Machine Learning (Assignment 3)

Prem Kumar Kamma

700756204

Github Link : https://github.com/PremKumarKamma/Assignment_3ML

1) Read the provided CSV file 'data.csv'

```
import pandas as pd
#Read  the csv file
data = pd.read_csv("data.csv")
data.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

2) Show the basic statistical description about the data.

```
#Shows statstical description of data
data.describe()
```

	Duration	Pulse	Maxpulse	Calories
count	169.000000	169.000000	169.000000	164.000000
mean	63.846154	107.461538	134.047337	375.790244
std	42.299949	14.510259	16.450434	266.379919
min	15.000000	80.000000	100.000000	50.300000
25%	45.000000	100.000000	124.000000	250.925000
50%	60.000000	105.000000	131.000000	318.600000
75%	60.000000	111.000000	141.000000	387.600000
max	300.000000	159.000000	184.000000	1860.400000

3) Check if the data has null values. a. Replace the null values with the mean

```
[4]  #Check if data has null values
     data.isnull().any()
```

```
Duration    False
Pulse       False
Maxpulse    False
Calories     True
dtype: bool
```

```
[5]  #Replace the null values with the mean
     data.fillna(data.mean(),  inplace=True)
     data.isnull().any()
```

```
Duration    False
Pulse       False
Maxpulse    False
Calories    False
dtype: bool
```

4) Select at least two columns and aggregate the data using: min, max, count, mean.

```
#Select Atleast two columns and aggregate the data using: min, max count, mean
data.agg({'Maxpulse':['min', 'max', 'count', 'mean'],'Calories':['min','max','count','mean']})
```

	Maxpulse	Calories
min	100.000000	50.300000
max	184.000000	1860.400000
count	169.000000	169.000000
mean	134.047337	375.790244

5) Filter the dataframe to select the rows with calories values between 500 and 1000.

```
#Filter the dataframe to select the rows with calories values between 500 and 1000.
data.loc[(data['Calories']>500)&(data['Calories']<1000)]
```

	Duration	Pulse	Maxpulse	Calories
51	80	123	146	643.1
62	160	109	135	853.0
65	180	90	130	800.4
66	150	105	135	873.4
67	150	107	130	816.0
72	90	100	127	700.0
73	150	97	127	953.2
75	90	98	125	563.2
78	120	100	130	500.4
90	180	101	127	600.1
99	90	93	124	604.1
103	90	90	100	500.4
106	180	90	120	800.3
108	90	90	120	500.3

6) Filter the dataframe to select the rows with calories values > 500 and pulse < 100.

```
#Filter the dataframe to select the rows with calories values > 500 and pulse < 100.
data.loc[(data['Calories']>500)&(data['Pulse']<100)]
```

	Duration	Pulse	Maxpulse	Calories
65	180	90	130	800.4
70	150	97	129	1115.0
73	150	97	127	953.2
75	90	98	125	563.2
99	90	93	124	604.1
103	90	90	100	500.4
106	180	90	120	800.3
108	90	90	120	500.3

7) Create a new “df_modified” dataframe that contains all the columns from df except for “Maxpulse”

```
#Create a new “df_modified” dataframe that contains all the columns from df except for “Maxpulse”.
df_modified = data[['Duration','Pulse','Calories']]
df_modified.head()
```

	Duration	Pulse	Calories
0	60	110	409.1
1	60	117	479.0
2	60	103	340.0
3	45	109	282.4
4	45	117	406.0

8) Delete the “Maxpulse” column from the main df dataframe

```
#Delete the “Maxpulse” column from the main df dataframe
del data['Maxpulse']
data.head()
```

	Duration	Pulse	Calories
0	60	110	409.1
1	60	117	479.0
2	60	103	340.0
3	45	109	282.4
4	45	117	406.0

9) . Convert the datatype of Calories column to int datatype.

```
#Convert the datatype of Calories column to int datatype.
data.dtypes
```

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

```
import numpy as np
data['Calories'] = data['Calories'].astype(np.int64)
data.dtypes
```

```
Duration    int64
Pulse       int64
Calories    int64
dtype: object
```