Here’s a comprehensive report of the analysis performed on the *Student Performance Factors* dataset, with detailed steps and encountered issues on R-studio

**Data Loading and Initial Exploration**

1. **Loading Data:** The dataset, *Student Performance Factors*, was loaded into R using read\_excel. Initial inspection via head(student\_factors, 10) showed the first 10 records, including variables like *Hours\_Studied*, *Attendance*, *Parental\_Involvement*, *Access\_to\_Resources*, among others.
2. **Summary Statistics:** The summary(student\_factors) function provided essential summary statistics:
   * *Hours\_Studied*: Mean of 19.98, ranging from 1 to 44.
   * *Attendance*: Mean of 79.98%, ranging from 60 to 100%.
   * *Exam\_Score*: Mean of 67.24, with values between 55 and 101.

**Data Manipulation and Filtering**

1. **Filtering Rows Based on Study Hours:** Attempting to filter rows with subset(student\_factors$Hours\_Studied > 10) raised an error due to incorrect syntax. The correct approach used filtered\_data <- subset(student\_factors, Hours\_Studied > 10), successfully creating a filtered dataset.
2. **Creating a Score Category Column:** To categorize *Exam\_Score*, we used:

R

Copy code

student\_factors$Exam\_Score <- cut(student\_factors$Exam\_Score, breaks = c(-Inf, 50, 70, Inf), labels = c("Low", "Medium", "High"))

This created a categorical variable based on score ranges.

**Data Visualization**

1. **Boxplot of Exam Scores by Study Time:** A boxplot visualized the distribution of *Exam\_Score* across study times using ggplot2:

R

Copy code

ggplot(student\_factors, aes(x = factor(Hours\_Studied), y = Exam\_Score)) +

geom\_boxplot(fill = "skyblue", color = "black") +

labs(title = "Distribution of Scores by Study Time", x = "Study Time (hours)", y = "Exam Score") +

theme\_minimal()

This provided insights into the score variability at different study hours.

1. **Histogram of Exam Scores:** The histogram initially faced an error as *Exam\_Score* was treated as a discrete variable. Converting it to numeric resolved this:

R

Copy code

student\_factors$Exam\_Score <- as.numeric(as.character(student\_factors$Exam\_Score))

ggplot(student\_factors, aes(x = Exam\_Score)) +

geom\_histogram(binwidth = 5, fill = "lightblue", color = "black") +

labs(title = "Distribution of Exam Scores Across Students", x = "Exam Score", y = "Number of Students") +

theme\_minimal(base\_size = 14)

Despite converting *Exam\_Score*, warnings indicated missing or infinite values. These need further exploration, possibly due to NA values in the *Exam\_Score* column.

**Statistical Analysis**

1. **Linear Regression Analysis:** Attempting to create a regression model with *Exam\_Score* as the dependent variable and *Hours\_Studied* as the independent variable (model <- lm(Exam\_Score ~ Hours\_Studied, data = student\_factors)) returned an error due to the presence of only NA values in *Exam\_Score*. This prevented model creation, necessitating data cleaning to handle missing values in *Exam\_Score* before re-running the regression.
2. **Correlation Calculation:** Computing the correlation between *Hours\_Studied* and *Exam\_Score* with cor(student\_factors$Hours\_Studied, student\_factors$Exam\_Score) resulted in NA, likely due to missing values in *Exam\_Score*.

**Conclusion and Recommendations**

* **Data Cleaning:** Address NA values in *Exam\_Score* and other columns to ensure statistical analyses can proceed without errors.
* **Further Analysis:** Once cleaned, reattempt the correlation and regression analyses to determine the relationship between *Hours\_Studied* and *Exam\_Score*.
* **Improvement in Visualization:** Additional transformations might be needed for categorical variables before plotting to avoid type-related issues.

This report summarizes the initial analysis steps, challenges encountered, and adjustments needed for future analysis on the *Student Performance Factors* dataset.

4o

Top of Form



Bottom of Form