VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CHENNAI

CSE3040

Exploratory Data Analysis

Digital Assignment - 1

On

Empowering Education: Exploring Student Data for a Brighter Future

By

Tarlana Vidya

23MIA1176

Sustainable Goal Number: 4

Sustainable Goal Name: Quality Education

About the SDG Goal:

SDG Goal 4 aims to ensure inclusive, equitable, and quality education for all, promoting lifelong learning opportunities. It focuses on accessibility, reducing dropout rates, and improving learning outcomes to build a better future.

Mapping the EDA Assignment to SDG Goal 4:

This EDA project analyses student data to uncover factors affecting academic performance, such as study time, gender, school type, and attendance. By identifying patterns in student success and challenges, this project supports data-driven strategies for improving education quality and reducing inequalities, aligning with SDG 4.

About the Dataset:

This dataset provides insights into the lives of 649 students with about 33 attributes, capturing their personal details, family background, academic journey, and lifestyle choices. It aims to analyse how various factors—such as study habits, parental support, and social activities—influence student performance.

By exploring this data, we can uncover what contributes to student success and the challenges they encounter in their education. This dataset serves as a valuable tool for predicting academic outcomes, identifying areas where students need support, and shaping more effective educational policies.

Dataset Link:

https://www.kaggle.com/datasets/larsen0966/student-performance-data-set

Loading the necessary libraries and finding the dimensions of the dataset:

```
> # Loading the necessary libraries
> library(ggplot2)
> library(reshape2)
> library(RColorBrewer)
> library(moments)
> library(dplyr)
> # Loading the dataset
> data <- read.csv("Student_dataset.csv")
> #Finding the dimensions of the dataset
> dimensions_of_dataset = dim(data)
> print(dimensions_of_dataset)
[1] 649  33
>
```

EDA Types

Univariate Analysis:

Non – Graphical Analysis:

Categorical

- Frequency Table

```
> #Frequency Table
> print(table(data$school))

 GP  MS
423 226
> #Counts how many students are in each school
```

Inference: This helps understand the distribution of students across schools, which is useful in assessing whether certain schools have more students and whether school size affects student performance, contributing to SDG Goal Quality Education.

Quantitative

- Measures of Central Tendency
    1. Mean

```
> #Mean
> print(mean(data$age))
[1] 16.74422
> #Calculates the average age of students
```

Inference: Helps to know the typical age of students. Understanding student age distribution is crucial for designing age-appropriate learning strategies, aligning with SDG Goal 4.

2. Median

```
> #Median
> print(median(data$age))
[1] 17
> #Finds the middle value of the age
```

Inference: Useful when data has extreme values that may affect the mean. A balanced age distribution can indicate inclusivity in education.

3. Mode

```
> #Mode
> mode_function <- function(x) {
+   return(names(sort(table(x), decreasing = TRUE)[1]))
+ }
> print(mode_function(data$age))
[1] "17"
> #Finds the most frequently occurring age
```

Inference: Shows the most common age among students, helping in designing tailored educational programs.

- Measures of Dispersion
  1. Variance

```
> #Variance
> print(var(data$age))
[1] 1.483859
> #Measures how the ages are spread out
```

Inference: A higher variance means greater variation in ages, which can indicate diverse student enrolment and potential gaps in educational accessibility.

2. Standard Deviation

```
> #Standard Deviation
> print(sd(data$age))
[1] 1.218138
> #Calculates the standard deviation of age
```

Inference: A low SD means most students have similar ages, indicating a uniform education structure. High SD might suggest educational disparity.

3. Interquartile Range

```
> #IQR
> print(IQR(data$age))
[1] 2
> #Finds the middle 50% range of ages
```

Inference: Helps understand age dispersion without being affected by outliers, ensuring education policies cater to a broad spectrum of students.

4. Range

```
> #Range
> print(range(data$age))
[1] 15 22
> #Shows the minimum and maximum ages
```

Inference: Provides the full span of student ages, useful in evaluating educational inclusivity across different age groups.

