**Assessment Test: Data Analytics (Excel and R Studio)**

**Dataset**: Student Performance Factors

**Part 1: Excel (50 marks)**

1. **Data Cleaning and Manipulation (10 marks)**

o Open the dataset in Excel and perform the following tasks:

▪ Remove any rows with missing values in the score column.

▪ Filter the dataset to show only students with sleep-hours greater than 7.

2. **Basic Functions and Calculations (10 marks)**

o Use formulas to answer the following:

▪ Calculate the average score of students whose school is public.

▪ Count the number of students in the dataset who are from the public School. 3. **Data Visualization (15 marks)**

o Create the following visualizations:

▪ A **bar chart** that shows the number of students grouped by their internet access (yes or no).

▪ A **scatter plot** showing the relationship between studytime and score. ▪ Add labels and a title to both charts.

4. **Pivot Tables (15 marks)**

o Create a Pivot Table to summarize the data:

▪ Show the average score for each combination of school and internet access.

▪ Use the pivot table to sort the results by the highest average score.

**Part 2: R Studio (50 marks)**

1. **Basic Data Loading and Exploration (10 marks)**

o Load the dataset in R using read.csv (). Perform the following:

▪ Display the first 10 rows of the dataset.

▪ Calculate and print the mean and median score of all students.

2. **Data Manipulation (15 marks)**

o Using the dataset, perform the following tasks:

▪ Filter the data to include only students with studytime of 10 or more

hours.

▪ Create a new column called score level that categorizes students based on their score:

▪ "High" for scores above 70.

▪ "Medium" for scores between 50 and 70.

▪ "Low" for scores below 50.

3. **Data Visualization (10 marks)**

o Using ggplot2, create the following visualizations:

▪ A **boxplot** showing the distribution of score for students grouped by studytime.

▪ A **histogram** showing the distribution of score across all students. ▪ Customize the plots with titles and axis labels.

4. **Statistical Analysis (15 marks)**

o Perform a simple linear regression using the following variables:

▪ **Dependent variable**: score

▪ **Independent variable**: studytime

▪ Plot the regression line on a scatter plot of studytime vs score.

▪ Calculate the correlation coefficient between score and studytime.

**Instructions:**

∙ **Time**: 1.5 hours

∙ **Software**: Microsoft Excel and R Studio

∙ **Dataset**: Student Performance Factors

∙ **Resources**: Online references are allowed.

http://127.0.0.1:31835/graphics/plot.png?width=569&height=365&randomizer=500738299**R STUDIO ANSWERS**

| #Basic Data Loading and Exploration  > library(readxl)  > View(Mubarak\_Daud\_StudentPerformanceFactors\_Ass3)  > head(Mubarak\_Daud\_StudentPerformanceFactors\_Ass3,10)  # A tibble: 10 × 25  Hours\_Studied Attendance Parental\_Involvement Access\_to\_Resources  *<dbl>* *<dbl>* *<chr>* *<chr>*  1 23 84 Low High  2 19 64 Low Medium  3 24 98 Medium Medium  4 29 89 Low Medium  5 19 92 Medium Medium  6 19 88 Medium Medium  7 29 84 Medium Low  8 25 78 Low High  9 17 94 Medium High  10 23 98 Medium Medium  # ℹ 21 more variables: Extracurricular\_Activities <chr>, Sleep\_Hours <dbl>,  # Previous\_Scores <dbl>, Motivation\_Level <chr>, Internet\_Access <chr>,  # Tutoring\_Sessions <dbl>, Family\_Income <chr>, Teacher\_Quality <chr>,  # School\_Type <chr>, Peer\_Influence <chr>, Physical\_Activity <dbl>,  # Learning\_Disabilities <chr>, Parental\_Education\_Level <chr>,  # Distance\_from\_Home <chr>, Gender <chr>, Exam\_Score <dbl>, `Average score` <dbl>,  # `Number of students in public` <chr>, ...23 <chr>, ...24 <chr>, ...25 <chr>  > mean\_score <- mean(Mubarak\_Daud\_StudentPerformanceFactors\_Ass3$Exam\_Score, na.rm = TRUE)  > View(mean\_score)  > median\_score <- median(Mubarak\_Daud\_StudentPerformanceFactors\_Ass3$Exam\_Score, na.rm = TRUE)  > View(median\_score)  > #Data Manipulation  > filtered\_data <- subset(Mubarak\_Daud\_StudentPerformanceFactors\_Ass3, Hours\_Studied >= 10)  > View(filtered\_data)  > Mubarak\_Daud\_StudentPerformanceFactors\_Ass3$score\_level <- with(Mubarak\_Daud\_StudentPerformanceFactors\_Ass3, ifelse(Exam\_Score > 70, "High", ifelse(Exam\_Score >= 50 & Exam\_Score <= 70, "Medium", "Low")))  > View(Mubarak\_Daud\_StudentPerformanceFactors\_Ass3)  > # Create a boxplot to show the distribution of Exam\_Score grouped by Hours\_Studied  > ggplot(Mubarak\_Daud\_StudentPerformanceFactors\_Ass3, aes(x = factor(Hours\_Studied), y = Exam\_Score)) +  + geom\_boxplot(fill = "lightblue", color = "darkblue") +  + labs(title = "Distribution of Exam Scores by Study Time",  + x = "Study Time (Hours)",  + y = "Exam Score") +  + theme\_minimal()  > # Histogram showing the distribution of Exam\_Score  > ggplot(Mubarak\_Daud\_StudentPerformanceFactors\_Ass3, aes(x = Exam\_Score)) +  + geom\_histogram(binwidth = 5, fill = "lightgreen", color = "black") +  + labs(title = "Distribution of Exam Scores Across All Students",  + x = "Exam Score",  + y = "Frequency") +  + theme\_minimal()  > # Linear regression (Dependent variable: Exam\_Score, Independent variable: Hours\_Studied)  > model <- lm(Exam\_Score ~ Hours\_Studied, data = Mubarak\_Daud\_StudentPerformanceFactors\_Ass3)  > # Summary of the linear regression model  > summary(model)  Call:  lm(formula = Exam\_Score ~ Hours\_Studied, data = Mubarak\_Daud\_StudentPerformanceFactors\_Ass3)  Residuals:  Min 1Q Median 3Q Max  -8.532 -2.243 -0.111 2.046 33.493  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 61.456984 0.149196 411.92 <2e-16 \*\*\*  Hours\_Studied 0.289291 0.007154 40.44 <2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 3.483 on 6605 degrees of freedom  Multiple R-squared: 0.1984, Adjusted R-squared: 0.1983  F-statistic: 1635 on 1 and 6605 DF, p-value: < 2.2e-16  > # Scatter plot of Hours\_Studied vs. Exam\_Score with the regression line  > ggplot(Mubarak\_Daud\_StudentPerformanceFactors\_Ass3, aes(x = Hours\_Studied, y = Exam\_Score)) +  + geom\_point(color = "blue") + # Scatter plot points  + geom\_smooth(method = "lm", color = "red", se = FALSE) + # Add the regression line  + labs(title = "Study Time vs. Exam Score",  + x = "Study Time (Hours)",  + y = "Exam Score") +  + theme\_minimal()  `geom\_smooth()` using formula = 'y ~ x'  > # Correlation between Exam\_Score and Hours\_Studied  > correlation\_coefficient <- cor(Mubarak\_Daud\_StudentPerformanceFactors\_Ass3$Exam\_Score, Mubarak\_Daud\_StudentPerformanceFactors\_Ass3$Hours\_Studied, use = "complete.obs")  > View(correlation\_coefficient) |
| --- |
|  |
| |  | | --- | |

**DATA VISUALIZATION**

http://127.0.0.1:31835/graphics/plot.png?width=569&height=365&randomizer=-2101701628