Statistical_Report[1][1] New.docx

Assignment

Class

University

Document Details

Submission ID

trn:oid:::1:3018946371

Submission Date

Sep 24, 2024, 1:15 PM UTC

Download Date

Sep 24, 2024, 1:16 PM UTC

File Name

Statistical_Report11_New.docx

File Size

5.3 MB

41 Pages

2,754 Words

15,608 Characters

0% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Detection Groups

1 AI-generated only 0%
Likely AI-generated text from a large-language model.

2 AI-generated text that was AI-paraphrased 0%
Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

Statistical Report

Introduction to the Statistical Report

In this report, we will discuss and deal with a few questions that depict the direction taken by the educational institutes as part of their business as well as the education and improvements within the education aspect of it for the students which will help navigate educational institutions such as schools, colleges ad universities to make better conscious decisions based on this statistical report for a better education for the world's future.

Key Question:

Evaluation of student performance in terms of Extra extracurricular activities and its effect on GPA and attendance rate.

Why?

Answering, this mentioned problem will not only help educational institutions find better ways to increase student performance in their educational institutions. Moreover, it will also be more attractive to other students who would want to join their educational institute based on the average student performances, extra-curricular activities, and social building of the educational institute, resulting in more business for the educational institutes and better student performance being the utmost root cause of it which can be marketed to the benefit of the educational institute examples of this can be seen in abundance with Harvard University, the University of Oxford. The University of Cambridge amongst many others is known for higher student performance rates and uses that to market themselves further. Examples of educational institutions marketing themselves for better extracurricular activities include Cornell University, and Baylor University amongst many others. Therefore, to market the educational institutes, the relationship in terms of increased or decreased student performance in relation to other curricula is key.

Additionally, the relationship between part-time jobs and student performance is key to making better policies or educational platforms for students, concerning existing governmental laws or educational institutional laws, for example, 20-hour working permits per week, etc. Therefore, this statistical report can be used as building grounds for not only better decisions made by educational institutions in increasing student performance levels for better governmental policies in aiding the educational institutions in increasing student performance levels. A known country for setting a higher precedent in this regard would be Finland, which has managed to secure the world's best educational system.

Furthermore, this statistical report will allow educational institutes to build on student performance by the help of the data set and will therefore also be aided in identifying the causes of low student performances or outlining promising gains in student performance when put in relation to extra-curricular activities, part-time jobs, etc. This, therefore, being the main "why?" we will be diving deep into with the dataset and concluding this statistical report on.

About the Dataset

This dataset is useful in evaluating students' performance in different schools in terms of their achievements & their demographics. The information that can be categorized as essential includes; student number, gender, age, grade level, and the result of a set of subject including; math, reading, and writing. Furthermore, it keeps track of attendance, as to how the students' presence in school relates to their performance.

Key features include:

Student ID: Individual codes for every child which prevent data distortions and allow following the development of a child through the years.

Gender and Age: Other personal characteristics, which may be used in evaluating performance patterns by various categories.

Grade Level: Special information about students and their academic level which is necessary to analyze their performance according to their age.

GPA: Diagnostic scores which can be general as well as subject-specific, and can give a complete insight of the overall and specific learning profiles needed for academic progress.

Attendance: Student enrollment information because they can help in determining the effectiveness of school attendance in the academic performance.

Major: The subjects like Education, science, business etc. Additional
Details: Extra curricular activities participation Part time job Information:
The students who do job or not.

Description of dataset

The main corpus includes data that characterized 500 students of different universities and contain 9 parameters. These variables consist of and range from the students' gender which is both male and female as well as age which lies between 18-24 years.

Furthermore, the dataset reveals the students' major fields of study like Science, Education, Business, Engineering, and Arts or popularity of the classes, and weekly study hours. It also provides a clear understanding on the ways and extents of association between the control variables such as the students' majors, age and gender and the dependant

variable that is the students' performance as reflected by the GPA.

Also, the data collected contain students' involvement in after-school activities and some employment, whether paid or unpaid.

In general, this dataset provides essential information for anyone interested in investing time and effort in studying educational performance and for creating effective strategies for increasing students' effectiveness. It assists in forecasting about the trends in the academic status of the students and variables impacting on it.

Besides, the feature-wise summary of dataset and function providing mean, median, 1st quartile, 3rd quartile, standard deviation, kurtosis etc is also described in code section.

R programming packages applied in this document include readxl, tidyverse, ggplot2, car, Rmarkdown, among others.

This dataset provides a detailed overview of student performance in various schools, focusing on academic achievements and demographic factors. It includes critical information such as student IDs, gender, age, grade levels, and scores in key subjects like mathematics, reading, and writing. Additionally, it captures attendance records, offering insights into how presence in school correlates with academic success. # Importing Dataset

```
library(readr)

student_performance_data <- read_csv("C:/Users/pc/OneDrive/Desktop/Programming/Rypdc_r_session_2024/ypdc_r_session/st
```

```
## Rows: 500 Columns: 9
## ── Column specification ────────────────────────────────────────────────────
## Delimiter: ","
## chr (4): Gender, Major, PartTimeJob, ExtraCurricularActivities
## dbl (5): StudentID, Age, StudyHoursPerWeek, AttendanceRate, GPA
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
View(student_performance_data)
```

Renaming and reading Dataset

```
df<-student_performance_data
head(df)
```

```
## # A tibble: 6 × 9
##   StudentID Gender   Age StudyHoursPerWeek AttendanceRate   GPA Major
##       <dbl> <chr>  <dbl>             <dbl>          <dbl> <dbl> <chr>
## 1         1 Male      24                37           90.8  3.47 Arts
## 2         2 Female    22                37           74.9  2.32 Education
## 3         3 Male      22                10           53.4  2.38 Business
## 4         4 Male      24                10           70.3  3.46 Science
## 5         5 Male      18                19           74.9  2.31 Education
## 6         6 Female    20                17           86.0  2.47 Business
## # ℹ 2 more variables: PartTimeJob <chr>, ExtraCurricularActivities <chr>
```

```
tail(df)
```

```
## # A tibble: 6 × 9
##   StudentID Gender   Age StudyHoursPerWeek AttendanceRate   GPA Major
##       <dbl> <chr>  <dbl>             <dbl>          <dbl> <dbl> <chr>
## 1       495 Male      24                23           67.1  2.11 Business
## 2       496 Male      22                37           76.6  2.97 Science
## 3       497 Male      23                11           56.3  3.2  Science
## 4       498 Female    20                 6           56.6  3.2  Science
## 5       499 Male      22                18           57.2  2.05 Business
```

Preprocessing Dataset

```
str(df)
```

```
## spc_tbl_ [500 × 9] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ StudentID               : num [1:500] 1 2 3 4 5 6 7 8 9 10 ...
##  $ Gender                  : chr [1:500] "Male" "Female" "Male" "Male" ...
##  $ Age                     : num [1:500] 24 22 22 24 18 20 19 18 19 24 ...
##  $ StudyHoursPerWeek       : num [1:500] 37 37 10 10 19 17 21 14 9 1 ...
##  $ AttendanceRate          : num [1:500] 90.8 74.9 53.4 70.3 74.9 ...
##  $ GPA                     : num [1:500] 3.47 2.32 2.38 3.46 2.31 2.47 3.93 2.51 3.32 3.96 ...
##  $ Major                   : chr [1:500] "Arts" "Education" "Business" "Science" ...
##  $ PartTimeJob             : chr [1:500] "Yes" "No" "No" "Yes" ...
##  $ ExtraCurricularActivities: chr [1:500] "No" "No" "No" "No" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   StudentID = col_double(),
##   ..   Gender = col_character(),
##   ..   Age = col_double(),
##   ..   StudyHoursPerWeek = col_double(),
##   ..   AttendanceRate = col_double(),
##   ..   GPA = col_double(),
##   ..   Major = col_character(),
##   ..   PartTimeJob = col_character(),
##   ..   ExtraCurricularActivities = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
## 6       500 Female    24               21          97.0  2.64 Engineering
## # ℹ 2 more variables: PartTimeJob <chr>, ExtraCurricularActivities <chr>
```

What is the length of dataset?

This dataset contains 500 rows and 9 variables.

What is Data pre-processing?

Data preprocessing is the concept of changing the raw data into a clean data set. The dataset is preprocessed in order to check missing values, noisy data, and other inconsistencies before executing it to the algorithm. Data must be in a format appropriate for analysis.

The col Sums() and sum(duplicated()) and many other functions in R programming are used for this purpose.

The unique() feature helps to observe uniqueness in different variables.

The unique feature helps to observe uniqueness in different variables.

```
unique(df$Gender)
```

```
## [1] "Male"   "Female"
```

```
unique(df$Age)
```

```
## [1] 24 22 18 20 19 23 21
```

```
unique(df$GPA)
```

```
##   [1] 3.47 2.32 2.38 3.46 2.31 2.47 3.93 2.51 3.32 3.96 2.75 2.04 2.49 3.66 3.73
##  [16] 3.54 2.23 3.59 3.84 3.20 3.88 3.65 3.33 3.15 2.70 3.21 2.64 3.99 3.50 2.18
##  [31] 2.58 2.22 3.30 3.53 3.85 2.88 3.80 3.76 2.63 3.74 2.59 2.66 3.26 3.22 3.13
##  [46] 2.76 3.10 3.19 3.09 2.89 2.03 3.17 2.34 3.29 3.52 3.00 3.08 3.90 3.69 3.83
##  [61] 3.63 2.30 3.05 2.95 2.79 3.49 3.43 3.04 3.64 2.86 3.79 2.14 3.91 3.31 2.39
##  [76] 2.96 2.33 3.23 3.34 3.86 3.06 3.16 3.44 2.80 3.95 2.25 3.72 3.60 2.37 2.85
##  [91] 2.17 2.01 3.28 3.03 2.35 2.72 3.81 3.11 2.60 3.70 3.62 3.07 2.36 2.62 2.46
## [106] 2.41 2.12 2.28 2.55 3.35 3.45 2.68 2.67 3.55 2.53 2.73 2.69 2.54 3.36 2.91
## [121] 2.10 3.92 3.27 3.51 2.82 2.00 2.77 2.24 2.19 3.58 3.01 2.13 2.45 2.48 3.02
## [136] 2.56 3.75 3.57 2.78 2.42 2.65 2.05 3.87 2.16 2.97 3.24 2.15 3.38 2.44 2.81
## [151] 2.71 2.50 2.27 2.84 3.67 2.40 2.09 2.74 2.20 3.89 3.56 2.99 2.90 3.78 2.21
## [166] 3.98 2.08 2.07 2.92 3.94 2.02 2.52 3.12 2.11 3.37 2.98 2.83 2.87 3.48 3.68
## [181] 2.94 2.43 2.06 3.41
```

```
unique(df$ExtraCurricularActivities)
```

```
## [1] "No"  "Yes"
```

```
unique(df$PartTimeJob)
```

```
## [1] "Yes" "No"
```

```
unique(df$Major)
```

```
## [1] "Arts"        "Education"   "Business"    "Science"     "Engineering"
```

```
# To check the empty cells/missing values in dataset
colSums(is.na(df))
```

```
##                StudentID                   Gender                      Age
##                        0                        0                        0
##         StudyHoursPerWeek           AttendanceRate                      GPA
##                        0                        0                        0
##                    Major             PartTimeJob ExtraCurricularActivities
##                        0                        0                        0
```

According to the results, there is no missing (Not Available) values in dataset.

