
state=['Maharashtra','Karnataka','Rajesthan','Bihar']
new_data={'Age':age,'State':state}
df=df.assign(**new_data)
print(df)
```

Output:



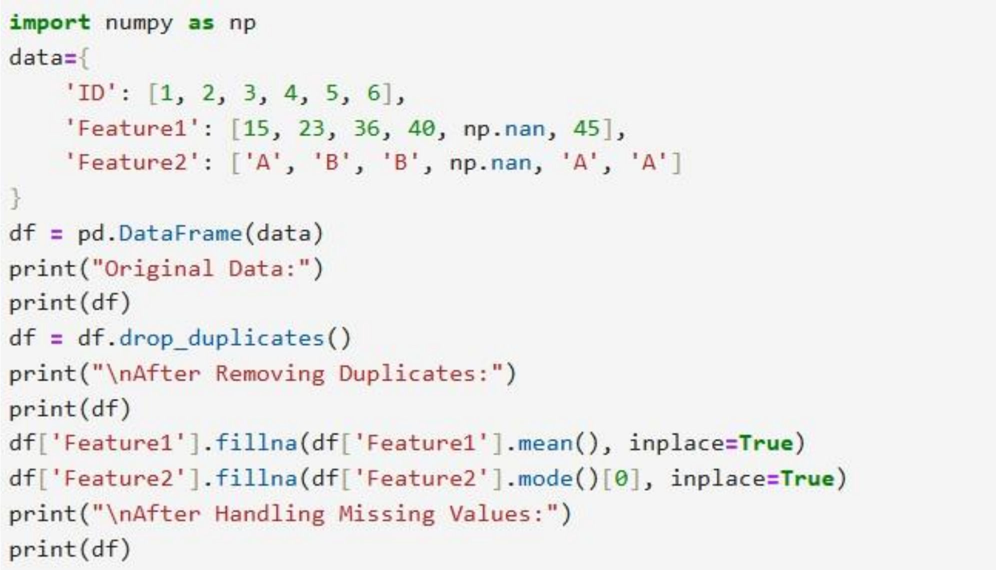
```
     Name  Height Qualification    Address  Age        State
0    Arya     5.8         BTech     Mumbai   19  Maharashtra
1    Amit     5.6         BTech  Bangalore   24    Karnataka
2  Shreya     4.9           BSC     Gujrat   20    Rajesthan
3  Kautil     5.5         BTech      Patna   21        Bihar
```

3] Removing duplicates and Handling missing values.

```
import numpy as np
data={
   'ID': [1, 2, 3, 4, 5, 6],
   'Feature1': [15, 23, 36, 40, np.nan, 45],
   'Feature2': ['A', 'B', 'B', np.nan, 'A', 'A']
}
df = pd.DataFrame(data)
print("Original Data:")
print(df)
df = df.drop_duplicates()
print("\nAfter Removing Duplicates:")
print(df)
df['Feature1'].fillna(df['Feature1'].mean(), inplace=True)
df['Feature2'].fillna(df['Feature2'].mode()[0], inplace=True)
```

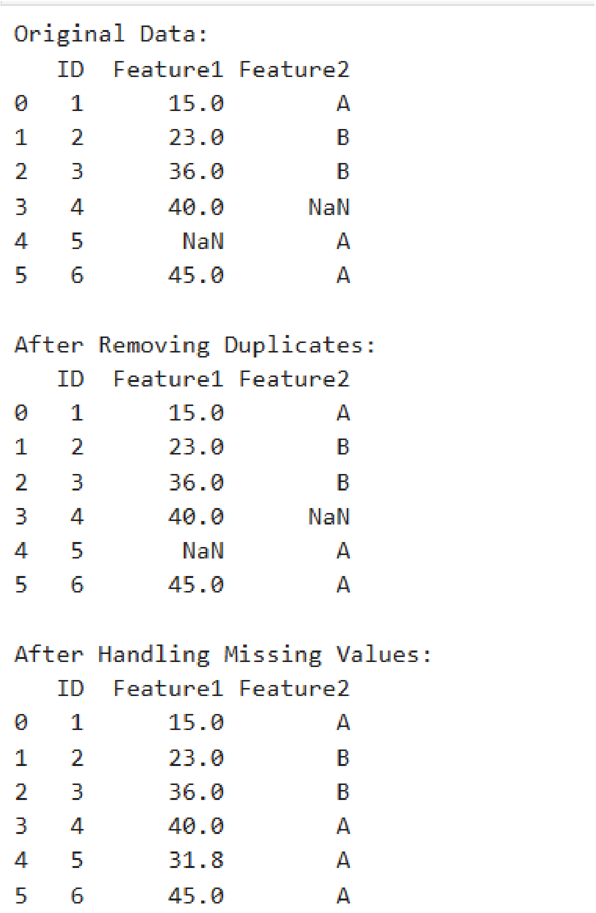
print("\nAfter Handling Missing Values:")

print(df)



