R Lab Practical Sheet 12

S.A.Abdulla - 22000021

February 13, 2025

Contents

1 Introduction

The objective of this assignment is to conduct a comprehensive analysis of a dataset using R, applying both statistical and machine learning techniques to derive meaningful insights. The chosen dataset provides a collection of variables that represent diverse characteristics, and the primary goal is to explore patterns, trends, and potential relationships within the data. This report will outline the methodologies used to process, analyze, and visualize the data, as well as the results obtained through the application of various statistical techniques and machine learning algorithms.

The dataset provided serves as a foundational resource for this report, which aims to conduct a thorough and insightful analysis. Although the dataset may already meet certain standards, it has been carefully examined using the full range of functionalities available through the R programming language, in conjunction with the powerful capabilities of the Jupyter Notebook application. This report leverages these tools to perform comprehensive data exploration, statistical analysis, and visualization, ultimately providing deeper insights and meaningful interpretations from the data.

By performing exploratory data analysis (EDA), calculating summary statistics, visualizing trends, and conducting hypothesis testing, this assignment seeks to uncover significant insights and draw conclusions from the data. Additionally, the report will include a discussion of the results, their implications, and potential limitations encountered during the analysis. Ultimately, the goal is to highlight the importance of data analysis techniques in extracting actionable knowledge from raw data.

1.1 Dataset

In this report, we analyze a dataset containing information about students, including demographic details, test preparation status, and their scores in math, reading, and writing subjects. The dataset provides insights into various attributes such as race/ethnicity, parental level of education, and the type of lunch provided. Having this as the base for our analysis with simple details we dive in to the real goal with the table structure given below with some data from the dataset as a exmaple.

Most of the vlaue in the table are dummy values there can be some misleading datas in the set don't condier about it that much keeping the concern on dataset to the minimum

ID	Race	Parental Education	Lunch	Preparation Course
1	Group B	Bachelor's degree	Standard	None
2	Group C	Some college	Standard	Completed
3	Group B	Master's degree	Standard	None
4	Group A	Associate's degree	Free/Reduced	None
5	Group C	Some college	Standard	None
6	Group B	Associate's degree	Standard	None

Table 1: Part 1 of the dataset: This section of the table displays the student ID, race/ethnicity, parental level of education, lunch type, and test preparation status.

Math Percentage	Reading Percentage	Writing Percentage	Sex
0.72	0.72	0.74	F
0.69	0.90	0.88	F
0.90	0.95	0.93	F
0.47	0.57	0.44	M
0.76	0.78	0.75	M
0.71	0.83	0.78	F

Table 2: Part 2 of the dataset: This section includes the math, reading, and writing percentages for each student, as well as the student's sex.

1.2 Dataset Overview

In this part we will conisder on the quaility of the dataset and spanning of the data set through multiple variables and a large number of rows as considered from R code given below

```
str(df)
#to see the lenght of the data set and number of columns it the data set
✓ 0.0s

'data.frame':   1000 obs. of  9 variables:
 $ X                         : int  0 1 2 3 4 5 6 7 8 9 ...
 $ race.ethnicity            : chr  "group B" "group C" "group B" "group A" ...
 $ parental.level.of.education: chr  "bachelor's degree" "some college" "master's degree" "associate's degree" ...
 $ lunch                     : chr  "standard" "standard" "standard" "free/reduced" ...
 $ test.preparation.course   : chr  "none" "completed" "none" "none" ...
 $ math.percentage           : num  0.72 0.69 0.9 0.47 0.76 0.71 0.88 0.4 0.64 0.38 ...
 $ reading.score.percentage  : num  0.72 0.9 0.95 0.57 0.78 0.83 0.95 0.43 0.64 0.6 ...
 $ writing.score.percentage  : num  0.74 0.88 0.93 0.44 0.75 0.78 0.92 0.39 0.67 0.5 ...
 $ sex                       : chr  "F" "F" "F" "M" ...
```

```
summary(df)   # To get a summary of data to take desicion based on the table structure and values
✓ 0.0s

       X         race.ethnicity     parental.level.of.education
 Min.   :  0.0   Length:1000        Length:1000
 1st Qu.:249.8   Class :character   Class :character
 Median :499.5   Mode  :character   Mode  :character
 Mean   :499.5
 3rd Qu.:749.2
 Max.   :999.0
    lunch           test.preparation.course math.percentage
 Length:1000        Length:1000             Min.   :0.0000
 Class :character   Class :character        1st Qu.:0.5700
 Mode  :character   Mode  :character        Median :0.6600
                                            Mean   :0.6609
                                            3rd Qu.:0.7700
                                            Max.   :1.0000
 reading.score.percentage writing.score.percentage     sex
 Min.   :0.1700           Min.   :0.1000           Length:1000
 1st Qu.:0.5900           1st Qu.:0.5775           Class :character
 Median :0.7000           Median :0.6900           Mode  :character
 Mean   :0.6917           Mean   :0.6805
 3rd Qu.:0.7900           3rd Qu.:0.7900
 Max.   :1.0000           Max.   :1.0000
```

