ML ASSIGNMENT -1

Likhitha Vempati

700757827

1. Read the provided CSV file 'data.csv'. https://drive.google.com/drive/folders/1h8C3mLsso-R-sIOLsvoYwPLzy2fJ4IOF?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.

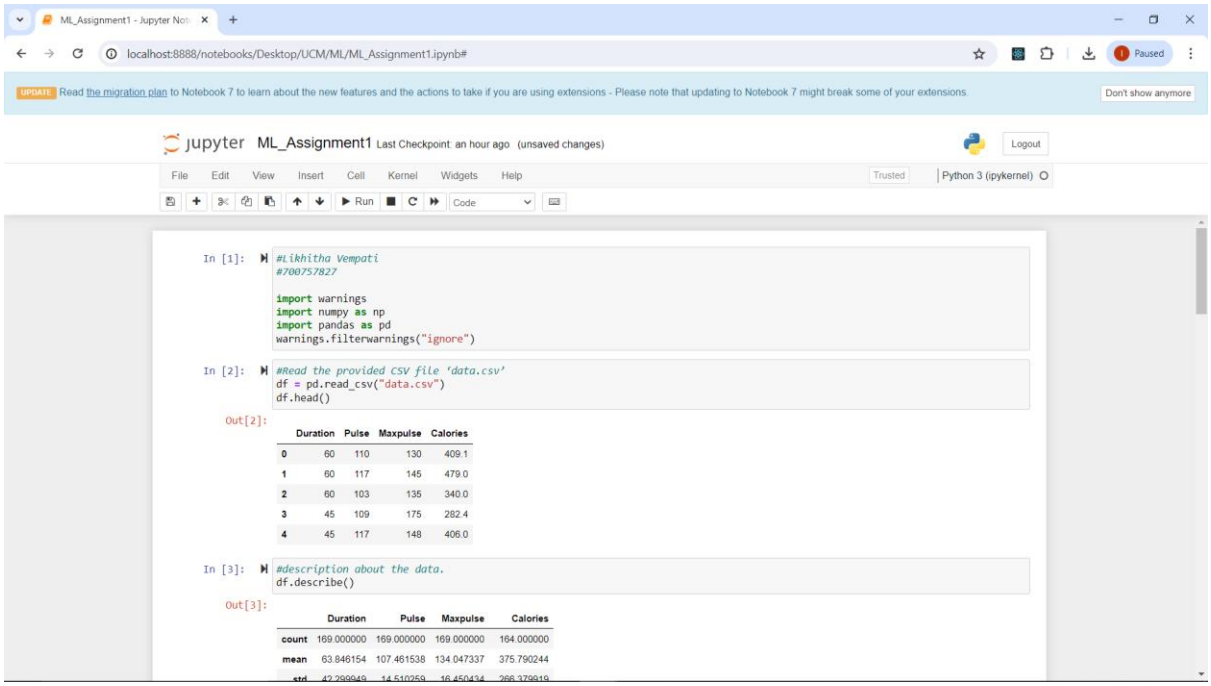
7. Create a new "df_modified" dataframe that contains all the columns from df except for "Maxpulse".

