GitHub Link -
https://github.com/sandy100061/MachineLearningAssignment/tree/main/Assignment_1

Video link-
https://drive.google.com/file/d/1AcggvcPQqpyzjLUH1qdvpnV6qQEBDHuk/view?usp=drive_link

1. Pandas

1. Read the provided CSV file 'data.csv'. https://drive.google.com/file/d/1-Ir3AXK1A77A-qCDu5gGkAxv-nbmWlHO/view?usp=sharing
2. Show the basic statistical description about the data.
3. Check if the data has null values. a. Replace the null values with the mean
4. Select at least two columns and aggregate the data using: min, max, count, mean.
5. Filter the dataframe to select the rows with calories values between 500 and 1000.
6. Filter the dataframe to select the rows with calories values > 500 and pulse < 100.
7. Create a new "df_modified" dataframe that contains all the columns from df except for "Maxpulse".
8. Delete the "Maxpulse" column from the main df dataframe
9. Convert the datatype of Calories column to int datatype.

```
#Read the provided CSV file 'data.csv'. https://drive.google.com/file/d/1-Ir3AXK1A77A-qCDu5gGkAxv-nbmWlHO/view?usp=sharing

import pandas as pd
df = pd.read_csv('/content/data.csv')
```

```
print(df)
```

2]

```
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]
```

```
df = pd.DataFrame(df)
```

```
#Show the basic statistical description about the data.
df=df.describe()
df
```

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

```
#Check if the data has null values.
df = pd.read_csv('/content/data.csv')
df.isnull()
```

Duration	Pulse	Maxpulse	Calories	
0	False	False	False	False
1	False	False	False	False
2	False	False	False	False
3	False	False	False	False
4	False	False	False	False
...	...	...	...	...
164	False	False	False	False

Duration	Pulse	Maxpulse	Calories	
165	False	False	False	False
166	False	False	False	False
167	False	False	False	False
168	False	False	False	False

169 rows × 4 columns

```
#checking is there any null value is there or not.
df.isnull().values.any()
```

True

```
# a. Replace the null values with the mean
new_df=df.fillna(df.mean())
```

```
new_df.isnull().values.any()
```

False

```
#4. Select at least two columns and aggregate the data using: min, max,
count, mean
# by using groupby function with aggregation to get mean, min and max
values
result = df.groupby('Duration').agg({'Calories': ['mean', 'min', 'max']})

print("Mean, min, and max values are")
print(result)
```

```
Mean, min, and max values are
             Calories
                 mean    min    max
Duration
15          87.350000   50.5  124.2
20         151.600000   50.3  229.4
25         244.200000  244.2  244.2
30         192.125000   86.2  319.2
45         273.236364  100.7  406.0
```

```
60          339.675000   215.2   486.0
75          325.400000   320.4   330.4
80          643.100000   643.1   643.1
90          541.800000   466.4   700.0
120         666.833333   500.0  1000.1
150         939.400000   816.0  1115.0
160         943.700000   853.0  1034.4
180         733.600000   600.1   800.4
210        1618.200000  1376.0  1860.4
270        1729.000000  1729.0  1729.0
300        1500.200000  1500.2  1500.2
```

```
#5. Filter the dataframe to select the rows with calories values between
500 and 1000.
df.query('Calories <= 1000 and Calories >= 500')
```

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
83	120	100	130	500.0
90	180	101	127	600.1
99	90	93	124	604.1
101	90	90	110	500.0
102	90	90	100	500.0
103	90	90	100	500.4

Duration	Pulse	Maxpulse	Calories	
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
df.query('Calories > 500 and 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”
df_modified=df.drop(columns=["Maxpulse"])
df_modified
```

	Pulse	Calories	
Duration			
0	60	110	409.1
1	60	117	479.0
2	60	103	340.0

Duration	Pulse	Calories	
3	45	109	282.4
4	45	117	406.0
...	...	...	...
164	60	105	290.8
165	60	110	300.0
166	60	115	310.2
167	75	120	320.4
168	75	125	330.4

169 rows × 3 columns

```
# 8. Delete the “Maxpulse” column from the main df dataframe
df.drop(columns=["Maxpulse"], axis=1, inplace=True)
df
```

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
...	...	...	...
164	60	105	290.8
165	60	110	300.0
166	60	115	310.2
167	75	120	320.4
168	75	125	330.4

169 rows × 3 columns

```
#9. Convert the datatype of Calories column to int datatype.
df=df.fillna(df.mean())
df = df.astype({'Calories':'int'})

print(df.dtypes)
```

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