# AMERICAN INTERNATIONAL UNIVERSITY-BANGLADESH

## Faculty of Science and Technology

# Mid Project

| | | | |
|---|---|---|---|
| Assignment Title: | Applying various data exploration and preprocessing techniques. | | |
| Assignment No: | Click here to enter text. | Date of Submission: | 27 April 2025 |
| Course Title: | Introduction to Data Science | | |
| Course Code: | Click here to enter text. | Section: | G |
| Semester: | Spring 2024-25 | Course Teacher: | **DR. ASHRAF UDDIN** |

**Declaration and Statement of Authorship:**

1. I/we hold a copy of this Assignment/Case-Study, which can be produced if the original is lost/damaged.
2. This Assignment/Case-Study is my/our original work and no part of it has been copied from any other student's work or from any other source except where due acknowledgement is made.
3. No part of this Assignment/Case-Study has been written for me/us by any other person except where such collaboration has been authorized by the concerned teacher and is clearly acknowledged in the assignment.
4. I/we have not previously submitted or currently submitting this work for any other course/unit.
5. This work may be reproduced, communicated, compared and archived for the purpose of detecting plagiarism.
6. I/we give permission for a copy of my/our marked work to be retained by the faculty for review and comparison, including review by external examiners.
7. I/we understand that Plagiarism is the presentation of the work, idea or creation of another person as though it is your own. It is a formofcheatingandisaveryseriousacademicoffencethatmayleadtoexpulsionfromtheUniversity. Plagiarized material can be drawn from, and presented in, written, graphic and visual form, including electronic data, and oral presentations. Plagiarism occurs when the origin of their arterial used is not appropriately cited.
8. I/we also understand that enabling plagiarism is the act of assisting or allowing another person to plagiarize or to copy my/our work.

* *Student(s) must complete all details except the faculty use part.*
** Please submit all assignments to your course teacher or the office of the concerned teacher.

Group Name/No.: 6

| No | Name | ID | Program | Signature |
|---|---|---|---|---|
| 1 | MD. TAHSIN HASIB | 22-46026-1 | BSc [CSE] | |
| 2 | **MUHTADI MANSIB** | 22-47083-1 | BSc [CSE] | |
| 3 | **MONISHA SARKAR** | 22-47063-1 | BSc [CSE] | |
| 4 | **MAHMUDA AKTER MUNNI** | 22-46495-1 | BSc [CSE] | |
| 5 | | | Choose an item. | |
| 6 | | | Choose an item. | |
| 7 | | | Choose an item. | |
| 8 | | | Choose an item. | |
| 9 | | | Choose an item. | |
| 10 | | | Choose an item. | |

| *Faculty use only* | | |
|---|---|---|
| FACULTY COMMENTS | **Marks Obtained** | |
| | **Total Marks** | |

## 1.Dataset Creation

The dataset used in this study, titled ***"Student Performance and Learning Style,"*** was obtained from Kaggle. It contains approximately **10,000 records** of students' academic and learning-related information. Each record represents a unique student and captures various attributes related to their demographic profile, academic performance, learning behavior, and preferences.

Dataset Link: https://www.kaggle.com/datasets/adilshamim8/student-performance-and-learning-style?select=student_performance_large_dataset.csv

Key features included in the dataset are:

- **Age**: The age of the student.
- **Gender**: The gender identity of the student.
- **Study_Hours_per_Week**: Average number of hours spent studying per week.
- **Exam_Score....**: Final exam score obtained by the student.
- **Attendance_Rate....**: Percentage of classes attended.
- **Assignment_Completion_Rate....**: Percentage of assignments completed.
- **Preferred_Learning_Style**: The student's most effective learning method (e.g., visual, auditory, kinesthetic).
- **Online_Courses_Completed**: Number of online courses completed by the student.

## 2. Data Preprocessing:

### Original Dataset Summary:

```
  Student_ID            Age           Gender         Study_Hours_per_Week Preferred_Learning_Style Online_Courses_Completed Participation_in_Discussions
 Length:10000       Min.   :18.00   Length:10000       Min.   : 5.00       Length:10000            Min.   : 0.00            Length:10000
 Class :character   1st Qu.:20.00   Class :character   1st Qu.:16.00       Class :character        1st Qu.: 5.00            Class :character
 Mode  :character   Median :23.00   Mode  :character   Median :27.00       Mode  :character        Median :10.00            Mode  :character
                    Mean   :23.48                      Mean   :27.13                               Mean   :10.01
                    3rd Qu.:27.00                      3rd Qu.:38.00                               3rd Qu.:15.00
                    Max.   :29.00                      Max.   :49.00                               Max.   :20.00
 Assignment_Completion_Rate.... Exam_Score....   Attendance_Rate.... Use_of_Educational_Tech Self_Reported_Stress_Level
 Min.   : 50.00                 Min.   : 40.00   Min.   : 50.00      Length:10000           Length:10000
 1st Qu.: 62.00                 1st Qu.: 55.00   1st Qu.: 62.00      Class :character       Class :character
 Median : 75.00                 Median : 70.00   Median : 75.00      Mode  :character       Mode  :character
 Mean   : 74.92                 Mean   : 70.19   Mean   : 75.09
 3rd Qu.: 88.00                 3rd Qu.: 85.00   3rd Qu.: 88.00
 Max.   :100.00                 Max.   :100.00   Max.   :100.00
 Time_Spent_on_Social_Media..hours.week. Sleep_Hours_per_Night Final_Grade
 Min.   : 0.00                           Min.   : 4.000        Length:10000
 1st Qu.: 7.00                           1st Qu.: 5.000        Class :character
 Median :15.00                           Median : 7.000        Mode  :character
 Mean   :14.94                           Mean   : 6.979
 3rd Qu.:23.00                           3rd Qu.: 9.000
 Max.   :30.00                           Max.   :10.000
```

**Fig-1: Original Dataset summary**

### Injection of missing values:

In order to simulate a real-world scenario where datasets often contain incomplete information, missing values were artificially introduced into selected columns of the dataset. In order to do that

a custom function named inject_na()is defined. This function takes a column from the data and a percentage value as input. It calculates how many data points need to be replaced with NA based on the specified percentage, randomly selects those positions within the column, and replaces the selected values with NA. The modified column is then returned.

The injection of missing values was performed on the following variables:

- **Age**: 10% of the values were randomly replaced with missing values.
- **Exam_Score....**: 10% of the entries were made missing.
- **Assignment_Completion_Rate....**: 15% of the data points were set to missing values.
- **Preferred_Learning_Style**: 10% missing values were introduced.
- **Study_Hours_per_Week**: 10% of the values were randomly converted to missing.

After injecting missing values, the head(df) function was used to preview the first few rows of the modified dataset:

| | Student_ID <chr> | Age <int> | Gender <chr> | Study_Hours_per_Week <int> | Preferred_Learning_Style <chr> | Online_Courses_Completed <int> | Participation_in_Discussions <chr> |
|---|---|---|---|---|---|---|---|
| 1 | S00001 | 18 | Female | 48 | Kinesthetic | 14 | Yes |
| 2 | S00002 | 29 | Female | 30 | Reading/Writing | 20 | No |
| 3 | S00003 | 20 | Female | 47 | *NA* | 11 | No |
| 4 | S00004 | *NA* | Female | 13 | Auditory | 0 | Yes |
| 5 | S00005 | 19 | Female | 24 | Auditory | 19 | Yes |
| 6 | S00006 | 28 | Female | 26 | *NA* | 5 | Yes |

6 rows | 1-8 of 15 columns

```
                    Student_ID                                  Age                               Gender 
                             0                                 1000                                    0 
            Study_Hours_per_Week                Preferred_Learning_Style           Online_Courses_Completed 
                          1000                                 1000                                    0 
    Participation_in_Discussions          Assignment_Completion_Rate....                       Exam_Score.... 
                             0                                 1500                                 1000 
              Attendance_Rate....                   Use_of_Educational_Tech          Self_Reported_Stress_Level 
                             0                                    0                                    0 
Time_Spent_on_Social_Media..hours.week.                Sleep_Hours_per_Night                          Final_Grade 
                             0                                    0                                    0 
```

**Fig-2: Example: NA is displayed as missing values**

### Sampling the Dataset

To manage computational efficiency and facilitate easier exploration during the early stages of analysis, a random subset of the dataset was selected. First, the random number generator was seeded using the set.seed(123) function. Setting a seed ensures that the random sampling process is **reproducible** — meaning the same sample will be generated each time the code is run, maintaining consistency in analysis and results. Following this, a random sampling technique was applied to the dataset using the sample() function. Specifically, **150 records** were randomly selected from the full dataset without replacement, ensuring that no data point was chosen more than once.

