

**Ex.No-5****Data Cleaning & Preparation (Excel file)****Aim**

To do the Data Cleaning and Preparation using Excel data.

**Description:**

Read the Excel file, do the data cleaning process and write the updated data set into excel file

1. Remove the white space using str.strip function
2. Fill the forward values for NaN using fillna(pad) method
3. Drop one particular column using drop function
4. Drop NaN rows using dropna function
5. Replace the values(s) using replace function
6. Extract the particular record based on the isin() function condition

**PROGRAM:**

```
import pandas as pd

pd.set_option('display.max_columns', 10)

print("Original data set from Excel file:\n")

df=pd.read_excel("d:\sample1.xlsx")

print(df)


f=df['NAME'].str.strip()

f.to_excel("D:\sample22.xlsx")


print("\nReplace value with Forward:\n")

k=df.fillna(method='pad')

print(k)

print("\nThe above updated data set will be stored in sample2.xlsx file....\n")

k.to_excel("d:\sample2.xlsx")


print("\nDrop one particular column and its values:\n")

k.drop(['TOTAL'],axis=1, inplace=True)

print(k)

print("\nThe above updated data set will be stored in sample3.xlsx file....\n")
```

```
k.to_excel("d:\sample3.xlsx")

print("\nDrop NaN rows:\n")
df=pd.read_excel("d:\sample4.xlsx")
print(df)
x=df.dropna()
print(x)
print("\nThe above updated data set will be stored in sample5.xlsx file....\n")
x.to_excel("d:\sample5.xlsx")

print("\nReplace values:\n")
n=pd.read_excel("d:\sample3.xlsx")
print("Original data set:\n");
print(n)
y=n.replace({49:50})
print("\nUpdated dataset with replaced values: {49:50}\n")
print(y)
print("\nThe above updated data set will be stored in sample6.xlsx file....\n")
y.to_excel("d:\sample6.xlsx")

print("Original data set from Excel file:\n")
df=pd.read_excel("d:\sample1.xlsx")
print(df)

print("\nExtract the particular record based on the isin() function condition:\n")
new=df['ENGLISH'].isin([49])
print(df[new])
```

**OUTPUT:**

Original data set from Excel file:

|   | ROLL NO | NAME    | ENGLISH | TAMIL | MATHS | SCIENCE | SOCIAL | TOTAL |
|---|---------|---------|---------|-------|-------|---------|--------|-------|
| 0 | 101     | DEEPA   | 50.0    | 67    | 50    | 67.0    | 50     | 284   |
| 1 | 102     | DINESH  | 56.0    | 89    | 56    | 89.0    | 56     | 346   |
| 2 | 103     | KAVIYA  | 80.0    | 80    | 80    | 80.0    | 80     | 400   |
| 3 | 104     | RACHEAL | 89.0    | 87    | 89    | 87.0    | 89     | 441   |
| 4 | 105     | RAJAN   | NaN     | 98    | 90    | 98.0    | 90     | 466   |
| 5 | 106     | RAMYA   | 67.0    | 76    | 67    | 76.0    | 67     | 353   |
| 6 | 107     | ROHAN   | 56.0    | 67    | 56    | 67.0    | 56     | 302   |
| 7 | 108     | ROHINI  | 57.0    | 65    | 57    | 65.0    | 57     | 301   |
| 8 | 109     | SANDHYA | 58.0    | 56    | 58    | 56.0    | 58     | 286   |
| 9 | 110     | SARANYA | 49.0    | 45    | 49    | NaN     | 49     | 237   |

| ROLL NO | NAME    | ENGLISH | TAMIL | MATHS | SCIENCE | SOCIAL | TOTAL |
|---------|---------|---------|-------|-------|---------|--------|-------|
| 101     | DEEPA   | 50      | 67    | 50    | 67      | 50     | 284   |
| 102     | DINESH  | 56      | 89    | 56    | 89      | 56     | 346   |
| 103     | KAVIYA  | 80      | 80    | 80    | 80      | 80     | 400   |
| 104     | RACHEAL | 89      | 87    | 89    | 87      | 89     | 441   |
| 105     | RAJAN   |         | 98    | 90    | 98      | 90     | 466   |
| 106     | RAMYA   | 67      | 76    | 67    | 76      | 67     | 353   |
| 107     | ROHAN   | 56      | 67    | 56    | 67      | 56     | 302   |
| 108     | ROHINI  | 57      | 65    | 57    | 65      | 57     | 301   |
| 109     | SANDHYA | 58      | 56    | 58    | 56      | 58     | 286   |
| 110     | SARANYA | 49      | 45    | 49    |         | 49     | 237   |

|   | NAME    |
|---|---------|
| 0 | DEEPA   |
| 1 | DINESH  |
| 2 | KAVIYA  |
| 3 | RACHEAL |
| 4 | RAJAN   |
| 5 | RAMYA   |
| 6 | ROHAN   |
| 7 | ROHINI  |
| 8 | SANDHYA |
| 9 | SARANYA |

