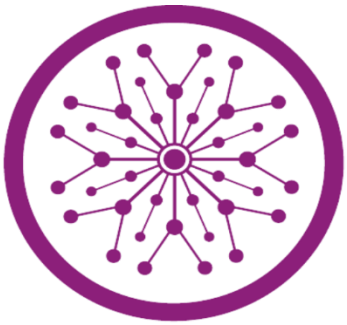


Name: Muhammad-Jawad-Haider

Roll No.: SU92-BSAIM-F24-056

Section: BSAI-3A

Subject: Artificial Intelligence

Department: Software Engineering

Project: Student Performance Analysis

Assignment: 1

Student Performance Analysis (Data Cleaning & EDA)

Objective:

The goal of this assignment is to explore and clean the dataset that contains information about students' marks, attendance, and study hours.
We will check the structure of the data, handle missing values, and prepare a clean version of the dataset for model training.

Dataset Description:

The dataset contains 7 columns and 1000 rows:

- Assignment Marks – marks obtained in assignments
- Quiz Marks – marks scored in quizzes
- Attendance Percentage – attendance of the student in percentage
- Mid Marks – marks in mid-term exam
- Final Marks – marks in final exam
- Study Hours – number of study hours per day
- Total Marks – total marks out of 100 (target variable)

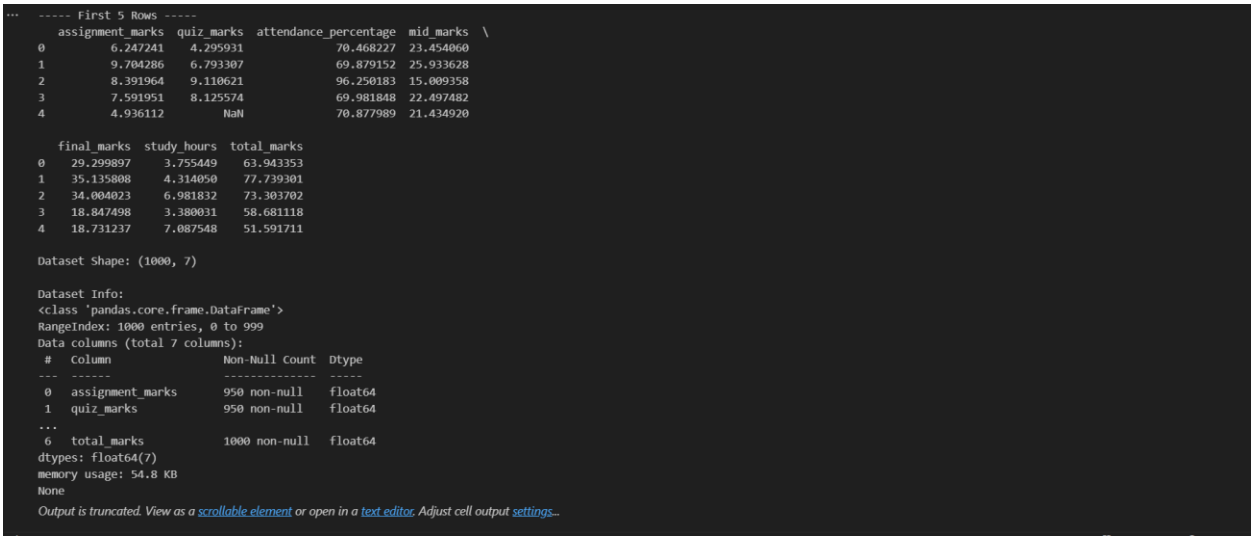
Steps Performed:

1. **Imported required libraries** – pandas, numpy, matplotlib, seaborn.
2. **Loaded the dataset** from the Excel file.
3. **Checked data information** using .head(), .shape(), and .info().
4. **Handled missing values** by replacing all NaN values with 0.
5. **Viewed summary statistics** using .describe().
6. **Plotted a correlation heatmap** to observe relationships between variables.
7. **Saved the clean dataset** as clean_student_data.csv for model training in Assignment 2.