## 3.Handling Missing Values

After introducing missing values into the dataset, appropriate imputation techniques were applied to handle them:

- **Median Imputation for Numerical Columns:**
  For numerical variables like Age, Exam_Score.... and Assignment_Completion_Rate.... missing values were replaced with the **median** of the respective columns. Median imputation is preferred for numerical data when the distribution may be skewed, as it is more robust to outliers compared to the mean.
- **Mode Imputation for Categorical Columns:**
  A custom get_mode() function was defined to calculate the mode (the most frequently occurring value) of a categorical variable. Missing values in the Preferred_Learning_Style column were imputed using the mode ensuring that the most common category fills in the missing entries.
- **Mean Imputation for Numerical Columns:**
  For the Study_Hours_per_Week column, missing values were replaced with the **mean**. Mean imputation is appropriate when the variable has a roughly symmetric distribution without extreme outliers.
- **Final Checks and Saving the Dataset:**
  After imputation, the dataset was checked using colSums(is.na(df)) to confirm that no missing values remained. The fully cleaned dataset was then saved to a new file named "StudentPerformance_Dataset_Cleaned.csv" using the write.csv() function.

After handling the missing values a check was performed to ensure that no values remain missing:

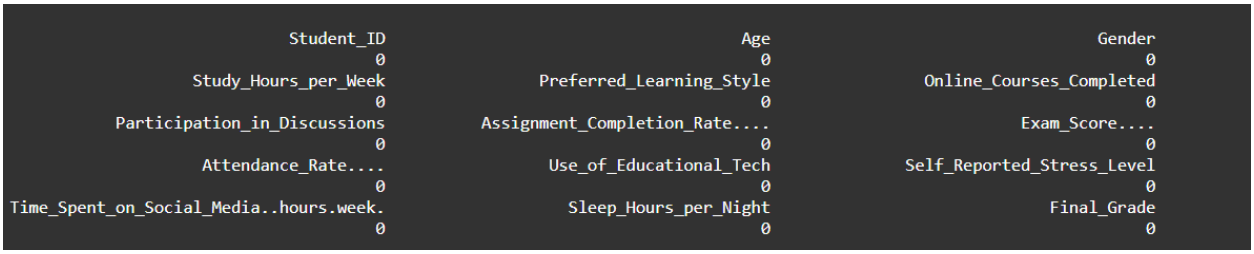



**Fig-3: Example: Number of missing values**

# 4. Data Exploration

## Univariate Analysis

### Histogram

The histograms show variables mostly measured on the ratio scale. Age clusters between 20–30 years, with a few reaching 80, showing a right-skewed distribution. Study Hours per Week mostly fall between 0–50 hours with rare cases up to 100 hours clustering at lower values. Online Courses Completed are fairly evenly spread from 0–20 with a slight decline as the number increases. Assignment Completion Rate (%) is centered around 60–80%, peaking near 70% suggesting a slight normal shape. Exam Score (%) is tightly clustered between 50–100% with a few extreme outliers near 200%. Attendance Rate (%) spreads evenly between 50–100% showing varied attendance patterns.
social media Time is concentrated between 0–30 hours with outliers up to 80, heavily right-skewed. Sleep Hours per Night mostly range between 5–9 hours peaking around 6 hours, showing moderate skewness.

Comparison:

- Age and Social Media Time show strong right skew.
- Study Hours drop sharply after 50, while Online Courses decrease more gradually.
- Assignment Completion and Exam Scores cluster around higher percentages, though exam scores have rare extreme values. Attendance is more evenly spread.
- Sleep Hours show a more balanced, near-normal distribution compared to others.

