CAB220 Portfolio 2

KA LONG LEE (N9845097)

28/09/2019

CAB220 Portfolio2

Overview This portfolio accounts for 20% of your overall grade of CAB220. Full mark of this portfolio is 20. The tasks in this portfolio are designed to assess your knowledge and skills in

- Descriptive statistical data analysis and visualisation • Statistical hypothesis testing • Linear regression
- Logistic regression

Data:

The fictitious data set for this portfolio includes the records of 2,550 first-year students of an Australian university in terms of case ID, Attrition, Degree Type, Achieved Credit Points, Attendance Type, Age, Failed Credit Points, International student, First in family in university, Gender, GPA, OP Score, Socio Economic Status, Teaching Period Admitted,and Faculty.

Working Environment Configuration:

```
# Import Library
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Setting up the working directory
# So that It can import external file
# Warning!!! -- Disable the next line, if you need to export the pdf report
#               Otherwise, you will need the next line to generates diagram later on
# setwd(dirname(rstudioapi::getActiveDocumentContext()$path))

# Import external files
# Most of the visualization function is stored in this file
# Please check it if you are interested in the code
source("data_visualization.R")

# Import Data
uniData <- read.csv("datasets/Portfolio_2_data.csv", header = TRUE) %>%
  select(2:15)
```

Task 1 Summarise the information in each variable (except case ID) using a table or an appropriate statistical graph

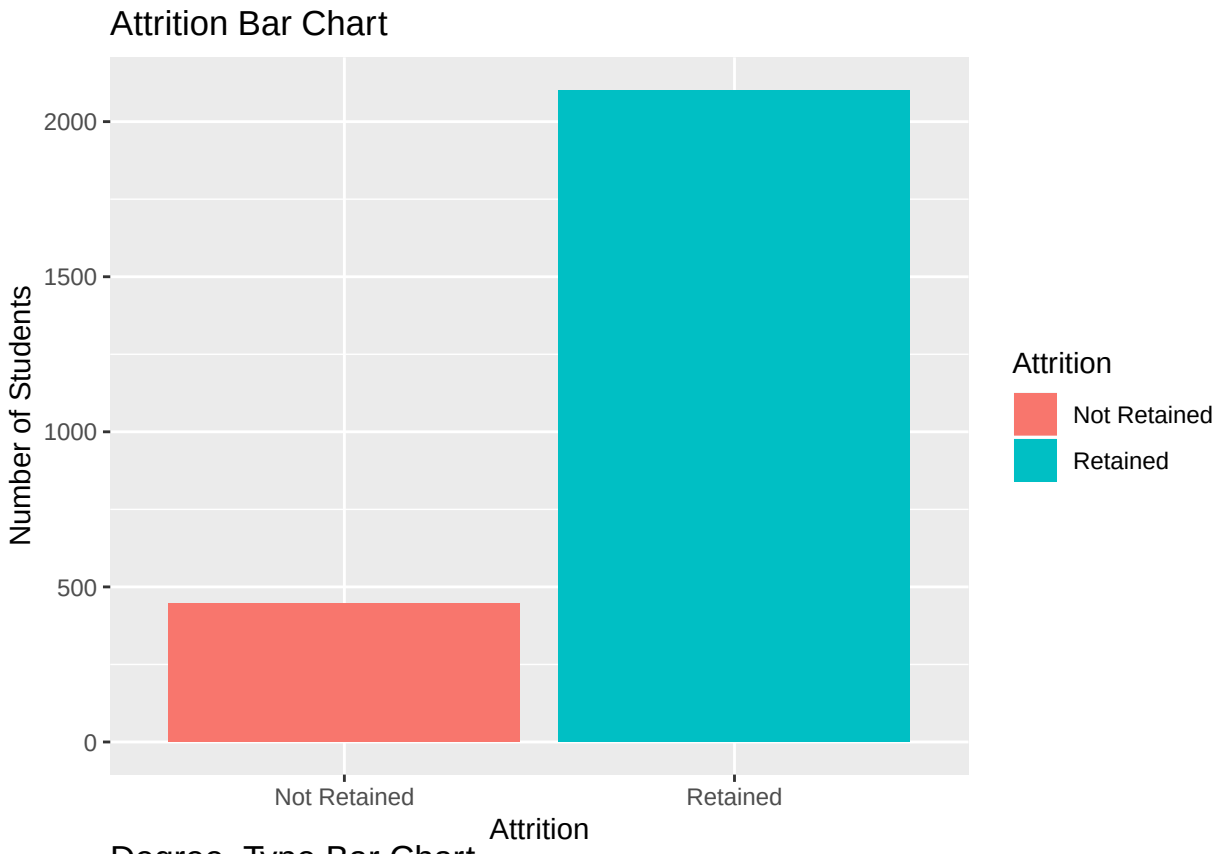
Summary each variables using a table

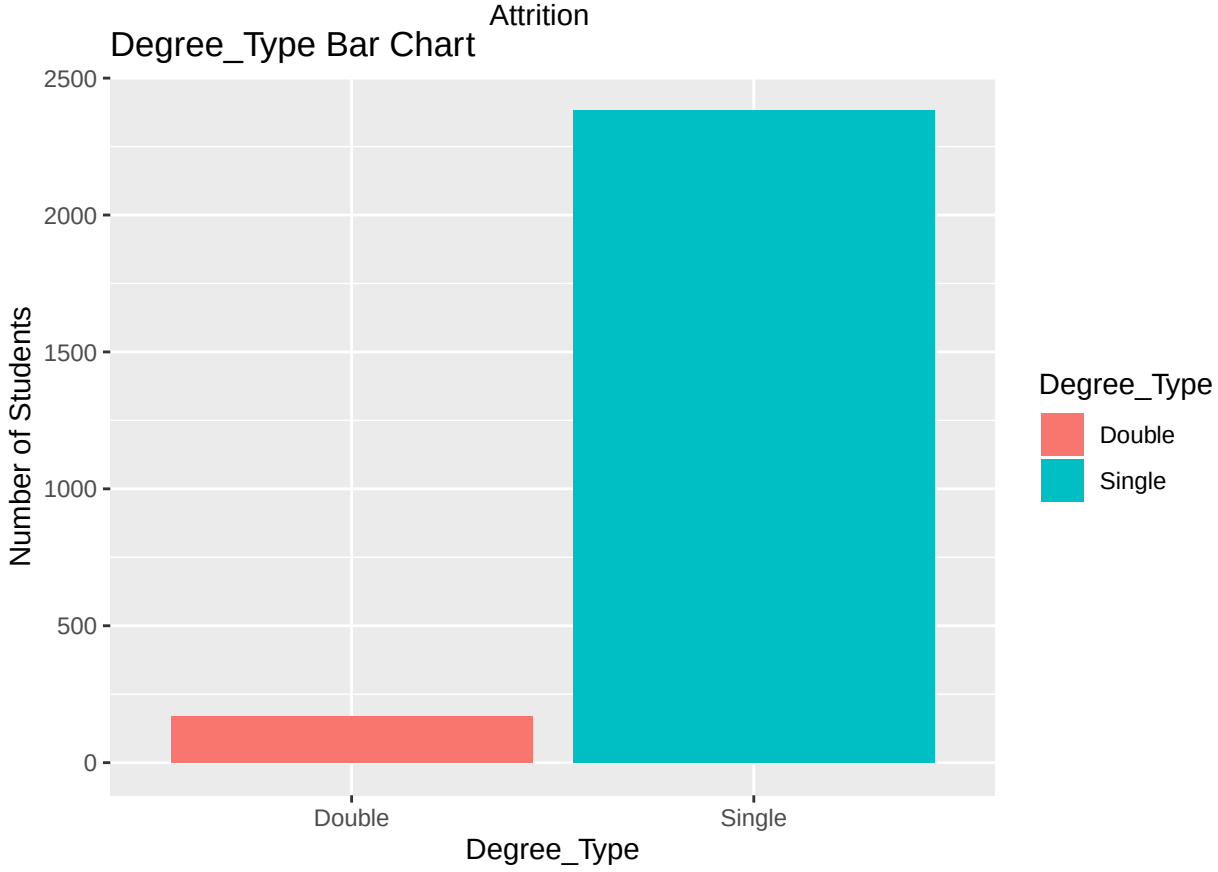
```r
summary(uniData)
```

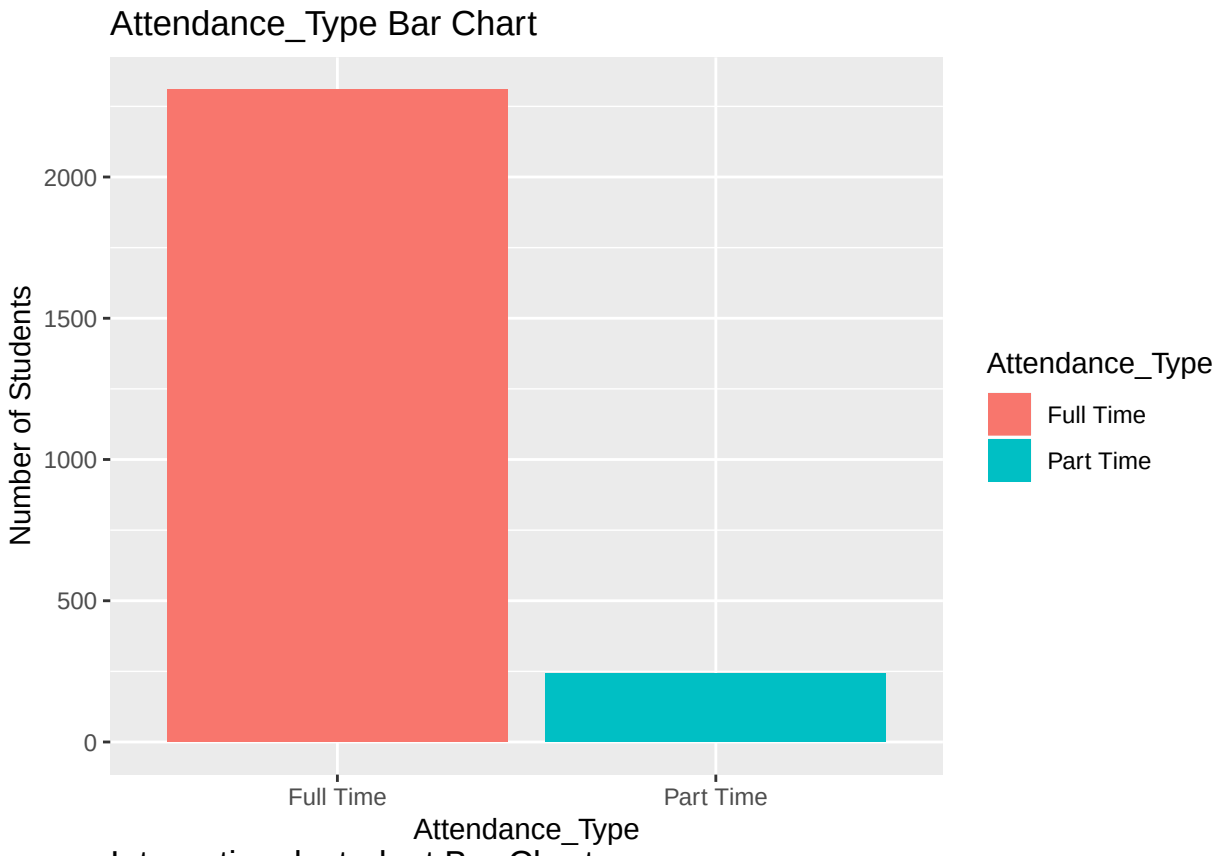
```
##         Attrition    Degree_Type   Achieved_Credit_Points  Attendance_Type
##  Not Retained: 448   Double: 169   Min.   :  0.00         Full Time:2308
##  Retained    :2102   Single:2381   1st Qu.: 60.00         Part Time: 242
##                                    Median : 96.00
##                                    Mean   : 92.97
##                                    3rd Qu.:108.00
##                                    Max.   :372.00
##       Age        Failed_Credit_Points International_student
##  Min.   :18.00   Min.   :  0.000      No :2223
##  1st Qu.:19.00   1st Qu.:  0.000      Yes: 327
##  Median :20.00   Median :  0.000
##  Mean   :22.74   Mean   :  8.033
##  3rd Qu.:23.00   3rd Qu.: 12.000
##  Max.   :86.00   Max.   :108.000
##  First_in_family Gender        GPA           OP_Score
##  No :1580        F:1254   Min.   :0.000   Min.   : 1.00
##  Yes: 970        M:1296   1st Qu.:4.130   1st Qu.: 6.00
##                           Median :4.880   Median : 9.00
##                           Mean   :4.549   Mean   :10.74
##                           3rd Qu.:5.630   3rd Qu.:15.00
##                           Max.   :7.000   Max.   :25.00
##  Socio_Economic_Status Teaching._Period_Admitted
##  High  : 771           SEM-1:2107
##  Low   : 463           SEM-2: 443
##  Medium:1316
##
##
##
##                  Faculty
##  CI Faculty          :430
##  Faculty of Education:158
##  Faculty of Health   :677
##  Faculty of Law      :244
##  QUT Business School :385
##  Sci and Eng Faculty :656
```