Replace value with Forward:

|   | ROLL NO | NAME    | ENGLISH | TAMIL | MATHS | SCIENCE | SOCIAL | TOTAL |
|---|---------|---------|---------|-------|-------|---------|--------|-------|
| 0 | 101     | DEEPA   | 50.0    | 67    | 50    | 67.0    | 50     | 284   |
| 1 | 102     | DINESH  | 56.0    | 89    | 56    | 89.0    | 56     | 346   |
| 2 | 103     | KAVIYA  | 80.0    | 80    | 80    | 80.0    | 80     | 400   |
| 3 | 104     | RACHEAL | 89.0    | 87    | 89    | 87.0    | 89     | 441   |
| 4 | 105     | RAJAN   | 89.0    | 98    | 90    | 98.0    | 90     | 466   |
| 5 | 106     | RAMYA   | 67.0    | 76    | 67    | 76.0    | 67     | 353   |
| 6 | 107     | ROHAN   | 56.0    | 67    | 56    | 67.0    | 56     | 302   |
| 7 | 108     | ROHINI  | 57.0    | 65    | 57    | 65.0    | 57     | 301   |
| 8 | 109     | SANDHYA | 58.0    | 56    | 58    | 56.0    | 58     | 286   |
| 9 | 110     | SARANYA | 49.0    | 45    | 49    | 56.0    | 49     | 237   |

The above updated data set will be stored in sample2.xlsx file....

|   | ROLL NO | NAME    | ENGLISH | TAMIL | MATHS | SCIENCE | SOCIAL | TOTAL |
|---|---------|---------|---------|-------|-------|---------|--------|-------|
| 0 | 101     | DEEPA   | 50      | 67    | 50    | 67      | 50     | 284   |
| 1 | 102     | DINESH  | 56      | 89    | 56    | 89      | 56     | 346   |
| 2 | 103     | KAVIYA  | 80      | 80    | 80    | 80      | 80     | 400   |
| 3 | 104     | RACHEAL | 89      | 87    | 89    | 87      | 89     | 441   |
| 4 | 105     | RAJAN   | 89      | 98    | 90    | 98      | 90     | 466   |
| 5 | 106     | RAMYA   | 67      | 76    | 67    | 76      | 67     | 353   |
| 6 | 107     | ROHAN   | 56      | 67    | 56    | 67      | 56     | 302   |
| 7 | 108     | ROHINI  | 57      | 65    | 57    | 65      | 57     | 301   |
| 8 | 109     | SANDHYA | 58      | 56    | 58    | 56      | 58     | 286   |
| 9 | 110     | SARANYA | 49      | 45    | 49    | 56      | 49     | 237   |

Drop one particular column and its values:

|   | ROLL NO | NAME    | ENGLISH | TAMIL | MATHS | SCIENCE | SOCIAL |
|---|---------|---------|---------|-------|-------|---------|--------|
| 0 | 101     | DEEPA   | 50.0    | 67    | 50    | 67.0    | 50     |
| 1 | 102     | DINESH  | 56.0    | 89    | 56    | 89.0    | 56     |
| 2 | 103     | KAVIYA  | 80.0    | 80    | 80    | 80.0    | 80     |
| 3 | 104     | RACHEAL | 89.0    | 87    | 89    | 87.0    | 89     |

```

4 105    RAJAN  89.0  98  90  98.0  90
5 106    RAMYA  67.0  76  67  76.0  67
6 107    ROHAN  56.0  67  56  67.0  56
7 108    ROHINI 57.0  65  57  65.0  57
8 109    SANDHYA 58.0  56  58  56.0  58
9 110    SARANYA 49.0  45  49  56.0  49

```

The above updated data set will be stored in sample3.xlsx file....

