ML_Assignment 3

Name: Divya Kudipudi

ld: 700765919

- 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.



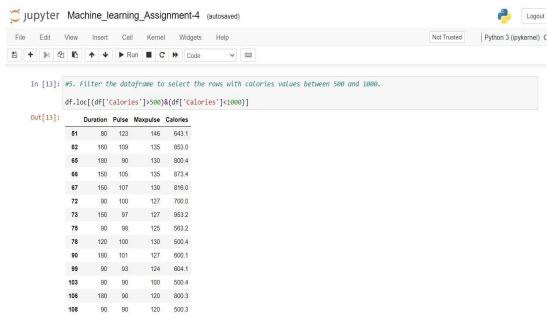
The code imports required libraries for working with tabular data and implementing algorithms, including numpy, pandas, seaborn, matplotlib, scipy, and scikit-learn. It also suppresses warnings. The provided CSV file 'data.csv' is read using pandas into a dataframe called 'df', and the first few rows of the dataframe are displayed using the 'head()' method

The code calls the 'describe()' method on the dataframe 'df', which returns basic statistical information about the data such as count, mean, standard deviation, minimum value, maximum value, and quartile values for each numerical column in the dataframe

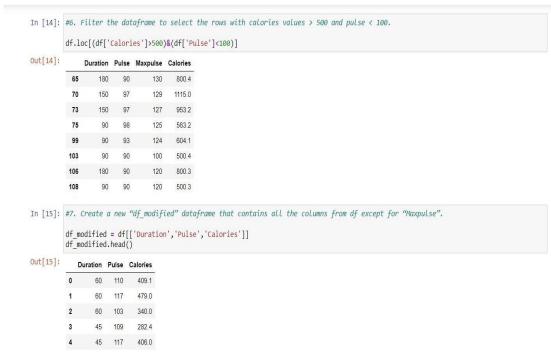
```
#3. Check if the data has null values.
   df.isnull().any()
Duration False
Pulse
           False
Maxpulse
           False
Calories
           True
dtype: bool
   #Replace the null values with the mean
   df.fillna(df.mean(), inplace=True)
   df.isnull().any()
Duration
          False
Pulse
           False
Maxpulse
          False
Calories
          False
dtype: bool
   #4. Select at least two columns and aggregate the data using: min, max, count, mean.
   df.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
```

The code first calls the 'isnull().any()' method on the dataframe 'df', which checks if there are any null values in the dataframe and returns a boolean value for each column indicating if it contains null values. Then, it fills the null values with the mean using the 'fillna()' method and the 'mean()' function, and checks again for null values using 'isnull().any()'.

Finally, the code selects two columns ('Maxpulse' and 'Calories') and aggregates the data using the 'agg()' method. The aggregation operations used are 'min', 'max', 'count', and 'mean'. The resulting dataframe shows the minimum, maximum, count, and mean values for each selected column.



The code filters the dataframe 'df' using the 'loc[]' method to select rows where the 'Calories' column has values between 500 and 1000 (exclusive). The resulting dataframe contains only the rows that satisfy the condition.



The code filters the dataframe 'df' to select rows where the 'Calories' column has values greater than 500 and the 'Pulse' column has values less than 100, using the 'loc[]' method. The resulting dataframe contains only the rows that satisfy the condition.

Then, the code creates a new dataframe 'df_modified' by selecting all columns from the original dataframe 'df' except for 'Maxpulse'. This is done by indexing the dataframe with a list of the desired column names. The resulting dataframe contains only the columns 'Duration', 'Pulse', and 'Calories',



and is assigned to the variable 'df_modified'. The code deletes the 'Maxpulse' column from the dataframe 'df' using the 'del' statement.

Then, the code converts the datatype of the 'Calories' column from float to integer using the 'astype()' method and the numpy 'int64' datatype. The resulting dataframe 'df' contains the modified 'Calories' column with integer datatype.

Finally, the 'dtypes' attribute is used to display the datatypes of all columns in the dataframe 'df'.

Github: https://github.com/Divyakudipudi7/Assignment-3

Video Link:

https://drive.google.com/file/d/1kDrPmkkIDk2ydvmPMvYkS0mbTS5bVALC/view?usp=share_link