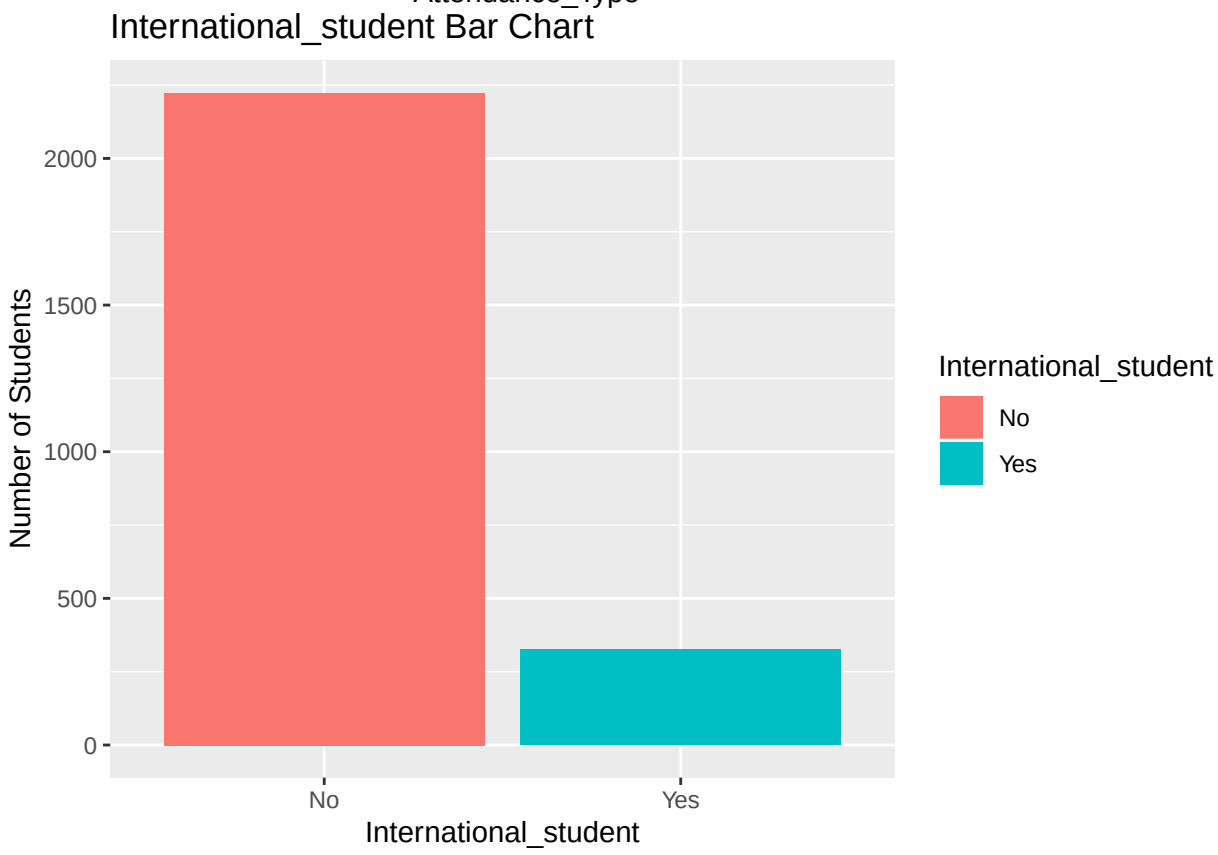
Summary each categorical data in uni dataframe using appropriate graphs

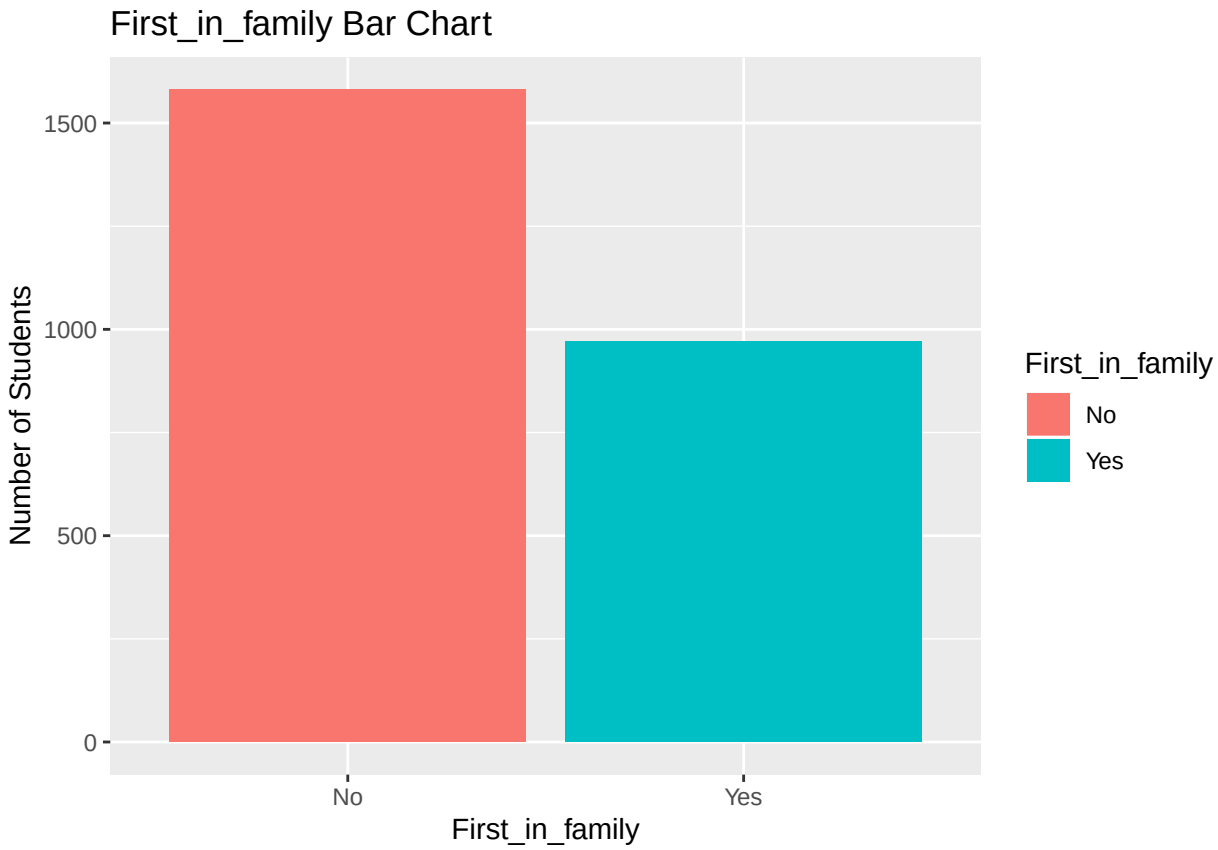
```r
visualize_categorical_data(uniData)
```