|   | ROLL NO | NAME    | ENGLISH | TAMIL | MATHS | SCIENCE | SOCIAL |
|---|---------|---------|---------|-------|-------|---------|--------|
| 0 | 101     | DEEPA   | 50      | 67    | 50    | 67      | 50     |
| 1 | 102     | DINESH  | 56      | 89    | 56    | 89      | 56     |
| 2 | 103     | KAVIYA  | 80      | 80    | 80    | 80      | 80     |
| 3 | 104     | RACHEAL | 89      | 87    | 89    | 87      | 89     |
| 4 | 105     | RAJAN   | 89      | 98    | 90    | 98      | 90     |
| 5 | 106     | RAMYA   | 67      | 76    | 67    | 76      | 67     |
| 6 | 107     | ROHAN   | 56      | 67    | 56    | 67      | 56     |
| 7 | 108     | ROHINI  | 57      | 65    | 57    | 65      | 57     |
| 8 | 109     | SANDHYA | 58      | 56    | 58    | 56      | 58     |
| 9 | 110     | SARANYA | 49      | 45    | 49    | 56      | 49     |

Drop NaN rows:

| Unnamed : 0 | ROLL NO | NAME    | ENGLISH | TAMIL | MATHS | SCIENCE | SOCIAL | TOTAL |
|-------------|---------|---------|---------|-------|-------|---------|--------|-------|
| 0           | 101     | DEEPA   | 50      | 67    | 50    | 67      | 50     | 284   |
| 1           | 102     | DINESH  | 56      | 89    | 56    | 89      | 56     | 346   |
| 2           | 103     | KAVIYA  | 80      | 80    | 80    | 80      | 80     | 400   |
| 3           | 104     | RACHEAL | 89      | 87    | 89    | 87      | 89     | 441   |
| 4           | 105     |         |         |       |       |         |        |       |
| 5           | 106     | RAMYA   | 67      | 76    | 67    | 76      | 67     | 353   |
| 6           | 107     | ROHAN   | 56      | 67    | 56    | 67      | 56     | 302   |
| 7           | 108     |         |         |       |       |         |        |       |
| 8           | 109     | SANDHYA | 58      | 56    | 58    | 56      | 58     | 286   |
| 9           | 110     | SARANYA | 49      | 45    | 49    | 45      | 49     | 237   |

Unnamed: 0 ROLL NO NAME ENGLISH TAMIL MATHS SCIENCE SOCIAL TOTAL



```

0    0   101  DEEPA   50.0  67.0  50.0   67.0   50.0 284.0
1    1   102  DINESH   56.0  89.0  56.0   89.0   56.0 346.0
2    2   103  KAVIYA   80.0  80.0  80.0   80.0   80.0 400.0
3    3   104  RACHEAL   89.0  87.0  89.0   87.0   89.0 441.0
5    5   106  RAMYA    67.0  76.0  67.0   76.0   67.0 353.0
6    6   107  ROHAN    56.0  67.0  56.0   67.0   56.0 302.0
8    8   109  SANDHYA   58.0  56.0  58.0   56.0   58.0 286.0
9    9   110  SARANYA   49.0  45.0  49.0   45.0   49.0 237.0

```

The above updated data set will be stored in sample5.xlsx file....

|   | Unnamed: 0 | ROLL NO | NAME    | ENGLISH | TAMIL | MATHS | SCIENCE | SOCIAL | TOTAL |
|---|------------|---------|---------|---------|-------|-------|---------|--------|-------|
| 0 | 0          | 101     | DEEPA   | 50      | 67    | 50    | 67      | 50     | 284   |
| 1 | 1          | 102     | DINESH  | 56      | 89    | 56    | 89      | 56     | 346   |
| 2 | 2          | 103     | KAVIYA  | 80      | 80    | 80    | 80      | 80     | 400   |
| 3 | 3          | 104     | RACHEAL | 89      | 87    | 89    | 87      | 89     | 441   |
| 5 | 5          | 106     | RAMYA   | 67      | 76    | 67    | 76      | 67     | 353   |
| 6 | 6          | 107     | ROHAN   | 56      | 67    | 56    | 67      | 56     | 302   |
| 8 | 8          | 109     | SANDHYA | 58      | 56    | 58    | 56      | 58     | 286   |
| 9 | 9          | 110     | SARANYA | 49      | 45    | 49    | 45      | 49     | 237   |

Replace values:

Original data set:

```

      Unnamed: 0  ROLL NO      NAME  ENGLISH  TAMIL  MATHS  SCIENCE  SOCIAL
0      0      101      DEEPA    50    67    50    67    50
1      1      102      DINESH    56    89    56    89    56
2      2      103      KAVIYA    80    80    80    80    80
3      3      104      RACHEAL    89    87    89    87    89
4      4      105      RAJAN     89    98    90    98    90
5      5      106      RAMYA     67    76    67    76    67
6      6      107      ROHAN     56    67    56    67    56

```

```

7      7      108      ROHINI      57      65      57      65      57
8      8      109      SANDHYA      58      56      58      56      58
9      9      110      SARANYA      49      45      49      56      49

```

Updated dataset with replaced values: {49:50}

```

Unnamed: 0  ROLL NO      NAME  ENGLISH  TAMIL  MATHS  SCIENCE  SOCIAL
0      0      101      DEEPA      50      67      50      67      50
1      1      102      DINESH      56      89      56      89      56
2      2      103      KAVIYA      80      80      80      80      80
3      3      104      RACHEAL      89      87      89      87      89
4      4      105      RAJAN      89      98      90      98      90
5      5      106      RAMYA      67      76      67      76      67
6      6      107      ROHAN      56      67      56      67      56
7      7      108      ROHINI      57      65      57      65      57
8      8      109      SANDHYA      58      56      58      56      58
9      9      110      SARANYA      50      45      50      56      50

```

The above updated data set will be stored in sample6.xlsx file....

|   | Unnamed : 0 | ROLL NO | NAME    | ENGLISH | TAMIL | MATHS | SCIENCE | SOCIAL |
|---|-------------|---------|---------|---------|-------|-------|---------|--------|
| 0 | 0           | 101     | DEEPA   | 50      | 67    | 50    | 67      | 50     |
| 1 | 1           | 102     | DINESH  | 56      | 89    | 56    | 89      | 56     |
| 2 | 2           | 103     | KAVIYA  | 80      | 80    | 80    | 80      | 80     |
| 3 | 3           | 104     | RACHEAL | 89      | 87    | 89    | 87      | 89     |
| 4 | 4           | 105     | RAJAN   | 89      | 98    | 90    | 98      | 90     |
| 5 | 5           | 106     | RAMYA   | 67      | 76    | 67    | 76      | 67     |
| 6 | 6           | 107     | ROHAN   | 56      | 67    | 56    | 67      | 56     |
| 7 | 7           | 108     | ROHINI  | 57      | 65    | 57    | 65      | 57     |
| 8 | 8           | 109     | SANDHYA | 58      | 56    | 58    | 56      | 58     |
| 9 | 9           | 110     | SARANYA | 50      | 45    | 50    | 56      | 50     |

Original data set from Excel file:

|   | ROLL NO | NAME    | ENGLISH | TAMIL | MATHS | SCIENCE | SOCIAL | TOTAL |
|---|---------|---------|---------|-------|-------|---------|--------|-------|
| 0 | 101     | DEEPA   | 50.0    | 67    | 50    | 67.0    | 50     | 284   |
| 1 | 102     | DINESH  | 56.0    | 89    | 56    | 89.0    | 56     | 346   |
| 2 | 103     | KAVIYA  | 80.0    | 80    | 80    | 80.0    | 80     | 400   |
| 3 | 104     | RACHEAL | 89.0    | 87    | 89    | 87.0    | 89     | 441   |
| 4 | 105     | RAJAN   | NaN     | 98    | 90    | 98.0    | 90     | 466   |
| 5 | 106     | RAMYA   | 67.0    | 76    | 67    | 76.0    | 67     | 353   |
| 6 | 107     | ROHAN   | 56.0    | 67    | 56    | 67.0    | 56     | 302   |
| 7 | 108     | ROHINI  | 57.0    | 65    | 57    | 65.0    | 57     | 301   |
| 8 | 109     | SANDHYA | 58.0    | 56    | 58    | 56.0    | 58     | 286   |
| 9 | 110     | SARANYA | 49.0    | 45    | 49    | NaN     | 49     | 237   |

Extract the particular record based on the isin() function condition:

|   | ROLL NO | NAME    | ENGLISH | TAMIL | MATHS | SCIENCE | SOCIAL | TOTAL |
|---|---------|---------|---------|-------|-------|---------|--------|-------|
| 9 | 110     | SARANYA | 49.0    | 45    | 49    | NaN     | 49     | 237   |

**Result:**

The programs were run successfully