Furthermore, there is no duplicated value in the dataset.

```
# Checking duplicate values in the dataset
sum(duplicated(df))
```

```
## [1] 0
```

Descriptive Statistics

In Descriptive statistics, we study Summary Statistics, Measures of Dispersion, Frequency distribution, Shape of distribution, Visualization of data, Correlation, & Normality Tests. ## Summary statistics

```
library(psych)
describe(df)
```

```
##                            vars   n   mean     sd median trimmed    mad   min
## StudentID                     1 500 250.50 144.48 250.50  250.50 185.32  1.00
## Gender*                       2 500   1.49   0.50   1.00    1.49   0.00  1.00
## Age                           3 500  20.96   2.00  21.00   20.94   2.97 18.00
## StudyHoursPerWeek             4 500  19.88  11.47  20.50   19.88  15.57  1.00
## AttendanceRate                5 500  74.99  14.57  75.73   75.03  18.16 50.01
## GPA                           6 500   2.99   0.56   3.00    2.98   0.74  2.00
## Major*                        7 500   2.91   1.37   3.00    2.89   1.48  1.00
## PartTimeJob*                  8 500   1.54   0.50   2.00    1.54   0.00  1.00
## ExtraCurricularActivities*    9 500   1.48   0.50   1.00    1.48   0.00  1.00
##                              max  range  skew kurtosis   se
## StudentID                 500.00 499.00  0.00    -1.21 6.46
## Gender*                     2.00   1.00  0.05    -2.00 0.02
## Age                        24.00   6.00  0.00    -1.24 0.09
## StudyHoursPerWeek          39.00  38.00 -0.03    -1.26 0.51
## AttendanceRate             99.97  49.96 -0.05    -1.15 0.65
## GPA                         3.99   1.99  0.01    -1.20 0.03
## Major*                      5.00   4.00  0.07    -1.25 0.06
## PartTimeJob*                2.00   1.00 -0.14    -1.98 0.02
## ExtraCurricularActivities*  2.00   1.00  0.08    -2.00 0.02
```

In this, the mean, median ,mode, 1st quartile, 3rd quartile, standard deviation, variance, skewness, kurtosis etc of dataset is described.

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ purrr     1.0.2
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ ggplot2::%+%()   masks psych::%+%()
## ✖ ggplot2::alpha() masks psych::alpha()
## ✖ dplyr::filter()  masks stats::filter()
## ✖ dplyr::lag()     masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

What is Descriptive Statistics?

In Descriptive statistics, we study Summary Statistics, Measures of Dispersion, Frequency distribution, Shape of distribution, Visualization of data, Correlation, & Normality Tests.

The describe () and summary () function is used to know the Summary Statistics.

Following table represents the overall summary of each column in the dataset:

```
summary(df)
```

```
##    StudentID        Gender               Age        StudyHoursPerWeek
##  Min.   :  1.0   Length:500         Min.   :18.00   Min.   : 1.00
##  1st Qu.:125.8   Class :character   1st Qu.:19.00   1st Qu.:10.00
##  Median :250.5   Mode  :character   Median :21.00   Median :20.50
##  Mean   :250.5                      Mean   :20.96   Mean   :19.88
##  3rd Qu.:375.2                      3rd Qu.:23.00   3rd Qu.:30.00
##  Max.   :500.0                      Max.   :24.00   Max.   :39.00
##  AttendanceRate       GPA           Major           PartTimeJob
##  Min.   :50.01   Min.   :2.000   Length:500         Length:500
##  1st Qu.:62.61   1st Qu.:2.487   Class :character   Class :character
##  Median :75.73   Median :3.000   Mode  :character   Mode  :character
##  Mean   :74.99   Mean   :2.985
##  3rd Qu.:87.22   3rd Qu.:3.480
##  Max.   :99.97   Max.   :3.990
##  ExtraCurricularActivities
##  Length:500
##  Class :character
##  Mode  :character
##
##
##
```

This is the complte Summary of each column in dataset 'df'

The table represents the minimum value, 1st quartile, mean, median, 3rd quartile, maximum value of each numeric column involving student ID, study hours per week, age, attendance rate, GPA.

What is Data Visualization?

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, maps, and other visual tools, data visualization makes it easier to understand patterns, trends, and insights in data. The main goal is to communicate complex data in a clear and visually appealing way, enabling quicker comprehension and better decision-making.

Common types of data visualizations include bar charts, histograms, line graphs, pie charts, scatter plots, boxplots, and infographics.

Data Visualization

Checking Distribution in data

Histogram & Boxplot

```
options(repr.plot.width = 16.0, repr.plot.height = 16.0)
```

```
library(ggplot2)
a <- ggplot(df) + geom_histogram(aes(x=Age), fill = 'orange', col = 'black', bins = 30, position = 'dodge') +
            labs(title = 'Age Histogram') +
            theme_classic()

b <- ggplot(df) + geom_boxplot(aes(x= Age), fill = 'purple') +
            labs(title = 'Age boxplot') +
            theme_classic()

c <- ggplot(df) + geom_histogram(aes(x=StudyHoursPerWeek), fill = 'orange', col = 'black', bins = 30, position = 'doc
            labs(title = 'StudyHoursPerWeek Histogram') +
            theme_classic()
d <- ggplot(df) + geom_boxplot(aes(x= StudyHoursPerWeek), fill = 'purple') +
            labs(title = 'StudyHoursPerWeek boxplot') +
            theme_classic()

e <- ggplot(df) + geom_histogram(aes(x=AttendanceRate), fill = 'orange', col = 'black', bins = 30, position = 'dodge'
            labs(title = 'AttendanceRate Histogram') +
            theme_classic()
```

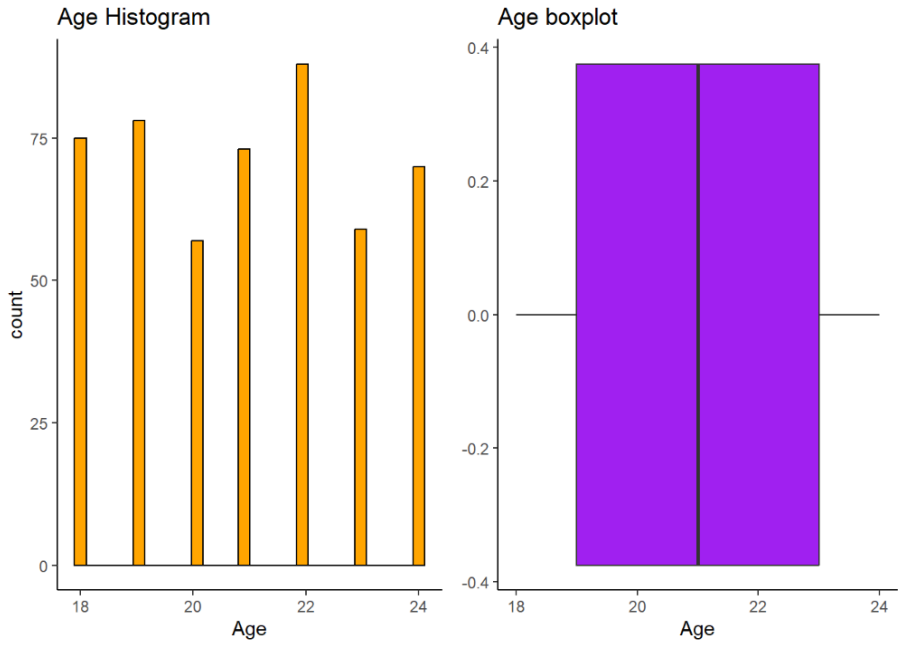
```
f <- ggplot(df) + geom_boxplot(aes(x= AttendanceRate), fill = 'purple') +
            labs(title = 'AttendanceRate boxplot') +
            theme_classic()

g <- ggplot(df) + geom_histogram(aes(x=GPA), fill = 'orange', col = 'black', bins = 30, position = 'dodge') +labs(tit
            theme_classic()

h <- ggplot(df) + geom_boxplot(aes(x= GPA), fill = 'purple') +
            labs(title = 'GPA boxplot') +
            theme_classic()
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
grid.arrange(a,b, ncol = 2)
```



These are the individual **Histograms** and **Boxplots** of numeric variables of data frame, which includes:

- Age
- Study hours per week
- Attendance Rate
- GPA

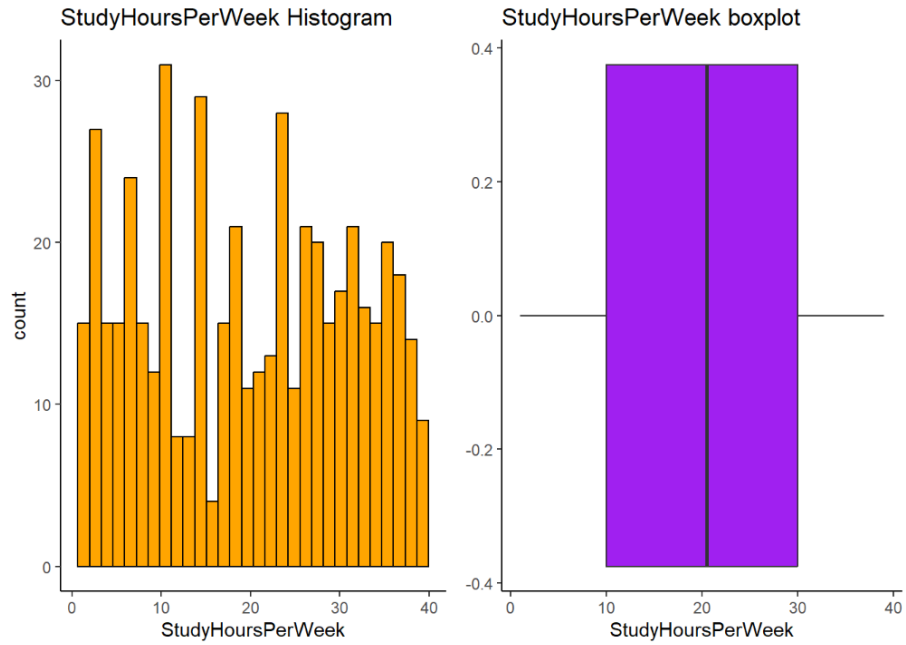
Histograms & its uses

A **histogram** is a type of bar chart that represents the distribution of a dataset. It shows the frequency of data points within specific ranges (called "bins" or "intervals"). Each bar in a histogram represents the count of data points that fall within that range, with the height of the bar indicating the frequency.

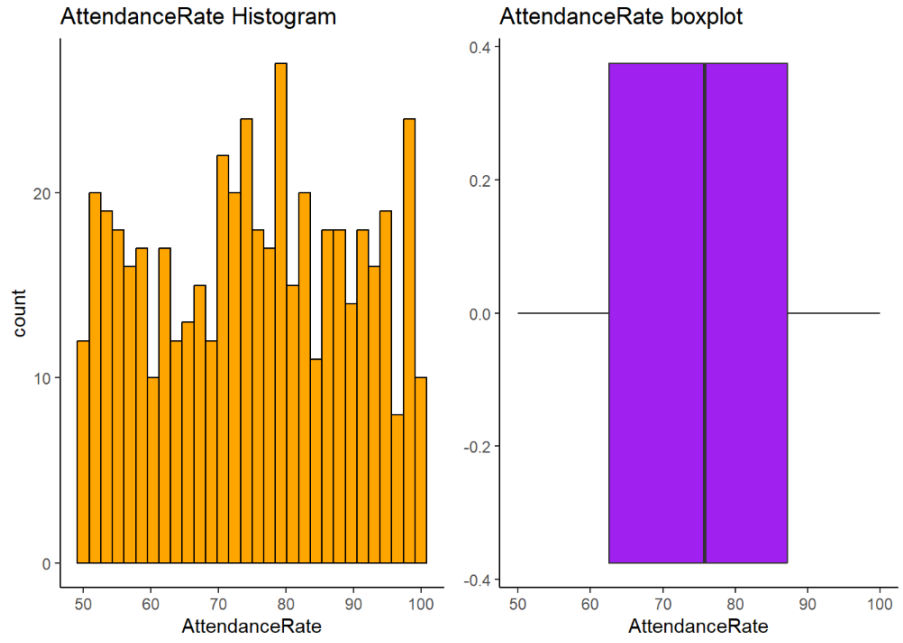
Uses of a Histogram:

1. **Understanding Distribution:** It helps to visualize the shape, spread, and central tendency of a dataset, such as whether the data is normally distributed, skewed, or has multiple peaks.
2. **Identifying Patterns:** Helps identify patterns, such as clusters of data points, gaps, or outliers.
3. **Comparing Data:** Used to compare the distribution of two or more datasets.
4. **Data Summarization:** Provides a quick summary of large datasets, making it easier to interpret and analyze.
5. **Decision-Making:** Helps in making informed decisions by revealing underlying trends and patterns in the data.

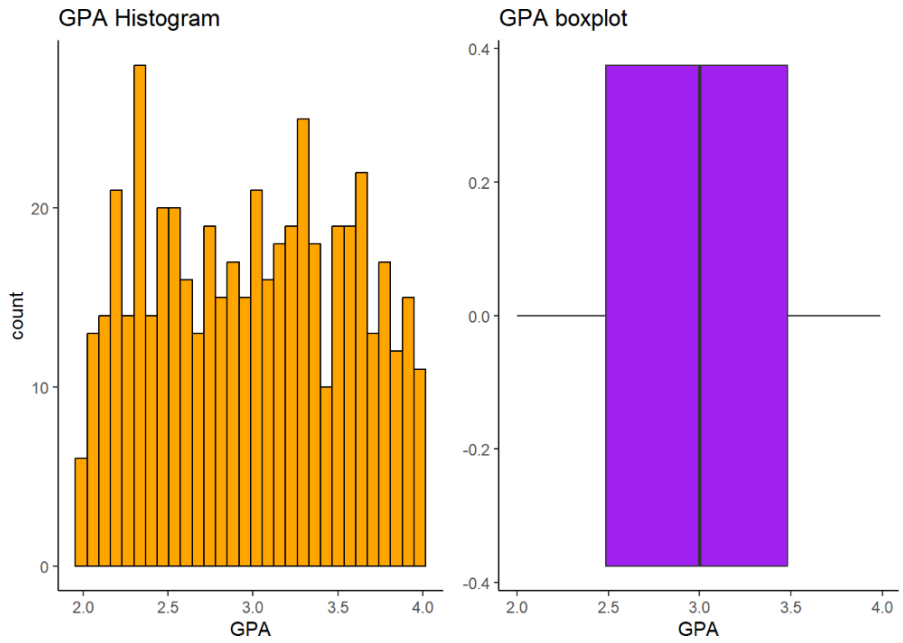
```
grid.arrange(c,d, ncol = 2)
```




```
grid.arrange(e,f, ncol = 2)
```



```
grid.arrange(g,h, ncol = 2)
```



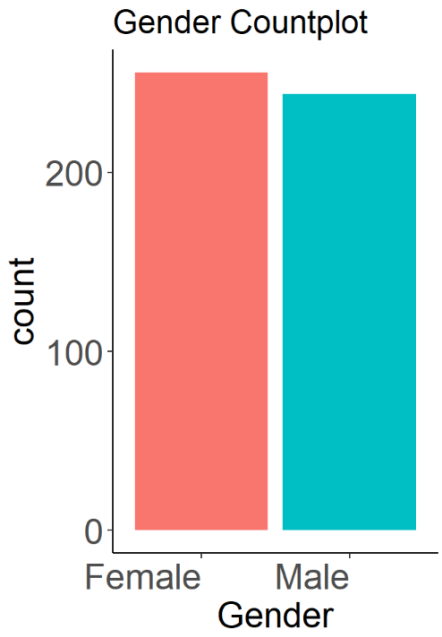
Count Plot

```
options(repr.plot.width = 20.0, repr.plot.height = 16.0)
```

```
library(ggplot2)
a <- ggplot(df) + geom_bar(aes(x= Gender , fill = Gender), position = 'dodge') +
    theme_classic() +
    ggtitle('Gender Countplot') +
    theme(
    legend.position = 'none',
    axis.title = element_text(size=20),
    axis.text.x = element_text(size = 20, hjust=1),
    axis.text.y = element_text(size = 20),
    title = element_text(size=15)
    )

b <- ggplot(df) + geom_bar(aes(x= Major , fill = Major), position = 'dodge') +
    theme_classic() +ggtitle('Major Countplot') +
    theme(
    legend.position = 'none',
    axis.title = element_text(size=20),
    axis.text.x = element_text(size = 20, hjust=1),
    axis.text.y = element_text(size = 20),
    title = element_text(size=15)
    )
```

```
c <- ggplot(df) + geom_bar(aes(x= PartTimeJob , fill = PartTimeJob), position = 'dodge') +
    theme_classic() +
    ggtitle('PartTimeJob Countplot') +
    theme(legend.position = 'none',
    axis.title = element_text(size=20),
    axis.text.x = element_text(size = 20, hjust=1),
    axis.text.y = element_text(size = 20),
    title = element_text(size=15)
    )

d <- ggplot(df) + geom_bar(aes(x= ExtraCurricularActivities , fill = ExtraCurricularActivities), position = 'dodge')
    theme_classic() +
    ggtitle('ExtraCurricularActivities Countplot') +
    theme(
    legend.position = 'none',
    axis.title = element_text(size=20),
    axis.text.x = element_text(size = 20, hjust=1),
    axis.text.y = element_text(size = 20),
    title = element_text(size=15)
    )
```

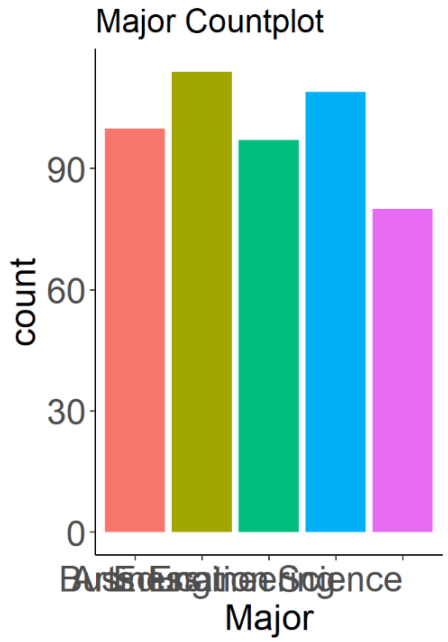
```
library(gridExtra)
grid.arrange(a, ncol = 2)
```



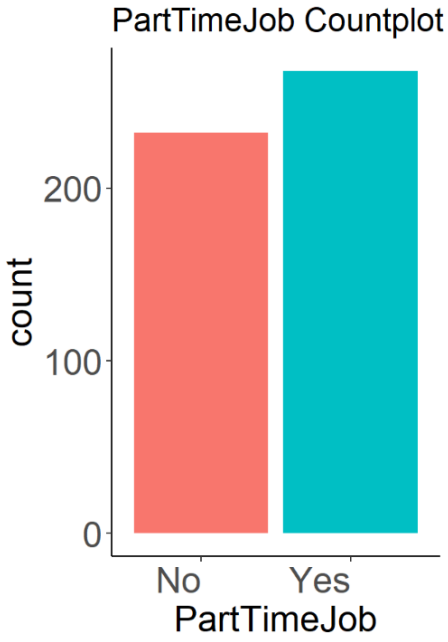
The **Count plots** are used to describe the counts of the character variable in the data frame, these include

- Gender
- Major
- Extra-curricular Activities
- Part time job

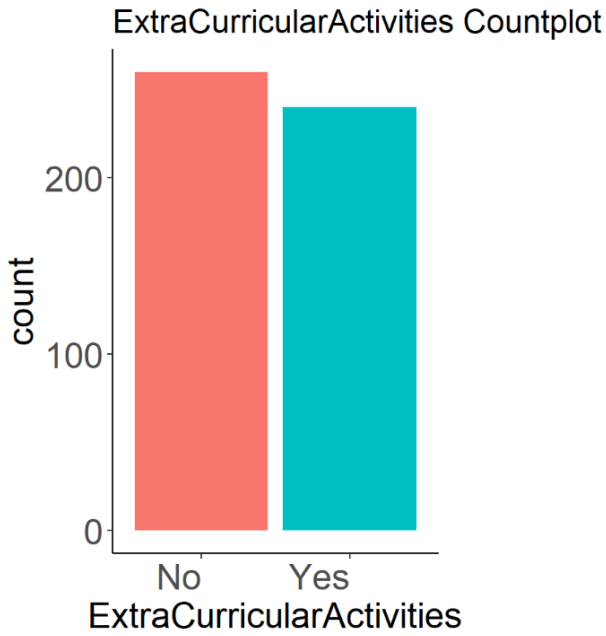
```
grid.arrange(b, ncol = 2)
```




```
grid.arrange(c, ncol = 2)
```



```
grid.arrange(d, ncol = 2)
```



The Count plot of **Gender** shows that there are more female students than the male students in the given data.

The Count plot of **Major** indicates that the students of business are highest in number besides other subjects such as arts, science, education, then comes the engineering students in 2nd place.

The Count plot of **Extra-curricular activities** shows that more students have not participated in such activities that are studying in different college & universities.

The Count plot of **Part time job** suggests that there are more students who do part time jobs besides their studies.

Through **Plotting,** primarily, I have observed the relationship of GPA with Part time job, Extra-curricular activities, and different Majors.

Secondarily, I have also analyzed the effect of Part time job, Extra-curricular activities, GPA, and different Majors on Attendance Rate of students.

PLOTTING

Plots are formed to see the relation between different variables.

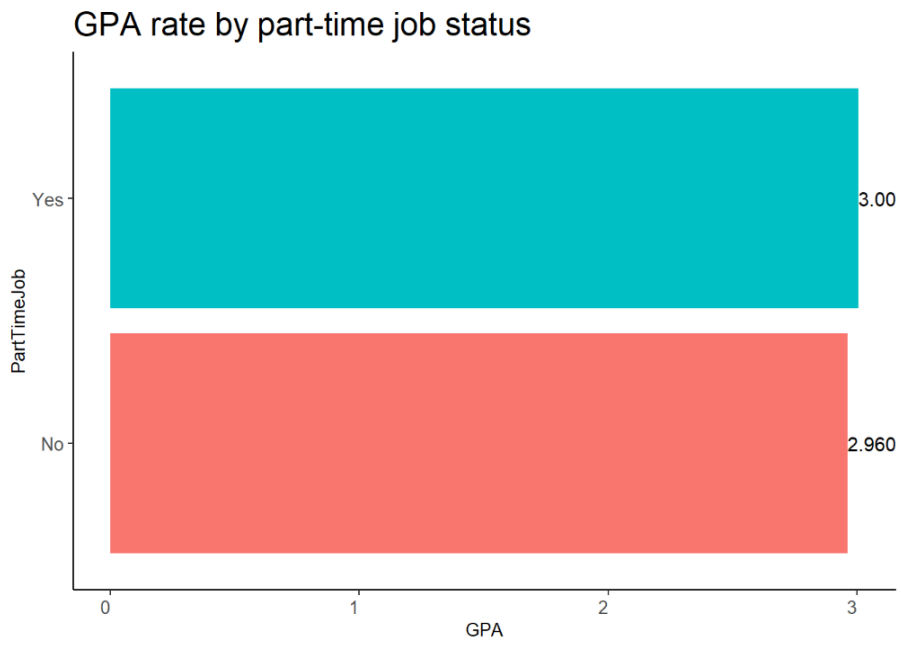
```
# Average GPA Rate by part-time job status
options(repr.plot.width = 16.0, repr.plot.height = 16.0)
```

```
library(tidyverse)
pt_ar <- df %>%
            group_by(PartTimeJob) %>%
            summarize(GPA = mean(GPA, na.rm=T))
print(pt_ar)
```

```
## # A tibble: 2 × 2
##   PartTimeJob   GPA
##   <chr>       <dbl>
## 1 No           2.96
## 2 Yes          3.01
```

It shows that the GPA of students is affected by part time job.

```
library(ggplot2)
ggplot(pt_ar) + geom_col(aes(x= GPA ,y = PartTimeJob , fill = PartTimeJob )) +
    geom_text(aes(x= GPA ,y = PartTimeJob , label = GPA, hjust=0)) +
    theme_classic() +
    ggtitle('GPA rate by part-time job status') +
    theme(
    legend.position = 'none',
    axis.title = element_text(size=10),
    axis.text.x = element_text(size = 10, hjust=1),
    axis.text.y = element_text(size = 10),
    title = element_text(size=15)
    )
```



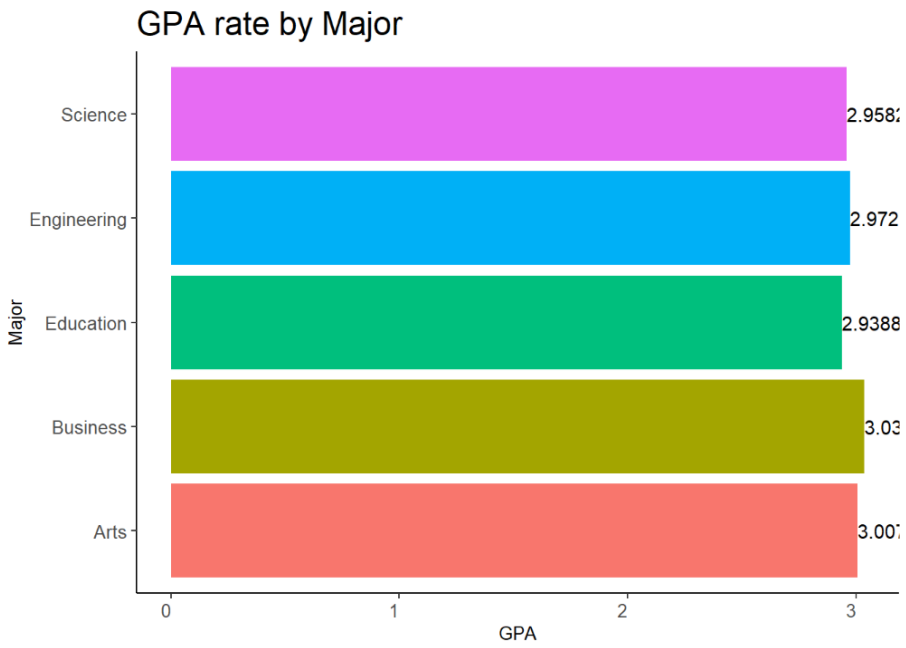
It shows that students who do part time jobs have high GPA as compare to those who don't do.

```
# Average GPA Rate by Major
library(tidyverse)
major_ar <- df %>%
            group_by(Major) %>%
     summarize(GPA = mean(GPA, na.rm=T)) %>%
  arrange(desc(GPA))
major_ar
```

```
## # A tibble: 5 × 2
##   Major         GPA
##   <chr>       <dbl>
## 1 Business     3.04
## 2 Arts         3.01
## 3 Engineering  2.97
## 4 Science      2.96
## 5 Education    2.94
```

It shows that students belonging to different Majors have different GPA .

```
library(ggplot2)
ggplot(major_ar) + geom_col(aes(x= GPA ,y = Major , fill = Major )) +
    geom_text(aes(x= GPA ,y = Major , label = GPA, hjust=0)) +
    theme_classic() +
    ggtitle('GPA rate by Major') +
    theme(
    legend.position = 'none',
    axis.title = element_text(size=10),
    axis.text.x = element_text(size = 10, hjust=1),
    axis.text.y = element_text(size = 10),
    title = element_text(size=15)
    )
```




```
# Average GPA by Extra curricular activities status
options(repr.plot.width = 16.0, repr.plot.height = 16.0)
```

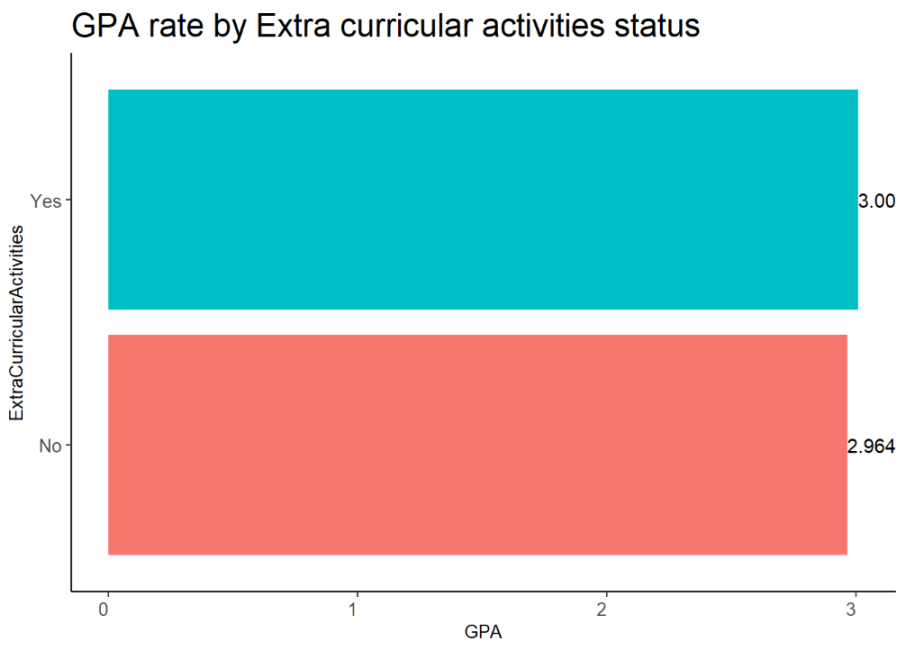
This shows that students of Business have secured higher GPA than students of other majors.

```
library(tidyverse)
ec_ar <- df %>%
            group_by(ExtraCurricularActivities) %>%
            summarize(GPA = mean(GPA, na.rm=T))
print(ec_ar)
```

```
## # A tibble: 2 × 2
##   ExtraCurricularActivities   GPA
##   <chr>                     <dbl>
## 1 No                         2.96
## 2 Yes                        3.01
```

It shows that GPA is affected by participation in Extracurricular activities.

```
library(ggplot2)
ggplot(ec_ar) + geom_col(aes(x= GPA ,y = ExtraCurricularActivities , fill = ExtraCurricularActivities )) +
    geom_text(aes(x= GPA ,y = ExtraCurricularActivities , label = GPA, hjust=0)) +
    theme_classic() +
    ggtitle('GPA rate by Extra curricular activities status') +
    theme(
    legend.position = 'none',
    axis.title = element_text(size=10),
    axis.text.x = element_text(size = 10, hjust=1),
    axis.text.y = element_text(size = 10),
    title = element_text(size=15)
    )
```




```
# Average AttendenceRate Rate by part-time job status
options(repr.plot.width = 16.0, repr.plot.height = 16.0)
```

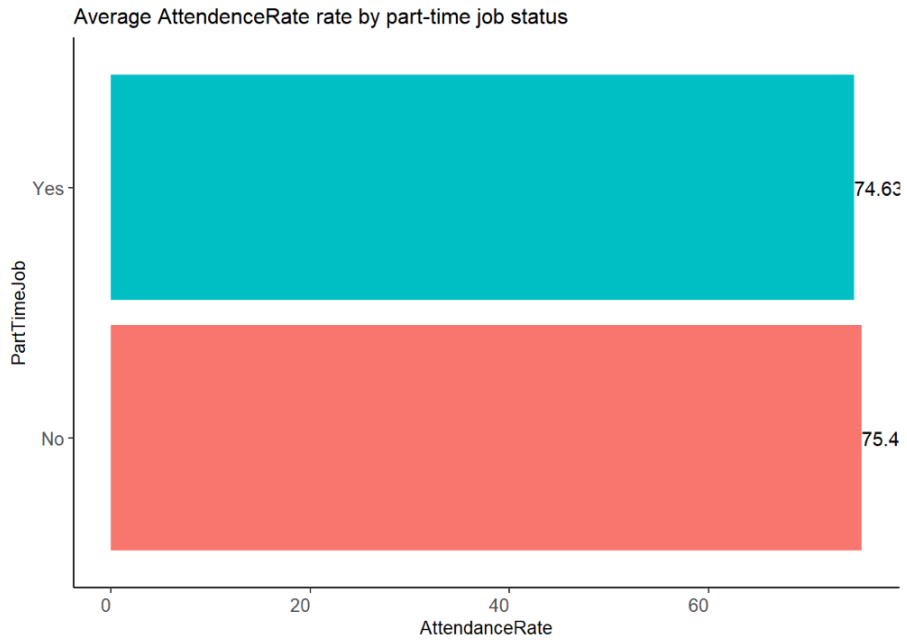
It indicates that students who participated in Extra Curricular activities have high GPA.

```
library(tidyverse)
pt_ar <- df %>%
            group_by(PartTimeJob) %>%
            summarize(AttendanceRate = mean(AttendanceRate, na.rm=T))
print(pt_ar)
```

```
## # A tibble: 2 × 2
##   PartTimeJob AttendanceRate
##   <chr>                <dbl>
## 1 No                    75.4
## 2 Yes                   74.6
```

It shows that Rate of Attendence is affected by Part time job.

```
library(ggplot2)
ggplot(pt_ar) + geom_col(aes(x= AttendanceRate ,y = PartTimeJob , fill = PartTimeJob )) +
    geom_text(aes(x= AttendanceRate ,y = PartTimeJob , label = AttendanceRate, hjust=0)) +
    theme_classic() +
    ggtitle('Average AttendenceRate rate by part-time job status') +
    theme(
    legend.position = 'none',
    axis.title = element_text(size=10),
    axis.text.x = element_text(size = 10, hjust=1),
    axis.text.y = element_text(size = 10),
    title = element_text(size=10)
    )
```

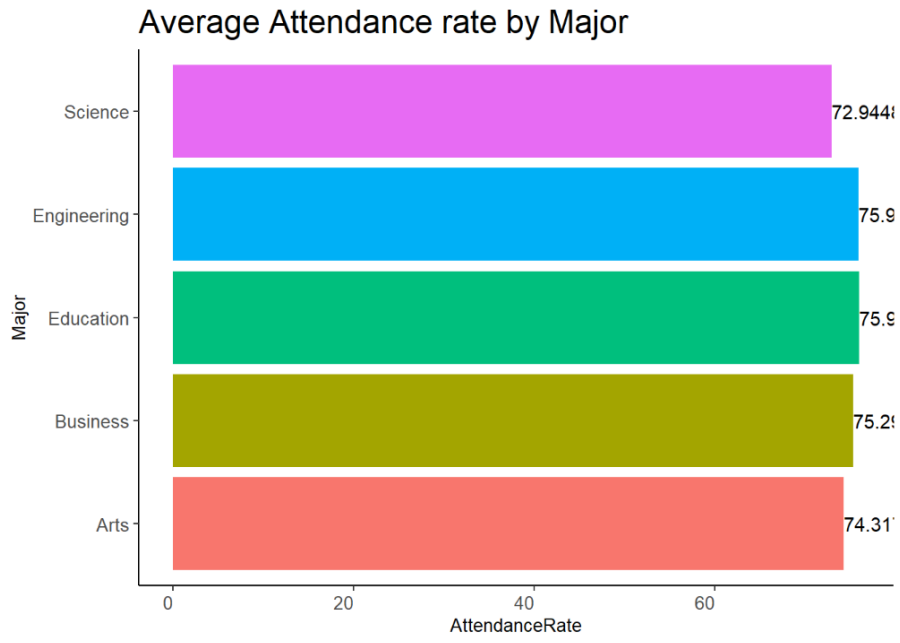


```
# Average [redacted] Rate by Major
library(tidyverse)
major_ar <- df %>%
            group_by(Major) %>%
    summarize(AttendanceRate = mean(AttendanceRate, na.rm=T)) %>%
  arrange(desc(AttendanceRate))
major_ar
```

```
## # A tibble: 5 × 2
##   Major       AttendanceRate
##   <chr>                <dbl>
## 1 Education             76.0
## 2 Engineering           75.9
## 3 Business              75.3
## 4 Arts                  74.3
## 5 Science               72.9
```

Attendance of different majors is different.

```
library(ggplot2)
ggplot(major_ar) + geom_col(aes(x= AttendanceRate ,y = Major , fill = Major )) +
    geom_text(aes(x= AttendanceRate ,y = Major , label = AttendanceRate, hjust=0)) +
    theme_classic() +
    ggtitle('Average Attendance rate by Major') +
    theme(
    legend.position = 'none',
    axis.title = element_text(size=10),
    axis.text.x = element_text(size = 10, hjust=1),
    axis.text.y = element_text(size = 10),
    title = element_text(size=15)
    )
```

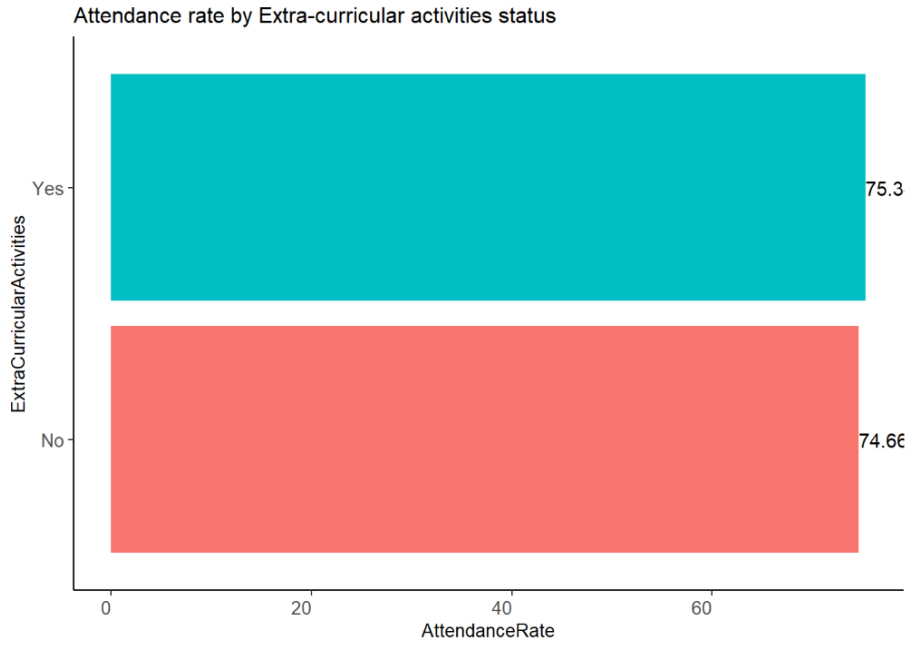


```
# Average Attendence rate by Extra curricular activities status
library(tidyverse)
ec_ar <- df %>%
           group_by(ExtraCurricularActivities) %>%
           summarize(AttendanceRate = mean(AttendanceRate, na.rm=T))
print(ec_ar)
```

```
## # A tibble: 2 × 2
##   ExtraCurricularActivities AttendanceRate
##   <chr>                             <dbl>
## 1 No                                 74.7
## 2 Yes                                75.3
```

Extracurricular activities do have impact on Attendance of students.

```
library(ggplot2)
ggplot(ec_ar) + geom_col(aes(x= AttendanceRate ,y = ExtraCurricularActivities , fill = ExtraCurricularActivities )) +
    geom_text(aes(x= AttendanceRate ,y = ExtraCurricularActivities , label = AttendanceRate, hjust=0)) +
    theme_classic() +
    ggtitle('Attendance rate by Extra-curricular activities status') +
    theme(
    legend.position = 'none',
    axis.title = element_text(size=10),
    axis.text.x = element_text(size = 10, hjust=1),
    axis.text.y = element_text(size = 10),
    title = element_text(size=10)
    )
```



Normality Check

Does Attendance Rate is affected by single factor?

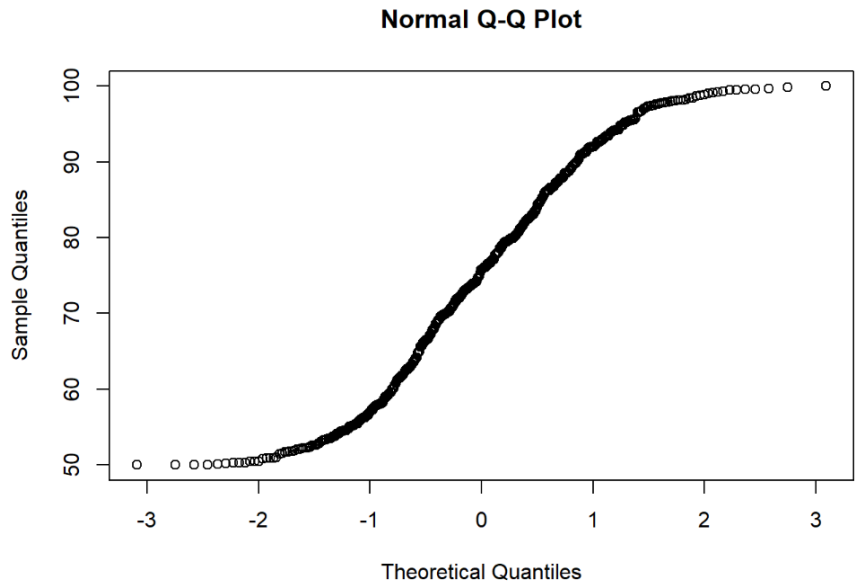
No, There are multiple factors that have impact on average attendance such as different Major subjects, Part time job status, participation in Extra Curricular activities, etc.

Does Extra Curricular Activities affect Attendance rate of students?

Yes, The students who participated in Extra Curricular activities have maintained a good Attendance Rate.

Checking Normality of Data by qq norm plots

```
# QQ-Norm plot of numeric variables
qqnorm(df$AttendanceRate)
```



QQ-NORM PLOT

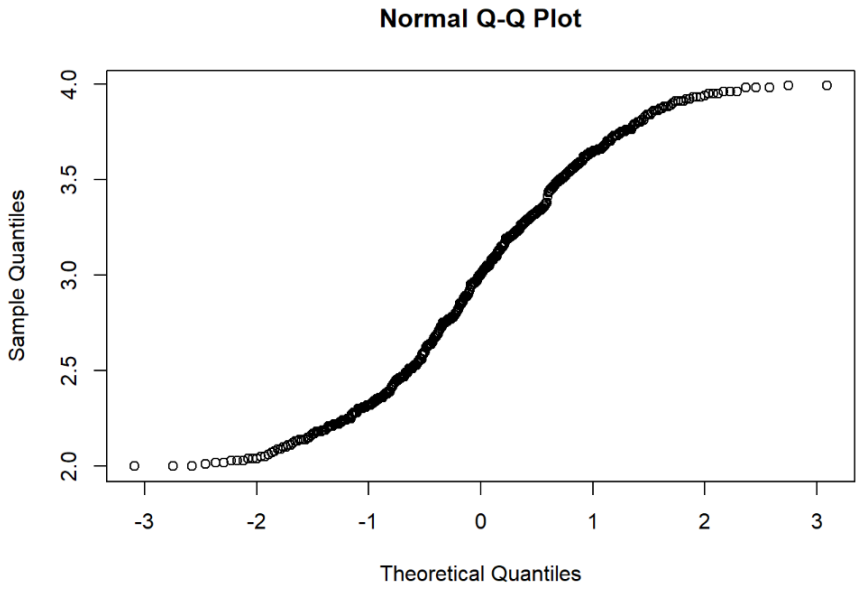
The qq-norm plot is used to check the normality of a variable. These qq-norm plots of 'Attendance Rate', 'GPA', and 'Study Hours per week' Show that data is not normally distributed, there is skewness in the data.

To normalize the data in these columns, I have applied methods like log transform, square root transformation, etc.