As provided Above in the image there are almot 9 variables and 1000 rows in the dataset which meets minimal requiremnets. By anlayizing the image given below we can see that there is suitable dataset for analysis to be done which having considerable amount of mean, median min and max

2 Methodology

The data preprocessing stage is a crucial part of any data analysis, ensuring that the dataset is cleaned, structured, and ready for analysis. Below are the steps followed in this report for preprocessing the dataset:

2.1 Reading dataset

To begin with, the dataset needs to be imported into the R environment. In this report, the dataset is loaded from a CSV file using the read.csv() function. This function reads the data into an R data frame, which is a table-like structure that makes it easy to perform subsequent operations.

```
df <- read.csv("student␣Performance␣new.csv")
```

2.2 Refining dataset

Missing data is a common issue in rel-world datasets, and it is essential to decide how to handle them before proceeding with analysis

```
df <- na.omit(df)
```

Duplicate records can occur due to errors in data collection or merging datasets. These duplicate entries need to be removed to ensure the integrity of the analysis

```
df <- df[!duplicated(df), ]
```

Need to install the needed package to display the ggplot in R

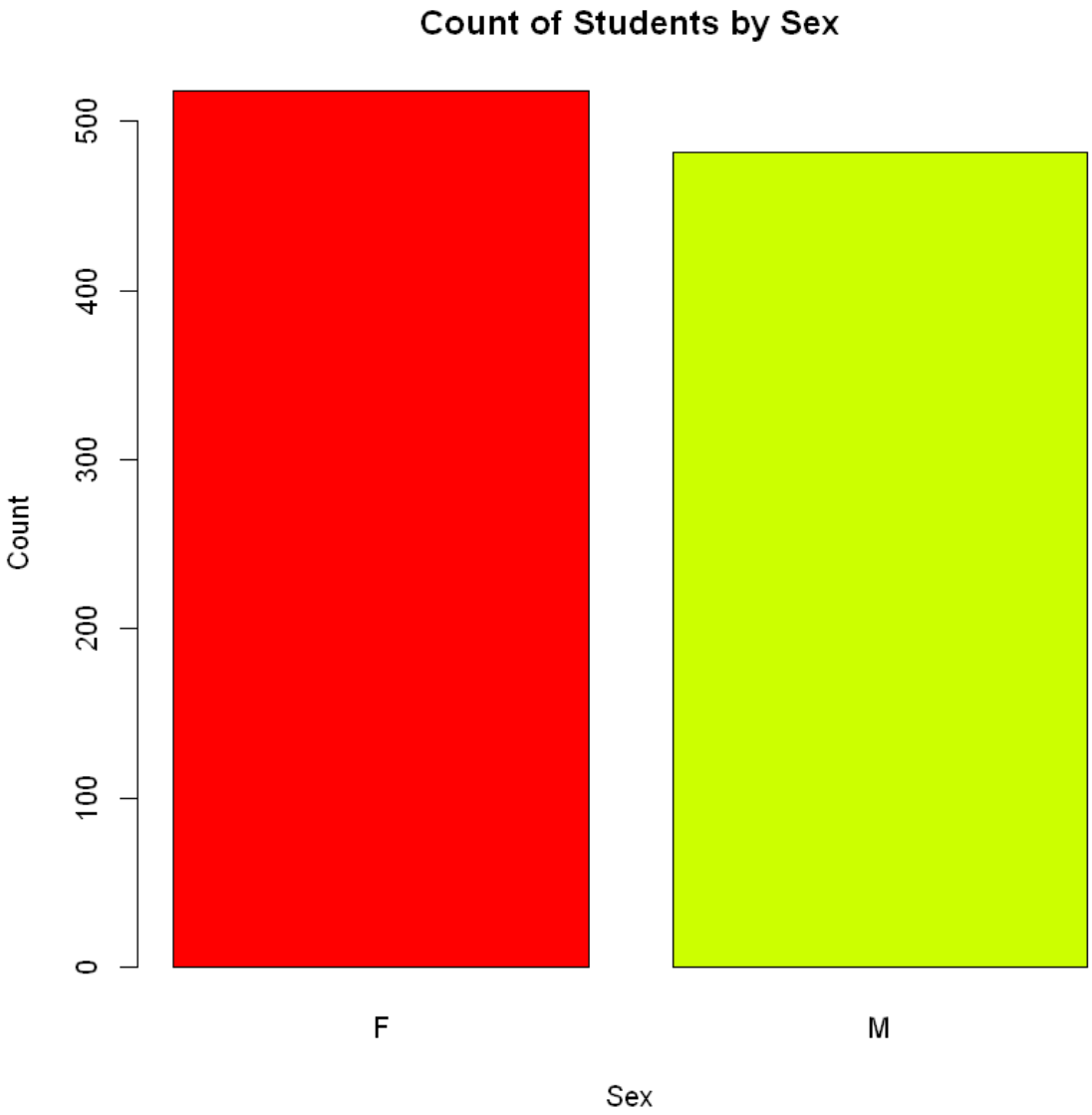
```
install.packages("ggplot2")
```

with this Methodolagys lest go into the Analysis part of the Lab Sheet

3 Results

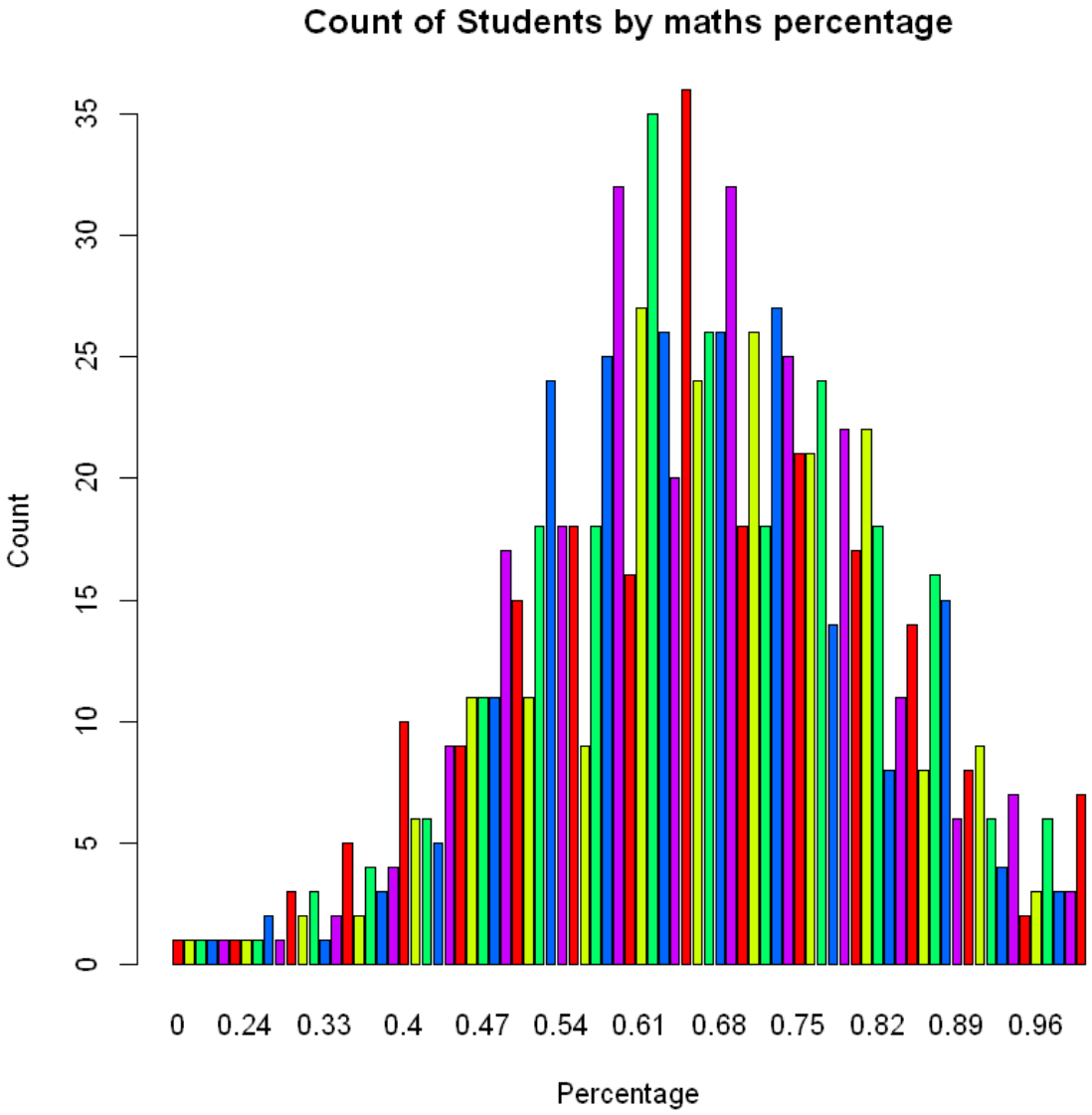
```
    barplot(table(df$sex),
        col = rainbow(5),
        main = "Count␣of␣Students␣by␣Sex",
        xlab = "Sex",
        ylab = "Count"
)
```

Given above code for the barplot shows the count of male and female students in the dataset. Each bar corresponds to one category of sex, and the height of the bar indicates the number of students in that category.



