```
import pandas as pd
```

```
df = pd.read_csv("dirtydata.csv")
df.head()
```

	Duration	Date	Pulse	Maxpulse	Calories
0	60	'2020/12/01'	110	130	409.1
1	60	'2020/12/02'	117	145	479.0
2	60	'2020/12/03'	103	135	340.0
3	45	'2020/12/04'	109	175	282.4
4	45	'2020/12/05'	117	148	406.0

```
shape=df.shape
print(shape)
```

```
(32, 5)
```

```
desc=df.describe()
print(desc)
```

```
          Duration       Pulse    Maxpulse    Calories
count   32.000000   32.000000   32.000000   30.000000
mean    68.437500  103.500000  128.500000  266.013333
std     70.039591    7.832933   12.998759  164.876415
min     30.000000   90.000000  101.000000 -300.000000
25%     60.000000  100.000000  120.000000  247.000000
50%     60.000000  102.500000  127.500000  282.200000
75%     60.000000  106.500000  132.250000  343.975000
max    450.000000  130.000000  175.000000  479.000000
```

```
type=df.dtypes
print(type)
```

```
Duration      int64
Date         object
Pulse         int64
Maxpulse      int64
Calories    float64
dtype: object
```

```
null=df.isnull()
print(null)
```

```
    Duration   Date  Pulse  Maxpulse  Calories
0      False  False  False     False     False
1      False  False  False     False     False
2      False  False  False     False     False
3      False  False  False     False     False
4      False  False  False     False     False
5      False  False  False     False     False
6      False  False  False     False     False
7      False  False  False     False     False
8      False  False  False     False     False
9      False  False  False     False     False
10     False  False  False     False     False
11     False  False  False     False     False
12     False  False  False     False     False
13     False  False  False     False     False
14     False  False  False     False     False
15     False  False  False     False     False
```

```
16     False  False  False     False     False
17     False  False  False     False     False
18     False  False  False     False      True
19     False  False  False     False     False
20     False  False  False     False     False
21     False  False  False     False     False
22     False   True  False     False     False
23     False  False  False     False     False
24     False  False  False     False     False
25     False  False  False     False     False
26     False  False  False     False     False
27     False  False  False     False     False
28     False  False  False     False      True
29     False  False  False     False     False
30     False  False  False     False     False
31     False  False  False     False     False
```

```
null=df.isnull().sum()
print(null)
```

```
Duration    0
Date        1
Pulse       0
Maxpulse    0
Calories    2
dtype: int64
```

```
df['Calories']=df['Calories'].abs()
x=df['Calories'].mean()
df['Calories'].fillna(x,inplace=True)
df['Calories']=df['Calories'].astype(int)
df
```

	Duration	Date	Pulse	Maxpulse	Calories
0	60	'2020/12/01'	110	130	409
1	60	'2020/12/02'	117	145	479
2	60	'2020/12/03'	103	135	340
3	45	'2020/12/04'	109	175	282
4	45	'2020/12/05'	117	148	406
5	60	'2020/12/06'	102	127	300
6	60	'2020/12/07'	110	136	374
7	450	'2020/12/08'	104	134	253
8	30	'2020/12/09'	109	133	195
9	60	'2020/12/10'	98	124	269
10	60	'2020/12/11'	103	147	329
11	60	'2020/12/12'	100	120	250

```
df.dropna(subset=['Date'],inplace=True)
df['Date']=pd.to_datetime(df['Date'])

df.loc[7,'Duration']=45

df
```

	Duration	Date	Pulse	Maxpulse	Calories
0	60	2020-12-01	110	130	409
1	60	2020-12-02	117	145	479
2	60	2020-12-03	103	135	340
3	45	2020-12-04	109	175	282
4	45	2020-12-05	117	148	406
5	60	2020-12-06	102	127	300
6	60	2020-12-07	110	136	374
7	45	2020-12-08	104	134	253
8	30	2020-12-09	109	133	195
9	60	2020-12-10	98	124	269

```
# duplicate=df.duplicated()
# print(duplicate)

duplicate=df.duplicated().sum()
print(duplicate)

df.drop_duplicates(inplace=True)
duplicate=df.duplicated().sum()
print(duplicate)
```

```
1
0
```

```
for x in df.index:
    if df.loc[x,'Duration']>45:
        df.loc[x,'Duration']=50
