**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 an age greater than 18.

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

o Use formulas to answer the following:

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

▪ Count the number of students in the dataset who are from the GP 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 2 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.

Answer for r

|  |
| --- |
| > library(readxl)  > student\_factors <- read\_excel("student factors.xlsx")  New names:  • `` -> `...4`  > View(student\_factors)  > head(student\_factors,10)  # A tibble: 10 × 20  Hours\_Studied Attendance Parental\_Involvement ...4 Access\_to\_Resources  *<dbl>* *<dbl>* *<chr>* *<chr>* *<chr>*  1 23 84 Low High No  2 19 64 Low Medium No  3 24 98 Medium Medium Yes  4 29 89 Low Medium Yes  5 19 92 Medium Medium Yes  6 19 88 Medium Medium Yes  7 29 84 Medium Low Yes  8 25 78 Low High Yes  9 17 94 Medium High No  10 23 98 Medium Medium Yes  # ℹ 15 more variables: 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>  > summary(student\_factors)  Hours\_Studied Attendance Parental\_Involvement ...4  Min. : 1.00 Min. : 60.00 Length:6607 Length:6607  1st Qu.:16.00 1st Qu.: 70.00 Class :character Class :character  Median :20.00 Median : 80.00 Mode :character Mode :character  Mean :19.98 Mean : 79.98  3rd Qu.:24.00 3rd Qu.: 90.00  Max. :44.00 Max. :100.00  Access\_to\_Resources Sleep\_Hours Previous\_Scores Motivation\_Level  Length:6607 Min. : 4.000 Min. : 50.00 Length:6607  Class :character 1st Qu.: 6.000 1st Qu.: 63.00 Class :character  Mode :character Median : 7.000 Median : 75.00 Mode :character  Mean : 7.029 Mean : 75.07  3rd Qu.: 8.000 3rd Qu.: 88.00  Max. :10.000 Max. :100.00  Internet\_Access Tutoring\_Sessions Family\_Income Teacher\_Quality  Length:6607 Min. :0.000 Length:6607 Length:6607  Class :character 1st Qu.:1.000 Class :character Class :character  Mode :character Median :1.000 Mode :character Mode :character  Mean :1.494  3rd Qu.:2.000  Max. :8.000  School\_Type Peer\_Influence Physical\_Activity  Length:6607 Length:6607 Min. :0.000  Class :character Class :character 1st Qu.:2.000  Mode :character Mode :character Median :3.000  Mean :2.968  3rd Qu.:4.000  Max. :6.000  Learning\_Disabilities Parental\_Education\_Level Distance\_from\_Home  Length:6607 Length:6607 Length:6607  Class :character Class :character Class :character  Mode :character Mode :character Mode :character        Gender Exam\_Score  Length:6607 Min. : 55.00  Class :character 1st Qu.: 65.00  Mode :character Median : 67.00  Mean : 67.24  3rd Qu.: 69.00  Max. :101.00  > mean(student\_factors$Exam\_Score,na.rm = T)  [1] 67.23566  > median(student\_factors$Exam\_Score,na.rm = T)  [1] 67  > # Filter students with study hours greater than 2  > filtered\_data <- subset(student\_factors$Hours\_Studied >10)  Error in subset.default(student\_factors$Hours\_Studied > 10) :  argument "subset" is missing, with no default  > # Assuming your data frame is 'student\_factors' and the column is 'Hours\_Studied'  > filtered\_data <- subset(student\_factors, Hours\_Studied > 10)  > View(filtered\_data)  > #create a column  > student\_factors$Exam\_Score <- cut(student\_factors$Exam\_Score,  + breaks = c(-Inf, 50, 70, Inf),  + labels = c("Low", "Medium", "High"))  > View(student\_factors)  > library(ggplot2)  > # Boxplot of Exam\_Score grouped by Hours\_Studied  > 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()  > # Histogram of Exam\_Score  > ggplot(student\_factors, aes(x = Exam\_Score)) +  + geom\_histogram(binwidth = 10, fill = "lightgreen", color = "black") +  + labs(title = "Score Distribution of Students", x = "Exam Score", y = "Frequency") +  + theme\_minimal()  **Error in `geom\_histogram()`:**  ! Problem while computing stat.  ℹ Error occurred in the 1st layer.  **Caused by error in `setup\_params()`:**  ! `stat\_bin()` requires a continuous x aesthetic.  ✖ the x aesthetic is discrete.  ℹ Perhaps you want `stat="count"`?  Run `rlang::last\_trace()` to see where the error occurred.  > # Load ggplot2  > library(ggplot2)  > # Ensure Exam\_Score is numeric (only needed if it's not already)  > student\_factors$Exam\_Score <- as.numeric(as.character(student\_factors$Exam\_Score))  Warning message:  NAs introduced by coercion  > # Create the histogram  > 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) + # Customize theme for clarity  + theme(plot.title = element\_text(hjust = 0.5)) # Center the title  Warning message:  Removed 6607 rows containing non-finite outside the scale range  (`stat\_bin()`).  > # Perform the linear regression  > model <- lm(Exam\_Score ~ Hours\_Studied, data = student\_factors)  Error in lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :  0 (non-NA) cases  > # View the summary of the regression model  > summary(model)  Error: object 'model' not found  > # Scatter plot with regression line  > ggplot(student\_factors, aes(x = Hours\_Studied, y = Exam\_Score)) +  + geom\_point(color = "blue") + # Scatter plot of studytime vs score  + geom\_smooth(method = "lm", se = FALSE, color = "red") + # Regression line  + labs(title = "Study Time vs Exam Score with Regression Line",  + x = "Study Time (Hours)",  + y = "Exam Score") +  + theme\_minimal()  `geom\_smooth()` using formula = 'y ~ x'  Warning messages:  1: Removed 6607 rows containing non-finite outside the scale range  (`stat\_smooth()`).  2: Removed 6607 rows containing missing values or values outside the scale  range (`geom\_point()`).  > # Calculate the correlation coefficient between Hours\_Studied and Exam\_Score  > cor(student\_factors$Hours\_Studied, student\_factors$Exam\_Score)  [1] NA |
|  |
| |  | | --- | | > | |

http://127.0.0.1:16181/graphics/4e01bda6-9461-40a6-a817-9ba9d54c2024.png http://127.0.0.1:16181/graphics/4e01bda6-9461-40a6-a817-9ba9d54c2024.png http://127.0.0.1:16181/graphics/4e01bda6-9461-40a6-a817-9ba9d54c2024.png http://127.0.0.1:16181/graphics/4e01bda6-9461-40a6-a817-9ba9d54c2024.png http://127.0.0.1:16181/graphics/4e01bda6-9461-40a6-a817-9ba9d54c2024.png http://127.0.0.1:16181/graphics/4e01bda6-9461-40a6-a817-9ba9d54c2024.png http://127.0.0.1:16181/graphics/4e01bda6-9461-40a6-a817-9ba9d54c2024.png http://127.0.0.1:16181/graphics/4e01bda6-9461-40a6-a817-9ba9d54c2024.png http://127.0.0.1:16181/graphics/4e01bda6-9461-40a6-a817-9ba9d54c2024.png http://127.0.0.1:16181/graphics/4e01bda6-9461-40a6-a817-9ba9d54c2024.png http://127.0.0.1:16181/graphics/4e01bda6-9461-40a6-a817-9ba9d54c2024.png http://127.0.0.1:16181/graphics/4e01bda6-9461-40a6-a817-9ba9d54c2024.png