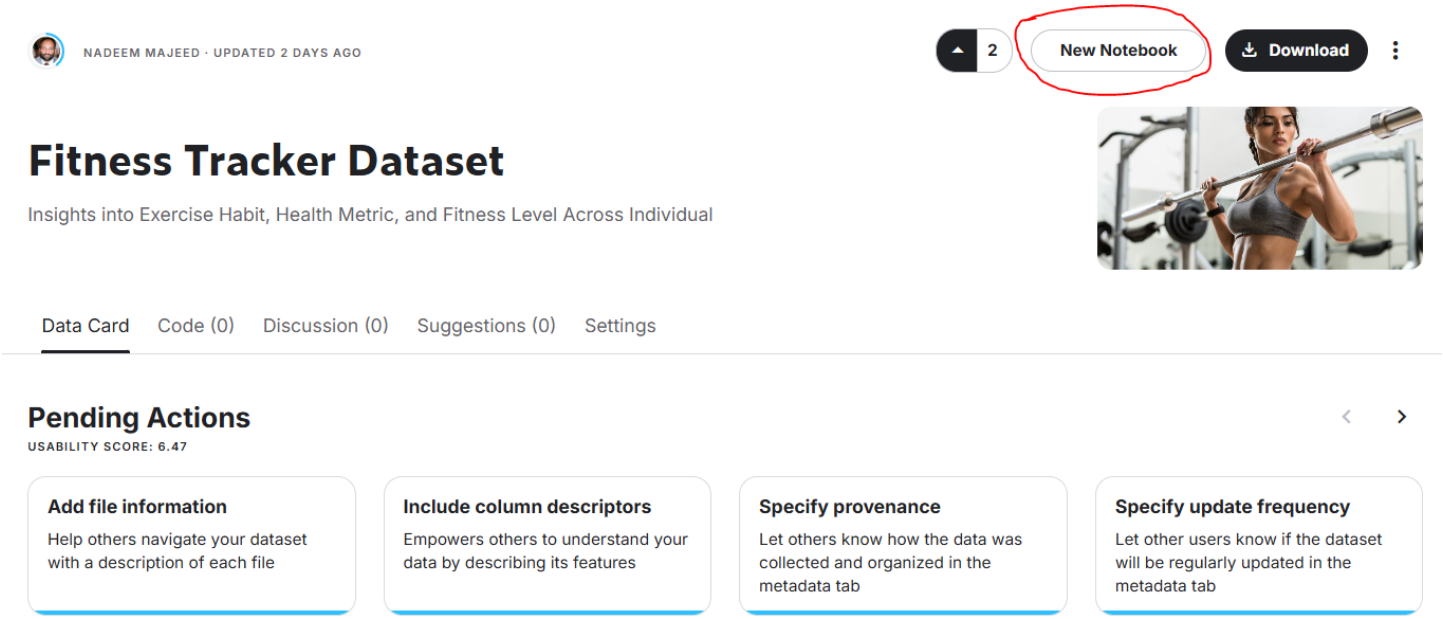
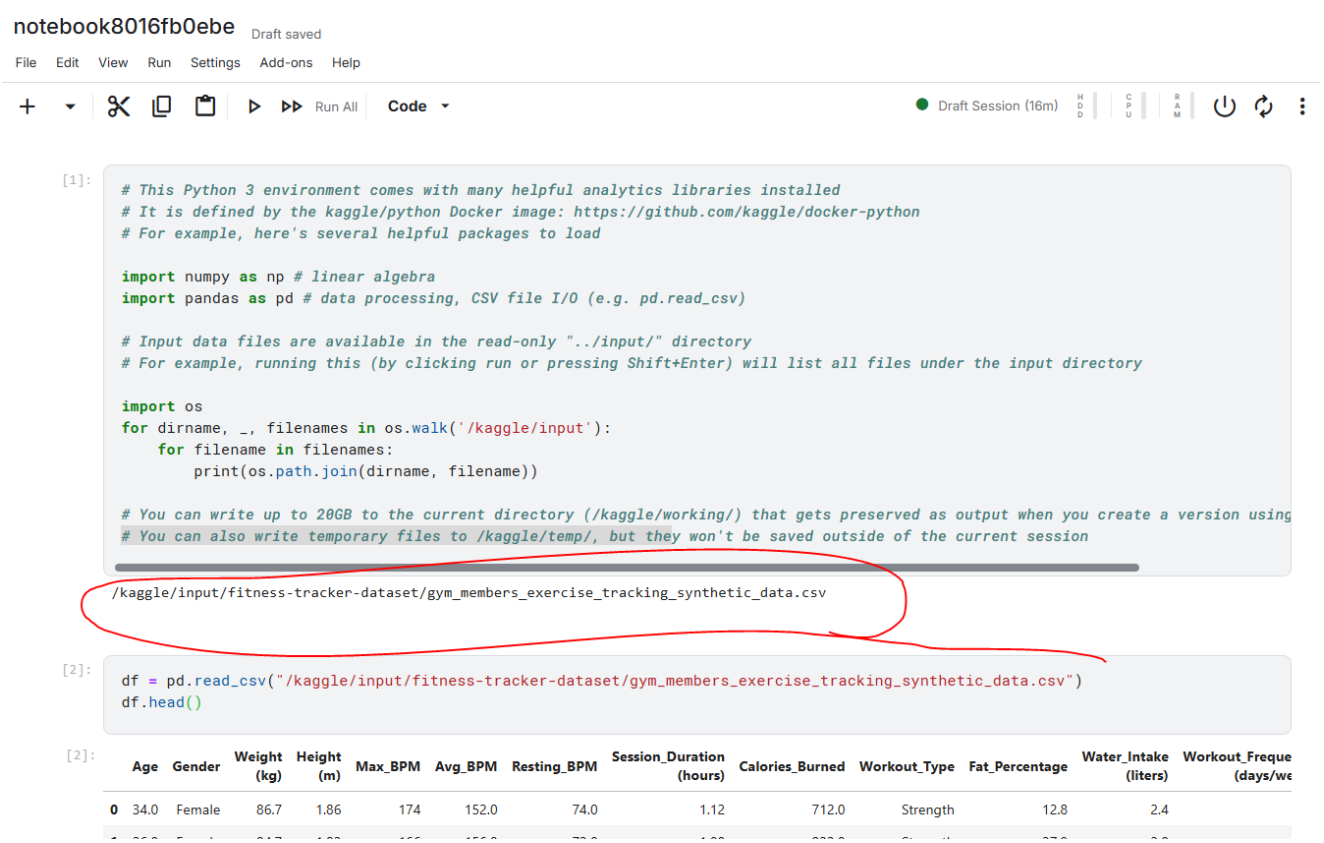
Assignment Pandas
Please use this dataset.