print(df)
```

```
    Duration       Date  Pulse  Maxpulse  Calories
0         50 2020-12-01    110       130       409
1         50 2020-12-02    117       145       479
2         50 2020-12-03    103       135       340
3         45 2020-12-04    109       175       282
4         45 2020-12-05    117       148       406
5         50 2020-12-06    102       127       300
6         50 2020-12-07    110       136       374
7         45 2020-12-08    104       134       253
8         30 2020-12-09    109       133       195
9         50 2020-12-10     98       124       269
10        50 2020-12-11    103       147       329
11        50 2020-12-12    100       120       250
13        50 2020-12-13    106       128       345
14        50 2020-12-14    104       132       379
15        50 2020-12-15     98       123       275
16        50 2020-12-16     98       120       215
17        50 2020-12-17    100       120       300
18        45 2020-12-18     90       112       304
19        50 2020-12-19    103       123       323
20        45 2020-12-20     97       125       243
21        50 2020-12-21    108       131       364
23        50 2020-12-23    130       101       300
24        45 2020-12-24    105       132       246
25        50 2020-12-25    102       126       334
26        50 2020-12-26    100       120       250
27        50 2020-12-27     92       118       241
28        50 2020-12-28    103       132       304
29        50 2020-12-29    100       132       280
30        50 2020-12-30    102       129       380
31        50 2020-12-31     92       115       243
```

NBA Dataset

```
from sklearn import preprocessing

desired_width = 320         # to display 10 columns
pd.set_option('display.width', desired_width)
pd.set_option('display.max_columns', 20)

df = pd.read_csv("nba.csv")
print(df)
```

```
                Name            Team  Number Position  Age  Height  Weight            College     Salary
0    Avery Bradley  Boston Celtics       0       PG   25   2-Jun     180              Texas  7730337.0
1      Jae Crowder  Boston Celtics      99       SF   25   6-Jun     235          Marquette  6796117.0
2     John Holland  Boston Celtics      30       SG   27   5-Jun     205  Boston University        NaN
3      R.J. Hunter  Boston Celtics      28       SG   22   5-Jun     185      Georgia State  1148640.0
4    Jonas Jerebko  Boston Celtics       8       PF   29  10-Jun     231                NaN  5000000.0
..             ...             ...     ...      ...  ...     ...     ...                ...        ...
452     Trey Lyles       Utah Jazz      41       PF   20  10-Jun     234           Kentucky  2239800.0
453   Shelvin Mack       Utah Jazz       8       PG   26   3-Jun     203             Butler  2433333.0
454      Raul Neto       Utah Jazz      25       PG   24   1-Jun     179                NaN   900000.0
455   Tibor Pleiss       Utah Jazz      21        C   26   3-Jul     256                NaN  2900000.0
456    Jeff Withey       Utah Jazz      24        C   26  Jul-00     231             Kansas   947276.0

[457 rows x 9 columns]
```

```
pos=df['Position'].unique()
print(pos)
```

```
['PG' 'SF' 'SG' 'PF' 'C']
```

```
label_Encode=preprocessing.LabelEncoder()
df['Pos']=label_Encode.fit_transform(df['Position'])
print(df)
```

```
                Name            Team  Number Position  Age  Height  Weight            College     Salary
0    Avery Bradley  Boston Celtics       0       PG   25   2-Jun     180              Texas  7730337.0
1      Jae Crowder  Boston Celtics      99       SF   25   6-Jun     235          Marquette  6796117.0
2     John Holland  Boston Celtics      30       SG   27   5-Jun     205  Boston University        NaN
3      R.J. Hunter  Boston Celtics      28       SG   22   5-Jun     185      Georgia State  1148640.0
4    Jonas Jerebko  Boston Celtics       8       PF   29  10-Jun     231                NaN  5000000.0
..             ...             ...     ...      ...  ...     ...     ...                ...        ...