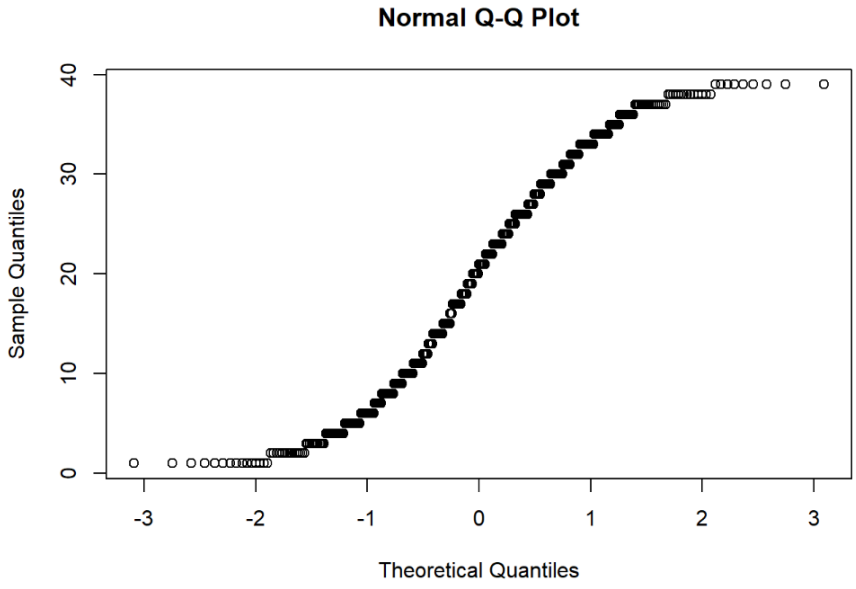
REASONS FOR NON NORMALITY IN DATA

- **Sampling Bias**: If the data collection process is flawed or biased, the entire dataset may not represent the underlying population accurately
- **Measurement Errors**: If there is an error in the instruments or methods used to collect data, the entire dataset could be inaccurate
- **Data Corruption**: If the dataset has been corrupted due to issues like hardware failures, data transmission errors, or intentional tampering, the whole dataset could be abnormal.
- **Outliers**: If the dataset consists mostly of outliers or extreme values, it might not represent normal conditions.
- **Incorrect Assumptions**: If the dataset was collected based on incorrect assumptions or models, the data might appear abnormal when analyzed.

```
qqnorm(df$GPA)
```




```
qqnorm(df$StudyHoursPerWeek)
```



According to the results of qqnorm plots, the data in columns(AttendanceRate, GPA, StudyHourPerWeek) is not normal. To normalize the data, I apply log transform, square root methods to these columns.

Inferential Statistics

Normality Test(Shapiro-Wilk Test)

```
# Shapiro-Wilk test
shapiro.test(df$AttendanceRate)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$AttendanceRate
## W = 0.95601, p-value = 4.681e-11
```

```
shapiro.test(df$GPA)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$GPA
## W = 0.95677, p-value = 6.169e-11
```

```
shapiro.test(df$StudyHoursPerWeek)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$StudyHoursPerWeek
## W = 0.94634, p-value = 1.767e-12
```

data: df$StudyHoursPerWeek
W = 0.94634, p-value = 1.767e-12

Checking Composition

Using “Levene test”, we can check the composition of data.

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
## The following object is masked from 'package:psych':
##
##     logit
```

Which test is used to check the composition of data?

Generally, Leven's Test is used for this purpose.

```
leveneTest(df$GPA~df$PartTimeJob, data = df)
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value  Pr(>F)
## group   1  4.1352 0.04253 *
##       498
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
leveneTest(df$GPA~df$ExtraCurricularActivities, data = df)
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  0.0151 0.9022
##       498
```

From the Levenes test p-value results, The Extracurricular activities data is homogenous.

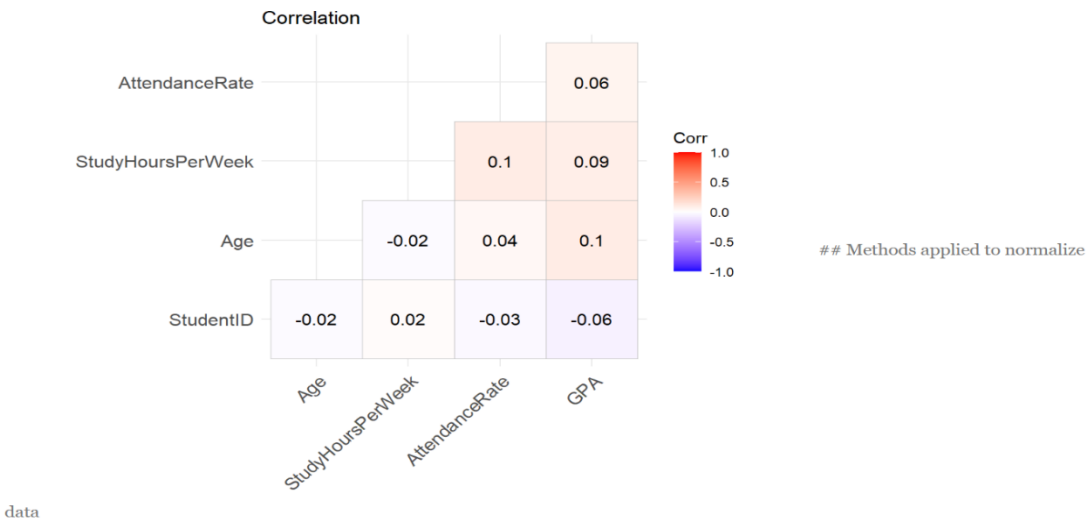
From Leven's Test p-value results, the Extra Curricular activities data column is homogenous.

Checking Correlation

```
num_cols <- {}

for(i in 1:length(df)){
    if(is.numeric(df[[i]]) == TRUE)
        num_cols <- c(num_cols ,colnames(df[i]))
}
```

```
# Correlation matrix Visualization
library(ggcorrplot)
ggcorrplot(cor(df[num_cols]) , type = "lower", lab=T, title = 'Correlation')
```

This Correlation Matrix shows that Attendance Rate and Study Hours per week are related closely. Afterwards, GPA is more likely to correlate with Study Hours per week.



Methods applied to normalize

data

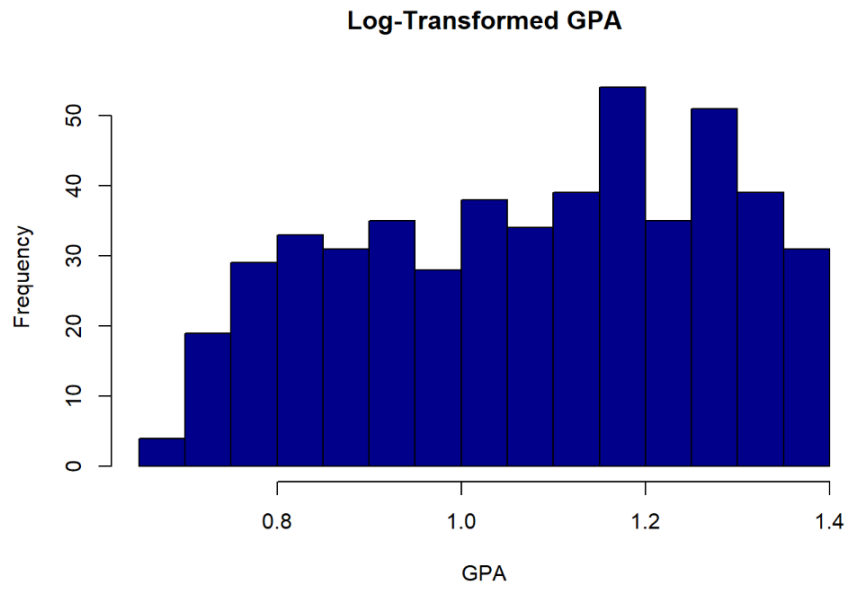
Methods to normalize the variables

- Log transformation
- Square-Root transformation

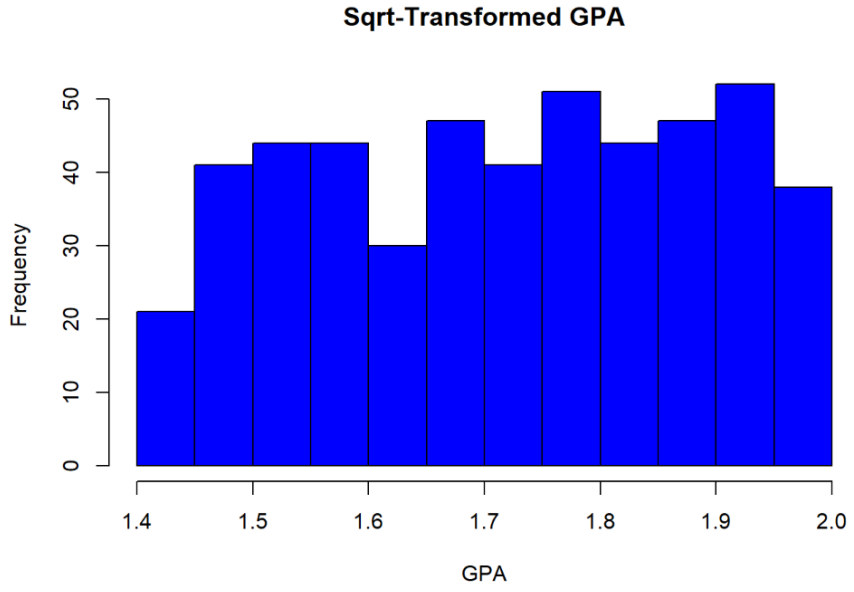
```
# Log transformation
df$GPA_log <- log(df$GPA)
df$StudyHoursPerWeek_log <- log(df$StudyHoursPerWeek + 1)  # Adding 1 to avoid log(0)
df$AttendanceRate_log <- log(df$AttendanceRate + 1)

# Square root transformation
df$GPA_sqrt <- sqrt(df$GPA)
df$StudyHoursPerWeek_sqrt <- sqrt(df$StudyHoursPerWeek)
df$AttendanceRate_sqrt <- sqrt(df$AttendanceRate)
```

```
# Check for normality of transformed data using histograms
hist(df$GPA_log, main="Log-Transformed GPA", xlab="GPA", col="darkblue")
```

The log and square root transformation of column GPA to normalize it. Since, the data is not normally distributed in it.




```
hist(df$GPA_sqrt, main="Sqrt-Transformed GPA", xlab="GPA", col="blue")
```



```
# Shapiro-Wilk test for transformed data
shapiro.test(df$GPA_log)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$GPA_log
## W = 0.95323, p-value = 1.744e-11
```

```
shapiro.test(df$GPA_sqrt)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$GPA_sqrt
## W = 0.95634, p-value = 5.269e-11
```

From above mentioned results, It is seen that data is not normalized to greater extent. So, I may proceed analysis with non-Parametric tests. ## Hypothesis Testing ### Possible Hypotheses for Data I have made multiple hypotheses for this dataset and testing is done accordingly. 1. Gender and GPA: There is a difference in GPA between male and female students. 2. Study Hours and GPA: Students who study more hours per week have a higher GPA. 3. Attendance and GPA: Higher attendance rates correlate with higher GPAs. 4. Part-Time Job and GPA: Students with a part-time job have lower GPAs compared to those without. 5. Extracurricular Activities and GPA: Participation in extracurricular activities influences GPA. 6. Major and GPA: There is a difference in GPA among students of different majors. 7. Age and GPA: Age has an effect on GPA. # Hypothesis Testing ## T-test 1. Gender and GPA:

Possible Hypotheses for Data

I have made multiple hypotheses for this dataset and testing is done accordingly.

1. Gender and GPA: There is a difference in GPA between male and female students.

2. Study Hours and GPA: Students who study more hours per week have a higher GPA.

3. Attendance and GPA: Higher attendance rates correlate with higher GPAs.

4. Part-Time Job and GPA: Students with a part-time job have lower GPAs compared to those without.

5. Extracurricular Activities and GPA: Participation in extracurricular activities influences GPA.

6. Major and GPA: There is a difference in GPA among students of different majors.

Hypothesis Testing

T-test

1. Gender and GPA:

```
df %>% group_by(Gender) %>% summarise(avg=mean(GPA)) %>% arrange(avg)
```

```
## # A tibble: 2 × 2
##   Gender   avg
##   <chr>  <dbl>
## 1 Male    2.97
## 2 Female  3.00
```

Test: Wilcoxon rank-sum test.

```
wilcox.test(GPA ~ Gender, data = df)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  GPA by Gender
## W = 31840, p-value = 0.707
## alternative hypothesis: true location shift is not equal to 0
```

```
t.test(GPA ~ Gender, data = df)
```

```
##
##  Welch Two Sample t-test
##
## data:  GPA by Gender
```

```
t.test(GPA ~ Gender, data = df)
```

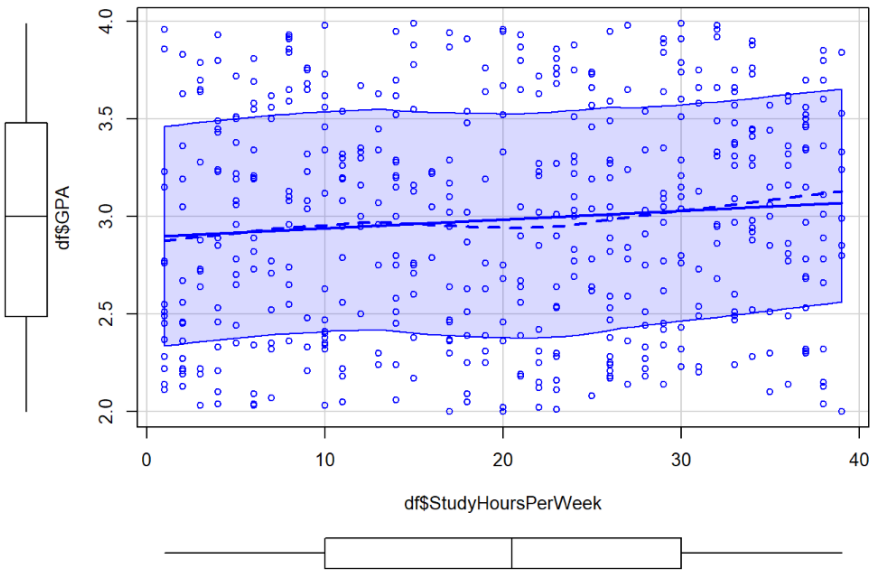
```
##
##  Welch Two Sample t-test
##
## data:  GPA by Gender
## t = 0.4135, df = 494.89, p-value = 0.6794
## alternative hypothesis: true difference in means between group Female and group Male is not equal to 0
## 95 percent confidence interval:
##  -0.07834842  0.12011712
## sample estimates:
## mean in group Female   mean in group Male
##             2.995352             2.974467
```

From the T-test results, null hypothesis is accepted. The Female students have secured higher GPA than Male students.

2.Study Hours and GPA:

```
scatterplot(df$StudyHoursPerWeek, df$GPA)
```

2.StudyHours and GPA:




```
# Pearson correlation
cor.test(df$StudyHoursPerWeek, df$GPA, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$StudyHoursPerWeek and df$GPA
## t = 2.055, df = 498, p-value = 0.0404
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.004042016 0.177959672
## sample estimates:
##       cor
## 0.0917001
```

```
# Spearman's rank correlation
cor.test(df$StudyHoursPerWeek, df$GPA, method = "spearman")
```

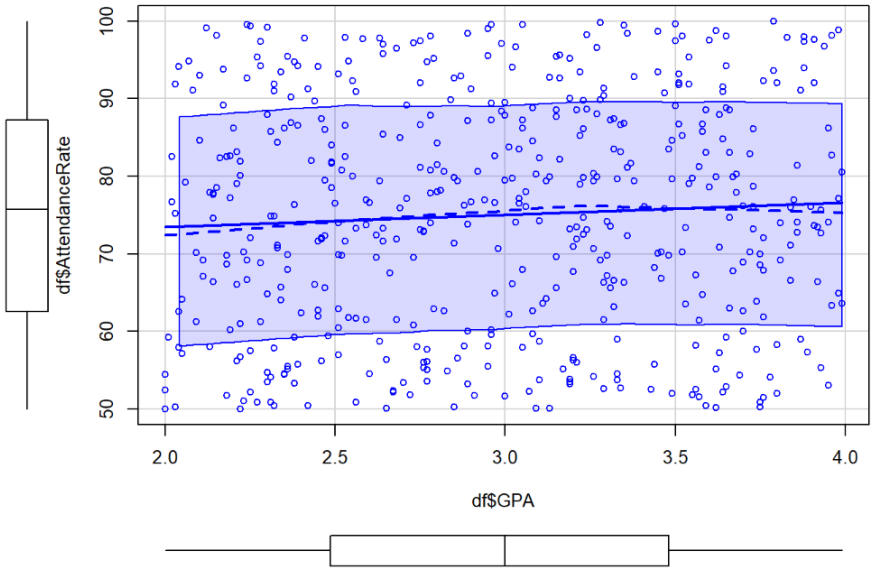
```
## Warning in cor.test.default(df$StudyHoursPerWeek, df$GPA, method = "spearman"):
## Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  df$StudyHoursPerWeek and df$GPA
## S = 18857414, p-value = 0.03399
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.0948405
```

From the spearman correlation test result, null hypothesis is rejected. The GPA of students not only dependent on study hours per week

3. Attendance and GPA

```
scatterplot(df$GPA, df$AttendanceRate)
```




```
# Pearson correlation
cor.test(df$AttendanceRate_sqrt, df$GPA, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$AttendanceRate_sqrt and df$GPA
## t = 1.3915, df = 498, p-value = 0.1647
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.02559681  0.14911032
## sample estimates:
##        cor
## 0.06223347
```

```
# Pearson correlation
cor.test(df$AttendanceRate_log, df$GPA, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$AttendanceRate_log and df$GPA
## t = 1.4198, df = 498, p-value = 0.1563
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.02433356  0.15034602
## sample estimates:
```

```
## 95 percent confidence interval:
##  -0.02433356  0.15034602
## sample estimates:
##        cor
## 0.06349251
```

```
# Spearman's rank correlation
cor.test(df$AttendanceRate, df$GPA, method = "spearman")
```

```
## Warning in cor.test.default(df$AttendanceRate, df$GPA, method = "spearman"):
## Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  df$AttendanceRate and df$GPA
## S = 19610616, p-value = 0.1902
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## 0.05868666
```

From the results of Spearman's rank test, null hypothesis is accepted. The students who attended the classes regularly and maintained required Attendance rate, have secured high GPA.

4. Part Time Job & GPA

```
# Independent t-test
t.test(GPA_log ~ PartTimeJob, data = df)
```

```
##
##  Welch Two Sample t-test
##
## data:  GPA_log by PartTimeJob
## t = -0.69174, df = 496.92, p-value = 0.4894
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -0.04579906  0.02194726
## sample estimates:
##  mean in group No mean in group Yes
##          1.068904          1.080830
```

```
t.test(df$GPA_sqrt ~ PartTimeJob, data = df)
```

```
##
##  Welch Two Sample t-test
##
## data:  df$GPA_sqrt by PartTimeJob
## t = -0.80258, df = 496.89, p-value = 0.4226
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -0.04055438  0.01703131
```

```
## sample estimates:
##  mean in group No mean in group Yes
##          1.713624          1.725385
```

```
# Wilcoxon rank-sum test (if data is not normally distributed)
wilcox.test(GPA ~ PartTimeJob, data = df)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  GPA by PartTimeJob
## W = 29698, p-value = 0.3884
## alternative hypothesis: true location shift is not equal to 0
```

Null Hypothesis is accepted The students who do part time job have low GPA.

5. Extra-curricular activities & GPA

```
# Independent t-test
t.test(GPA_log ~ ExtraCurricularActivities, data = df)
```

```
##
##  Welch Two Sample t-test
##
## data:  GPA_log by ExtraCurricularActivities
## t = -0.84072, df = 494.13, p-value = 0.4009
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -0.04862701  0.01948303
## sample estimates:
##  mean in group No mean in group Yes
##          1.068302          1.082874
```

```
t.test(GPA_sqrt ~ ExtraCurricularActivities, data = df)
```

```
##
##  Welch Two Sample t-test
##
## data:  GPA_sqrt by ExtraCurricularActivities
## t = -0.8577, df = 493.87, p-value = 0.3915
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -0.04159253  0.01631405
## sample estimates:
##  mean in group No mean in group Yes
##          1.713861          1.726500
```

```
# Wilcoxon rank-sum test (if data is not normally distributed)
wilcox.test(GPA ~ ExtraCurricularActivities, data = df)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  GPA by ExtraCurricularActivities
## W = 29786, p-value = 0.381
## alternative hypothesis: true location shift is not equal to 0
```

Null hypothesis is accepted The students who participated in ExtraCurricular Activities have positive effect on their GPA.

6. Major & GPA:

```
df %>% group_by(Major) %>% summarise(avg=mean(GPA)) %>% arrange(avg)
```

```
## # A tibble: 5 × 2
##   Major         avg
##   <chr>       <dbl>
## 1 Education    2.94
## 2 Science      2.96
## 3 Engineering  2.97
## 4 Arts         3.01
## 5 Business     3.04
```

```
# ANOVA
aov_result <- aov(GPA_log ~ Major, data = df)
summary(aov_result)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Major         4  0.075 0.01869   0.497  0.738
## Residuals   495 18.611 0.03760
```

```
aov_result <- aov(GPA_sqrt ~ Major, data = df)
summary(aov_result)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Major         4  0.054 0.01338   0.492  0.741
## Residuals   495 13.450 0.02717
```