```
import numpy as np
data={
    'ID': [1, 2, 3, 4, 5, 6],
    'Feature1': [15, 23, 36, 40, np.nan, 45],
    'Feature2': ['A', 'B', 'B', np.nan, 'A', 'A']
}
df = pd.DataFrame(data)
print("Original Data:")
print(df)
df = df.drop_duplicates()
print("\nAfter Removing Duplicates:")
print(df)
df['Feature1'].fillna(df['Feature1'].mean(), inplace=True)
df['Feature2'].fillna(df['Feature2'].mode()[0], inplace=True)
print("\nAfter Handling Missing Values:")
print(df)
```

Output:

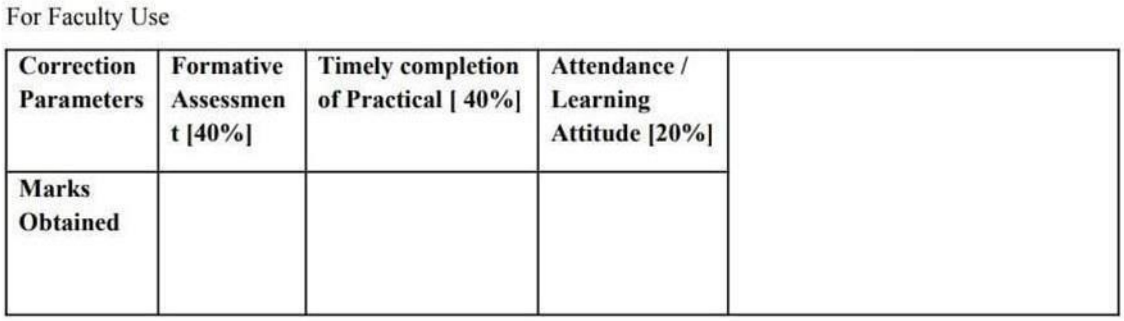


```
Original Data:
   ID  Feature1 Feature2
0   1      15.0        A
1   2      23.0        B
2   3      36.0        B
3   4      40.0      NaN
4   5       NaN        A
5   6      45.0        A

After Removing Duplicates:
   ID  Feature1 Feature2
0   1      15.0        A
1   2      23.0        B
2   3      36.0        B
3   4      40.0      NaN
4   5       NaN        A
5   6      45.0        A

After Handling Missing Values:
   ID  Feature1 Feature2
0   1      15.0        A
1   2      23.0        B
2   3      36.0        B
3   4      40.0        A
4   5      31.8        A
5   6      45.0        A
```

Conclusion: By meticulously preparing the rows (samples) and columns (features) of our data, we ensure that our machine learning models are trained on high-quality, relevant information. This leads to more accurate, robust, and reliable predictions, ultimately maximizing the value derived from our machine learning efforts. Effective data preparation is often the difference between a mediocre model and a highly successful one.



For Faculty Use

Correction Parameters	Formative Assessment [40%]	Timely completion of Practical [40%]	Attendance / Learning Attitude [20%]	
Marks Obtained				