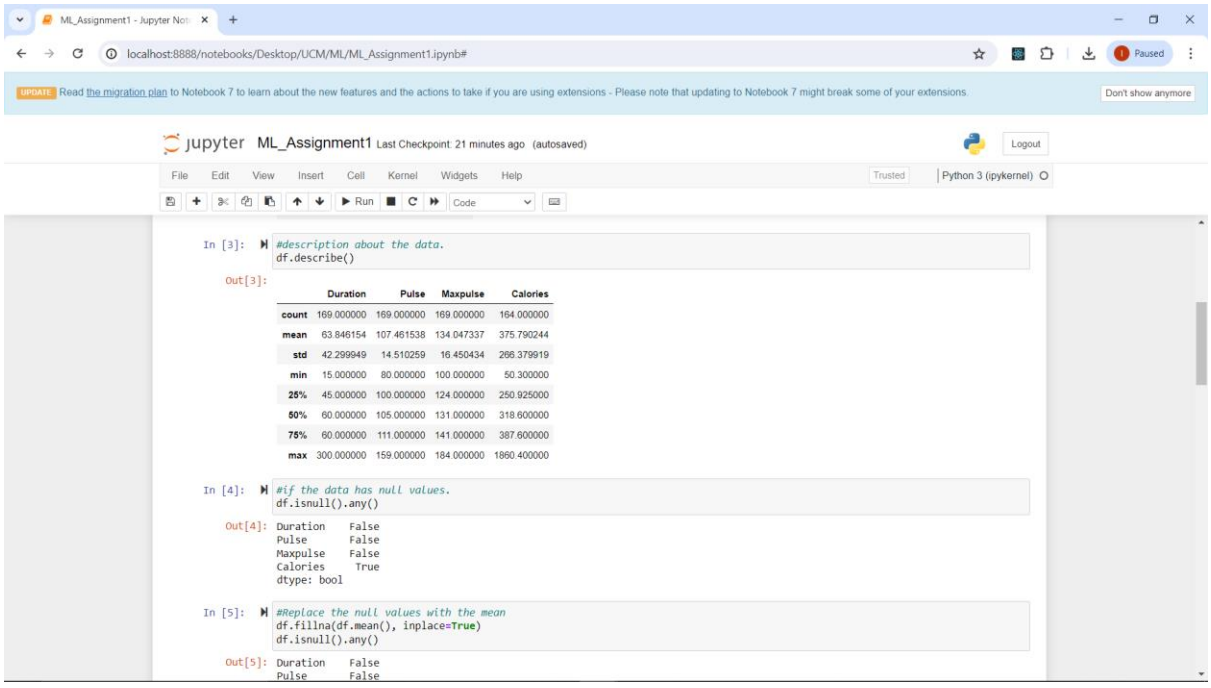
8. Delete the "Maxpulse" column from the main df dataframe

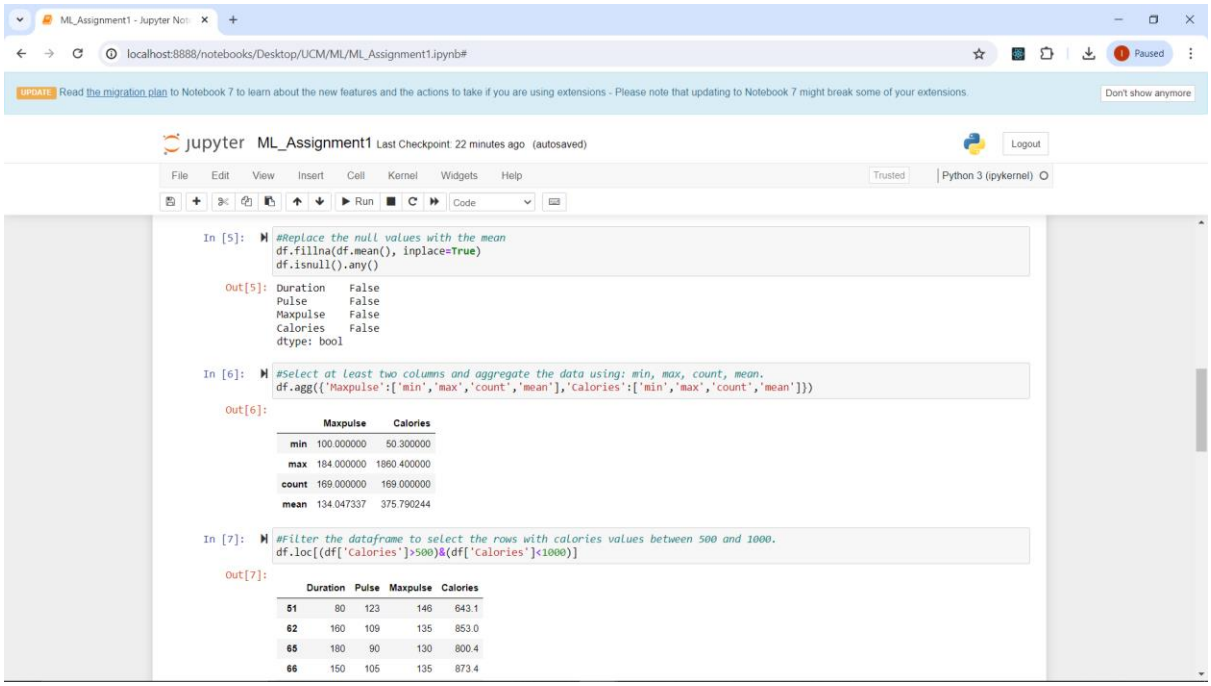
9. Convert the datatype of Calories column to int datatype.

