

In [ ]:

```

1  # string ''' '''
2  '''
3  There are several ways to handle or clean up data for the data to be used in datasets
4  data inconsistency means data is in different formats in multiple tables
5  which results in unreliable or meaningless information
6
7  so for handling data we need consistency in dataset
8  we can
9  1. remove whitespaces ' mca 2 ' with 'mca2'
10 2. replace capital 'Abc' with 'abc' or 'ABC'
11 3. remove . ' dot ' from end of string 'python.' with 'python'
12
13 e.g
14
15 before : ' Python .'
16 after  : 'PYTHON'
17
18 '''

```

In [1]:

```

1 import pandas as pd
2
3 filename = input('Enter filename here : ') # enter filename here
4
5 df = pd.read_csv(filename, sep=',')
6 df

```

Enter filename here : inconsistent.csv

Out[1]:

	rno	name	subject	marks
0	5	Umesh Bilade	Ai with Python	60
1	25	Shivam Limbhare	DIP	60
2	33	Sankalp Oswal	Ai with Python	60
3	43	Mahesh Patil	Ai with Python	60
4	44	Manjit Patil	Ai with Python	60
5	49	Rupesh Patil	Ai with Python	60
6	59	Umesh S onar	Ai with Python	60
7	56	Bhaves Shete	Ai with Python	60
8	35	U m a r K h a n	Ai with Python	24
9	0	Naam mein kya rakha hain	nahi hai	0

In [54]:

```
1 before = df['name']  
2 before
```

Out[54]:

```
0          Umesh Bilade  
1          Shivam Limbhare  
2          Sankalp Oswal  
3          Mahesh Patil  
4          Manjit Patil  
5          Rupesh Patil  
6          Umesh S onar  
7          Bhavesh Shete  
8          U m a r Kh a n  
9          Naam mein kya rakha hain  
Name: name, dtype: object
```

In [55]:

```
1 df1 = df['rno'].unique() # displays only unique values / duplicates as single value  
2 df1
```

Out[55]:

```
array([ 5, 25, 33, 43, 44, 49, 59, 56, 35,  0], dtype=int64)
```

In [56]:

```
1 df2 = df['name'].str.upper() # UPPERCASE FORMAT  
2 df2
```

Out[56]:

```
0          UMESH BILADE  
1          SHIVAM LIMBHARE  
2          SANKALP OSWAL  
3          MAHESH PATIL  
4          MANJIT PATIL  
5          RUPESH PATIL  
6          UMESH S ONAR  
7          BHAVESH SHETE  
8          U M A R KH A N  
9          NAAM MEIN KYA RAKHA HAIN  
Name: name, dtype: object
```

In [57]:

```
1 df3 = df['name'].str.lower() # Lowercase format
2 df3
```

Out[57]:

```
0          umesh bilade
1        shivam limbhare
2      sankalp oswal
3        mahesh patil
4        manjit patil
5        rupesh patil
6        umesh s onar
7        bhavesh shete
8      u m a r kh   a n
9  naam mein kya rakha hain
Name: name, dtype: object
```

In [58]:

```
1 df4 = df['name'].str.strip() # removes starting and ending spaces
2 df4
```

Out[58]:

```
0          Umesh Bilade
1        Shivam Limbhare
2        Sankalp Oswal
3        Mahesh Patil
4        Manjit Patil
5        Rupesh Patil
6        Umesh S onar
7        Bhavesh Shete
8      U m a r Kh   a n
9  Naam mein kya rakha hain
Name: name, dtype: object
```

In [59]:

```
1 df4 = df['name'].str.rstrip() #
2 df4
```

Out[59]:

```
0          Umesh Bilade
1        Shivam Limbhare
2        Sankalp Oswal
3        Mahesh Patil
4        Manjit Patil
5        Rupesh Patil
6        Umesh S onar
7        Bhavesh Shete
8      U m a r Kh   a n
9  Naam mein kya rakha hain
Name: name, dtype: object
```

In [60]:

```
1 df4 = df['name'].str.lstrip() #
2 df4
```

Out[60]:

```
0      Umesh Bilade
1      Shivam Limbhare
2      Sankalp Oswal
3      Mahesh Patil
4      Manjit Patil
5      Rupesh Patil
6      Umesh S onar
7      Bhavesh Shete
8      U m a r Kh a n
9      Naam mein kya rakha hain
Name: name, dtype: object
```

In [61]:

```
1 df5 = df['name'].str.replace(' ', '')
2 after = df5.str.upper()
3 after
```

Out[61]:

```
0      UMESHBILADE
1      SHIVAMLIMBHARE
2      SANKALPOSWAL
3      MAHESHPATIL
4      MANJITPATIL
5      RUPESHPATIL
6      UMESHSONAR
7      BHAVESHSHETE
8      UMARKHAN
9      NAAMMEINKYARAKHAHAIN
Name: name, dtype: object
```

In [62]:

```
1 df6 = df['name'].str.replace(' ', '_')
2 df6
```

Out[62]:

```
0      __Umesh_Bilade__
1      Shivam_Limbhare
2      Sankalp_Oswal__
3      __Mahesh_Patil
4      Manjit_Patil__
5      __Rupesh_Patil
6      __Umesh__S_onar
7      Bhavesh__Shete
8      U_m_a_r_Kh__a_n_
9      Naam_mein_kya_rakha_hain
Name: name, dtype: object
```

In [63]:

```
1 import re # regular expression
2
3 df4 = re.sub(' +', ' ',str(df[['name','subject']])) # removes duplicate whilte spaces fr
```

In [64]:

```
1 print(df4) # df4 is now a string object and not a dataframe
```

```
name subject
0 Umesh Bilade Ai with Python
1 Shivam Limbhare DIP
2 Sankalp Oswal Ai with Python
3 Mahesh Patil Ai with Python
4 Manjit Patil Ai with Python
5 Rupesh Patil Ai with Python
6 Umesh S onar Ai with Python
7 Bhavesh Shete Ai with Python
8 U m a r Kh a n Ai with Python
9 Naam mein kya rakha hain nahi hai
```

In [65]:

```
1 type(df4)
```

Out[65]:

str

In [66]:

```
1 type(df)
```

Out[66]:

pandas.core.frame.DataFrame

In [70]:

```
1 print(df4)
```

```
name subject
0 Umesh Bilade Ai with Python
1 Shivam Limbhare DIP
2 Sankalp Oswal Ai with Python
3 Mahesh Patil Ai with Python
4 Manjit Patil Ai with Python
5 Rupesh Patil Ai with Python
6 Umesh S onar Ai with Python
7 Bhavesh Shete Ai with Python
8 U m a r Kh a n Ai with Python
9 Naam mein kya rakha hain nahi hai
```

In [71]:

```
1 print('Before\n')
2 before
```

Before

Out[71]:

```
0          Umesh Bilade
1          Shivam Limbhare
2          Sankalp Oswal
3          Mahesh Patil
4          Manjit Patil
5          Rupesh Patil
6          Umesh S onar
7          Bhavesh Shete
8          U m a r Kh a n
9          Naam mein kya rakha hain
Name: name, dtype: object
```

In [72]:

```
1 print('After\n')
2 after
```

After

Out[72]:

```
0          UMESHBILADE
1          SHIVAMLIMBHARE
2          SANKALPOSWAL
3          MAHESHPATIL
4          MANJITPATIL
5          RUPESHPATIL
6          UMESHSONAR
7          BHAVESHSHETE
8          UMARKHAN
9          NAAMMEINKYARAKHAHAIN
Name: name, dtype: object
```