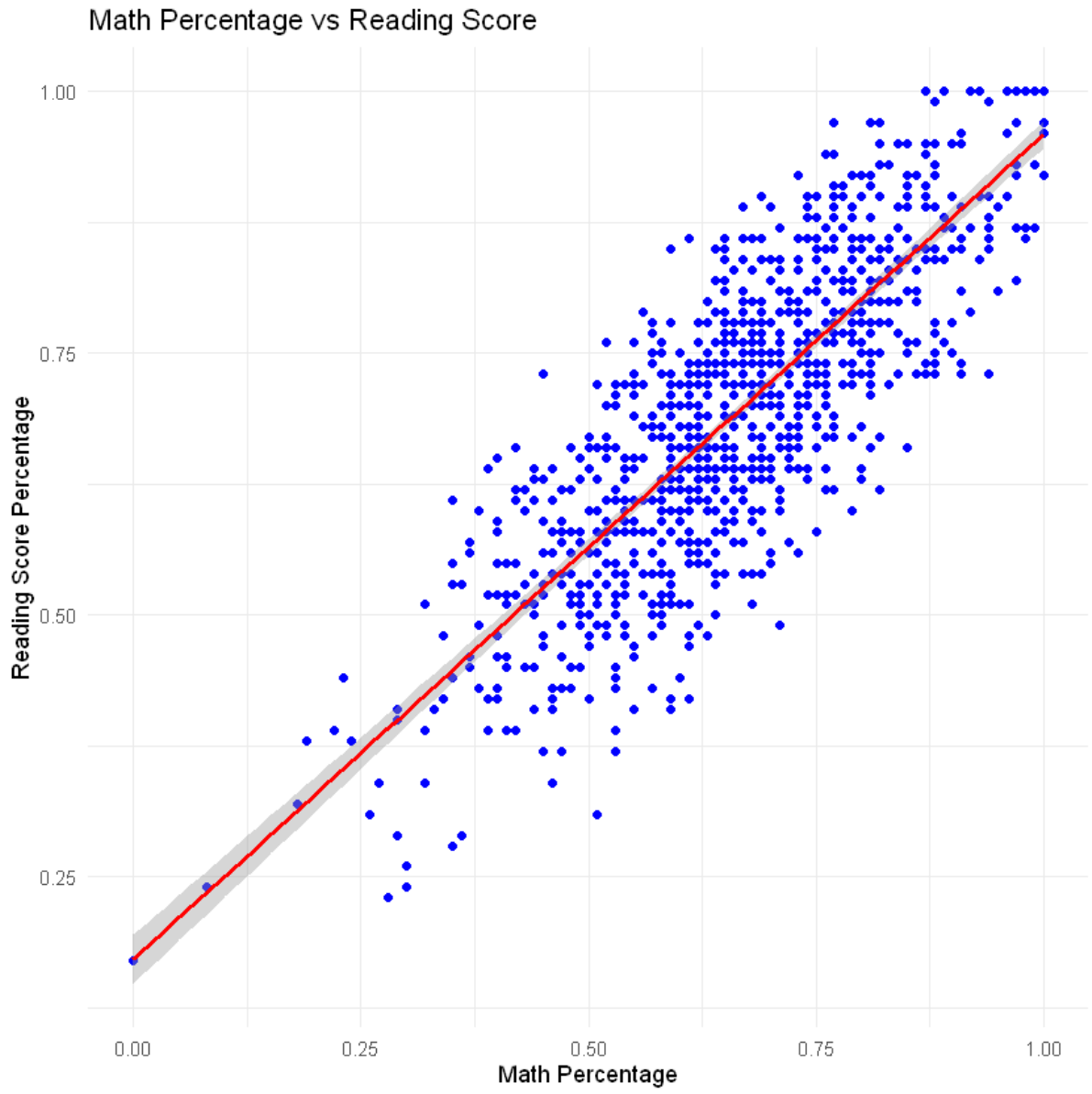
```
    barplot(table(df$math.percentage),
    col = rainbow(5),
    main = "Count␣of␣Students␣by␣maths␣percentage",
    xlab = "Percentage",
    ylab = "Count"
)
```

This barplot shows the frequency of students who fall within different math percentage ranges. Each bar represents a specific percentage range, and its height reflects how many students scored within that range