- Measures of Symmetry
    1. Skewness

```
> #Skewness
> skewness(data$G3)
[1] -0.910798
> #Measures the asymmetry of the data.
```

Inference**:** If skewness is positive, the G3 tend to be concentrated on the lower end, with a few students having very high scores. If it's negative, the distribution is weighted towards higher scores, with few students getting low grades.

- Measures of Peakiness
    1. Kurtosis

```
> #Kurtosis
> kurtosis(data$G3)
[1] 5.682123
> #Measures the "tailedness" of the data.
```

Inference: If kurtosis is high (greater than 3), it suggests that the final grades have a lot of extreme values (outliers). If it's low (less than 3), the distribution is more uniform, with fewer extreme scores.

Graphical Analysis:

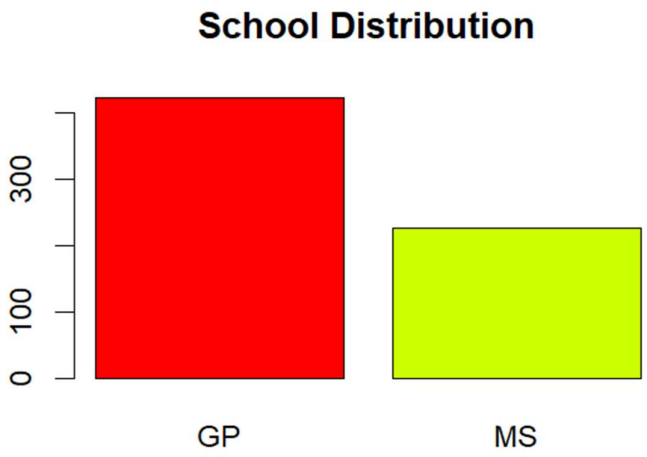
Categorical

- Bar Plot

Code:

```
> #Barplot
> barplot(table(data$school), col = rainbow(5), main = "School Distribution")
> #Creates a bar plot for school counts
>
```

Plot:



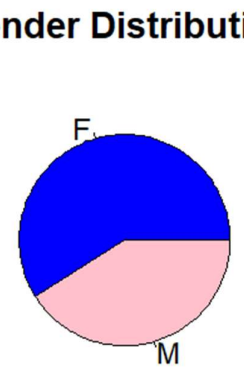
Inference: Helps compare the number of students in each school visually, assisting in assessing whether resources and teachers are allocated properly in schools, aligning with SDG Goal 4.

- Pie Chart

Code:

```
> #Pie Chart
> pie(table(data$sex), col = c("blue", "pink"), main = "Gender Distribution")
> #Creates a pie chart for gender
>
```

Plot:



Inference: Helps visualize gender proportions. Gender balance in education is essential for ensuring equal learning opportunities, supporting SDG Goal 5: Gender Equality as well.

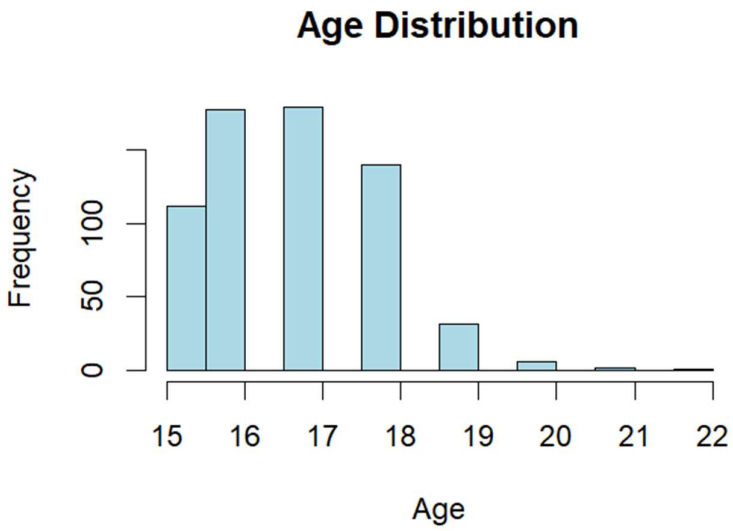
Quantitative

- Histogram:

Code:

```
> #Histogram
> hist(data$age, col = "lightblue", main = "Age Distribution", xlab = "Age")
> #Creates a histogram for age
>
```

Plot:



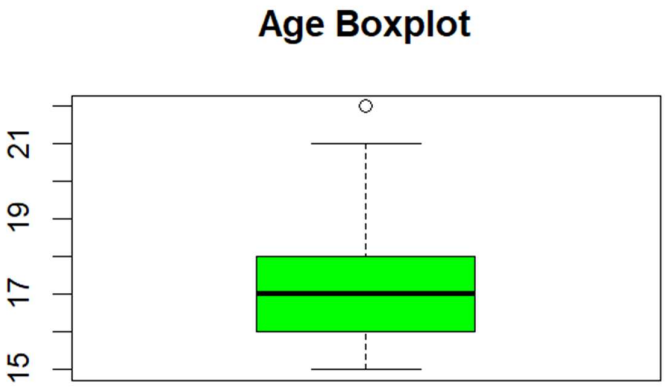
Inference: Helps see how ages are distributed, assisting in curriculum planning and support programs.

- Box Plot:

Code:

```
> #Box Plot
> boxplot(data$age, col = "green", main = "Age Boxplot")
> #Creates a boxplot for age
>
```

Plot:



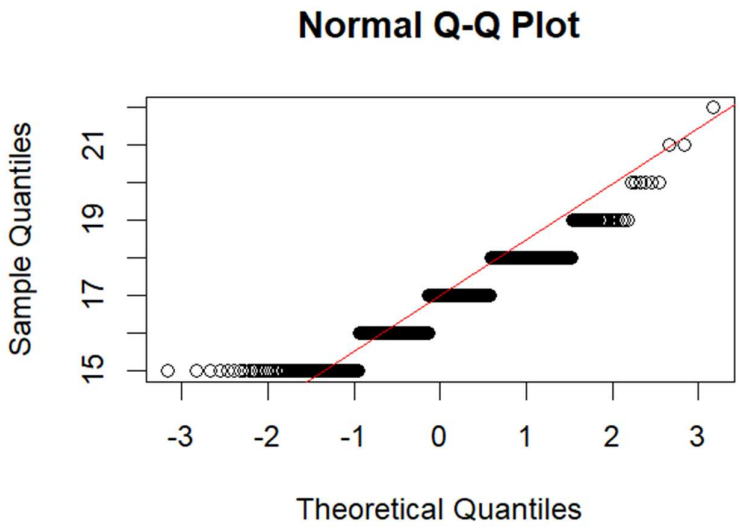
Inference: Helps identify outliers and distribution shape. Outliers in student ages may indicate delayed education or dropouts, a critical issue related to SDG Goal 4.

- Quantile Plot (Normal Plot)

Code:

```
> #Quantile Plot
> qqnorm(data$age)
> #Creates a normal probability plot
> qqline(data$age, col = "red")
> #Adds a reference line to check normality
>
```

Plot:



Inference: Helps assess if age data follows a normal distribution. A non-normal distribution might indicate age gaps due to educational barriers.