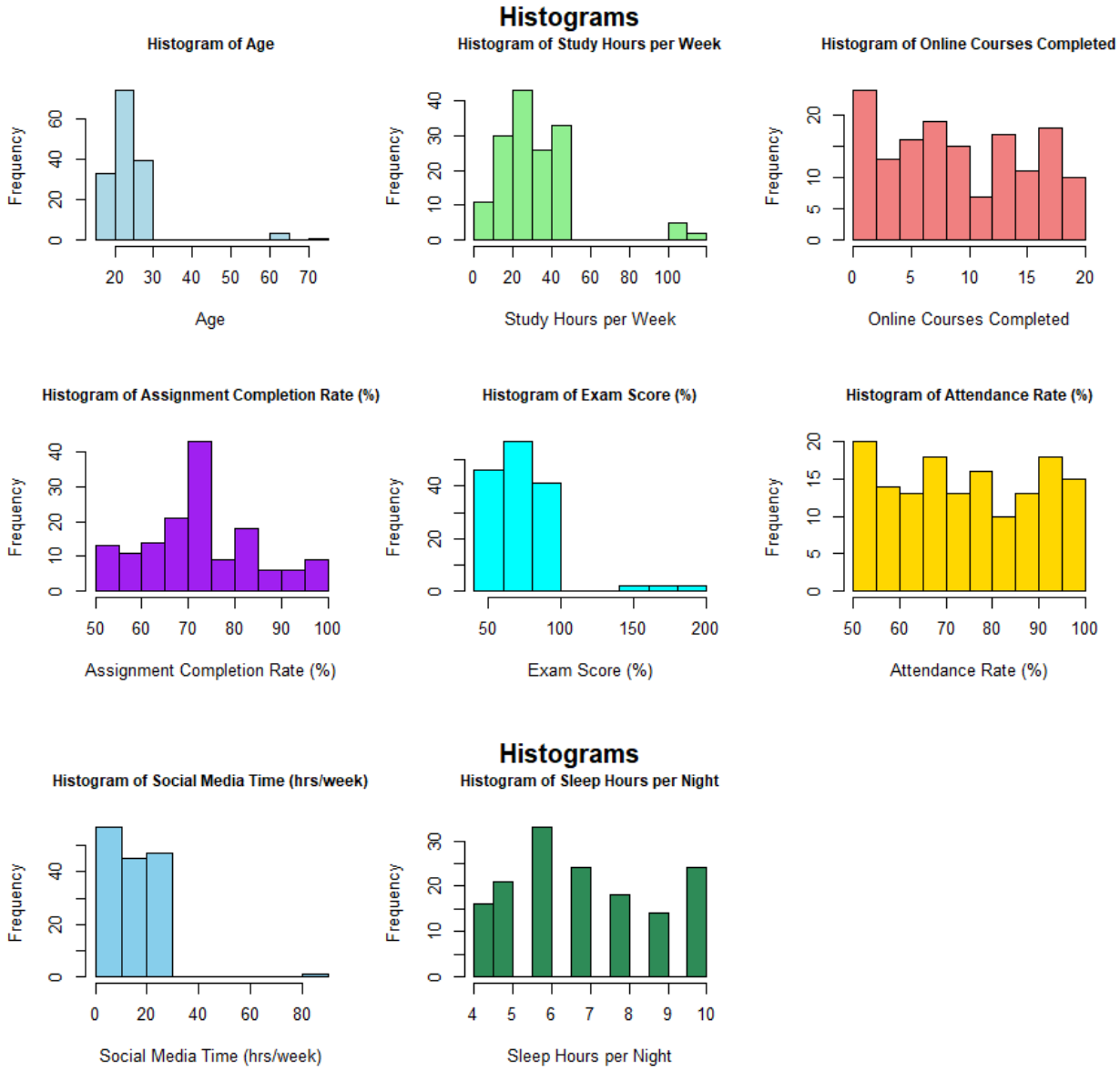


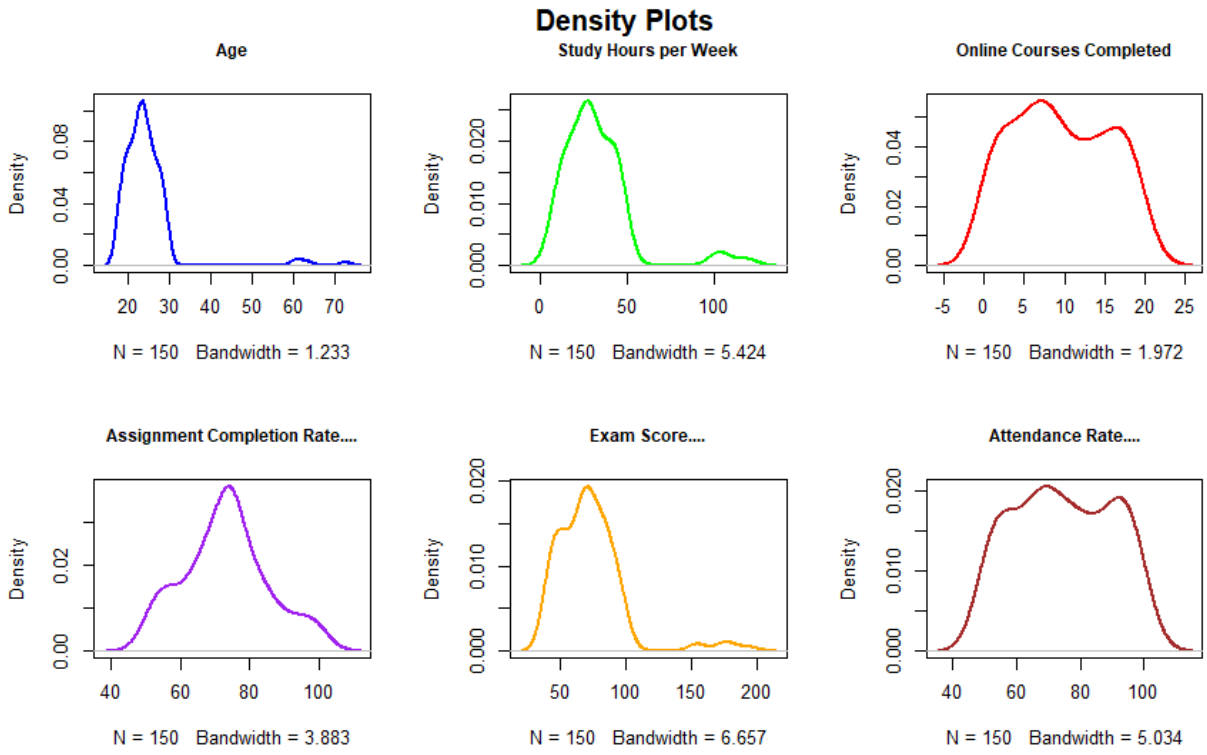

**Fig-4: Univariate Analysis: Histogram**

**Density plot**

The density plots represent continuous, ratio-level data with equal intervals. Age peaks around 20–30 years and shows a right skew. Study Hours per Week clusters between 10–40 hours, peaking at 20–30, also right-skewed. Online Courses Completed is more evenly spread with small peaks at 5 and 15 courses. Assignment Completion Rate (%) mostly falls between 60–90%, peaking near 70%. Exam Score (%) centers between 50–100%, peaking at 75%, with rare outliers nearing 200%. Attendance Rate (%) spreads broadly between 50–100% with minor peaks. Time Spent on Social Media heavily clusters below 30 hours with outliers exceeding 80 hours. Sleep Hours per Night display a bimodal pattern (or two peaks) with peaks at 6 and 8 hours.

Comparison:

- Age, Study Hours and Social Media Time show right-skewed clustering at lower values.
- Online Courses Completed and Attendance Rate are more uniform with multiple small peaks.
- Assignment Completion Rate and Exam Score are bell-shaped but slightly skewed.
- Sleep Hours are unique with a clear bimodal distribution(two peaks point).



