Data Frame with Pandas Library

1. Install Numpy and Pandas
2. Create dictionary and dataframe
3. Creat labels
4. read file to dataframes
5. DataFrame operations
6. Plotting data
7. save to file

In [2]:

```
import pandas as pd

a = [1, 7, 2]
print(a)
```

```
[1, 7, 2]
```

In [3]:

```
myvar = pd.Series(a)

print(myvar)
```

```
0    1
1    7
2    2
dtype: int64
```

In [4]:

```
myvar = pd.Series(a, index = ["x", "y", "z"])
print(myvar)
```

```
x    1
y    7
z    2
dtype: int64
```

In [7]:

```
calories = {"day1": 420, "day2": 380, "day3": 390}

myvar1 = pd.Series(calories)
print(myvar1)

myvar2 = pd.Series(calories, index = ["day1", "day2"])
print(myvar2)
```

```
day1    420
day2    380
day3    390
dtype: int64
day1    420
day2    380
dtype: int64
```

In [8]:

```
#creating dataFrame
mydataset = {
  'cars': ["BMW", "Volvo", "Ford"],
  'passings': [3, 7, 2]
}

df = pd.DataFrame(mydataset)

print(df)
```

```
    cars  passings
0    BMW         3
1  Volvo         7
2   Ford         2
```

In [12]:

```
#DataFrame Access
#Access row index:
print(df.loc[1])
```

```
cars        Volvo
passings        7
Name: 1, dtype: object
```

In [13]:

```
#specify a lsit of index
print(df.loc[[0, 2]])
```

```
   cars  passings
0   BMW         3
2  Ford         2
```

In [14]:

```
print(df.loc[0:2])
```

```
    cars  passings
0    BMW         3
1  Volvo         7
2   Ford         2
```

In [15]:

```
#assign names for index
data = {
  "calories": [420, 380, 390],
  "duration": [50, 40, 45]
}

df = pd.DataFrame(data)
print(df)

df = pd.DataFrame(data, index = ["day1", "day2", "day3"])
print(df)

print(df.index)
```

```
   calories  duration
0       420        50
1       380        40
2       390        45
      calories  duration
day1       420        50
day2       380        40
day3       390        45
Index(['day1', 'day2', 'day3'], dtype='object')
```

In [16]:

```
#access by index name
print(df.loc["day2"])
```

```
calories    380
duration     40
Name: day2, dtype: int64
```

In [18]:

```
#Reading data from a file
df = pd.read_csv('dataFile.csv')
print(df)
```

```
     Duration  Pulse  Maxpulse  Calories
0          60    110       130     409.1
1          60    117       145     479.0
2          60    103       135     340.0
3          45    109       175     282.4
4          45    117       148     406.0
..        ...    ...       ...       ...
164        60    105       140     290.8
165        60    110       145     300.0
166        60    115       145     310.2
167        75    120       150     320.4
168        75    125       150     330.4

[169 rows x 4 columns]
```

In [19]:

```
#print entire data
#print(df.to_string())

#first 10 rows
print(df.head())
```

```
   Duration  Pulse  Maxpulse  Calories
0        60    110       130     409.1
1        60    117       145     479.0
2        60    103       135     340.0
3        45    109       175     282.4
4        45    117       148     406.0
```

In [20]:

```
#last 5 rows
print(df.tail())
```

```
     Duration  Pulse  Maxpulse  Calories
164        60    105       140     290.8
165        60    110       145     300.0
166        60    115       145     310.2
167        75    120       150     320.4
168        75    125       150     330.4
```

In [21]:

```
#print information about the data
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 169 entries, 0 to 168
Data columns (total 4 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   Duration  169 non-null    int64  
 1   Pulse     169 non-null    int64  
 2   Maxpulse  169 non-null    int64  
 3   Calories  164 non-null    float64
dtypes: float64(1), int64(3)
memory usage: 5.4 KB
None
```

In [24]:

```
print(df[0:20])
```

```
    Duration  Pulse  Maxpulse  Calories
0         60    110       130     409.1
1         60    117       145     479.0
2         60    103       135     340.0
3         45    109       175     282.4
4         45    117       148     406.0
5         60    102       127     300.0
6         60    110       136     374.0
7         45    104       134     253.3
8         30    109       133     195.1
9         60     98       124     269.0
10        60    103       147     329.3
11        60    100       120     250.7
12        60    106       128     345.3
13        60    104       132     379.3
14        60     98       123     275.0
15        60     98       120     215.2
16        60    100       120     300.0
18        60    103       123     323.0
19        45     97       125     243.0
20        60    108       131     364.2
```

In [25]:

```
#print(new_df.to_string())
#print(df.tail().to_string())
#drop rows with empty cells
new_df = df.dropna(inplace = True)

#remove rows that has empty values in specific column
df.dropna(subset=['Calories'], inplace = True)

#fill N/A cells with a specific values
new_df = df.fillna(130)
df["Calories"].fillna(130, inplace = True)
```

In [ ]:

In [26]:

```
#calculate the mean for a column
x = df["Calories"].mean()
print(x)

#calculate median, maximum, minimum, mode, etc.....
```

```
375.79024390243916
```

In [27]:

```
#iterate on entire dataFrame usind index
#delete rows that has Duration > 120
for x in df.index:
  if df.loc[x, "Duration"] > 120:
    df.drop(x, inplace = True)
```

In [28]:

```
#calulate each values in column
i = df["Duration"].value_counts()
print(i)

#access a specific cell
#df.at[i, "Duration"] = ...
#df["Duration"].iat[i] = ...
```

```
60     76
45     33
30     16
20      9
90      8
120     3
15      2
75      2
25      1
80      1
Name: Duration, dtype: int64
```

refer to the following link for pandas documentation:

https://pandas.pydata.org/docs/reference/frame.html (https://pandas.pydata.org/docs/reference/frame.html)

In [ ]: