**Report: Data Analytics (Excel and R Studio)**

**Introduction:** The analysis aims to explore the relationship between various student performance factors, specifically focusing on study time and exam scores. The dataset, Mubarak\_Daud\_StudentPerformanceFactors\_Ass3, contains information on various factors, including hours studied, attendance, parental involvement, access to resources, and exam scores. In this report, we perform basic data loading, exploratory data analysis, data manipulation, and regression modeling to uncover patterns and insights.

**Data Loading and Exploration:** The dataset is loaded using the readxl library in R, which allows us to access and view the data. The View() function provides a glimpse into the entire dataset, while the head() function is used to display the first 10 rows for initial inspection. The dataset consists of 25 variables, including numerical data such as Hours\_Studied and Exam\_Score, as well as categorical data like Parental\_Involvement, Motivation\_Level, and Gender.

The following are a few of the columns in the dataset:

* Hours\_Studied: Hours spent studying by the student (numeric).
* Attendance: Percentage of classes attended by the student (numeric).
* Parental\_Involvement: Level of involvement by the student's parents (categorical).
* Exam\_Score: The score obtained in the exam (numeric).

**Statistical Measures:**

1. **Mean and Median of Exam Scores:**
   * **Mean Exam Score:** The average exam score across all students is calculated using the mean() function with the argument na.rm = TRUE to exclude missing values. The mean score is found to be approximately **61.85**.
   * **Median Exam Score:** The median score, which represents the middle value in the distribution of exam scores, is calculated using the median() function and is found to be **62.00**. The median is close to the mean, suggesting a roughly symmetric distribution of exam scores.

**Data Manipulation:**

1. **Filtering Data Based on Hours Studied:** We filter the dataset to include only those students who studied for at least 10 hours using the subset() function. This allows us to focus on the subset of data where students are more likely to show performance influenced by their study habits. The filtered\_data contains all students who studied for 10 or more hours.
2. **Categorizing Exam Scores into Levels:** A new variable, score\_level, is created to categorize students' exam scores into three levels: "High" (for scores greater than 70), "Medium" (for scores between 50 and 70), and "Low" (for scores below 50). The ifelse() function is used to perform this categorization. This variable is useful for understanding how many students fall into each performance category.

**Data Visualization:**

1. **Boxplot of Exam Scores by Hours Studied:** A boxplot is created using ggplot2 to visualize the distribution of exam scores grouped by the number of hours studied. The boxplot illustrates how exam scores tend to increase as the number of study hours rises, highlighting that higher study time is associated with better exam performance.

The boxplot uses the geom\_boxplot() function and is customized with colors to make the visualization clear. The x axis represents the study time (in hours), while the y axis represents the exam score.

1. **Histogram of Exam Scores:** A histogram is plotted to show the overall distribution of exam scores across all students. The histogram, created with geom\_histogram(), uses a binwidth of 5 to group exam scores. The histogram reveals that most students scored between 50 and 70, with fewer students achieving very high or low scores.

**Regression Analysis:**

1. **Linear Regression of Exam Score on Hours Studied:** A linear regression model is built to investigate the relationship between Hours\_Studied (independent variable) and Exam\_Score (dependent variable). The model is created using the lm() function, and the results of the model summary are as follows:
   * **Intercept (β₀):** 61.46
   * **Slope (β₁):** 0.29
   * **R-squared:** 0.1984
   * **p-value:** < 2.2e-16

The regression model shows that for each additional hour of study, the exam score increases by approximately 0.29 points. The p-value for the slope is less than 0.05, indicating that the relationship between hours studied and exam score is statistically significant.

The R-squared value of 0.1984 means that about 19.84% of the variation in exam scores can be explained by the number of hours studied, suggesting that while study time plays a role in performance, other factors may also influence exam scores.

1. **Scatter Plot of Hours Studied vs. Exam Score with Regression Line:** A scatter plot is created to show the relationship between Hours\_Studied and Exam\_Score, with a red regression line overlaid. The plot reveals a positive trend, confirming that as study time increases, exam scores tend to improve. The line represents the best fit according to the linear regression model.

**Correlation:** To quantify the strength of the linear relationship between study time and exam scores, the Pearson correlation coefficient is calculated. The correlation coefficient is found to be approximately **0.445**, which indicates a moderate positive relationship between hours studied and exam scores.

**Conclusion:** This exploratory analysis reveals that there is a positive correlation between hours studied and exam scores, as confirmed by both the regression model and the correlation coefficient. The data also shows that most students fall into the "Medium" performance category. Visualizations, such as boxplots and histograms, further support the idea that study time positively impacts exam scores. However, the relatively low R-squared value suggests that other factors, such as parental involvement, motivation, and resources, may also influence exam performance. Further analysis of these factors would be useful in gaining a deeper understanding of student performance.