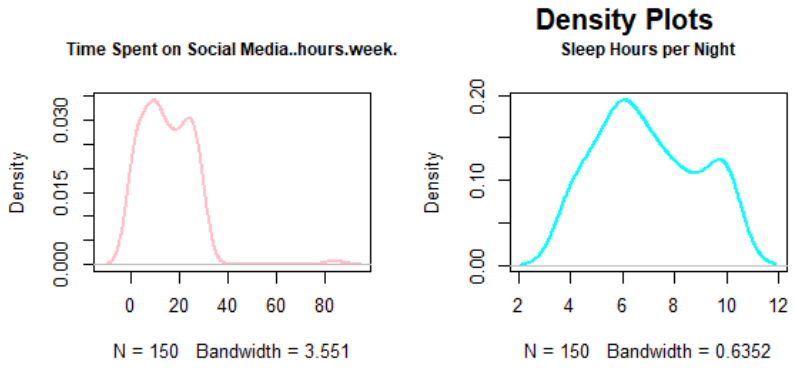


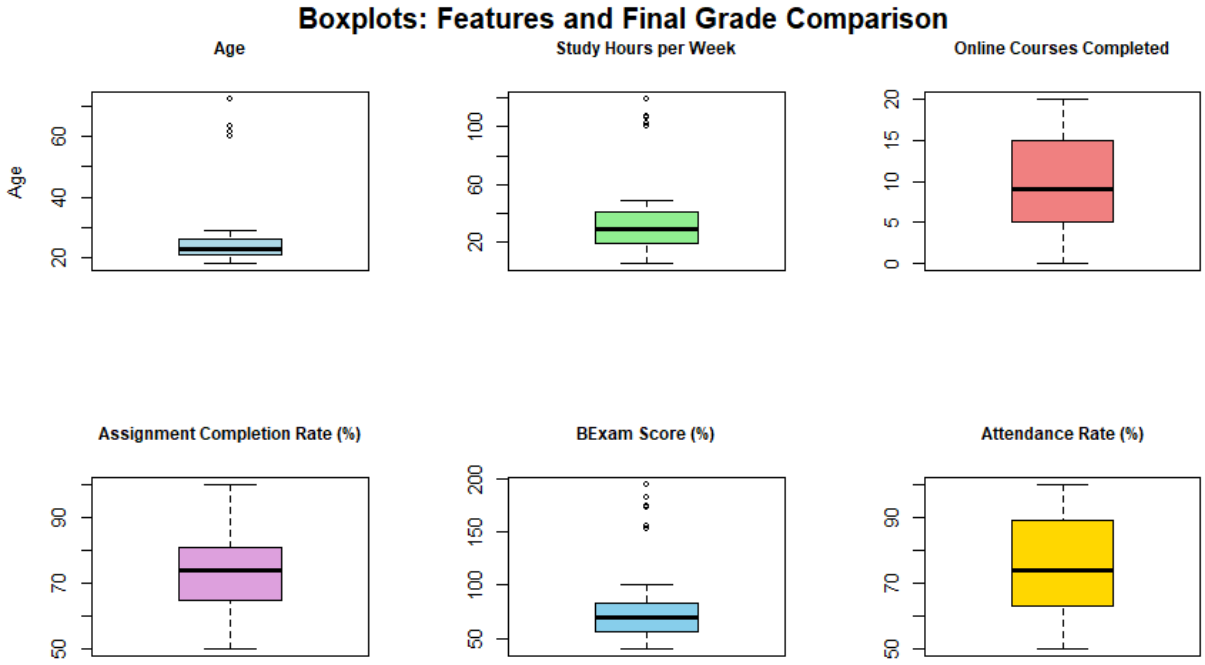

**Fig-5: Density plot**

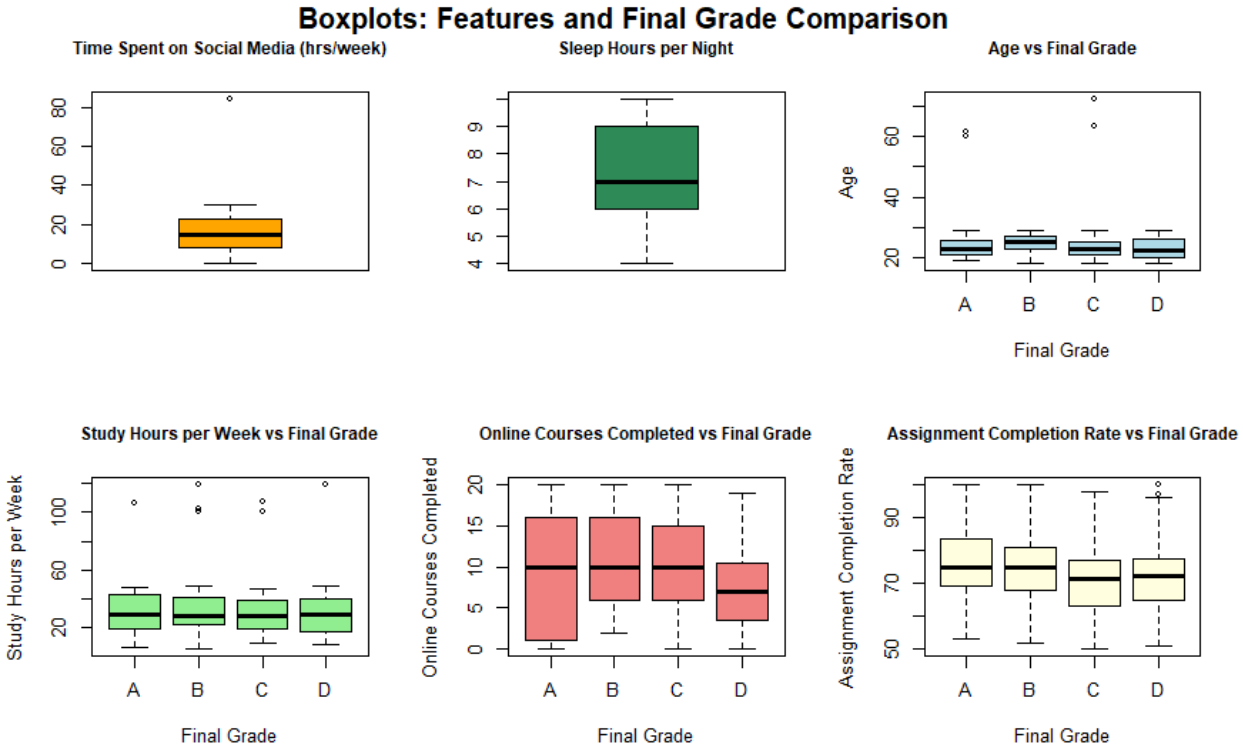
**Box Plot**

The boxplot shows that most students are young, study moderately per week, complete several online courses, and generally perform well in assignments, exams, and attendance, though there are a few outliers. When comparing these features to final grades, students with **higher grades (A)** tend to have higher study hours, online course completions, assignment completion rates, exam scores, and attendance rates. Lower grades (D) are associated with less study, fewer courses completed, lower assignment and attendance rates, and lower exam scores. Time spent on social media is slightly higher for students with lower grades, while sleep hours don't show a strong direct pattern with grades.

Comparison:

- Study Hours: More study hours → significantly better grades.
- Online Courses Completed: Completing more courses → better performance.
- Assignment Completion Rate: Higher completion rate → higher grades.
- Exam Scores: Strongest indicator of final grade.
- Attendance Rate: Better attendance → better grades.


Boxplots: Features and Final Grade Comparison
Age
Study Hours per Week
Online Courses Completed
Assignment Completion Rate (%)
BExam Score (%)
Attendance Rate (%)
Age


Boxplots: Features and Final Grade Comparison
Time Spent on Social Media (hrs/week)
Sleep Hours per Night
Age vs Final Grade
Age
A
B
C
D
Final Grade
Study Hours per Week vs Final Grade
Study Hours per Week
Online Courses Completed vs Final Grade
Online Courses Completed
Assignment Completion Rate vs Final Grade
Assignment Completion Rate

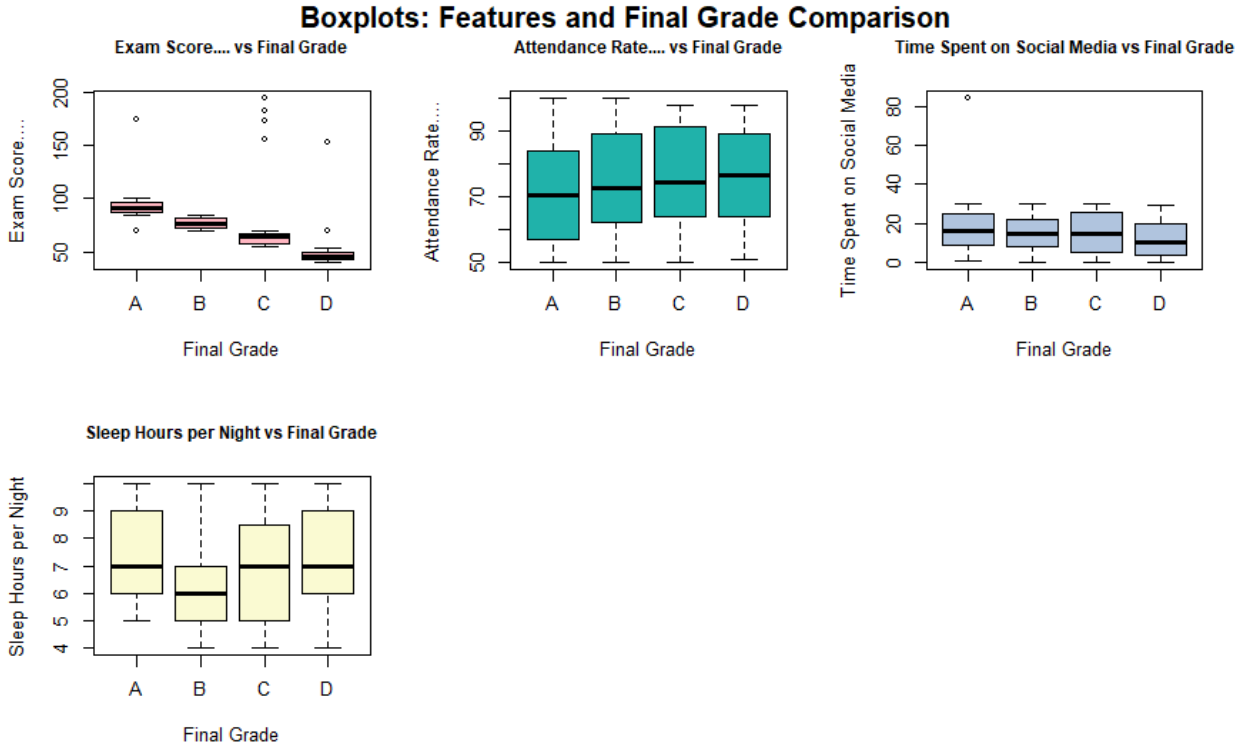



**Fig-6: Univariate Analysis: Boxplot**

**Bar Plot**

The bar plots show different information about students. Most students are between 18–22 years old. There are more male students than female students with very few choosing "Other." Many students study about 2–5 hours each day. Auditory and visual learning styles are the most common among students, while reading/writing is less popular. Most students have completed between one and three online courses. A large number of students also take part in discussions and use educational technology. When it comes to stress levels most students feel a medium amount of stress. For final grades, the most common grade is "B," followed by "C," "D" and "A."

Comparison

- Gender: More males than females overall.
- Learning Style: Auditory and visual are the most preferred.
- Age and Study Hours: Shown clearly, highlighting that most are young and study a few hours daily.
- Stress Levels: Medium stress is the most common feeling among students.
- Final Grades: "B" grade is achieved by most students.

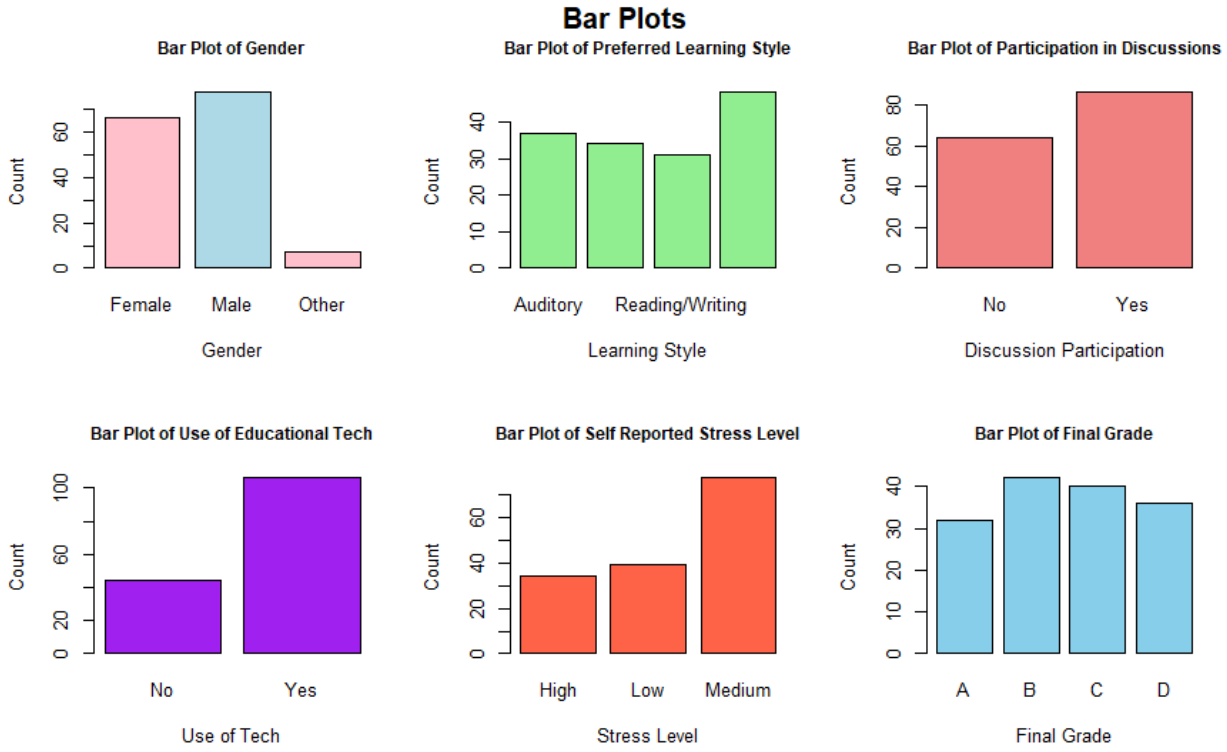
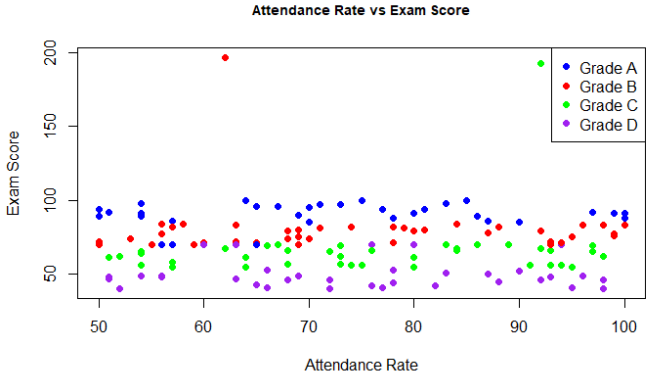