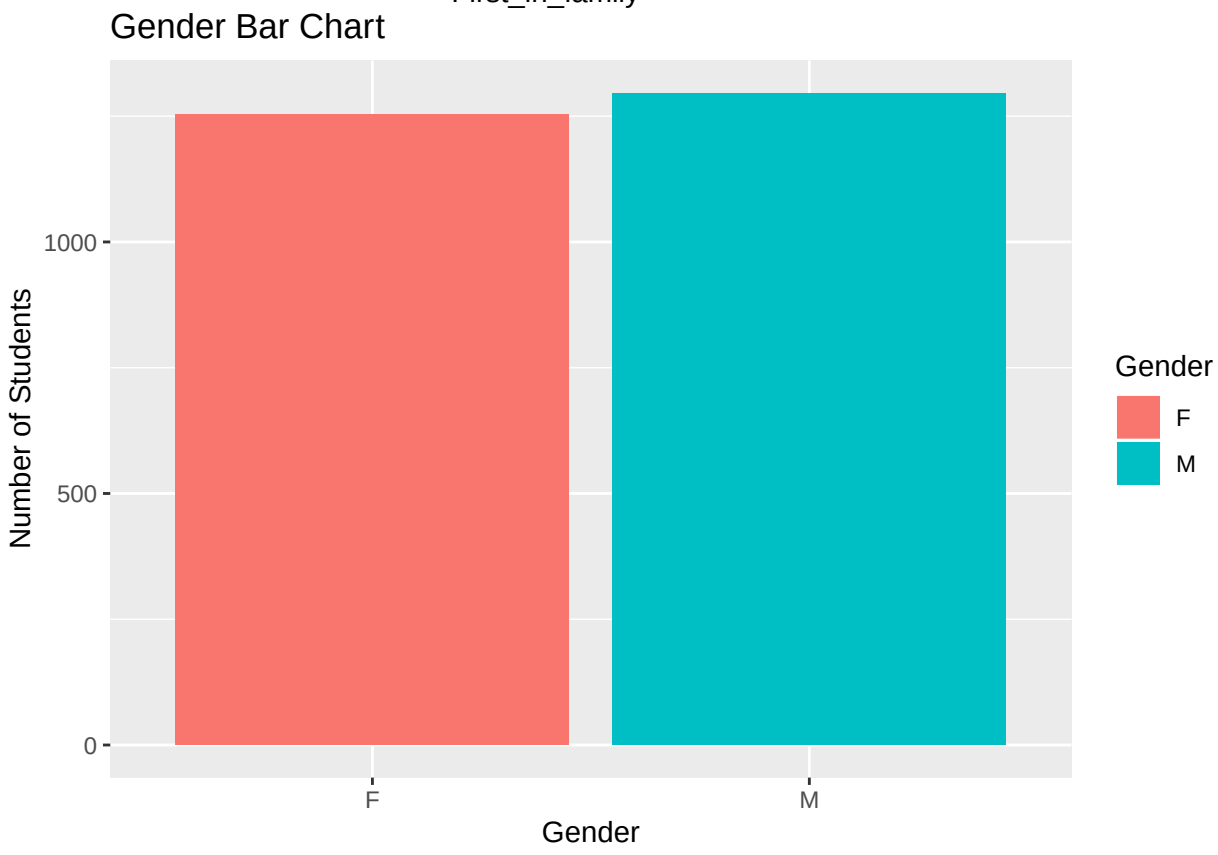

Attrition Bar Chart
2000
1500
1000
500
0
Number of Students
Not Retained
Retained
Attrition
Attrition
Not Retained
Retained

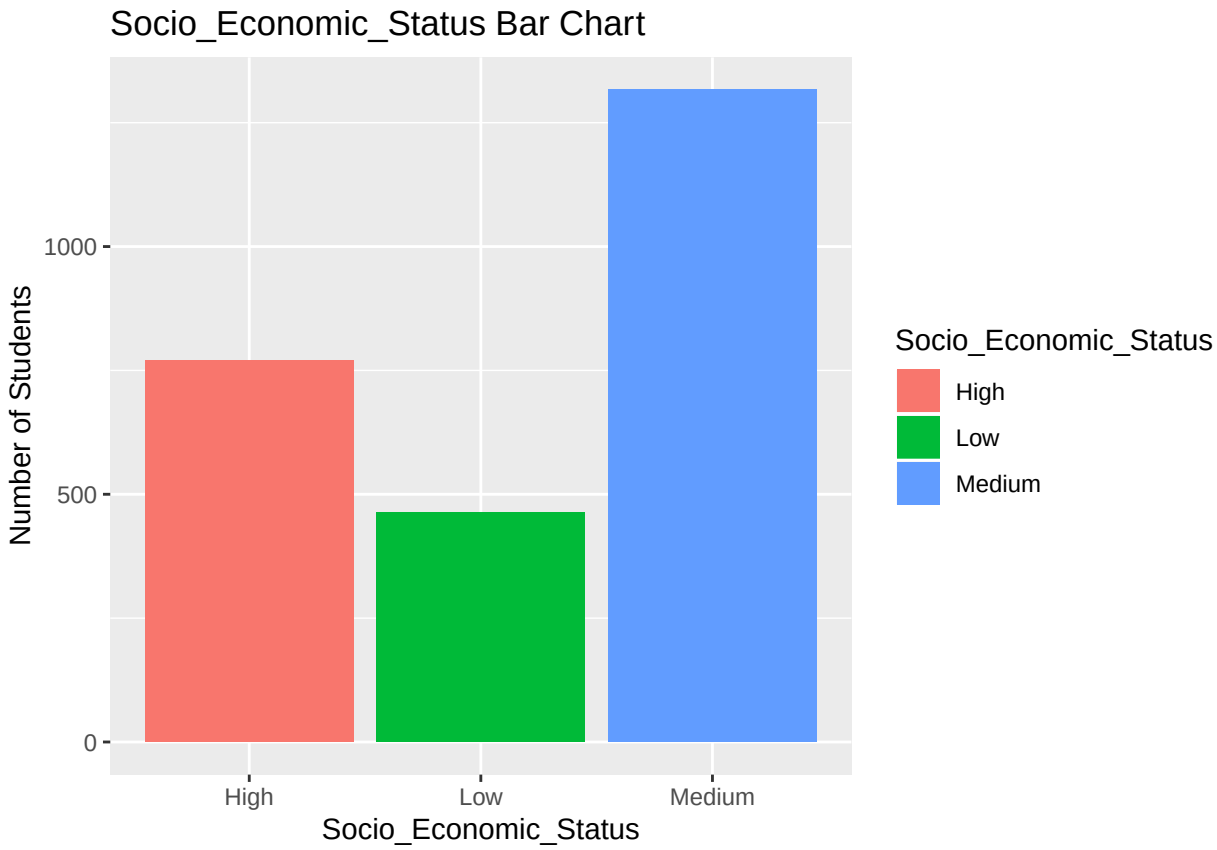

Degree_Type Bar Chart
2500
2000
1500
1000
500
0
Number of Students
Double
Single
Degree_Type
Degree_Type
Double
Single

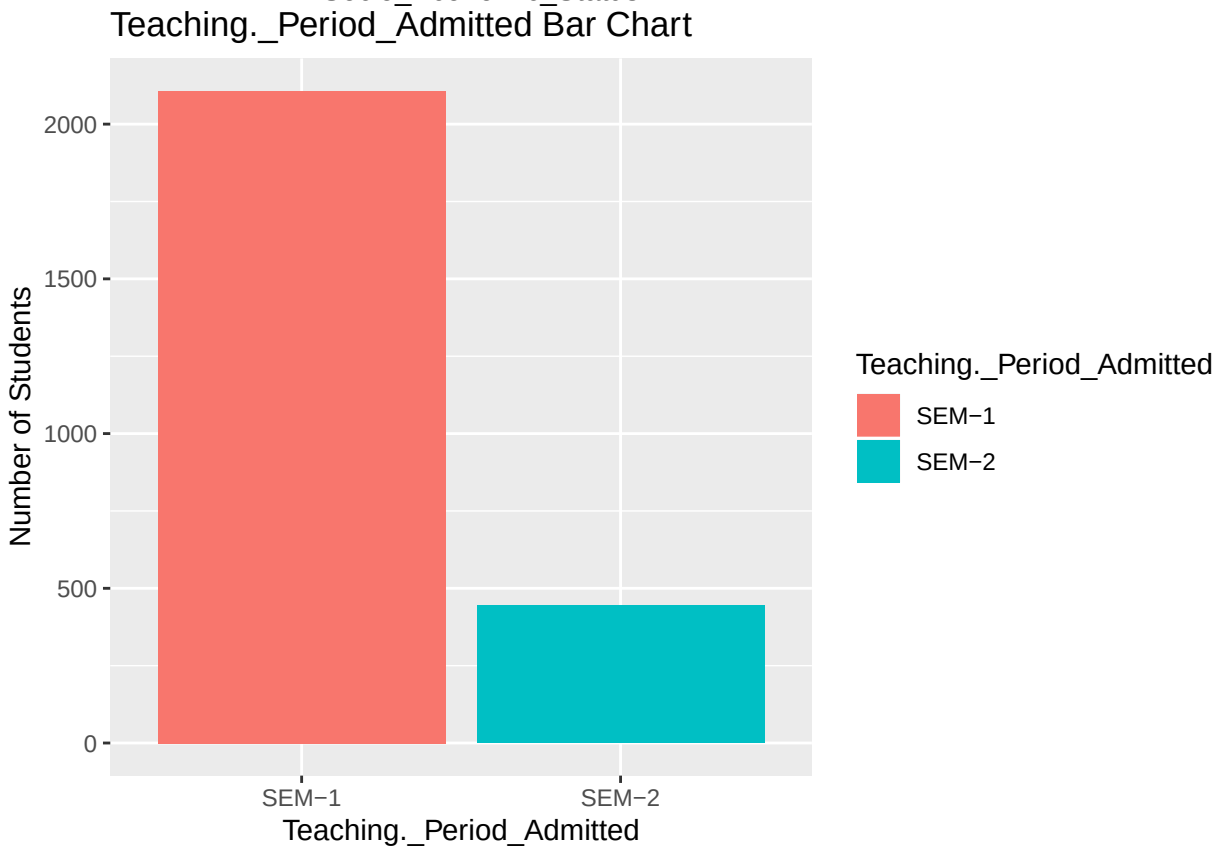

Attendance_Type Bar Chart
2000
1500
1000
500
0
Number of Students
Full Time
Part Time
Attendance_Type
Attendance_Type
Full Time
Part Time

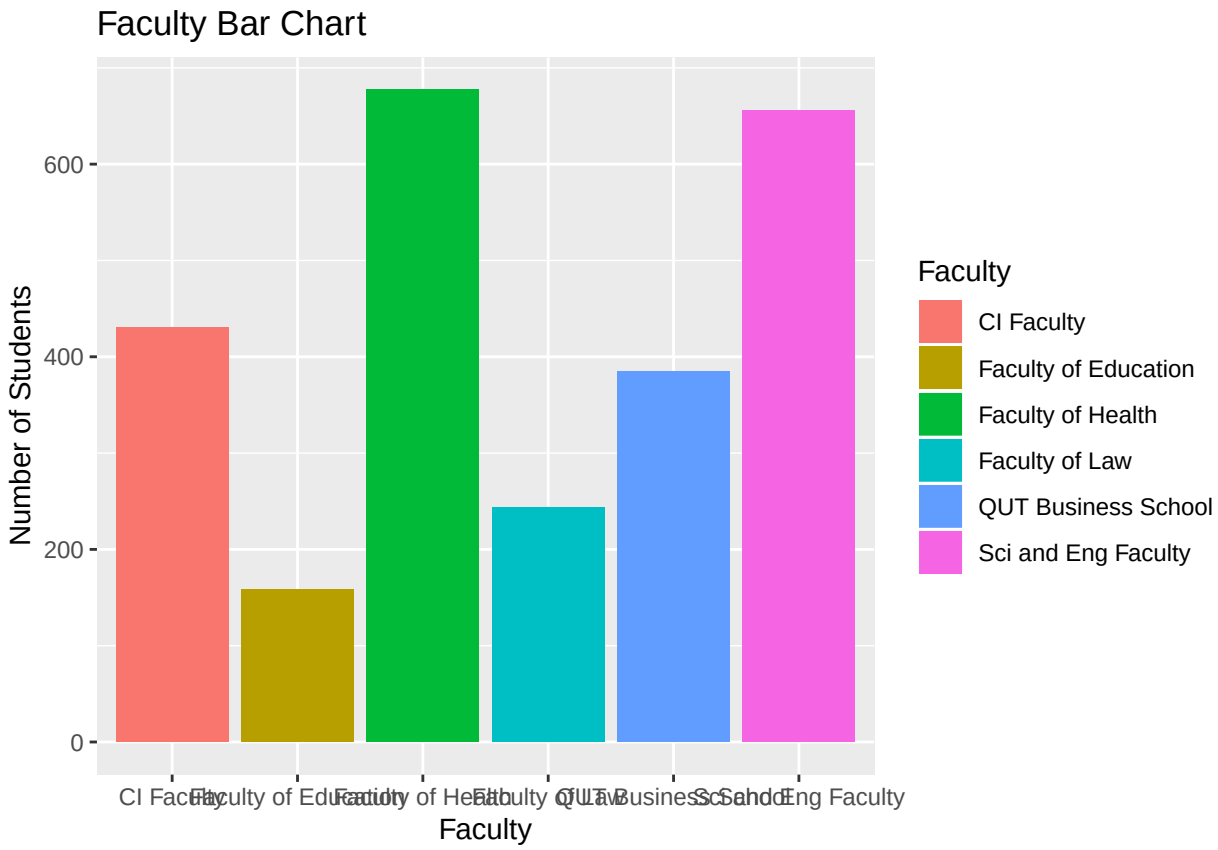

International_student Bar Chart
2000
1500
1000
500
0
Number of Students
No
Yes
International_student
International_student
No
Yes


First_in_family Bar Chart
1500
1000
500
0
Number of Students
No
Yes
First_in_family
First_in_family
No
Yes


Gender Bar Chart
1000
500
0
Number of Students
F
M
Gender
Gender
F
M


Socio_Economic_Status Bar Chart
Number of Students
1000
500
0
High
Low
Medium
Socio_Economic_Status
Socio_Economic_Status
High
Low
Medium


Teaching._Period_Admitted Bar Chart
Number of Students
2000
1500
1000
500
0
SEM−1
SEM−2
Teaching._Period_Admitted
Teaching._Period_Admitted
SEM−1
SEM−2



The function operated above generates 9 bar charts illustrating the distribution of each categorical variables in the dataframe. The summary of each charts are listed below.

1. The distribution of the attrition of the students

It is evident that Students in retained attrition are approximately four times more than students in not retained attrition.

2. The distribution of degree type among students

Almost 93% students are doing single degree and the rest are doing double degree.

3. The attendance type distribution among students

Not surprisingly, most of the students are studying full-time in university and only ten percent of students are a part-time student.

4. The distribution of first in family in all the students

5. The distribution of gender among students

It is interesting that gender in the university is evenly distributed. It doesn't have a huge statistical outliner.

6. The economic status of each students.

Half proportion of the students are in medium-income family. Approximately 30 percent of students are in high-income family, while around 18% of students was concerning their income.

7. The distribution of the period students admitted to university

The chart shows that approximately 80 percent of the students joined university in semester 1, while only 20 percent students joined university in semester 2.

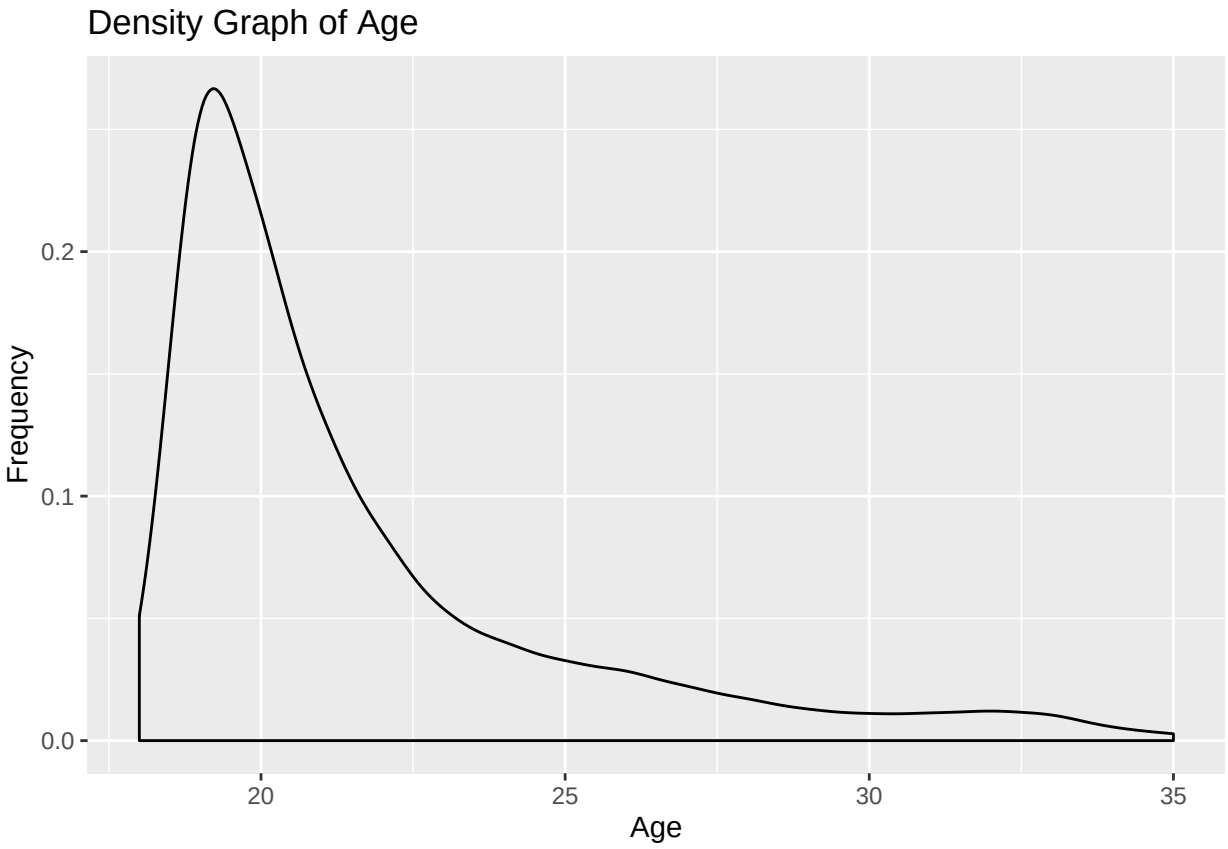
8. The distribution of students in each faculty

Both Faculty of Health, Science and Engineering contains the most amount of student, while CI Faculty and Business School contains the second most amount of student. Faculty of Education, however, has the least amount of student enrolled in the recorded period.

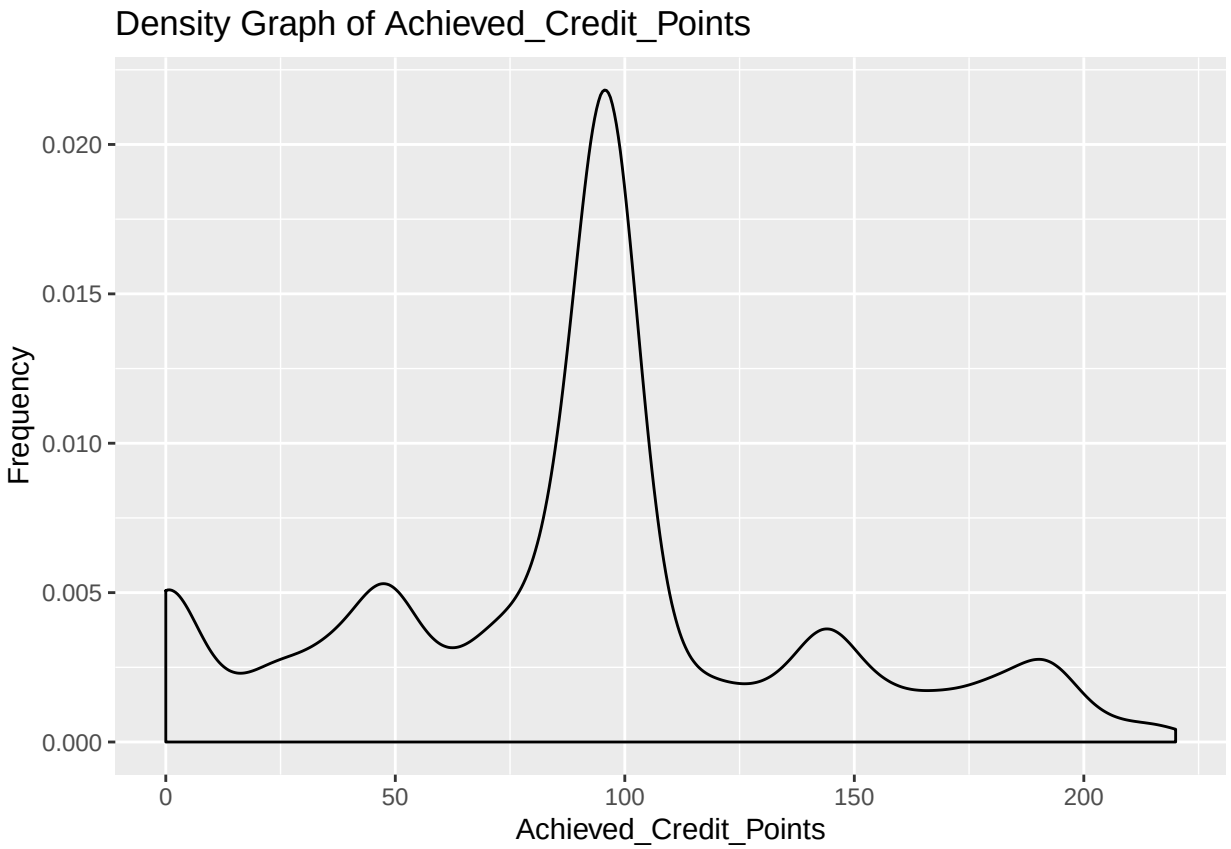
Summary each numerical data using appropriate graphs