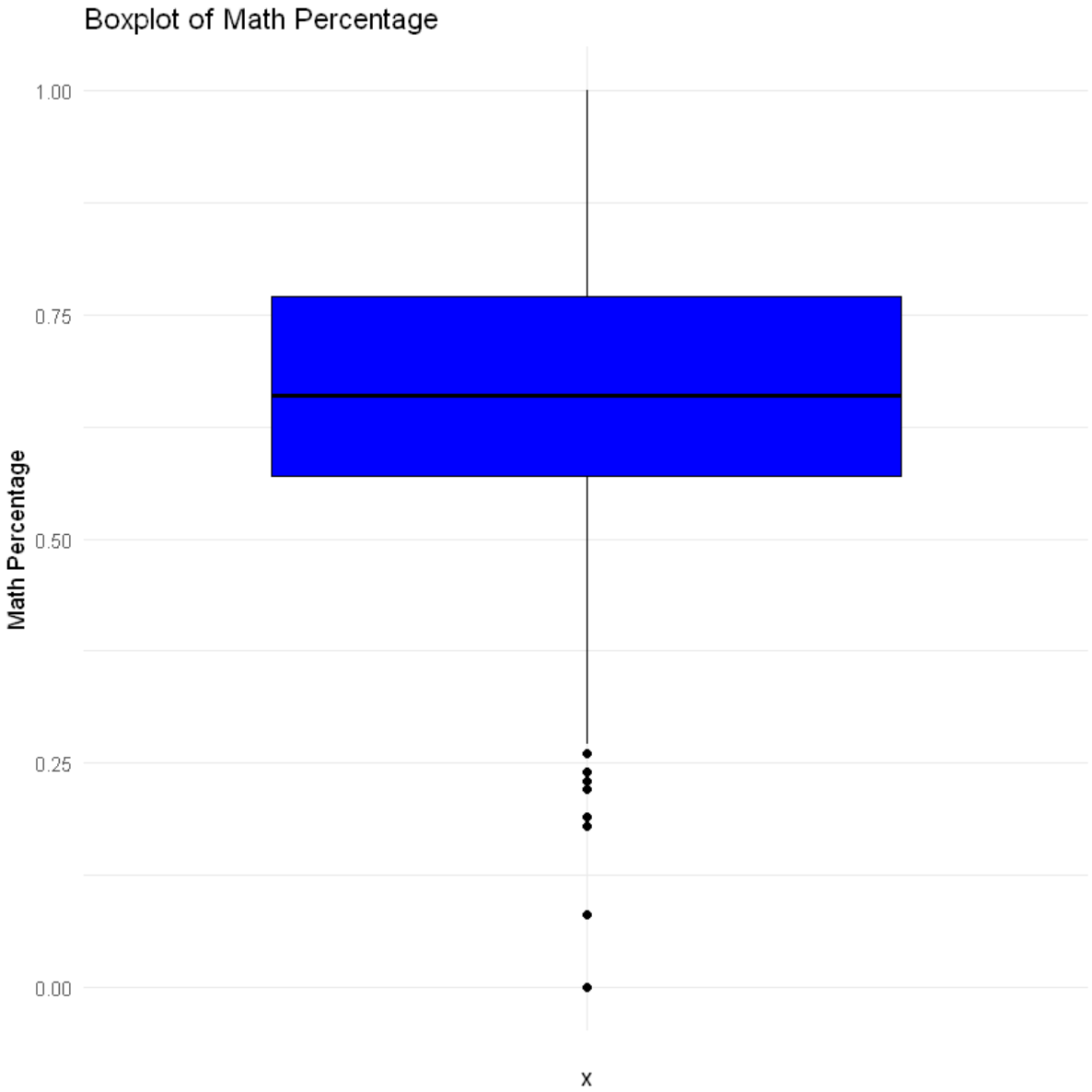
```
ggplot(df, aes(x = math.percentage, y = reading.score.percentage)) +
geom_point(color = "blue") +            #plots to see where the data are
geom_smooth(method = "lm", color = "red") + # Linear line for easire
    analysis
labs(
    title = "Math Percentage vs Reading Score", #title
    x = "Math Percentage",                      #X-Title
    y = "Reading Score Percentage"              #Y-Tilte
) +
theme_minimal()
```

plot shows individual data points, each representing a student's math percentage on the x-axis and their corresponding reading score percentage on the y-axis. The red line is a fitted linear regression line, helping to illustrate the trend in the relationship between the two variables



```
ggplot(df, aes(x = "", y = math.percentage)) +
geom_boxplot(fill = "blue", color = "black") +
labs(
    title = "Boxplot of Math Percentage",
    y = "Math Percentage"
) +
theme_minimal()
```

shows the spread and central tendency of the math percentage data. The box itself represents the interquartile range (IQR), which is where the middle 50



```
mean_math <- mean(df$math.percentage, na.rm = TRUE)
cat("Mean of Math Percentage:", mean_math, "\n")

median_math <- median(df$math.percentage, na.rm = TRUE)
cat("Median of Math Percentage:", median_math, "\n")

variance_math <- var(df$math.percentage, na.rm = TRUE)
cat("Variance of Math Percentage:", variance_math, "\n")

sd_math <- sd(df$math.percentage, na.rm = TRUE)
cat("Standard Deviation of Math Percentage:", sd_math, "\n")
```

as a analysis to the above data set we have gathered some info on these as below from the code sagements from the above codespace