- Stem and Leaf Plot

```
> #Stem and Leaf Plot
> stem(data$age)

  The decimal point is at the |

  15 | 00000000000000000000000000000000000000000000000000000000000000000000000+32
  15 |
  16 | 00000000000000000000000000000000000000000000000000000000000000000000000+97
  16 |
  17 | 00000000000000000000000000000000000000000000000000000000000000000000000+99
  17 |
  18 | 00000000000000000000000000000000000000000000000000000000000000000000000+60
  18 |
  19 | 00000000000000000000000000000000
  19 |
  20 | 000000
  20 |
  21 | 00
  21 |
  22 | 0

> #organizes numerical data by splitting values into stems
> #(leading digits) and leaves (trailing digits)
>
```

Inference: By analysing the plot, we can determine if the data is symmetrical, skewed, or clustered around certain values. If the leaves are evenly spread, the distribution is uniform, while an uneven spread suggests skewness. The plot also helps in detecting outliers and identifying common age groups within the dataset.

Multivariate Analysis (Bivariate Analysis)**:**

Non – Graphical Analysis:

Categorical:

- Cross – Tabulation

```
> #Cross Tabulation
> print(table(data$sex, data$school))

     GP  MS
  F 237 146
  M 186  80
> #Creates a table showing school vs gender
> |
```

Inference: Helps understand gender distribution across different schools. Gender balance is essential in ensuring equitable access to education, supporting SDG Goals 4 and 5.

Quantitative:

- Co-variance:

```
> #Co-variance
> print(cov(data$G1, data$G3))
[1] 7.329234
> #Calculates covariance between first and final grades
> |
```

Inference: Shows if grades move together (positive or negative relationship). A strong relationship suggests that early academic performance is a good predictor of future success, emphasizing the importance of strong foundational learning.

- Co-relation:

```
> #Correlation
> print(cor(data$G1, data$G3))
[1] 0.8263871
> #Calculates correlation between first and final grades
> |
```

Inference: A high correlation suggests early grades predict final grades well. This can help in early intervention strategies for struggling students, ensuring quality education and reducing dropout rates (SDG Goal 4).

Graphical:

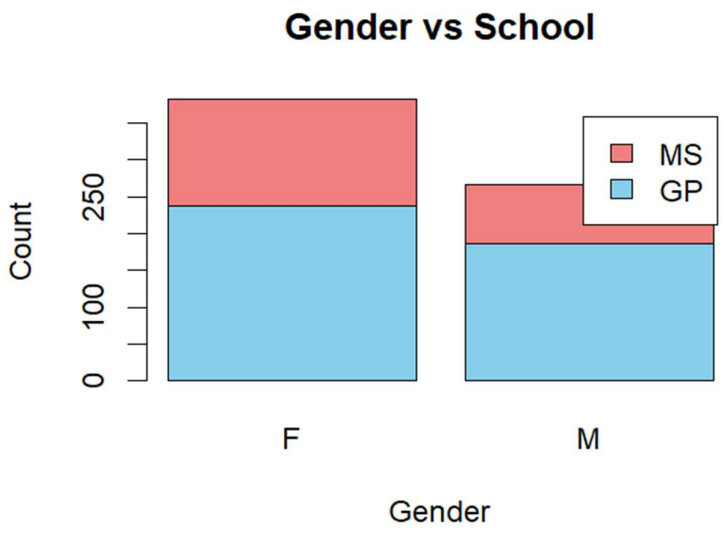
Categorical:

- Stacked Bar Plot

Code:

```
> #Stacked Bar Plot
> barplot(table(data$school, data$sex), beside = FALSE, col = c("skyblue", "lightcoral"),
+ main = "Gender vs School", xlab = "Gender", ylab = "Count",
+ legend = rownames(table(data$school, data$sex)))
> #visualizes two categorical variables by stacking segments within bars
>
```

Plot:



Inference: The stacked bar plot shows the gender distribution across schools, highlighting variations in representation among them.