452     Trey Lyles       Utah Jazz      41       PF   20  10-Jun     234           Kentucky  2239800.0
453   Shelvin Mack       Utah Jazz       8       PG   26   3-Jun     203             Butler  2433333.0
454      Raul Neto       Utah Jazz      25       PG   24   1-Jun     179                NaN   900000.0
455   Tibor Pleiss       Utah Jazz      21        C   26   3-Jul     256                NaN  2900000.0
456    Jeff Withey       Utah Jazz      24        C   26  Jul-00     231             Kansas   947276.0

[457 rows x 10 columns]
```

```
df['Position'].replace(['SG','PF','PG','SF','C'],[1,2,3,4,5],inplace=True)
print(df)
```

```
                Name            Team  Number  Position  Age  Height  Weight            College     Salary
0    Avery Bradley  Boston Celtics       0         3   25   2-Jun     180              Texas  7730337.0
1      Jae Crowder  Boston Celtics      99         4   25   6-Jun     235          Marquette  6796117.0
2     John Holland  Boston Celtics      30         1   27   5-Jun     205  Boston University        NaN
3      R.J. Hunter  Boston Celtics      28         1   22   5-Jun     185      Georgia State  1148640.0
4    Jonas Jerebko  Boston Celtics       8         2   29  10-Jun     231                NaN  5000000.0
..             ...             ...     ...       ...  ...     ...     ...                ...        ...
452     Trey Lyles       Utah Jazz      41         2   20  10-Jun     234           Kentucky  2239800.0
453   Shelvin Mack       Utah Jazz       8         3   26   3-Jun     203             Butler  2433333.0
```

```
454      Raul Neto       Utah Jazz      25         3   24   1-Jun     179              NaN   900000.0
455   Tibor Pleiss       Utah Jazz      21         5   26   3-Jul     256              NaN  2900000.0
456    Jeff Withey       Utah Jazz      24         5   26  Jul-00     231           Kansas   947276.0

[457 rows x 10 columns]
```

```
cat=pd.cut(df.Age,bins=[19,25,30,35,45],labels=['A','B','C','D'])
df.insert(5,'AgeGroup',cat)
print(df)
```

```
              Name            Team  Number Position  Age AgeGroup  Height  Weight            College
0    Avery Bradley  Boston Celtics       0       PG   25        A   2-Jun     180              Texas  7
1      Jae Crowder  Boston Celtics      99       SF   25        A   6-Jun     235          Marquette  6
2     John Holland  Boston Celtics      30       SG   27        B   5-Jun     205  Boston University
3      R.J. Hunter  Boston Celtics      28       SG   22        A   5-Jun     185      Georgia State  1
4    Jonas Jerebko  Boston Celtics       8       PF   29        B  10-Jun     231                NaN  5
..             ...             ...     ...      ...  ...      ...     ...     ...                ...