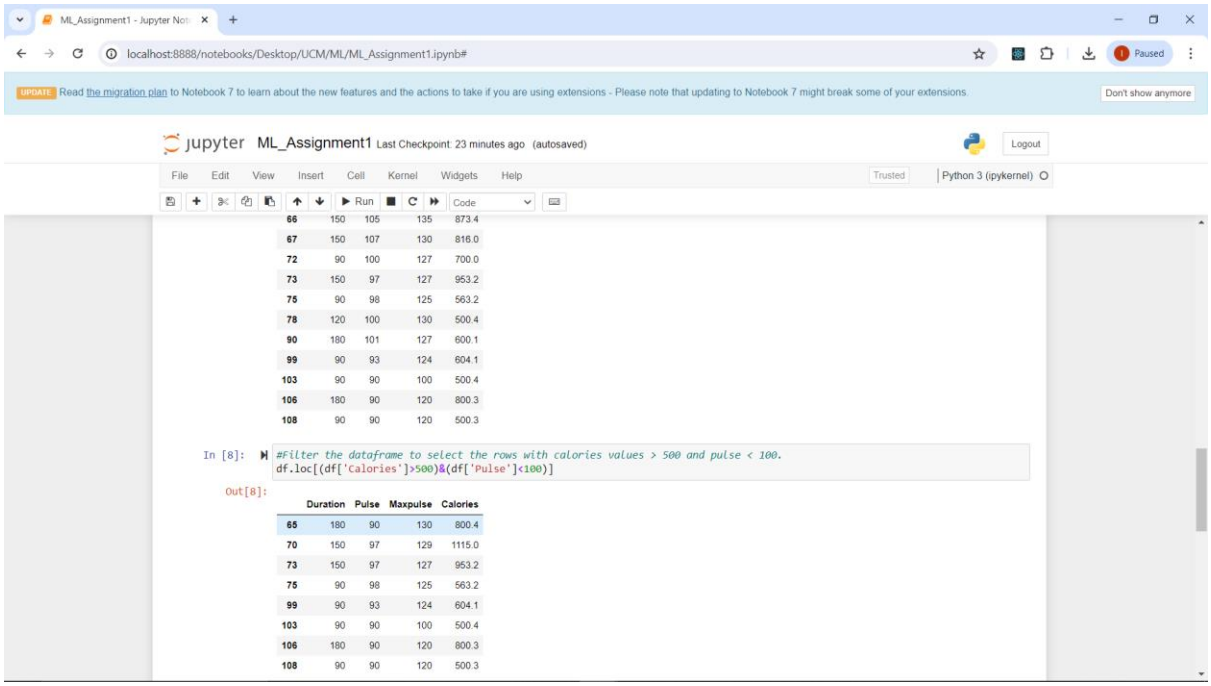
Github link: https://github.com/Likhitha78270/ML_Assignment1/tree/main