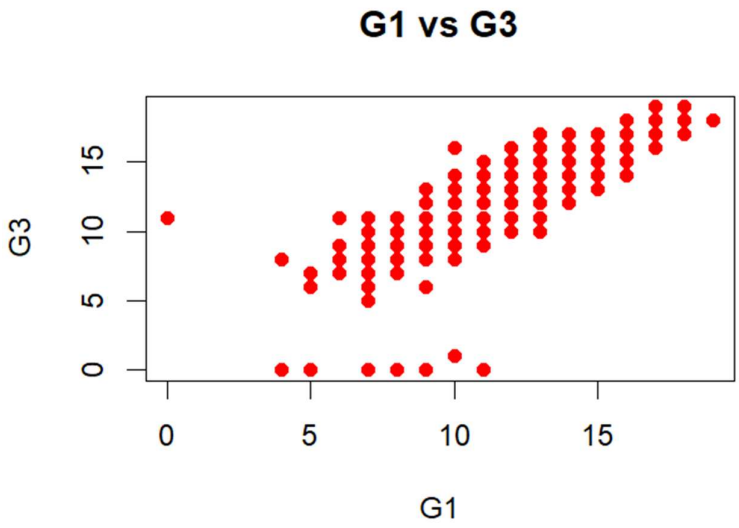
Quantitative:

- Scatter Plot

Code:

```
> #Scatter Plot
> plot(data$G1, data$G3, col = "red", pch = 16, main = "G1 vs G3", xlab = "G1", ylab = "G3")
> #Creates a scatter plot for grades
>
```

Plot:



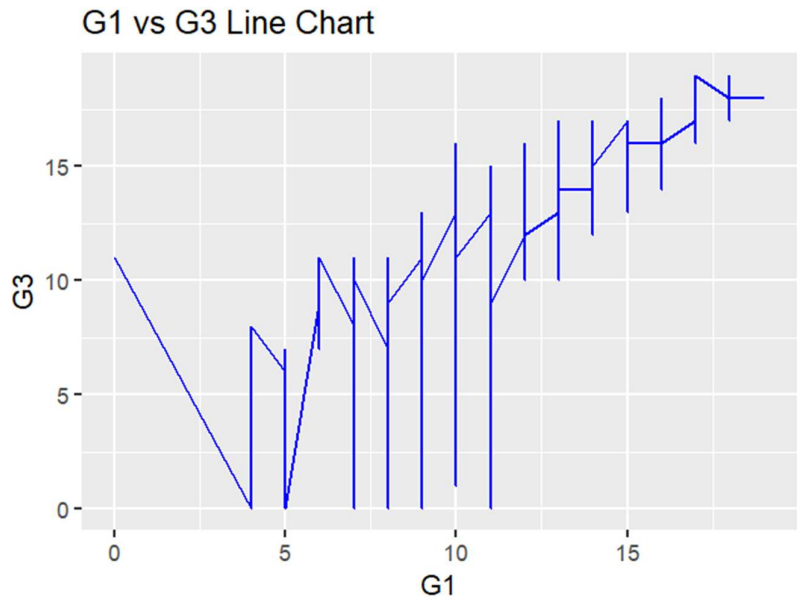
Inference: Helps see if early grades (G1) influence final grades (G3). Early interventions for weak students can improve final outcomes, aligning with SDG Goal 4.

- Line Chart

Code:

```
> #Line Chart
> ggplot(data, aes(x = G1, y = G3)) + geom_line(color = "blue") +
+ ggtitle("G1 vs G3 Line Chart")
> #Creates a line chart for grades
>
```

Plot:



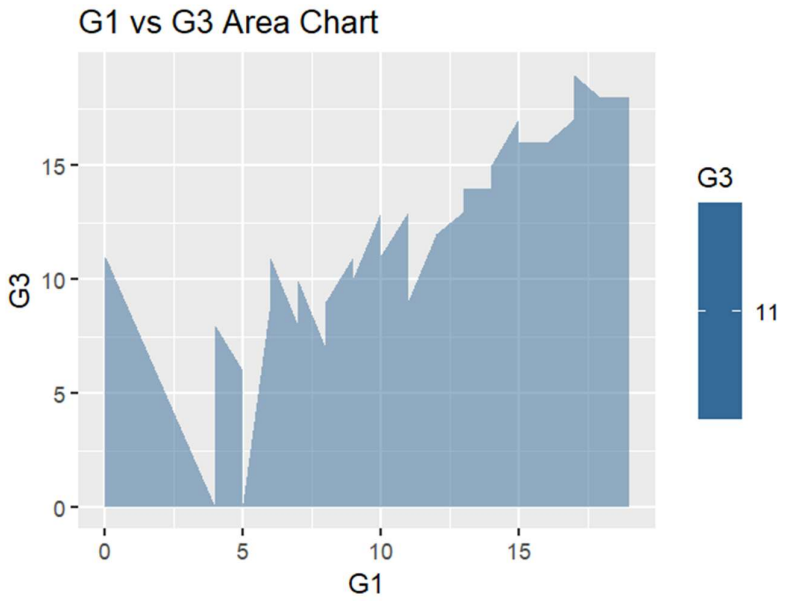
Inference: Helps track changes in grades over time, allowing for trend analysis in student performance.

- Area Graph

Code:

```
> #Area Graph
> ggplot(data, aes(x = G1, y = G3, fill = G3)) + geom_area(alpha = 0.5) +
+ ggtitle("G1 vs G3 Area Chart")
> #Creates an area graph for grades
> |
```

Plot:



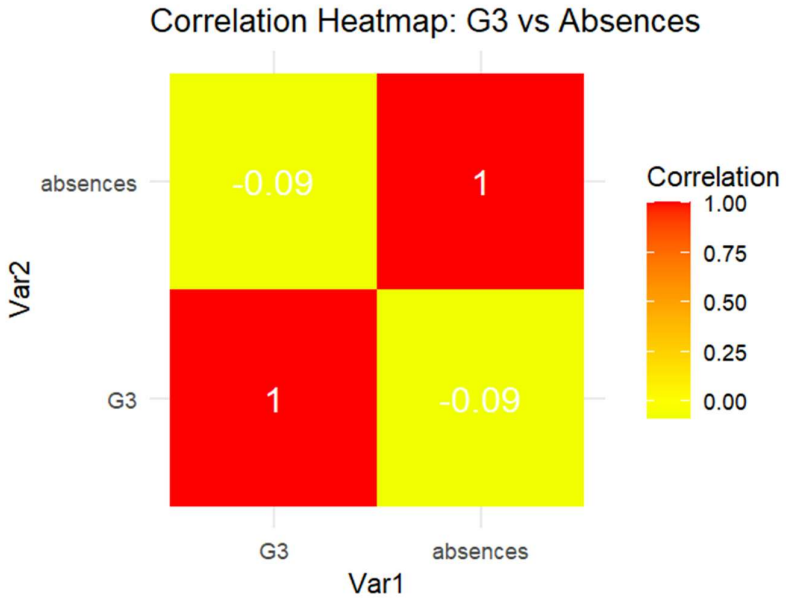
Inference: Helps visualize grade differences more clearly. Identifying performance dips allows for better academic intervention, supporting SDG Goal 4.

- Heat Map (Correlation Matrix):

Code:

```
> #Heatmap
> #Selecting two numeric variables for finding the correlation
> cor_matrix <- cor(data[c("G3", "absences")], use = "complete.obs")
> #Converting the data into long format for heatmap
> melted_cor <- melt(cor_matrix)
> #Creating heatmap
> ggplot(melted_cor, aes(Var1, Var2, fill = value)) + geom_tile() +
+ geom_text(aes(label = round(value, 2)), color = "white", size = 5) +
+ scale_fill_gradient2(low = "green", high = "red", mid = "yellow", midpoint = 0) +
+ theme_minimal() +
+ labs(title = "Correlation Heatmap: G3 vs Absences", fill = "Correlation")
> |
```

Plot:



Inference: The R code calculates the correlation between G3 (final grade) and absences using a heatmap. A negative correlation suggests that students with more absences tend to have lower final grades.