https://www.kaggle.com/datasets/nadeemajeedch/fitness-tracker-dataset

login with your Kaggle account. And click the new Notebook.



You will get the following Notebook page.

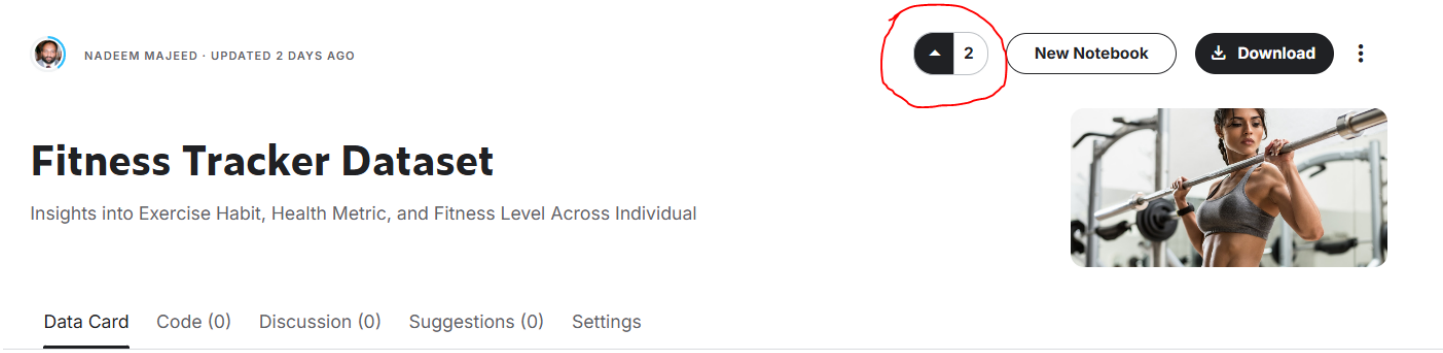


On execution of the cell, you will get the path of the file. Use this path to load the data file.

df= pd.read_csv("/kaggle/input/fitness-tracker-dataset/gym_members_exercise_tracking_synthetic_data.csv")

Click here and change the file name.