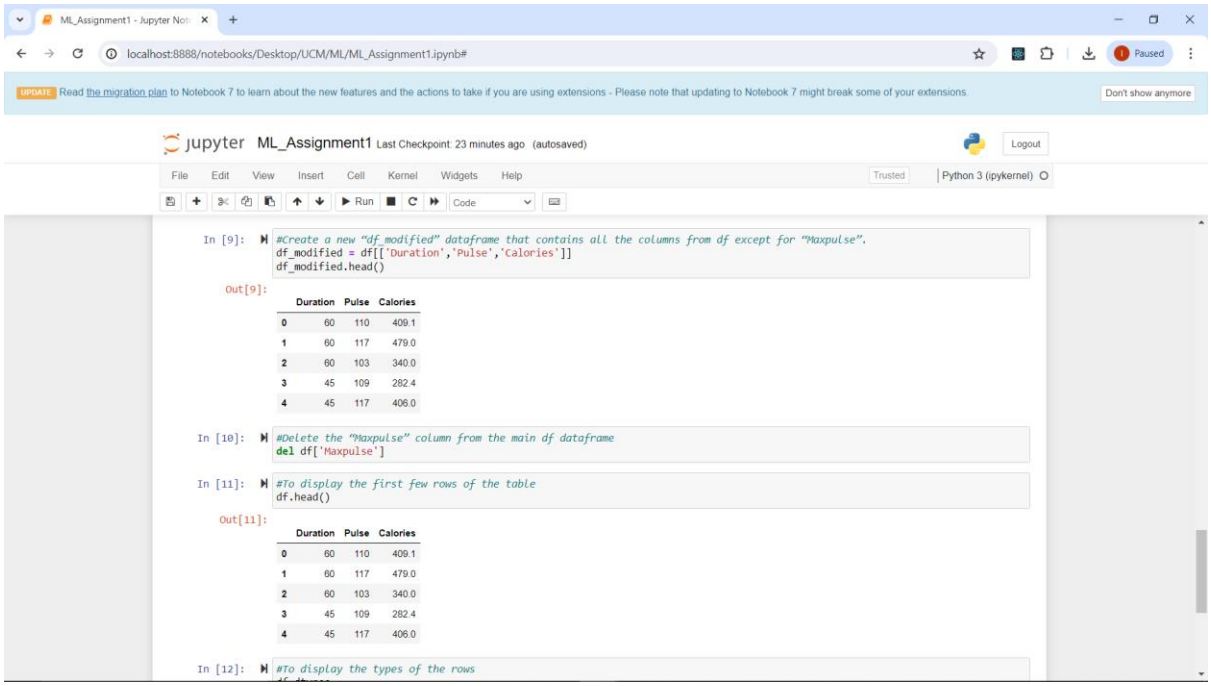
Execution video link: https://drive.google.com/file/d/1QAYnYfo54XLcSEfkPXL1FV-g4U0wtwXZ/view?usp=sharing