```
# Kruskal-Wallis test ( data is not normally distributed)
kruskal.test(GPA ~ Major, data = df)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  GPA by Major
## Kruskal-Wallis chi-squared = 1.958, df = 4, p-value = 0.7435
```

Hypothesis 1: Null hypothesis is accepted. The GPA of two different genders is different. The female have higher GPA than male students.

Hypothesis 2: The study Hours of student don't affect much on the score of students. There might be other factors that allow students to get good scores.

Hypothesis 3: Null Hypothesis is accepted. The Higher attendance rates correlate with higher GPAs.

Hypothesis 4: Null Hypothesis is accepted. Students with a part-time job have lower GPAs compared to those without.

Hypothesis 5: Null Hypothesis is accepted. Participation in extracurricular activities positively influences GPA.

Hypothesis 6: The results of Kruskal-wallis test shows that there is a difference in GPA among students of different majors.

The students of different Major have different GPAs.

Modeling

```
library(tidymodels)
```

```
## ── Attaching packages ─────────────────────────────────────── tidymodels 1.2.0 ──
```

```
## ✓ broom        1.0.6     ✓ rsample      1.2.1
## ✓ dials        1.3.0     ✓ tune         1.2.1
## ✓ infer        1.0.7     ✓ workflows    1.1.4
## ✓ modeldata    1.4.0     ✓ workflowsets 1.1.0
## ✓ parsnip      1.2.1     ✓ yardstick    1.3.1
## ✓ recipes      1.1.0
```

```
## ── Conflicts ────────────────────────────────────── tidymodels_conflicts() ──
## ✖ ggplot2::%+%()       masks psych::%+%()
## ✖ scales::alpha()      masks ggplot2::alpha(), psych::alpha()
## ✖ gridExtra::combine() masks dplyr::combine()
## ✖ scales::discard()    masks purrr::discard()
## ✖ dplyr::filter()      masks stats::filter()
## ✖ recipes::fixed()     masks stringr::fixed()
## ✖ dplyr::lag()         masks stats::lag()
## ✖ car::recode()        masks dplyr::recode()
## ✖ car::some()          masks purrr::some()
## ✖ yardstick::spec()    masks readr::spec()
```

```
## • Dig deeper into tidy modeling with R at https://www.tmwr.org
```

```
library(finetune)
library(xgboost)
```

```
##
## Attaching package: 'xgboost'
```

```
## The following object is masked from 'package:dplyr':
##
##     slice
```

```
library(mltools)
```

```
##
## Attaching package: 'mltools'
```

```
## The following objects are masked from 'package:yardstick':
##
##     mcc, rmse
```

```
## The following object is masked from 'package:tidyr':
##
##     replace_na
```

```
library(themis)
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following objects are masked from 'package:yardstick':
##
##     precision, recall, sensitivity, specificity
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(bonsai)
library(withr)
library(Metrics)
```

```
##
## Attaching package: 'Metrics'
```

```
## The following objects are masked from 'package:caret':
```

```
library(bonsai)
library(withr)
library(Metrics)
```

```
##
## Attaching package: 'Metrics'
```

```
## The following objects are masked from 'package:caret':
##
##     precision, recall
```

```
## The following objects are masked from 'package:mltools':
##
##     mse, msle, rmse, rmsle
```

```
## The following objects are masked from 'package:yardstick':
##
##     accuracy, mae, mape, mase, precision, recall, rmse, smape
```

```
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

Preprocessing

```
df <- df[,-1]
```

```
df$Gender <- ifelse(df$Gender == 'Male',1,0)
df$Major <- as.factor(df$Major)
df$PartTimeJob <- ifelse(df$PartTimeJob == 'Yes',1,0)
df$ExtraCurricularActivities <- ifelse(df$ExtraCurricularActivities == 'Yes','1','0')
```

```
encoding_data <- one_hot(as.data.table(df[,-12]))
encoding_data$ExtraCurricularActivities <- as.factor(df$ExtraCurricularActivities)
```

Hypothesis structured for model

ExtraCurricularActivities have impact on student performance and Attendence rate. ## Train Test Split

```
idx <- sample(1:nrow(df), 0.7 * nrow(encoding_data))
train <- encoding_data[idx,]
test <- encoding_data[-idx,]
```

Simple XGB Model

```
xgb_rec <-
  recipe(ExtraCurricularActivities ~ ., data = train) %>%
      step_dummy(all_nominal_predictors()) %>%
      step_YeoJohnson(all_numeric_predictors())
```

```
xgb_spec <- boost_tree() %>%
  set_engine('xgboost' ,
             nthread  = future::availableCores()) %>%
  set_mode('classification')
```

```
xgb_wf <- workflow() %>%
  add_recipe(xgb_rec) %>%
  add_model(xgb_spec)

xgb_wf
```

```
## == Workflow ============================================================
## Preprocessor: Recipe
## Model: boost_tree()
##
## -- Preprocessor ----------------------------------------------------------
## 2 Recipe Steps
##
## • step_dummy()
```

```
## == Workflow ============================================================
## Preprocessor: Recipe
## Model: boost_tree()
##
## -- Preprocessor ----------------------------------------------------------
## 2 Recipe Steps
##
## • step_dummy()
## • step_YeoJohnson()
##
## -- Model -----------------------------------------------------------------
## Boosted Tree Model Specification (classification)
##
## Engine-Specific Arguments:
##   nthread = future::availableCores()
##
## Computational engine: xgboost
```

```
model_xgb =
  xgb_wf %>%
  fit(train) %>%
  with_seed(7, .)
```

```
model_xgb
```

```
## == Workflow [trained] ==================================================
```

```
## == Workflow [trained] ==========================================================
## Preprocessor: Recipe
## Model: boost_tree()
##
## -- Preprocessor ----------------------------------------------------------------
## 2 Recipe Steps
##
## • step_dummy()
## • step_YeoJohnson()
##
## -- Model -----------------------------------------------------------------------
## ##### xgb.Booster
## raw: 34.2 Kb
## call:
##   xgboost::xgb.train(params = list(eta = 0.3, max_depth = 6, gamma = 0,
##     colsample_bytree = 1, colsample_bynode = 1, min_child_weight = 1,
##     subsample = 1), data = x$data, nrounds = 15, watchlist = x$watchlist,
##     verbose = 0, nthread = c(system = 8), objective = "binary:logistic")
## params (as set within xgb.train):
##   eta = "0.3", max_depth = "6", gamma = "0", colsample_bytree = "1", colsample_bynode = "1", min_child_weight =
"1", subsample = "1", nthread = "8", objective = "binary:logistic", validate_parameters = "TRUE"
## xgb.attributes:
##   niter
## callbacks:
##   cb.evaluation.log()
## # of features: 16
## niter: 15
## nfeatures : 16
## evaluation_log:
```

```
## evaluation_log:
##      iter training_logloss
##     <num>            <num>
##         1        0.6147456
##         2        0.5650724
## ---
##        14        0.3337149
##        15        0.3212019
```

```
pred_xgb <- predict(model_xgb, test)
```

```
cm_xgb <- confusionMatrix(pred_xgb$.pred_class, test$ExtraCurricularActivities)
cm_xgb
```

Interpretations

1. **Accuracy:** The model correctly predicted the class 46.67% of the time. This is not particularly high, suggesting the model struggles with accurate predictions.
2. **95% CI:** This is the confidence interval for the accuracy, indicating that we are 95% confident that the true accuracy of the model lies between 38.49% and 54.98%. This wide range further suggests uncertainty in the model's performance.
3. **No Information Rate (NIR):** The NIR is the accuracy you would achieve by always predicting the most frequent class. Here, it is higher than the model's accuracy, which means the model is performing worse than a naive prediction strategy that always chooses the most frequent class.
4. **P-Value [Acc > NIR]: 0.9578**

   This is the p-value for testing whether the model's accuracy is better than the NIR. A p-value of 0.9578 is very high, meaning there is no significant difference between the model's accuracy and the NIR. This suggests the model isn't better than random guessing.

5. **Kappa: -0.0619** Kappa is a measure of agreement between the predicted and actual classes, accounting for chance. A negative Kappa indicates that the model's predictions are worse than random guessing.
6. **Sensitivity (Recall): 0.5000**

   The model correctly identified 50% of the true positives. Sensitivity indicates the model's ability to correctly identify instances of the positive class.

7. **Specificity: 0.4375**

   The model correctly identified 43.75% of the true negatives. Specificity measures the ability to correctly identify instances of the negative class.

8. **Positive Predictive Value (Precision): 0.4375**

   Of all the instances that the model predicted as positive, 43.75% were actually positive.

9. **Negative Predictive Value: 0.5000**

   Of all the instances that the model predicted as negative, 50% were actually negative.

10. **Balanced Accuracy: 0.4688**

    Balanced accuracy is the average of sensitivity and specificity. It accounts for imbalanced classes, but here it's still low, confirming the poor performance of the model.

Overall Assessment:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 35 45
##          1 35 35
##
##                Accuracy : 0.4667
##                  95% CI : (0.3849, 0.5498)
##     No Information Rate : 0.5333
##     P-Value [Acc > NIR] : 0.9570
##
##                   Kappa : -0.0619
##
##  Mcnemar's Test P-Value : 0.3143
##
##             Sensitivity : 0.5000
##             Specificity : 0.4375
##          Pos Pred Value : 0.4375
##          Neg Pred Value : 0.5000
##              Prevalence : 0.4667
##          Detection Rate : 0.2333
##    Detection Prevalence : 0.5333
##       Balanced Accuracy : 0.4688
##
##        'Positive' Class : 0
##
```

Does Students Performance is affected by single factor?

No, It is dependent on multiple factors, and different factors plays it's own role in improving the overall performance of students.

Key Findings

I have performed complete Statistical analysis on this dataset of ‘Student’s Performance’ which involves Descriptive Statistics & Inferential Statistics.

The Hypothesis I have chosen are also tested.

In addition to this, I have designed a simple XGBoost model that suggests the impact of Extra Curricular Activities on performance of different students.

Conclusion

In this document, I have thoroughly explored and analyzed the Student’s performance dataset, providing valuable insights into Effect of Extracurricular Activities, part time Job, Study hours per week of students, & their Attendance Rate on their GPA. Key findings are Extra Curricular Activities have strong impact on student GPA and Attendance Rate. The students who do part time jobs beside studies are unable to maintain their GPA and Attendance rate. The analysis serves a foundation for exploring and analyzing factors that impact student performance. By examining this data, I identify trends, uncover relationships between different variables, and develop strategies to enhance educational outcomes. Furthermore, I have designed the simple XGBOOST model that predicted the Extra Curricular activities impact.

Recommendations

In my opinion, In order to improve the Student’s performance irrespective of their gender and Major’s, Educational Institutes should encourage the participation of students in Extra Curricular activities and provide fund benefits so that students don’t have to do Part time jobs. Hence, increasing student performance and productivity in the educational aspect as guided by this statistical report.

Limitations

The distribution of some variables is not normal, it might affect the results of my analysis.

Future Work

While this analysis offers a comprehensive overview of the Student Performance data, several avenues for future work and exploration remain. There might be other factors that also affect the academic activity of students. Furthermore, if more precise and accurate data is collected further from more students and education institutes, then it will increase the chance of accurate analysis. Additionally, the creation of societies within the education institute that monitor the performance of students can enhance the future work. Overall, continuous data collection and analysis will be crucial for adapting to changing education system and ensuring a seamless performance experience for students.

GitHub link: https://github.com/Herr-Mahin/Applied-Statistical-Modelling.git