Conclusion:

After cleaning, the dataset has no missing values and is ready to be used for building a regression model. The correlation plot shows that assignments, mid-term, and final marks strongly affect the total marks.

- **Basic Info:**



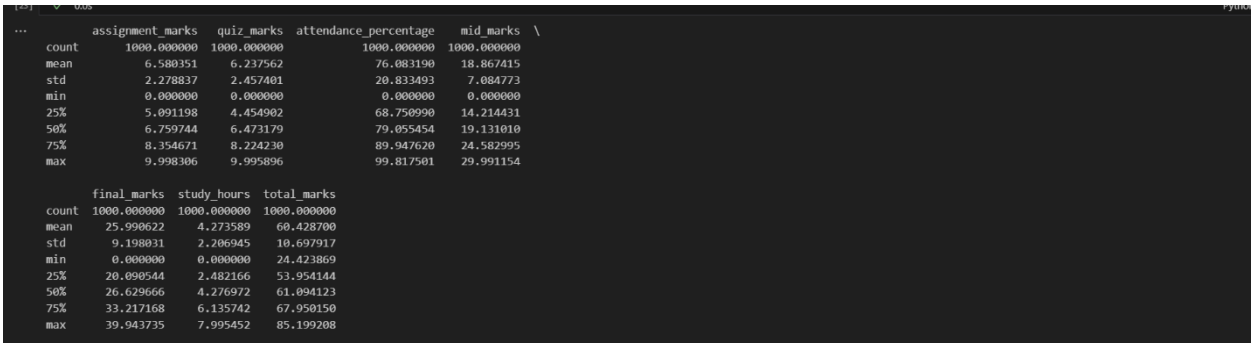
```
----- First 5 Rows -----
   assignment_marks  quiz_marks  attendance_percentage  mid_marks  \
0          6.247241    4.295931              70.468227  23.454060
1          9.704286    6.793307              69.879152  25.933628
2          8.391964    9.110621              96.250183  15.009358
3          7.591951    8.125574              69.981848  22.497482
4          4.936112         NaN              70.877989  21.434920

   final_marks  study_hours  total_marks
0    29.299897     3.755449    63.943353
1    35.135808     4.314050    77.739301
2    34.004023     6.981832    73.303702
3    18.847498     3.380031    58.681118
4    18.731237     7.087548    51.591711

Dataset Shape: (1000, 7)

Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 7 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   assignment_marks      950 non-null    float64
 1   quiz_marks            950 non-null    float64
...
 6   total_marks           1000 non-null   float64
dtypes: float64(7)
memory usage: 54.8 KB
None
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
```