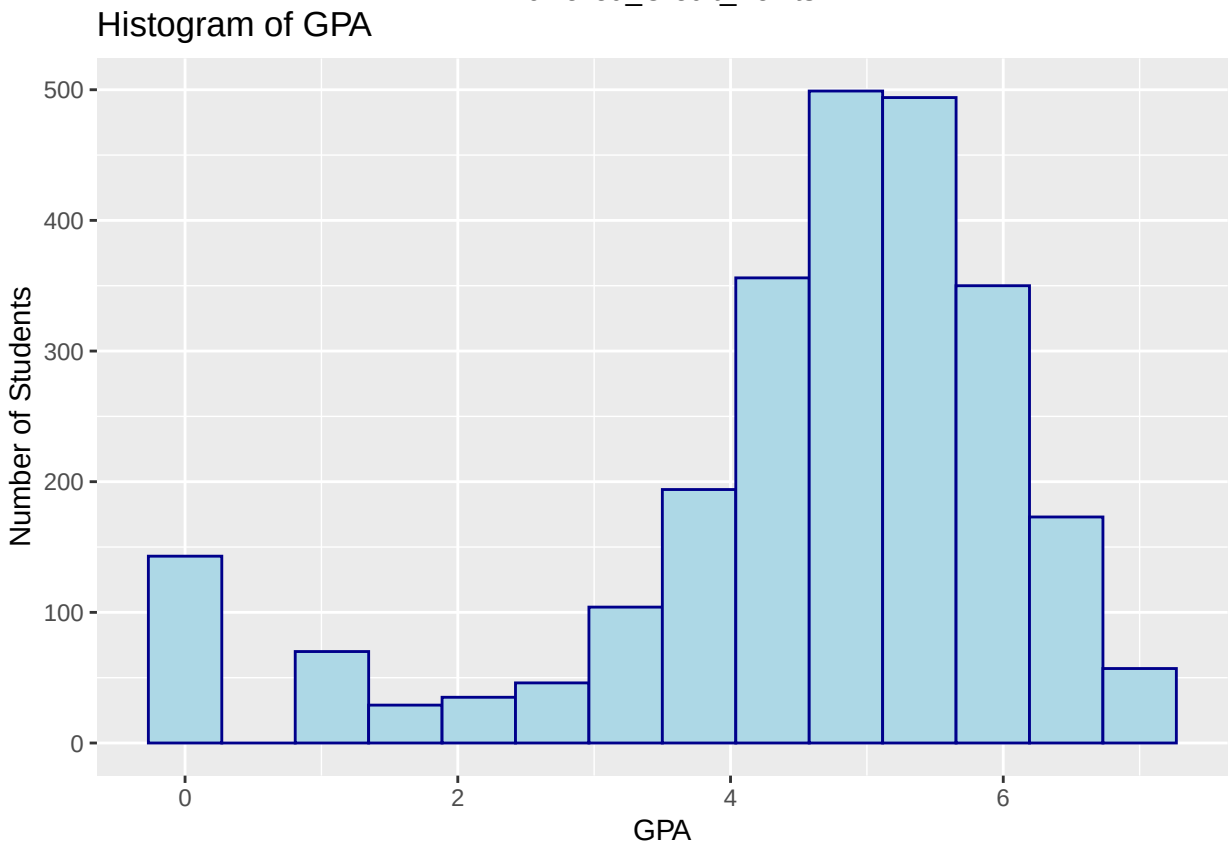
```
# A function print out each appropriate graphs
visualize_numerical_data(uniData)
```

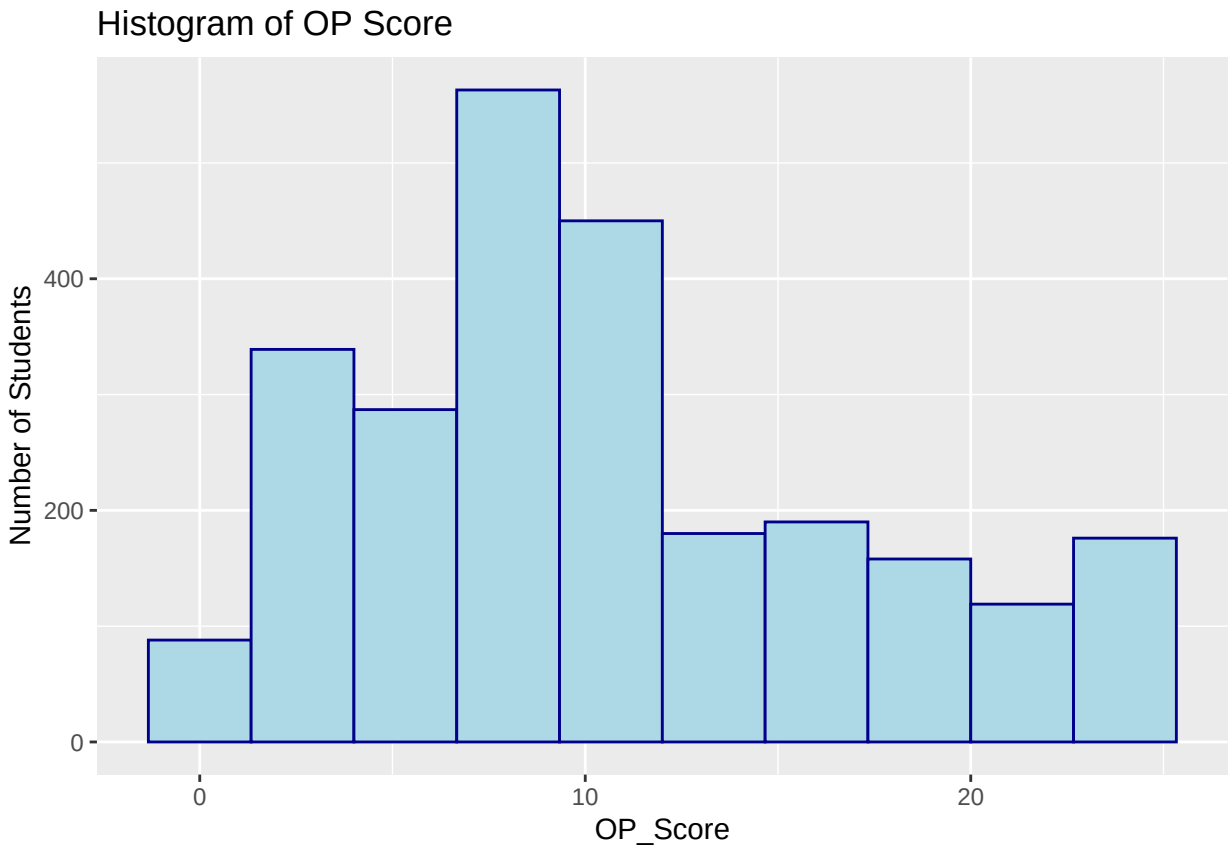
```
## Warning: Removed 132 rows containing non-finite values (stat_density).
```

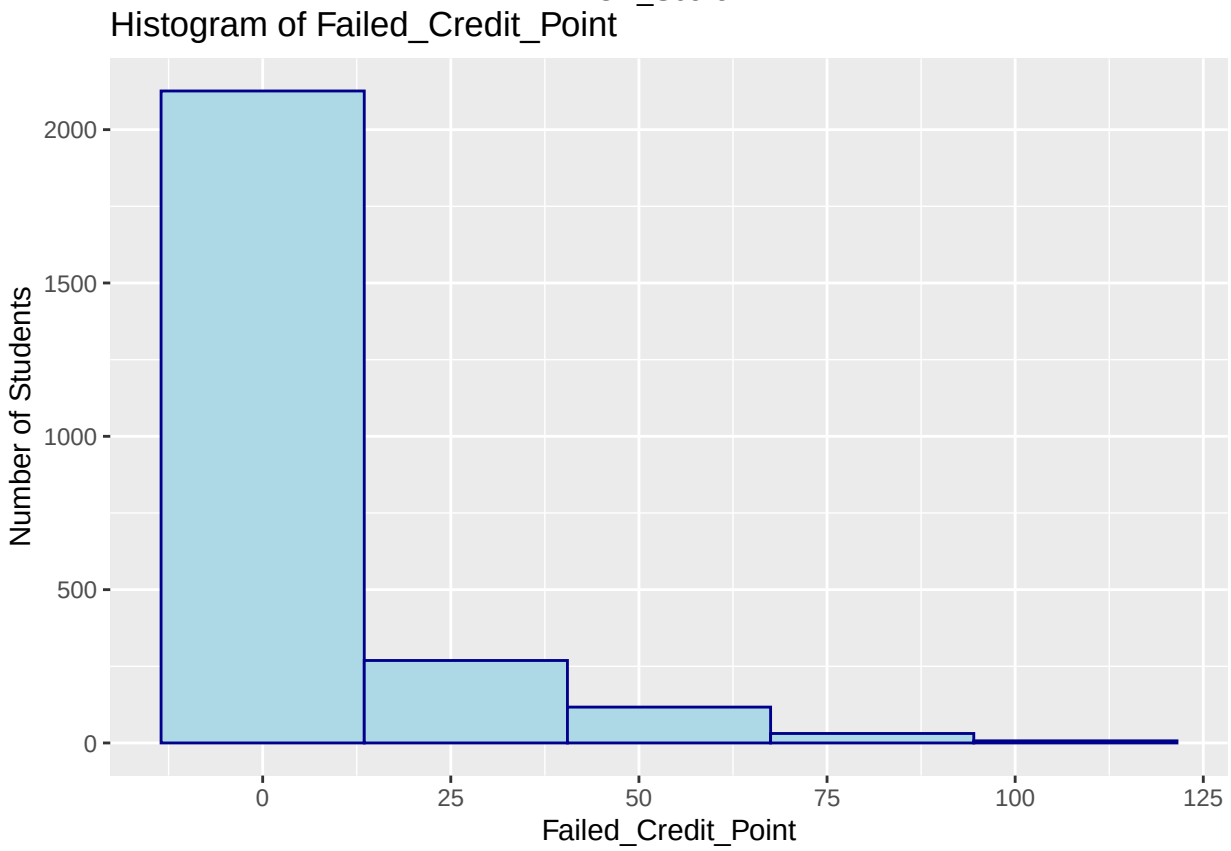



```
## Warning: Removed 46 rows containing non-finite values (stat_density).
```


Density Graph of Achieved_Credit_Points
Frequency
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
Achieved_Credit_Points


Histogram of GPA
Number of Students
500
400
300
200
100
0
0
2
4
6
GPA


Histogram of OP Score
Number of Students
400
200
0
0
10
20
OP_Score


Histogram of Failed_Credit_Point
Number of Students
2000
1500
1000
500
0
0
25
50
75
100
125
Failed_Credit_Point

Task 2 Compare average GPA between male and female students using a graph, conduct a statistical test, and interpret its results

Summary GPA for male

```
male_data <- uniData %>%
  filter(Gender == "M")

summary(male_data$GPA)
```