1. **Mean of Math Percentage:** 0.66089
2. **Median of Math Percentage:** 0.66
3. **Variance of Math Percentage:** 0.0229919
4. **Standard Deviation of Math Percentage:** 0.1516308

4 Discussion: Interpretation and Significance of Results

The dataset provided a comprehensive view of student performance in three subjects: math, reading, and writing. After performing various statistical analyses, the following key insights were observed:

- **Mean and Median of Math Percentages:** The mean math percentage of the students is approximately 66.1%, while the median is 66%. This indicates that the distribution of math scores is nearly symmetric. Given that the mean and median are so close, it suggests that the data does not contain extreme outliers or a skewed distribution. This is also supported by the relatively low variance of 0.0229919, indicating that most students' math scores are clustered around the average.
- **Variance and Standard Deviation of Math Scores:** The variance and standard deviation of math scores (0.0229919 and 0.1516, respectively) indicate that while the majority of students have scores around the mean, there are some students with scores significantly different from the average. The standard deviation of 0.1516 suggests moderate variability in math scores.
- **Correlation Between Math and Reading Scores:** A linear regression analysis between math percentages and reading scores revealed a positive correlation. The scatter plot, combined with a fitted regression line, shows a clear trend where students who perform well in math also tend to perform better in reading. This suggests that enhancing math scores might also lead to improvements in reading performance, possibly due to shared cognitive skills or academic support across subjects.

- **Boxplot Analysis:** The boxplot for math percentages showed no significant outliers, with the bulk of the data falling within the interquartile range. This reinforces the observation that most students' scores are clustered around the average, and there are no extreme outliers skewing the dataset.

The analysis revealed several important trends that can be useful in educational settings. For example, a targeted intervention aimed at improving math performance could potentially also improve reading scores. Additionally, it's worth noting that the low variance in math scores suggests that most students are performing at similar levels, indicating that efforts should be directed at boosting overall student performance across the board.

5 Conclusion: Summary and Potential Future Work

The analysis of the dataset provided a deep dive into the academic performance of students, particularly focusing on math, reading, and writing scores. Key findings such as the mean, median, variance, and correlation between subjects have revealed some significant patterns in student performance.

Key Takeaways:

- The math scores are normally distributed with a small spread, indicating consistency in performance across students.
- A positive correlation between math and reading scores suggests that interventions in math may also have a positive impact on other subjects.
- The overall performance data supports the need for tailored educational programs to enhance the academic performance of students, especially in math and reading.

Potential Future Work:

- **Further Statistical Testing:** Future analyses could include ANOVA or t-tests to explore the differences in scores based on categorical variables like gender, race, and parental education level.
- **Machine Learning Models:** It would be interesting to apply machine learning models, such as decision trees or random forests, to predict student performance based on the available features. This could provide insights into the most important factors affecting academic success.
- **Longitudinal Analysis:** A longitudinal analysis could explore how students' performance evolves over time, especially in response to educational interventions.
- **Exploration of More Variables:** Additional data, such as socioeconomic status or teacher quality, could be included in future analyses to further refine the understanding of student performance.

The findings of this analysis can be used to inform education policymakers and school administrators, helping them to make data-driven decisions regarding curriculum development and student support programs.

6 References

1. Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

2. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer.

3. UCI Machine Learning Repository (2018). *Student Performance Data Set*. [Link to Dataset].

4. R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.