**Fig-7: Barplot**

## Multivariate Analysis

### Scatter Plot




The scatter plot titled **"Attendance Rate vs Exam Score"** examines the relationship between attendance rate and exam scores. The x-axis represents attendance rate, ranging from 50% to 100%, while the y-axis represents exam scores, ranging from 0 to 200. Four grades are depicted using distinct colors: Grade A (blue), Grade B (red), Grade C (green), and Grade D (purple).

The data shows that higher attendance rates generally correspond to higher exam scores. Grade A points cluster near the upper right, indicating both high scores and high attendance. As

attendance rates decrease, the data points shift towards lower exam scores, as represented by Grades B, C, and D. This suggests a positive correlation between consistent attendance and better academic performance.

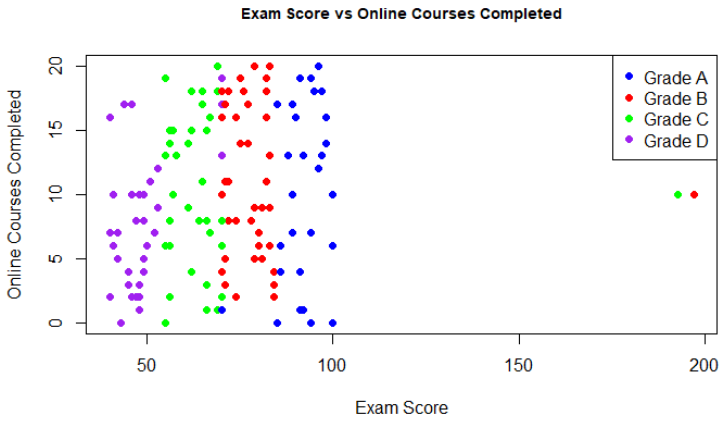



**Fig-8: Scatter Plots**

A scatter plot visually depicts the relationship between exam scores and the number of online courses completed. The x-axis represents exam scores, while the y-axis denotes the count of courses completed. Each point is color-coded to indicate grades: blue for Grade A, red for Grade B, green for Grade C, and purple for Grade D. The plot shows a clear trend where students with higher exam scores tend to complete more online courses. Grade A students cluster towards the upper right, reflecting high scores and extensive course completion, while Grades B, C, and D progressively shift towards lower scores and fewer courses completed. This visualization highlights a positive correlation between online course participation and academic performance.

**Correlation Matrix**

The 8x8 correlation matrix presents the pairwise correlation coefficients between eight variables: Age, Study Hours per Week, Online Courses Completed, Assignment Completion Rate, Exam Score, Attendance Rate, Time Spent on Social Media and Sleep Hours per Night. The matrix is symmetric, with the diagonal representing the correlation of each variable with itself (always 1.00). The off-diagonal values indicate the strength and direction of the linear relationships between the variables. Most correlations are weak, with values close to zero, implying little to no linear relationship between most pairs of variables. For example, the correlation between Assignment Completion Rate and Exam Score is 0.17, suggesting a weak positive relationship, while the correlation between Time Spent on Social Media and Sleep Hours is -0.13, showing a weak negative relationship. Other pairs exhibit similarly weak correlations, with values close to zero.

This matrix reflects a general lack of strong linear relationships between the variables, with a few exceptions showing mild associations.

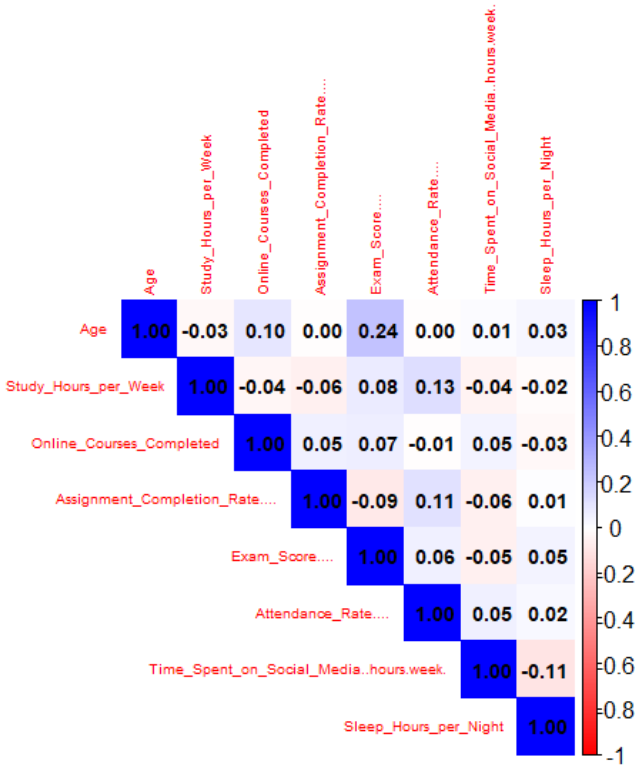



**Fig-9: Correlation Matrix**

**Pair Plot**

The pair plot presents all numerical variables in a dataset, with each data point colored based on the students' final grade. The variables compared include Age, Study Hours per Week, Courses Completed, Content Completion, Exam Scores, Attendance Rate, Time on Social Media and Sleep Hours per Night. Each subplot displays the relationship between two different variables while the diagonal shows the distribution of individual variables. The use of different colors for final grades allows for a clear visual distinction between performance groups. This plot helps to identify patterns, trends and possible correlations between variables, offering insights into how factors like study habits, attendance and lifestyle choices might influence academic outcomes.

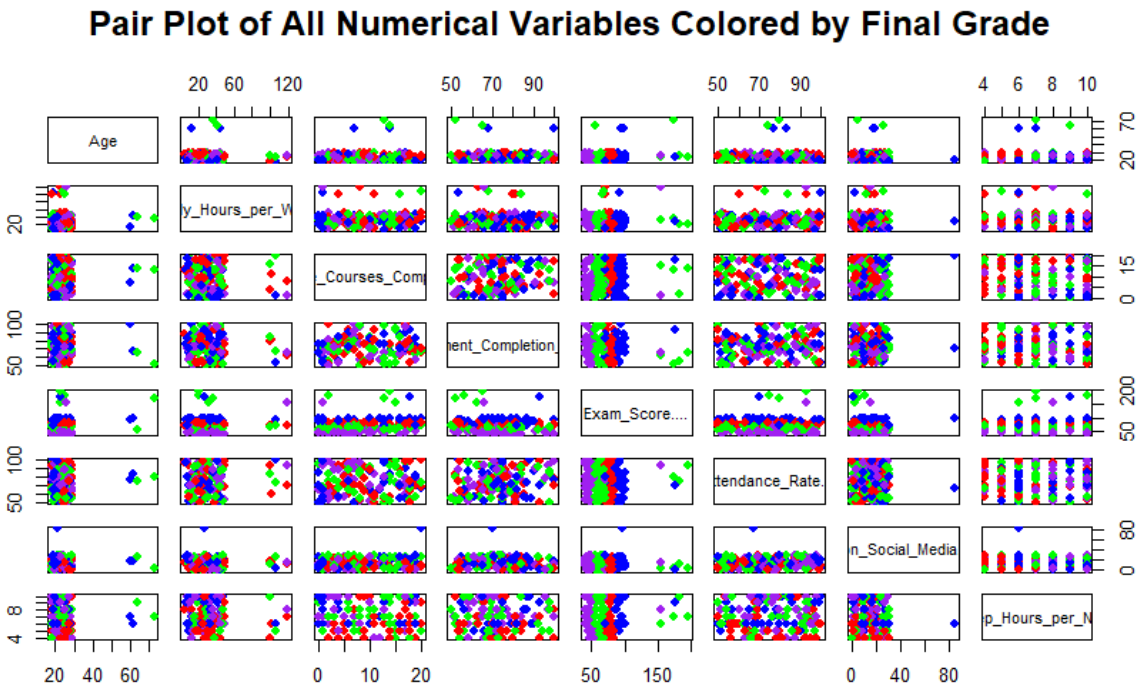



**Fig-10: Pair Plot**

## 5. Data Type Conversion:

To prepare the dataset several categorical variables are converted into numeric formats using conditional statements (ifelse() function). This transformation is necessary because many models require numeric inputs and cannot handle character or factor types directly.