Don't forget to vote for the dataset. 😊



Tasks: Preprocessing and EDA Steps

Step 1: Load the Data

- Import the necessary libraries (`pandas`, `numpy`, `matplotlib`, `seaborn`).
- Load the dataset into a pandas DataFrame using `pd.read_csv()` or from the provided file.
- Display the first few rows using `.head()` to understand the structure.

Step 2: Inspect the Data

1. **Shape**: Use `.shape` to check the number of rows and columns.
2. **Columns**: Use `.columns` to list column names.
3. **Info**: Use `.info()` to examine data types and non-null counts for each column.
4. **Description**: Use `.describe()` to summarize numeric columns (mean, min, max, etc.).

Step 3: Identify and Handle Missing Values

1. Use `.isnull().sum()` to check the number of missing values per column.
2. Visualize missing data using a heatmap (`sns.heatmap`) to identify patterns.
3. Handle missing values:
    - **Age, Weight (kg), Height (m), and Numeric Columns**:
        - Impute missing values using the mean or median.
    - **Gender and Workout_Type**:
        - Impute missing values using the mode (most frequent value).
    - Document and justify your imputation strategy.

Step 4: Check for Duplicates

1. Use `.duplicated().sum()` to check for duplicate rows.
2. Remove duplicates, if any, using `.drop_duplicates()`.

Step 5: Validate Data

1. **Numeric Columns**:
    - Check for invalid entries (e.g., special characters like `?` in `Max_BPM`).
    - Convert columns like `Max_BPM` to numeric using `pd.to_numeric()` with `errors='coerce'`.
    - Replace invalid values with `NaN` and impute as needed.
2. **Categorical Columns**:
    - Use `.unique()` to check for inconsistencies in `Gender`, `Workout_Type`, etc.
    - Standardize inconsistent values (e.g., `Male`, `M` → `Male`).

Step 6: Create New Features

1. **BMI Validation**:
    - Verify if the `BMI` column is consistent with the formula:
      BMI=Weight (kg)Height (m)2\text{BMI} = \frac{\text{Weight (kg)}}{\text{Height (m)}^2}BMI=Height (m)2Weight (kg)
    - Recalculate BMI where inconsistencies are found.
2. **Workout Intensity**:
    - Create a new column: `Workout_Intensity = Avg_BPM / Max_BPM`.

Step 7: Explore Data Distributions

1. Plot histograms for numeric columns (`Age`, `Calories_Burned`, etc.) to understand their distributions.
2. Use box plots to check for outliers in columns like `Age`, `BMI`, and `Calories_Burned`.
3. Examine the distribution of categorical columns (`Gender`, `Workout_Type`) using bar plots.

Step 8: Handle Outliers

1. Use the IQR method to detect outliers in numeric columns.
2. Decide whether to:
    - Remove outliers.
    - Transform them (e.g., log transformation).
    - Cap them (e.g., set to a specific threshold).

Step 9: Analyze Relationships

1. **Correlation**:
    - Use `.corr()` and visualize with a heatmap to find relationships between numeric columns (e.g., `Calories_Burned`, `Session_Duration`).
2. **Categorical vs. Numeric**:
    - Compare `Calories_Burned` and `Workout_Type` using a bar plot.
    - Analyze differences in `BMI` across `Gender` using a box plot.
3. **Multi-Variable**:
    - Use pair plots (`sns.pairplot`) to analyze relationships between key metrics (e.g., `BMI`, `Calories_Burned`, `Workout_Frequency`).

Step 10: Encode Categorical Variables

1. Convert `Gender` and `Workout_Type` to numeric formats:
    - Use one-hot encoding (`pd.get_dummies()`) or label encoding.

Step 11: Normalize Numeric Columns

1. Normalize or standardize columns with large ranges (`Calories_Burned`, `Session_Duration`, etc.) if needed for further analysis or machine learning.

Step 12: Summarize Findings

1. Highlight key insights from the data:
    - Trends in `Calories_Burned` based on `Workout_Type` and `Gender`.
    - Correlation between `Session_Duration` and `Calories_Burned`.
    - Any notable differences in `BMI` across `Workout_Frequency`.

If you think any other related task, you can add in notebook.

Happy Learning 😊