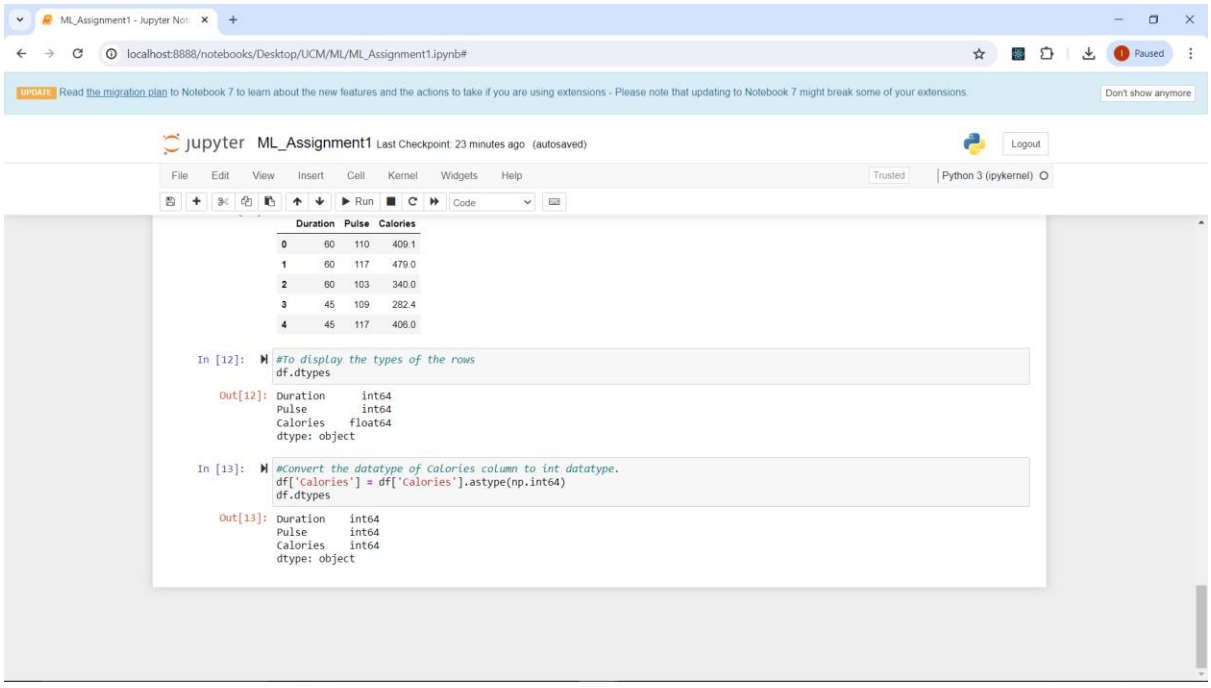

ML_Assignment1 - Jupyter Not
localhost:8888/notebooks/Desktop/UCM/ML/ML_Assignment1.ipynb#
Paused
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.
Don't show anymore
jupyter ML_Assignment1 Last Checkpoint: an hour ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
In [1]:
#Likhitha Vempati
#700757827
import warnings
import numpy as np
import pandas as pd
warnings.filterwarnings("ignore")
In [2]:
#Read the provided CSV file 'data.csv'
df = pd.read_csv("data.csv")
df.head()
Out[2]:
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 [3]:
#description about the data.
df.describe()
Out[3]:
Duration Pulse Maxpulse Calories
count 169.000000 169.000000 169.000000 164.000000
mean 63.846154 107.461538 134.047337 375.790244


ML_Assignment1 - Jupyter Not
localhost:8888/notebooks/Desktop/UCM/ML/ML_Assignment1.ipynb#
Paused
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.
Don't show anymore
jupyter ML_Assignment1 Last Checkpoint: 21 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
In [3]:
#description about the data.
df.describe()
Out[3]:
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
In [4]:
#if the data has null values.
df.isnull().any()
Out[4]:
Duration False
Pulse False
Maxpulse False
Calories True
dtype: bool
In [5]:
#Replace the null values with the mean
df.fillna(df.mean(), inplace=True)
df.isnull().any()
Out[5]:
Duration False
Pulse False


ML_Assignment1 - Jupyter Not
localhost:8888/notebooks/Desktop/UCM/ML/ML_Assignment1.ipynb#
Paused
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.
Don't show anymore
jupyter ML_Assignment1 Last Checkpoint: 22 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
In [5]:
#Replace the null values with the mean
df.fillna(df.mean(), inplace=True)
df.isnull().any()
Out[5]:
Duration False
Pulse False
Maxpulse False
Calories False
dtype: bool
In [6]:
#Select at least two columns and aggregate the data using: min, max, count, mean.
df.agg({'Maxpulse':['min','max','count','mean'],'Calories':['min','max','count','mean']})
Out[6]:
Maxpulse Calories
min 100.000000 50.300000
max 184.000000 1860.400000
count 169.000000 169.000000
mean 134.047337 375.790244
In [7]:
#filter the dataframe to select the rows with calories values between 500 and 1000.
df.loc[(df['Calories']>500)&(df['Calories']<1000)]
Out[7]:
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


ML_Assignment1 - Jupyter Not
localhost:8888/notebooks/Desktop/UCM/ML/ML_Assignment1.ipynb#
Paused
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.
Don't show anymore
jupyter ML_Assignment1 Last Checkpoint: 23 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
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
In [8]:
#filter the dataframe to select the rows with calories values > 500 and pulse < 100.
df.loc[(df['Calories']>500)&(df['Pulse']<100)]
Out[8]:
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





Here the program imports required libraries for visualization, data processing, machine learning, handling errors. A dataset is loaded and represented in the form of DataFrame. To provide the brief description of the data, the code shows top five rows of the DataFrame with describe() method. For numerical columns in DataFrame, it computes descriptive statistics like count, mean, standard deviation. Program checks for any missing/null values and returns a Boolean value(True/False) .The code replaces the mean value for any null data. The code selects two columns namely Maxpulse and Calories to combine them using min, max, count and mean from the DataFrame. And DataFrame is filtered according to criteria selecting rows where Calories column >500 and <1000 or where Calories >500 and pulse <100. The duration, pulse, and calories columns from the original DataFrame are the sole columns in the newly formed DataFrame, df_modified. This altered DataFrame's initial

few rows are shown. The DataFrame's 'Maxpulse' column gets removed. The code shows datatype and converts the 'Calories' column's data type to a 64-bit integer type (int64).