i. **Gender Conversion**: The Gender variable is encoded as 1 for Male and 0 for Female. Missing values are set to NA. The new column Gender_Numeric stores these numeric values.

ii. **Preferred Learning Style**: The Preferred Learning Style is mapped as follows: Visual = 1, Auditory = 2, and Kinesthetic = 3. Missing or unrecognized values are set to NA. The transformed values are saved in Preferred_Learning_Style_Numeric.

iii. **Participation in Discussions**: For Participation in Discussions, Yes is mapped to 1 and No to 0, with missing values as NA. The new column Participation_in_Discussions_Numeric holds these numeric values.

iv. **Use of Educational Technology**: The Use of Educational Technology variable is encoded as Yes = 1 and No = 0, with missing or "Other" responses as NA. This transformation is stored in Use_of_Educational_Tech_Numeric.

v. **Self-Reported Stress Level**: The Self-Reported Stress Level is encoded as Low = 1, Moderate = 2, and High = 3. Undefined responses are set to NA, and the transformed data is stored in Self_Reported_Stress_Level_Numeric.

These transformations create new numeric columns alongside the original ones, allowing flexibility to either use the original categorical form or the newly numeric-encoded version depending on modeling needs.

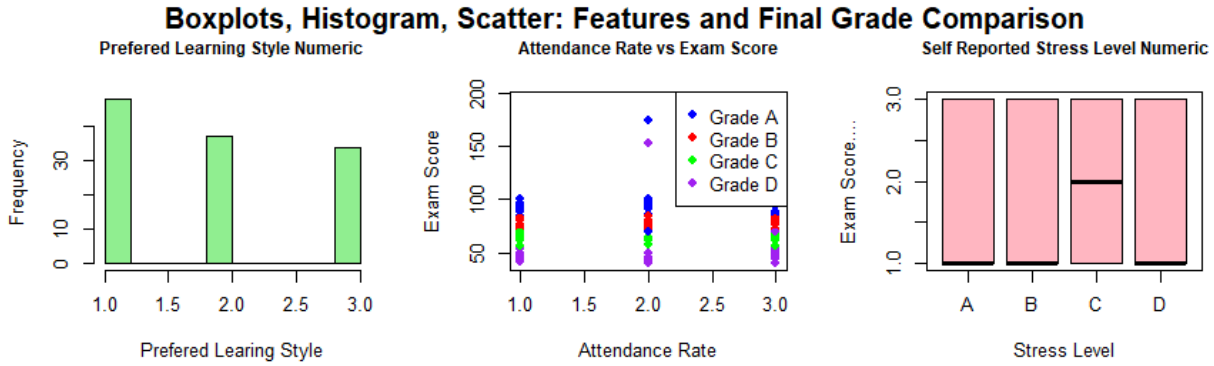



**Fig-11: Data Type Conversion**

## 6. Data Transformation

**Scaling** and **normalization** are two fundamental techniques used in data preprocessing to adjust the range and distribution of numerical variables, ensuring that they are suitable for further analysis and modeling.

**Scaling:**

Z-score standardization was performed using the **scale()** function in R. This method adjusts each selected variable to have a mean of 0 and a standard deviation of 1. The following transformations were made:

- **Study_Hours_per_Week** was scaled and stored as **StudyHours_scaled**
- **Exam_Score....** was scaled and stored as **ExamScore_scaled**
- **Attendance_Rate....** was scaled and stored as **Attendance_scaled**

- **Online_Courses_Completed** was scaled and stored as **OnlineCourses_scaled**

The **scale()** function automatically computes the mean and standard deviation for each variable and applies the standardization formula: where **x** is the original value

$$Z = (x - mean)/standard\ deviation$$

**Normalization (Min-Max Scaling):**

**Normalization** (specifically min-max normalization) rescales the values of a feature to a fixed range, typically between 0 and 1. Min-max normalization was manually performed to rescale the same set of variables to a range between 0 and 1. For each variable, the normalization formula applied was:

$$Xnorm = (X - Xmin)/(Xmax - Xmin)$$

where ***Xmin*** and ***Xmax*** represent the minimum and maximum values of the respective variable.

The transformations were as follows:

- **Study_Hours_per_Week** was normalized and stored as **StudyHours_norm**
- **Exam_Score....** was normalized and stored as **ExamScore_norm**
- **Attendance_Rate....** was normalized and stored as **Attendance_norm**
- **Online_Courses_Completed** was normalized and stored as **OnlineCourses_norm**

The na.rm = TRUE parameter was included to ensure that any missing values (NA) were ignored during the calculation of minimum and maximum values.