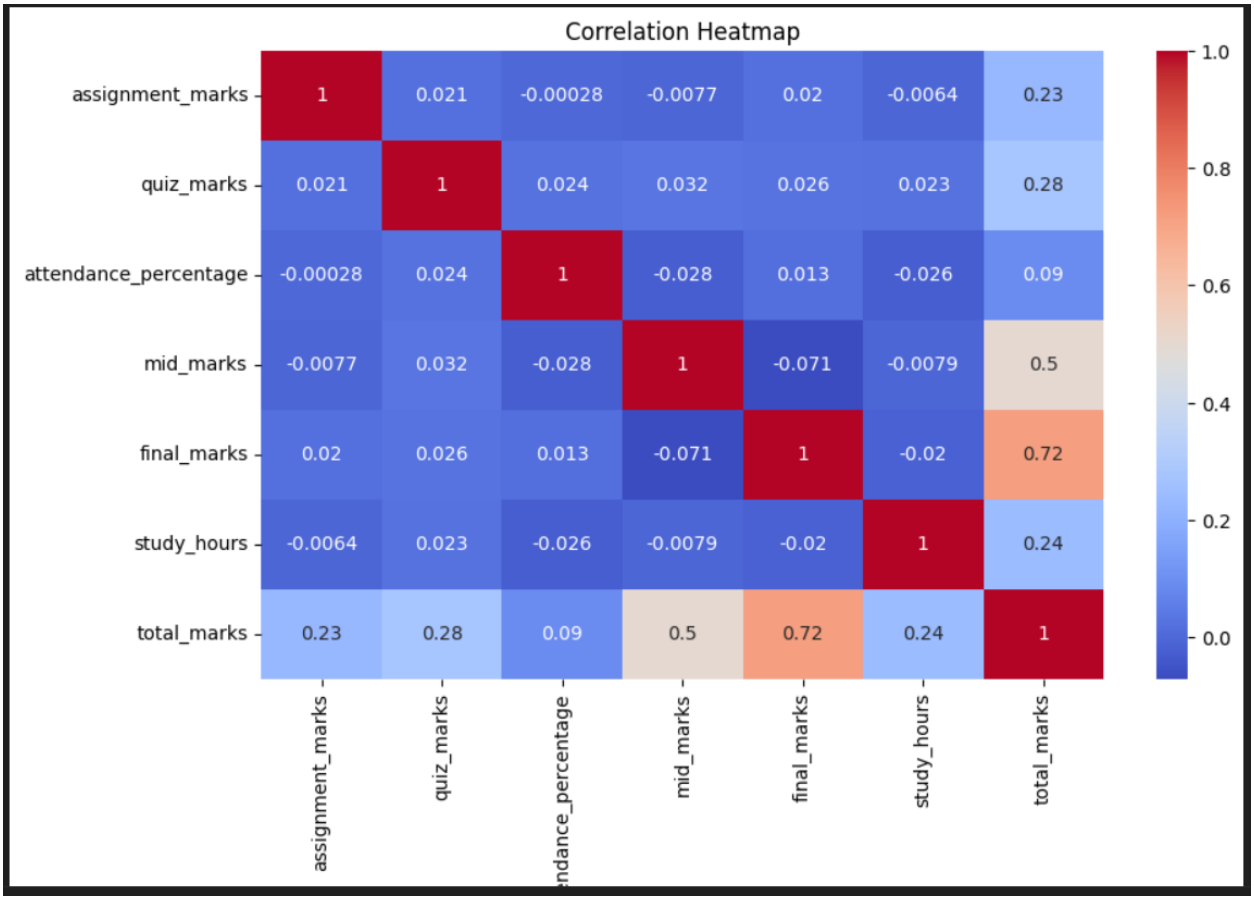
- **Statical Summary:**



```
       assignment_marks   quiz_marks  attendance_percentage    mid_marks  \
count       1000.000000  1000.000000            1000.000000  1000.000000
mean           6.580351     6.237562              76.083190    18.867415
std            2.278837     2.457401              20.833493     7.084773
min            0.000000     0.000000               0.000000     0.000000
25%            5.091198     4.454902              68.750990    14.214431
50%            6.759744     6.473179              79.055454    19.131010
75%            8.354671     8.224230              89.947620    24.582995
max            9.998306     9.995896              99.817501    29.991154

       final_marks  study_hours  total_marks
count  1000.000000  1000.000000  1000.000000
mean     25.990622     4.273589    60.428700
std       9.198031     2.206945    10.697917
min       0.000000     0.000000    24.423869
25%      20.090544     2.482166    53.954144
50%      26.629666     4.276972    61.094123
75%      33.217168     6.135742    67.950150
max      39.943735     7.995452    85.199208
```

- **Heatmap:**


Correlation Heatmap
assignment_marks
quiz_marks
attendance_percentage
mid_marks
final_marks
study_hours
total_marks
1 0.021 -0.00028 -0.0077 0.02 -0.0064 0.23
0.021 1 0.024 0.032 0.026 0.023 0.28
-0.00028 0.024 1 -0.028 0.013 -0.026 0.09
-0.0077 0.032 -0.028 1 -0.071 -0.0079 0.5
0.02 0.026 0.013 -0.071 1 -0.02 0.72
-0.0064 0.023 -0.026 -0.0079 -0.02 1 0.24
0.23 0.28 0.09 0.5 0.72 0.24 1
1.0
0.8
0.6
0.4
0.2
0.0