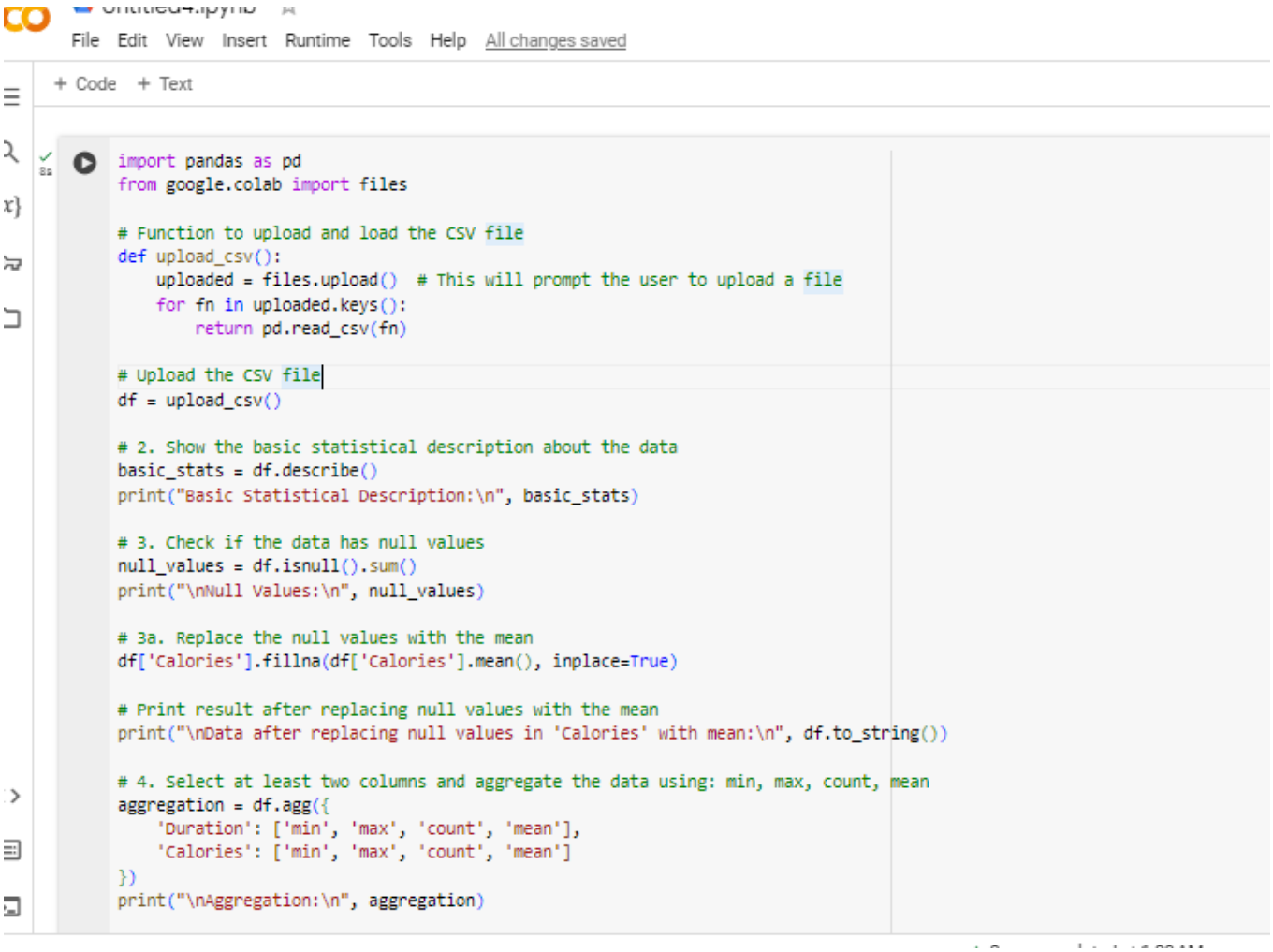
Summer 2024: Machine Learning (Assignment 3)

Ramya Sadhineni 700757305

GitHub link: https://github.com/RamyaSadhineni/Repo CRN:30562

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.

Code snippets :



```
+ Code + Text

# 5. Filter the dataframe to select the rows with calories values between 500 and 1000
df_filtered_1 = df[(df['Calories'] >= 500) & (df['Calories'] <= 1000)]
print("\nFiltered DataFrame (Calories between 500 and 1000):\n", df_filtered_1)

# 6. Filter the dataframe to select the rows with calories values > 500 and pulse < 100
df_filtered_2 = df[(df['Calories'] > 500) & (df['Pulse'] < 100)]
print("\nFiltered DataFrame (Calories > 500 and Pulse < 100):\n", df_filtered_2)

# 7. Create a new "df_modified" dataframe that contains all the columns from df except for "Maxpulse"
df_modified = df.drop(columns=['Maxpulse'])
print("\nModified DataFrame without Maxpulse:\n", df_modified.to_string())

# 8. Delete the "Maxpulse" column from the main df dataframe
df.drop(columns=['Maxpulse'], inplace=True)

# Print result after dropping Maxpulse
print("\nMain DataFrame after dropping Maxpulse:\n", df.to_string())

# 9. Convert the datatype of Calories column to int datatype
df['Calories'] = df['Calories'].astype(int)

# Print the final dataframe
print("\nMain DataFrame after converting 'Calories' to int:\n", df.to_string())
```

```
Choose Files data.csv
• data.csv(text/csv) - 2858 bytes, last modified: 6/11/2024 - 100% done
Saving data.csv to data (2).csv
Basic Statistical Description:
          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
```

Explanation :

This code is designed to be executed in a Google Colab notebook and performs several data manipulation tasks on a CSV file. It begins by importing the necessary libraries (`pandas` for data manipulation and `files` from `google.colab` for file upload). The `upload_csv` function is defined to prompt the user to upload a CSV file and load it into a pandas DataFrame. The script then proceeds to display the basic statistical description of the data and checks for null values. It replaces any null values in the "Calories" column with the column's mean and prints the entire DataFrame to show the changes. Subsequent steps involve aggregating data for selected columns, filtering rows based on specific conditions, creating a modified DataFrame without the "Maxpulse" column, deleting the "Maxpulse" column from the main DataFrame, and converting the "Calories" column to an integer type. Each significant step prints the resulting DataFrame to show the modifications applied.

Summary of Steps:

1. Upload a CSV file.
2. Display basic statistics of the data.
3. Check and replace null values in the "Calories" column with the mean.
4. Aggregate data using min, max, count, and mean for selected columns.
5. Filter the DataFrame based on specific "Calories" and "Pulse" conditions.
6. Create a new DataFrame excluding the "Maxpulse" column and modify the main DataFrame to drop this column.
7. Convert the "Calories" column to integer type and display the final DataFrame.