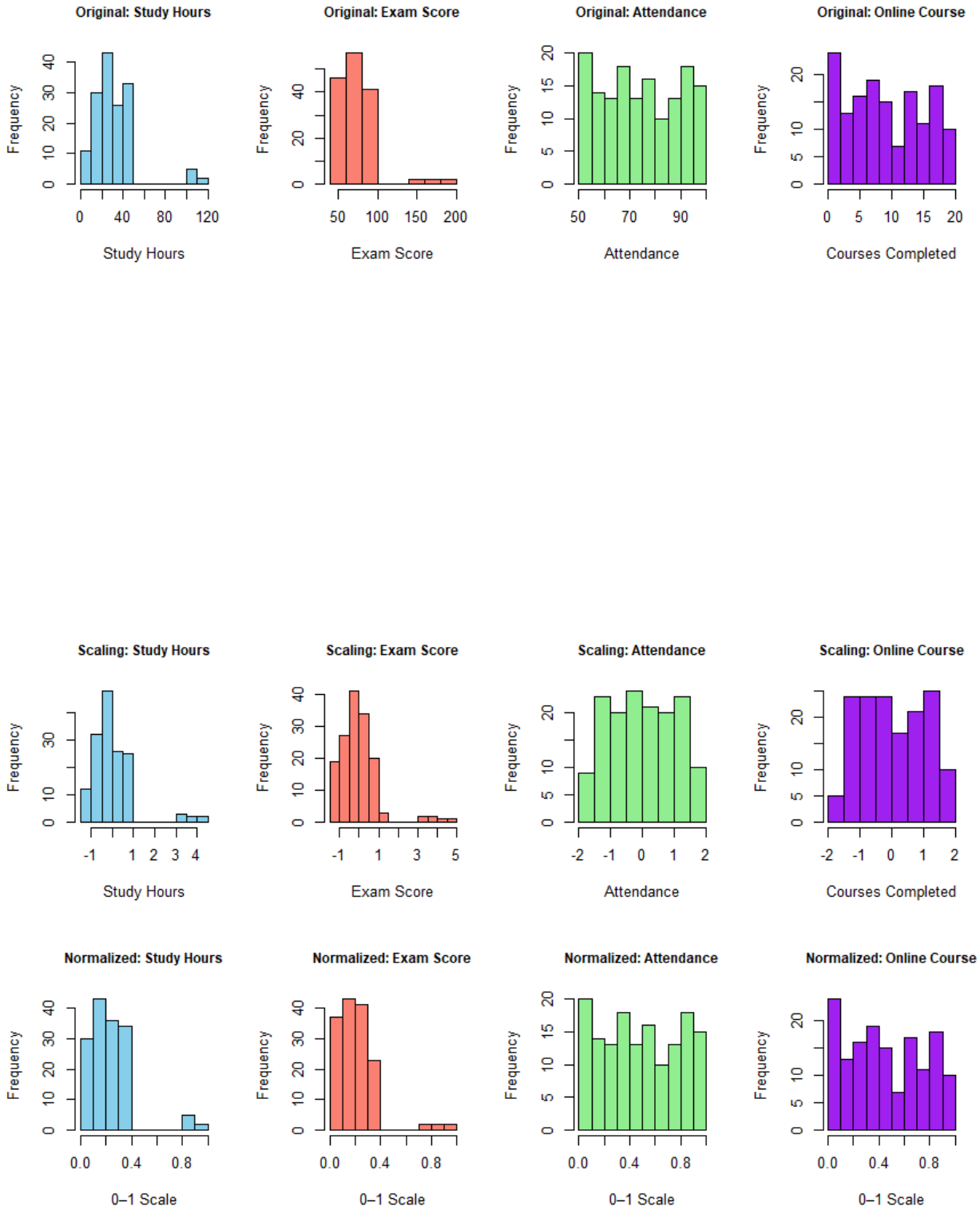



**Fig-12: Histogram of original vs scaled and normalized**

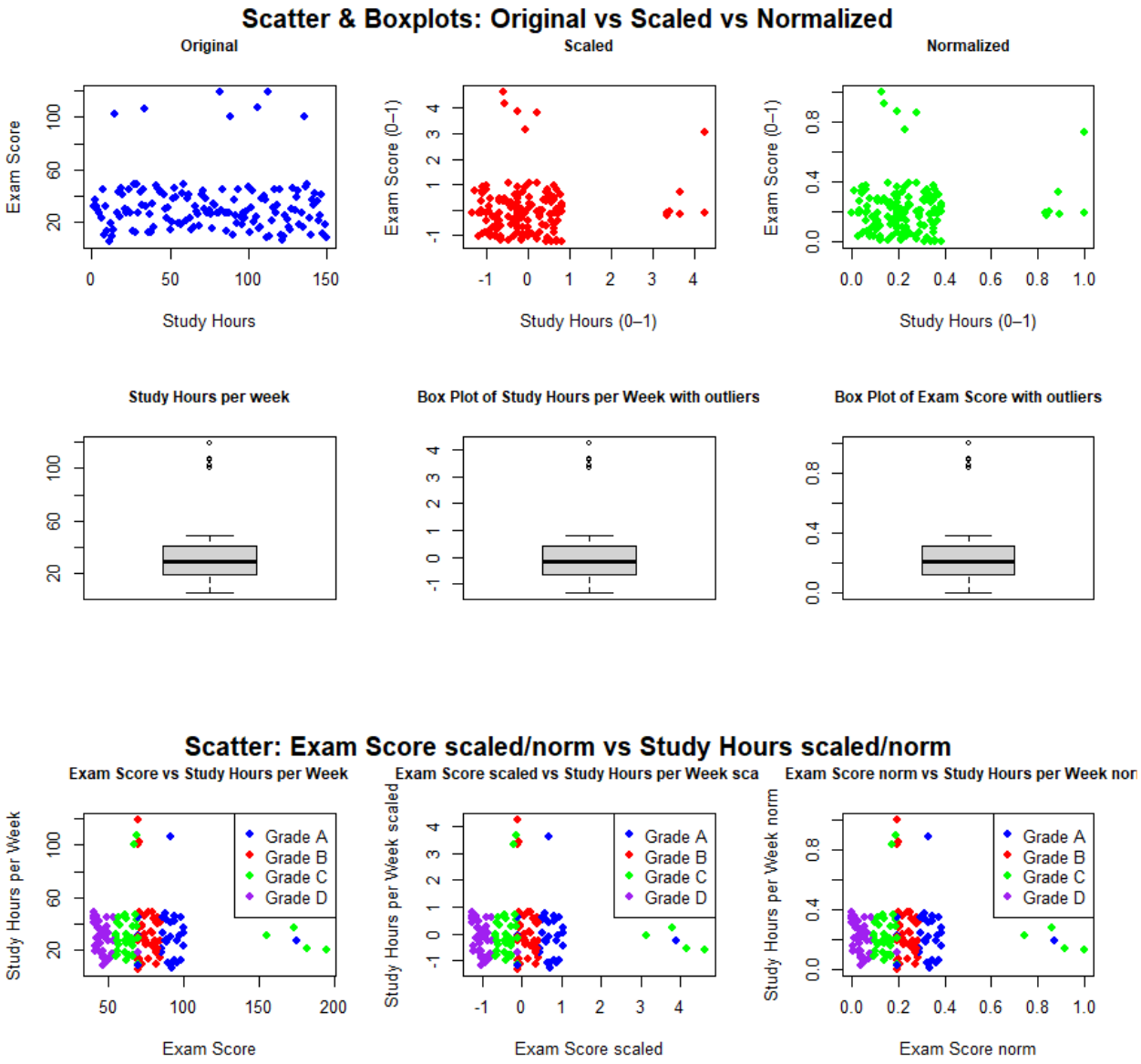



**Fig-13: Scatter & Box plot of original vs scaled and normalized**

## 7. Outliers Detection

A data point that substantially deviates from the usual or anticipated pattern within a dataset is referred to as an outlier in data science. These points have the potential to influence the outcomes of data analysis and can be much higher or lower than the other values.

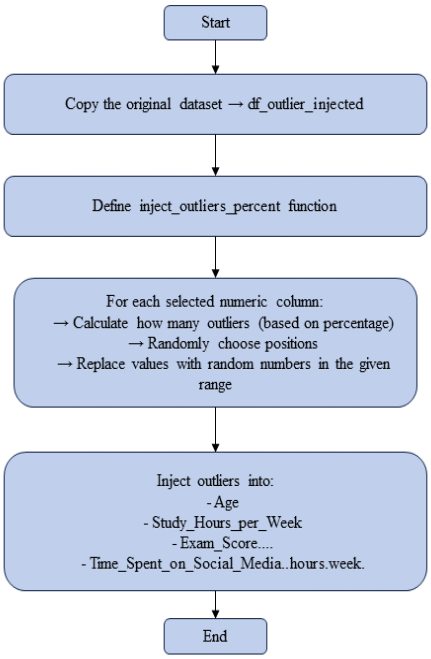



*Figure 14: Injection of Outliers flowchart.*

**Identification of Outliers by Boxplot Diagram:**

A boxplot is a simple graphical representation that summarizes the distribution of a dataset. It displays important statistical measures like the minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum. The main body of the plot, called the "box," shows the middle 50% of the data (the interquartile range or IQR). Lines, known as "whiskers," extend from the box to cover the range of the data within 1.5 times the IQR from the quartiles.

Outliers are detected based on the spread of the data. Any data points that lie **below** *(Q1 - 1.5 × IQR)* or **above** *(Q3 + 1.5 × IQR)* are flagged as potential outliers. In a boxplot, these outliers are often displayed as individual points or small dots outside the whiskers. This makes it very easy to visually spot unusually high or low values in the dataset without having to manually calculate thresholds.

