Assignment No.1 : Data Wrangling - 1

In [12]:
```
import numpy as np
import pandas as pd
```

In [13]:
```
df = pd.read_csv('data.csv')
```

In [14]:
```
df.head()
```

Out[14]:

	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

In [15]:
```
df.shape
```

Out[15]:
```
(32, 5)
```

In [16]:
```
df.describe()
```

Out[16]:

	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

In [21]:
```
features_with_nan = [feat for feat in df.columns if df[feat].isnull().sum() > 0 and df[feat].dtype != 'O']


for feat in df.columns:
    print('{} has {} % missing values'.format(feat,df[feat].isnull().mean()))
```

```
Duration has 0.0 % missing values
Date has 0.03125 % missing values
Pulse has 0.0 % missing values
Maxpulse has 0.0 % missing values
Calories has 0.0625 % missing values
```

In [22]:
```
features_with_nan
```

Out[22]:
```
['Calories']
```

In [26]:
```
for feat in features_with_nan:
    mean_value = df[feat].mean()
    df[feat] = df[feat].fillna(mean_value)
```

In [27]:
```
df[features_with_nan].isnull().sum()
```

Out[27]:
```
Calories    0
dtype: int64
```

In [28]:
```
for feat in df.columns:
    print('{} has {} data type'.format(feat,df[feat].dtypes))
```

```
Duration has int64 data type
Date has object data type
Pulse has int64 data type
Maxpulse has int64 data type
Calories has float64 data type
```

In [31]:
```
df['Calories'] = df['Calories'].astype('int64')
```

In [32]:
```
for feat in df.columns:
    print('{} has {} data type'.format(feat,df[feat].dtypes))
```

```
Duration has int64 data type
Date has object data type
Pulse has int64 data type
Maxpulse has int64 data type
Calories has int64 data type
```

Assignment No.1 : Data Wrangling - 1

In [37]:
```
df['Calories'] = np.where((df['Calories'] < 0), -(df['Calories']),df['Calories'])
```

In [44]:
```
(df['Calories'] < 0).value_counts()
```

Out[44]:
```
False    32
Name: Calories, dtype: int64
```

In [39]:
```
df.describe()
```

Out[39]:

	Duration	Pulse	Maxpulse	Calories
count	32.000000	32.000000	32.000000	32.000000
mean	68.437500	103.500000	128.500000	302.125000
std	70.039591	7.832933	12.998759	64.552429
min	30.000000	90.000000	101.000000	195.000000
25%	60.000000	100.000000	120.000000	250.000000
50%	60.000000	102.500000	127.500000	282.000000
75%	60.000000	106.500000	132.250000	341.250000
max	450.000000	130.000000	175.000000	479.000000