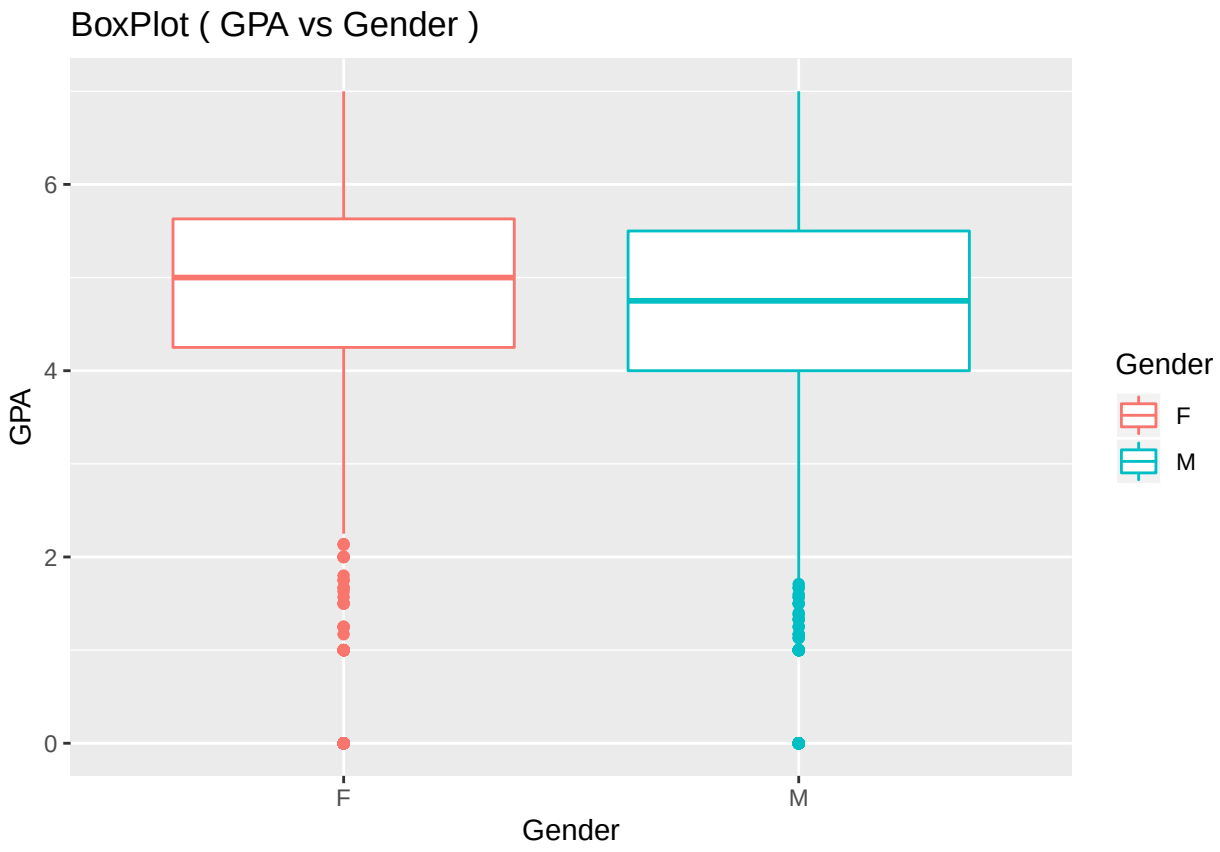
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   4.000   4.750   4.472   5.500   7.000
```

Summary GPA for Female

```
female_data <- uniData %>%
  filter(Gender == "F")

summary(female_data$GPA)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   4.250   5.000   4.629   5.630   7.000
```

```
# Compare average GPA between Male and Female
# Conduct a statistical Test
# Interpret its results
visualize_boxplot_gpa_vs_gender(uniData)
```



T-Test & Variance

```
# T Test
t.test(uniData$GPA ~ uniData$Gender)
```

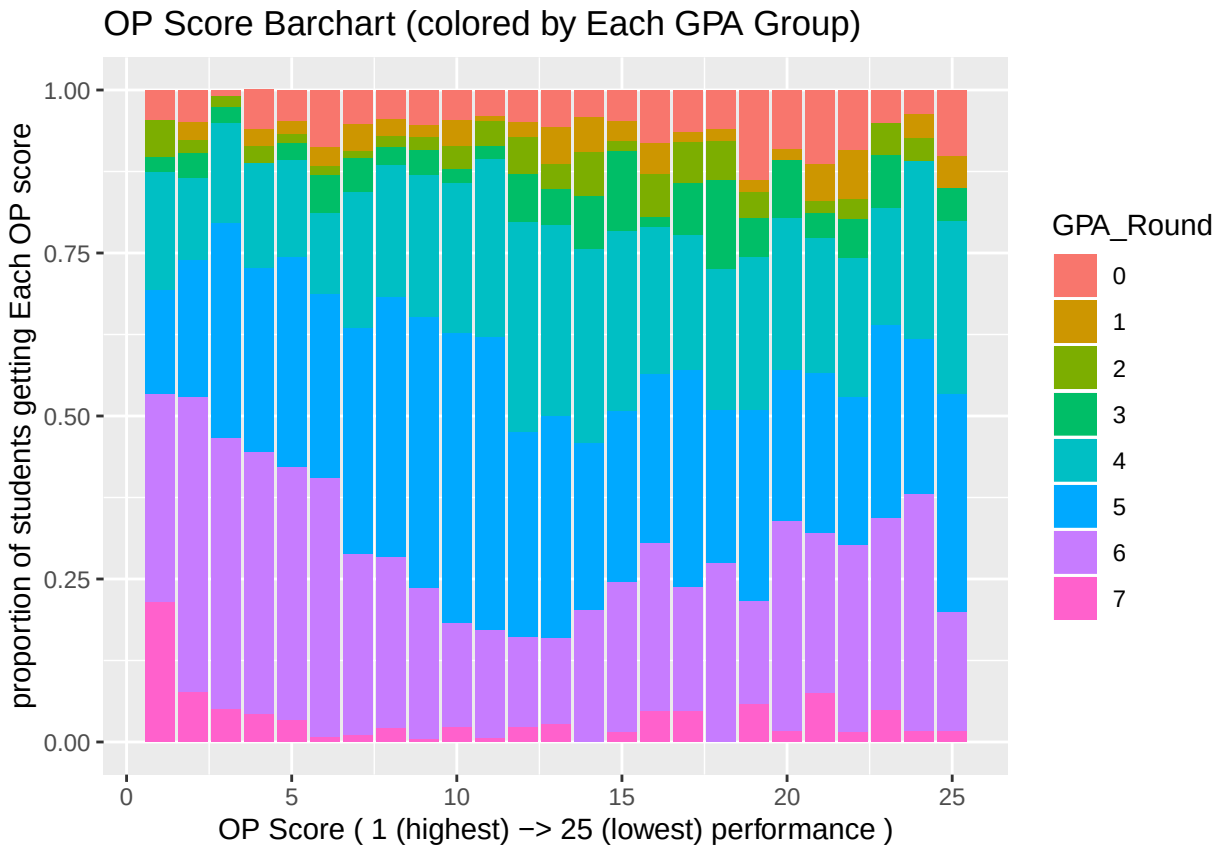
```
##
##  Welch Two Sample t-test
##
## data:  uniData$GPA by uniData$Gender
## t = 2.4454, df = 2539.7, p-value = 0.01453
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.03111718 0.28297210
## sample estimates:
## mean in group F mean in group M
##        4.629282        4.472238
```

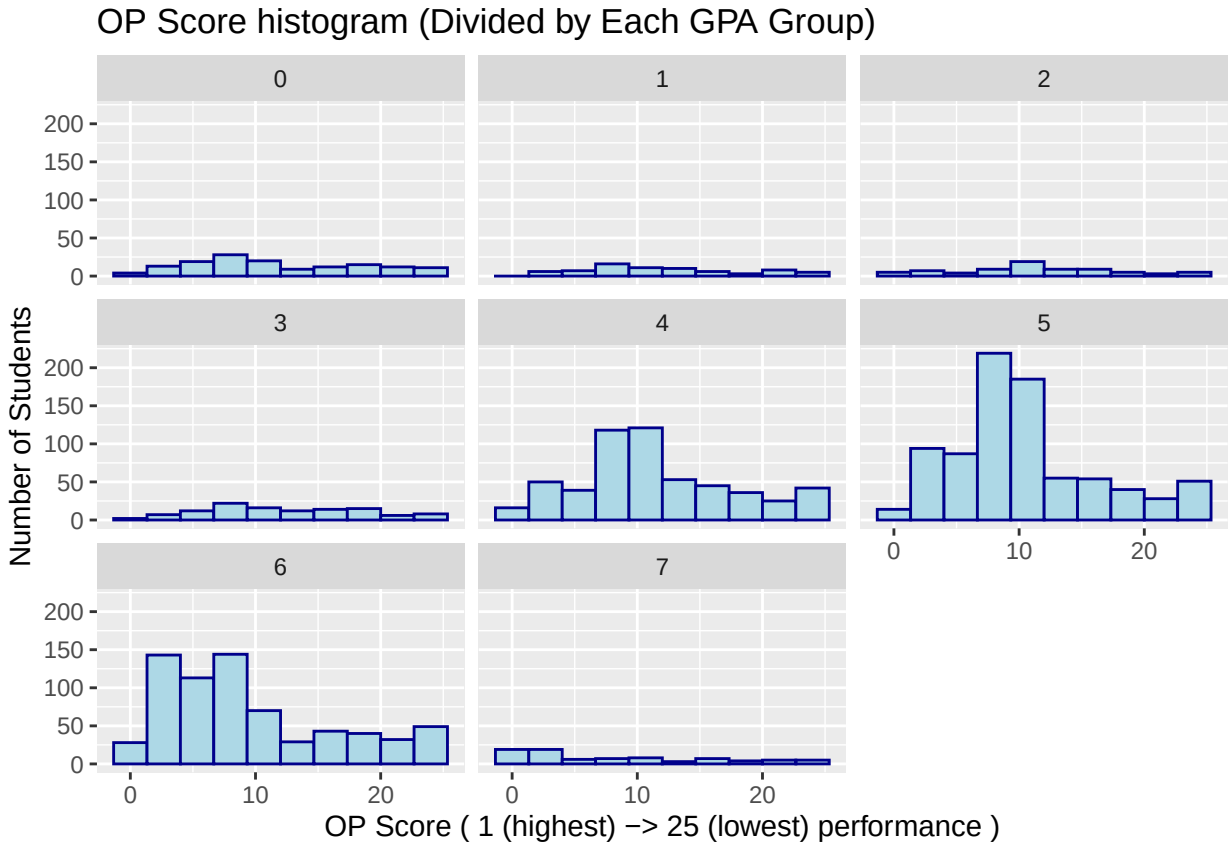
```
# Variance
var.test(uniData$GPA ~ uniData$Gender)
```

```
##
##  F test to compare two variances
##
## data:  uniData$GPA by uniData$Gender
## F = 1.0496, num df = 1253, denom df = 1295, p-value = 0.3873
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9404454 1.1716026
## sample estimates:
## ratio of variances
##           1.049627
```

Task3 Explore the relationship between OP Score and GPA using a graph, describe the relationship