452     Trey Lyles       Utah Jazz      41       PF   20        A  10-Jun     234          Kentucky  2
453   Shelvin Mack       Utah Jazz       8       PG   26        B   3-Jun     203            Butler  2
454      Raul Neto       Utah Jazz      25       PG   24        A   1-Jun     179               NaN
455   Tibor Pleiss       Utah Jazz      21        C   26        B   3-Jul     256               NaN  2
456    Jeff Withey       Utah Jazz      24        C   26        B  Jul-00     231            Kansas

[457 rows x 10 columns]
```

Alcohol Dataset

```
df = pd.read_csv("A1_ALCHOHOL.csv")
df.head()
```

	Country	Alcohol	Deaths	Heart	Liver
0	Australia	2.5	785	211.0	15.300000
1	Austria	3.0	863	167.0	45.599998
2	Belg. and Lux.	2.9	883	131.0	20.700001
3	Canada	2.4	793	NaN	16.400000
4	Denmark	2.9	971	220.0	23.900000

```
print(df.columns)
df.columns = [c.strip() for c in df.columns]
print(df.columns)
```

```
Index(['Country', ' Alcohol ', 'Deaths ', 'Heart', ' Liver'], dtype='object')
Index(['Country', 'Alcohol', 'Deaths', 'Heart', 'Liver'], dtype='object')
```

```
df['Deaths'] = df['Deaths'].abs()
df['Alcohol'] = df['Alcohol'].abs()

df.loc[10,'Alcohol']=7
print(df)
```

```
           Country  Alcohol  Deaths  Heart      Liver
0        Australia     2.50     785  211.0  15.300000
1          Austria     3.00     863  167.0  45.599998
2   Belg. and Lux.     2.90     883  131.0  20.700001
3           Canada     2.40     793    NaN  16.400000
4          Denmark     2.90     971  220.0  23.900000
5          Finland     0.80     970  297.0  19.000000
```

```
6           France     9.10     751   11.0  37.900002
7          Iceland     0.80     743  211.0  11.200000
8          Ireland     0.70    1000  300.0   6.500000
9           Israel     0.60     834  183.0  13.700000
10           Italy     7.00     775  107.0  42.200001
11           Japan     1.50     680   36.0  23.200001
12     Netherlands     1.80     773  167.0   9.200000
13     New Zealand     1.90     916  266.0   7.700000
14          Norway     0.08     806  227.0  12.200000
15           Spain     6.50     724    NaN        NaN
16          Sweden     1.60     743  207.0  11.200000
17     Switzerland     5.80     693  115.0  20.299999
18              UK     1.30     941  285.0  10.300000
19              US     1.20     926  199.0  22.100000
20    West Germany     2.70     861  172.0  36.700001
21           India     2.95     750  171.0  20.270000
```

```
x=df['Liver'].median()
df['Liver'].fillna(x,inplace=True)
df['Liver']=df['Liver'].astype(int)
print(df)
```

```
              Country  Alcohol  Deaths  Heart  Liver
0        Australia     2.50     785  211.0     15
1          Austria     3.00     863  167.0     45
2   Belg. and Lux.     2.90     883  131.0     20
3           Canada     2.40     793    NaN     16
4          Denmark     2.90     971  220.0     23
5          Finland     0.80     970  297.0     19
6           France     9.10     751   11.0     37
7          Iceland     0.80     743  211.0     11
8          Ireland     0.70    1000  300.0      6
9           Israel     0.60     834  183.0     13
10           Italy     7.00     775  107.0     42
11           Japan     1.50     680   36.0     23
12     Netherlands     1.80     773  167.0      9
13     New Zealand     1.90     916  266.0      7
14          Norway     0.08     806  227.0     12
15           Spain     6.50     724    NaN     19
16          Sweden     1.60     743  207.0     11
17     Switzerland     5.80     693  115.0     20
18              UK     1.30     941  285.0     10
19              US     1.20     926  199.0     22
20    West Germany     2.70     861  172.0     36
21           India     2.95     750  171.0     20
```

```
df.dropna(subset=['Heart'],inplace=True)
print(df)
```

```
              Country  Alcohol  Deaths  Heart  Liver
0        Australia     2.50     785  211.0     15
1          Austria     3.00     863  167.0     45
2   Belg. and Lux.     2.90     883  131.0     20
4          Denmark     2.90     971  220.0     23
5          Finland     0.80     970  297.0     19
6           France     9.10     751   11.0     37
7          Iceland     0.80     743  211.0     11
8          Ireland     0.70    1000  300.0      6
9           Israel     0.60     834  183.0     13
10           Italy     7.00     775  107.0     42
11           Japan     1.50     680   36.0     23
12     Netherlands     1.80     773  167.0      9
13     New Zealand     1.90     916  266.0      7
14          Norway     0.08     806  227.0     12
16          Sweden     1.60     743  207.0     11
17     Switzerland     5.80     693  115.0     20
18              UK     1.30     941  285.0     10
19              US     1.20     926  199.0     22
```

```
20    West Germany      2.70      861  172.0      36
21           India      2.95      750  171.0      20
```

```
import matplotlib.pyplot as plt
df.boxplot(column=['Alcohol'])
plt.show()
```