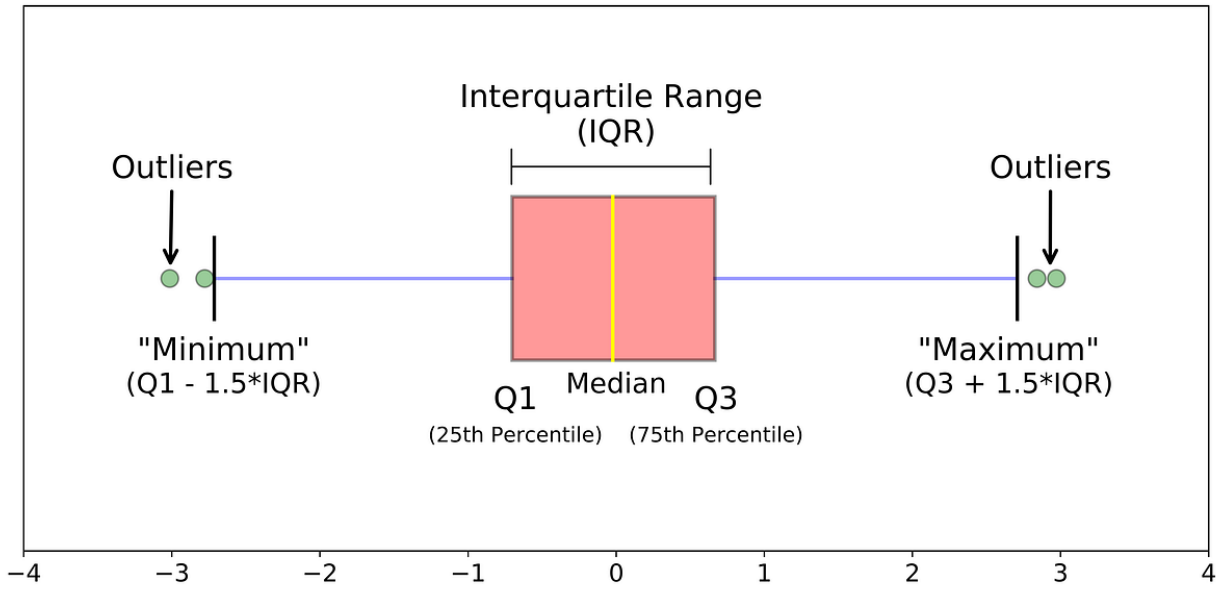



*Figure 15: How boxplot detects outliers*

**Outliers presented in the sample dataset:**

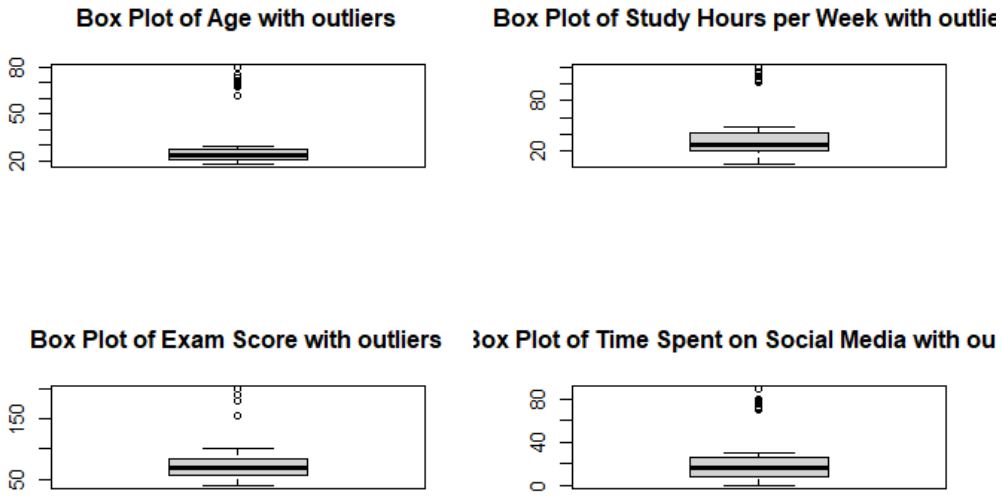


*Figure 16: Outliers presented by boxplot diagram.*

**Age:** The box plot displays the distribution of ages, with the central box representing the interquartile range (IQR) and whiskers showing the typical range. Outliers, observed above the upper whisker, indicate individuals with ages significantly higher than the norm.

**Study Hours per Week:** The plot highlights the concentration of study hours across individuals. The IQR encapsulates most of the data, while outliers above the whisker signify unusually high study hours, likely reflecting exceptional academic dedication or unique circumstances.

**Exam Scores:** This plot visualizes the variation in exam scores. While the majority fall within the IQR, outliers above the whisker represent individuals achieving notably high scores, which may indicate extraordinary performance or special factors influencing their results.

**Time Spent on social media:** The distribution of weekly social media hours is depicted in this box plot. Outliers above the upper whisker suggest some individuals spend significantly more time online than others, potentially pointing to heavy usage or distinctive behavioral patterns.

## 8. Outliers Handling by Using IQR Method

To remove outliers from the dataset, the Interquartile Range (IQR) method was applied. A custom function remove_outliers was created to handle this process automatically for any selected numeric column. The function calculates the **first quartile (Q1)** and **third quartile (Q3)** for the given column, and then finds the IQR by subtracting Q1 from Q3. Using the standard formula, it determines the **lower bound** and the **upper bound**. Data points falling outside these bounds are considered outliers and are removed from the dataset

A loop was then used to apply this function across all selected numeric features. To ensure safety, a copy of the dataset was made first so that the original data (with injected outliers) remained unaffected.Finally, the dataset was cleaned by keeping only the rows where values lie within the acceptable range.

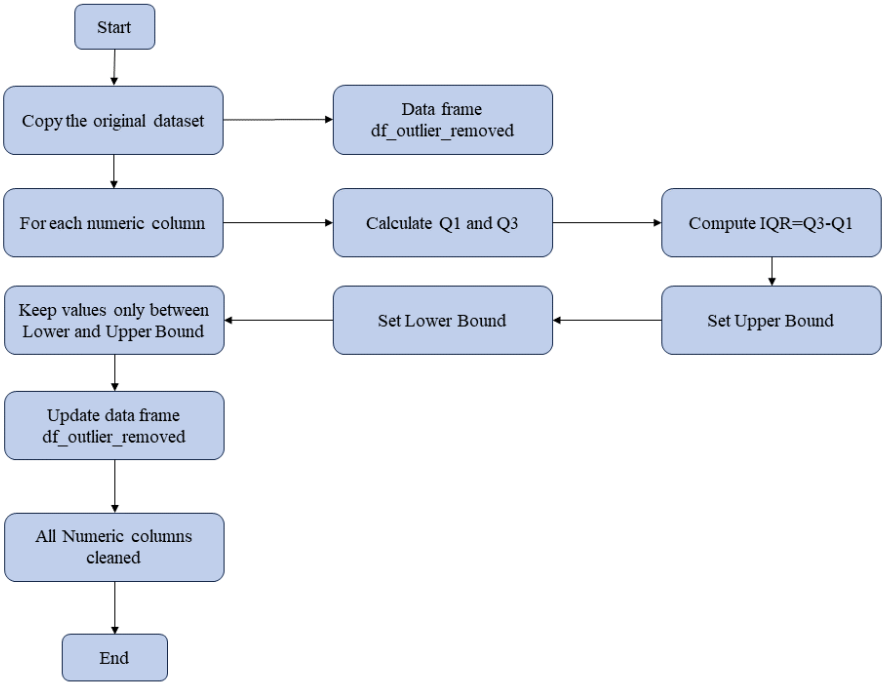



*Figure 17: Outliers removal flowchart.*

**i) Interquartile Range (IQR) Calculation:**

$$\boldsymbol{IQR}=Q3-Q1,$$

*Here,* **Q1** (First Quartile) means the value below which 25% of the data fall and **Q3** (Third Quartile) means the value below which 75% of the data fall.

**ii) Lower Bound Calculation:**

$$\boldsymbol{Lower\ Bound} = Q1 - 1.5 * IQR$$

**iii) Upper Bound Calculation:**

$$\boldsymbol{Upper\ Bound} = Q3 + 1.5 * IQR$$

**iv) Outlier Condition:**

*Value < Lower Bound* ***OR*** *Value > Upper Bound*

A data point is considered an outlier if it is less than the **Lower Bound,** or it is greater than the **Upper Bound.**

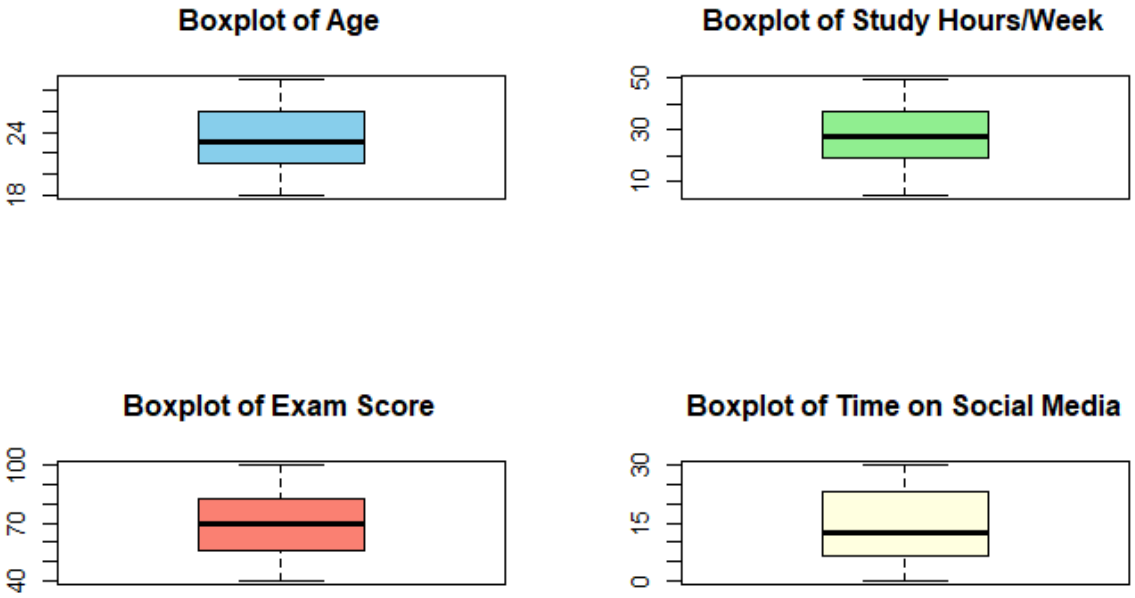



*Figure 18: Boxplots after handling the outliers.*