```
visualize_relationship_op_and_gpa(uniData)
```


OP Score Barchart (colored by Each GPA Group)
proportion of students getting Each OP score
1.00
0.75
0.50
0.25
0.00
0
5
10
15
20
25
OP Score (1 (highest) −> 25 (lowest) performance)
GPA_Round
0
1
2
3
4
5
6
7


OP Score histogram (Divided by Each GPA Group)
0
1
2
3
4
5
6
7
Number of Students
200
150
100
50
0
0
10
20
OP Score (1 (highest) −> 25 (lowest) performance)

Bar chart (OP Score VS GPA)

The first bar chart displayed the relationship between OP score and GPA. Each bar indicates every students achieves in the OP exam, while each bar is filled by 8 different colors which indicates how these students performs in the university. The GPA score is rounded to the nearest integer, for instance, 3.67 will be rounded to 4 and 6.18 will be rounded to 6.

Most of the students, who get the lowerest OP exam, tends to performs better in the university. Approximately 50% of students, who get 1 OP score, archieved above GPA 6 when they are studying in university. In contrast, about 40% of student, who get 25 OP score, archieved below GPA 4 which means failed the study in university.

In conclusion, if students get the lower OP scores tends to performs better in the university.

Task 4 Linear Regression

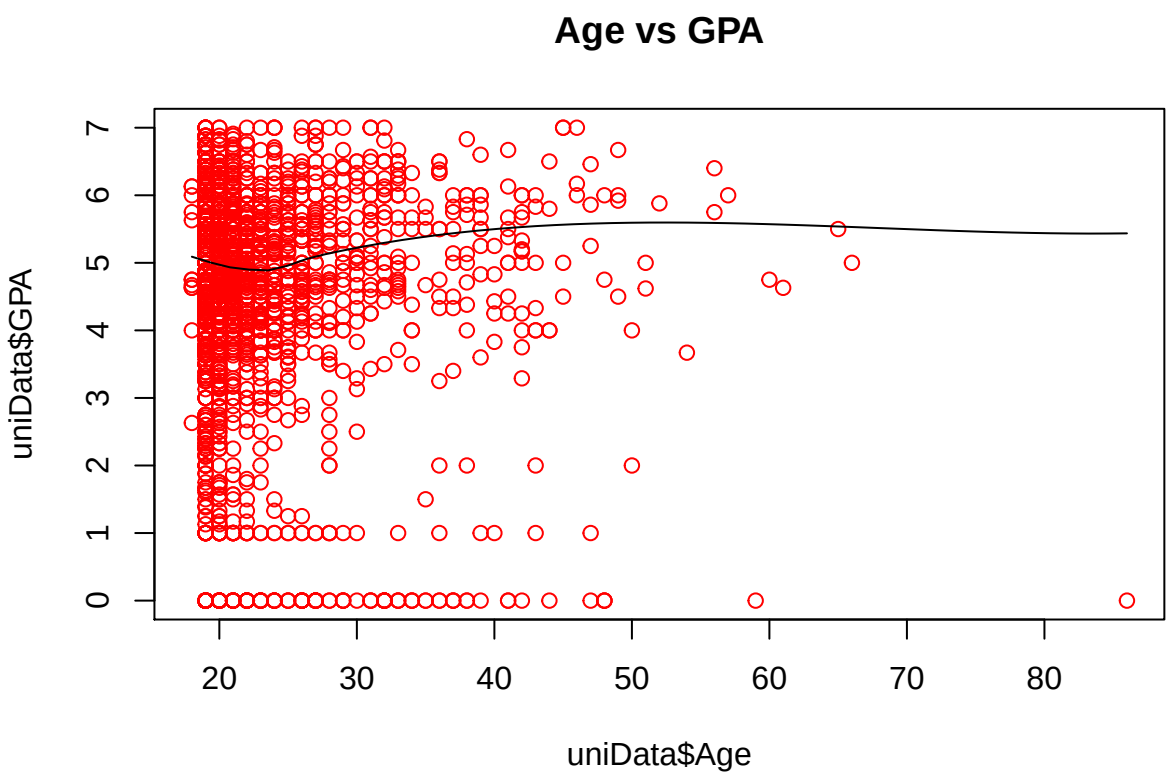
Develop a linear regression model of GPA using the given data. You need to describe your choice of predictors, examine your model's assumptions, assess model fit, and interpret the final model's regression coefficients.

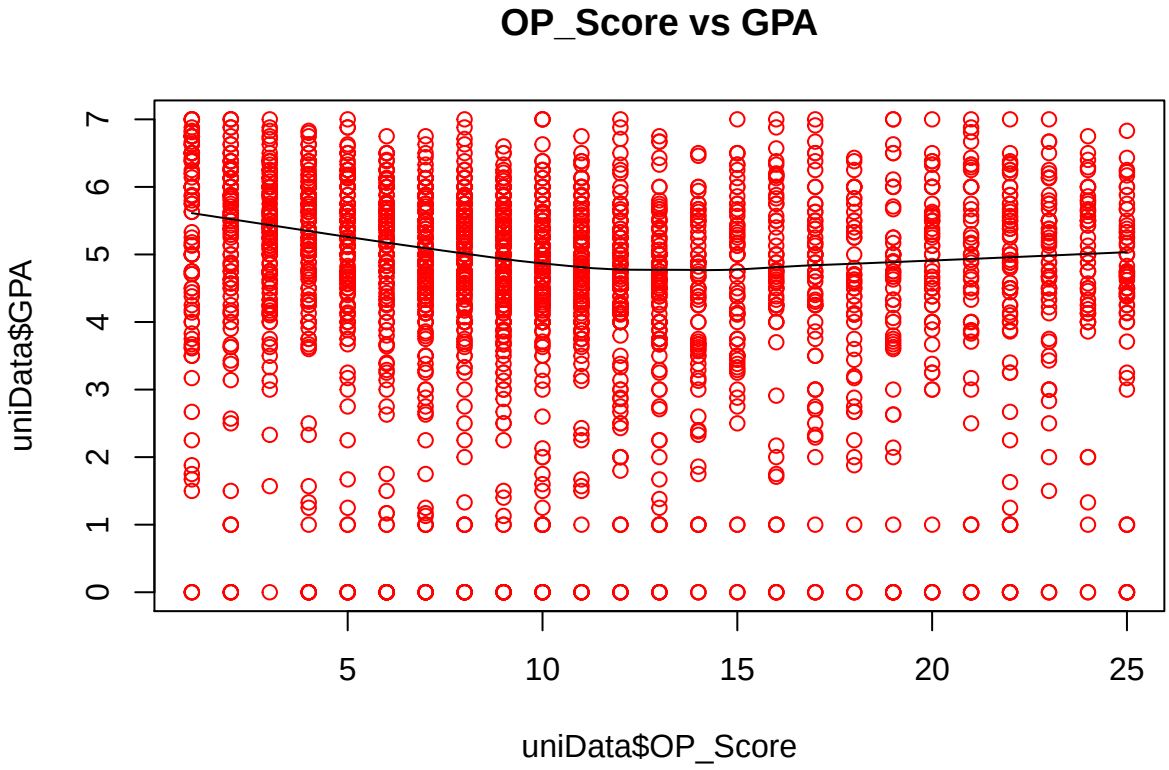
Analyse Each numerical data its relation related to GPA

Correlation between each numerical data and GPA

1. Age : -0.0641342
2. OP_Score : -0.129619
3. Achieved_Credit_Points : 0.4920035
4. Failed_Credit_Points : -0.473419

```
visualize_scatterplots_Vs_GPA(uniData)
```

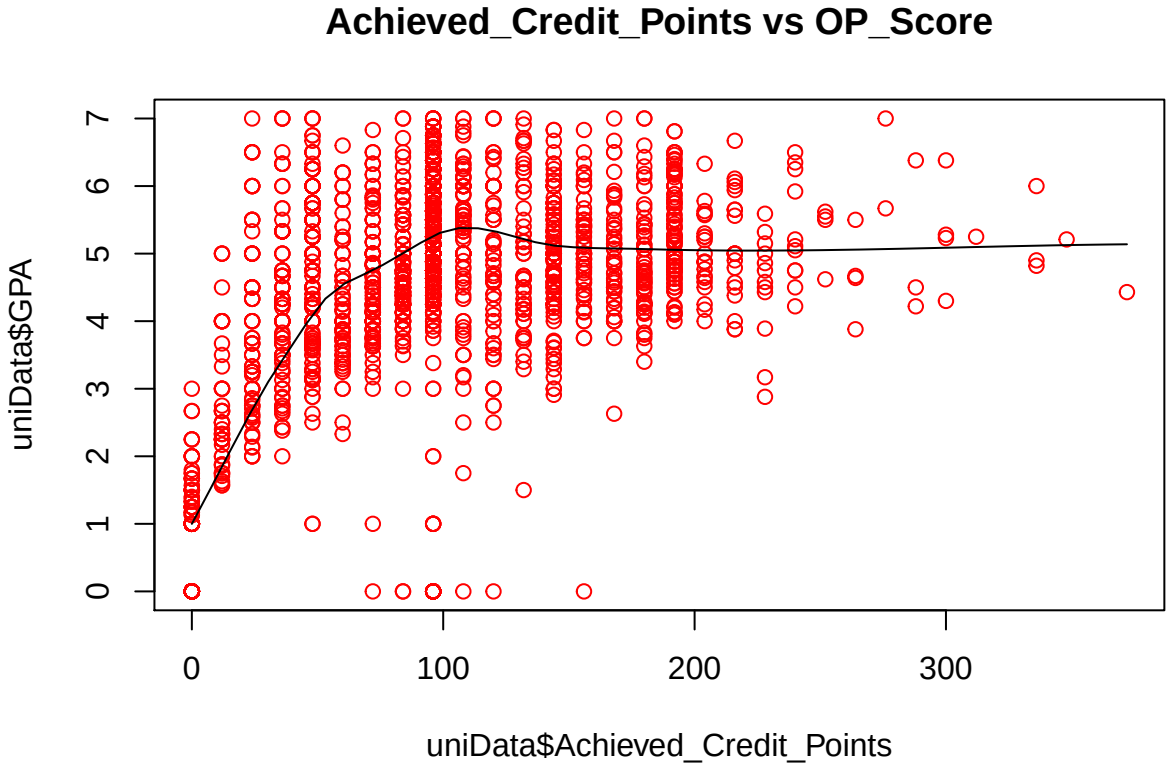


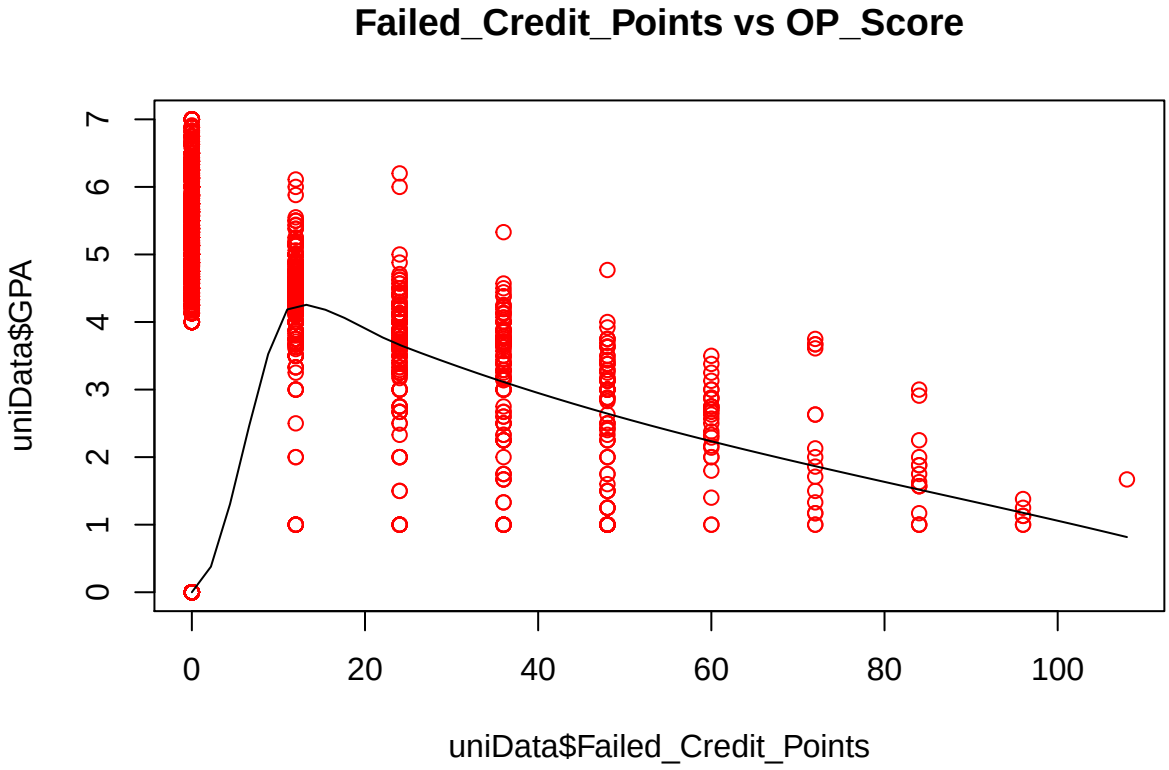



```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : at -0.54

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : radius 0.2916

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : all data on boundary of neighborhood. make span bigger

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : pseudoinverse used at -0.54

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : neighborhood radius 0.54

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : reciprocal condition number 1

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : zero-width neighborhood. make span bigger

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : There are other near singularities as well. 144

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : at -0.54

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : radius 0.2916

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : all data on boundary of neighborhood. make span bigger

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : pseudoinverse used at -0.54
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : neighborhood radius 0.54

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : reciprocal condition number 1

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : zero-width neighborhood. make span bigger

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : There are other near singularities as well. 144

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : at -0.54

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : radius 0.2916

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : all data on boundary of neighborhood. make span bigger

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : pseudoinverse used at -0.54

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : neighborhood radius 0.54

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : reciprocal condition number 1

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : zero-width neighborhood. make span bigger

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : There are other near singularities as well. 144

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : at -0.54

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : radius 0.2916

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : all data on boundary of neighborhood. make span bigger

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : pseudoinverse used at -0.54

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : neighborhood radius 0.54

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : reciprocal condition number 1

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : zero-width neighborhood. make span bigger

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : There are other near singularities as well. 144

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : at -0.54
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : radius 0.2916

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : all data on boundary of neighborhood. make span bigger

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : pseudoinverse used at -0.54

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : neighborhood radius 0.54

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : reciprocal condition number 1

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : zero-width neighborhood. make span bigger

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## FALSE, : There are other near singularities as well. 144
```





Spiting dataframe into training set and test set

```
# Data Preprocessing Library
library(caTools)

set.seed(2)

split <- sample.split(uniData, SplitRatio = 0.7)
train <- subset(uniData, split==TRUE)
test <- subset(uniData, split==FALSE)
```

Training Linear Regression Model

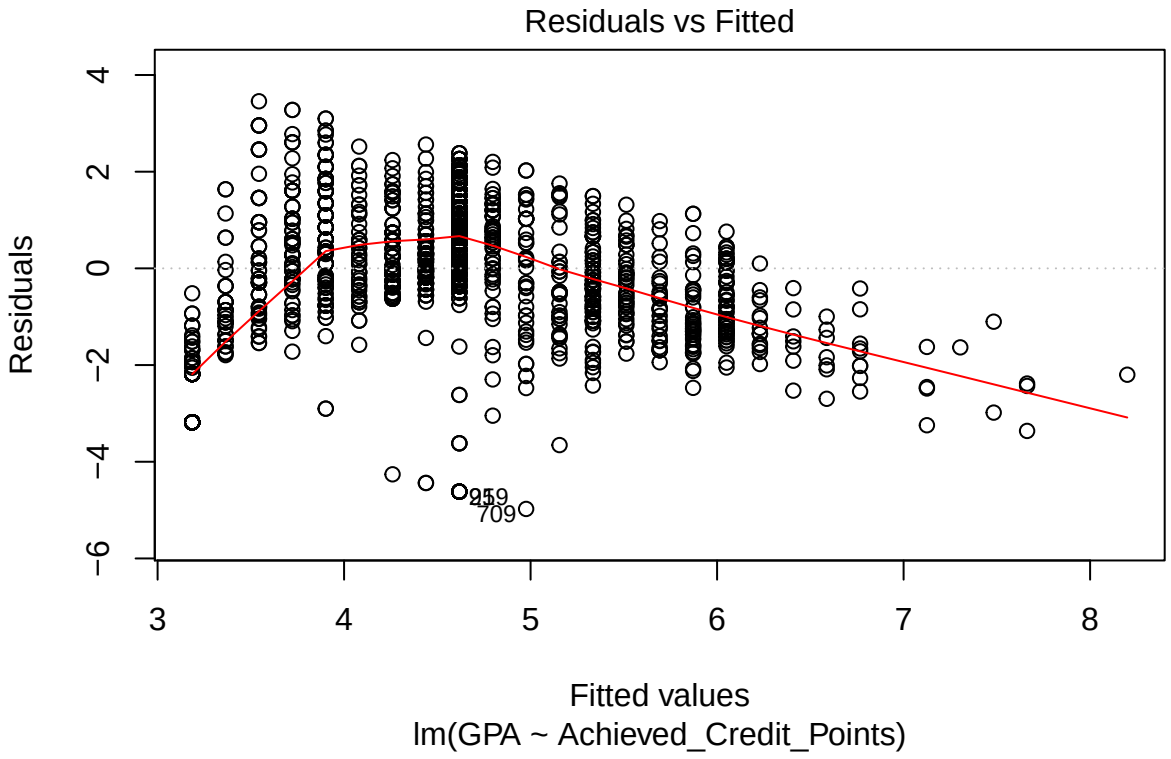
```
linear_model <- lm(GPA ~ Achieved_Credit_Points, data=train)
summary(linear_model)
```

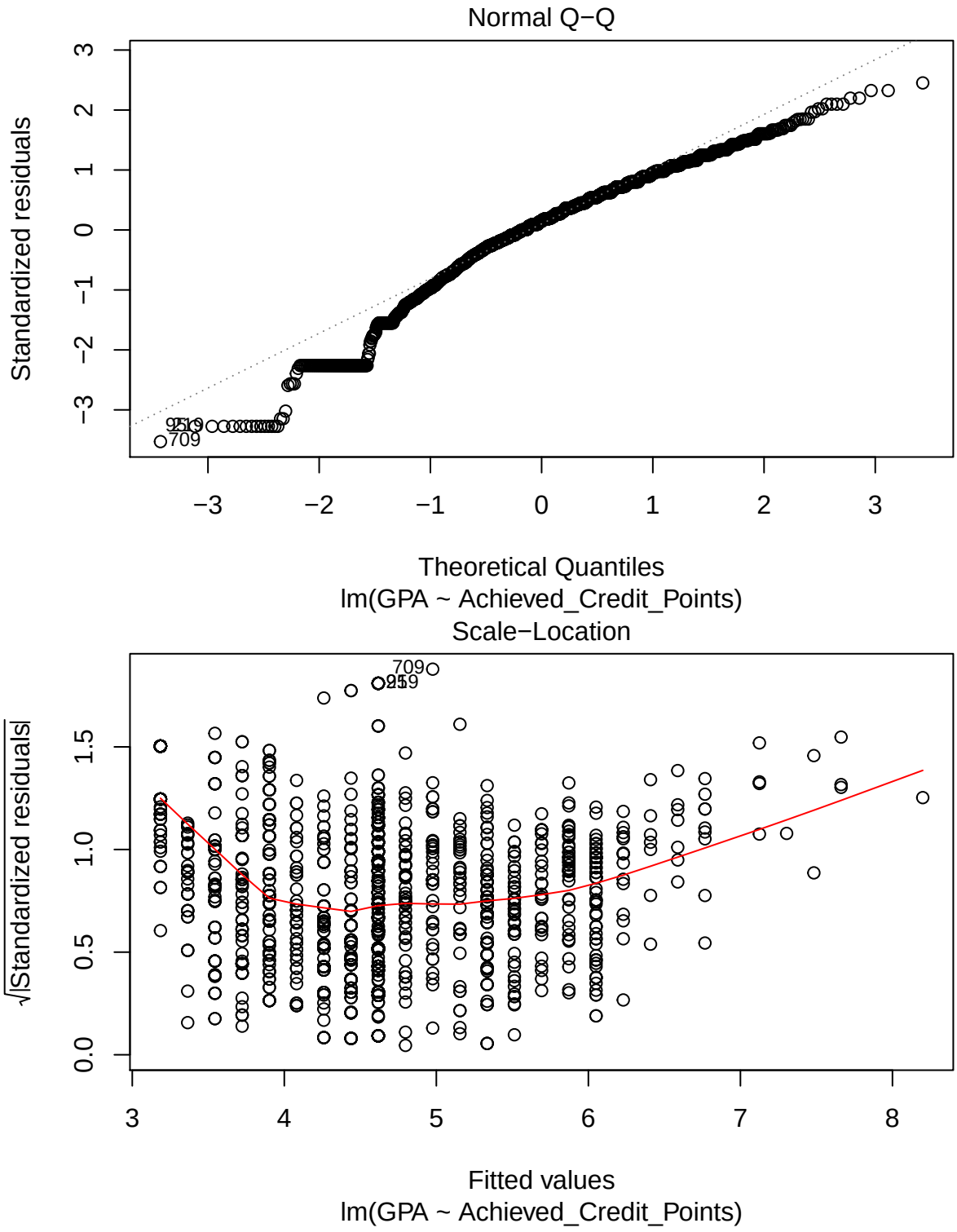
```
##
## Call:
## lm(formula = GPA ~ Achieved_Credit_Points, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9761 -0.7239  0.2029  1.0120  3.4564
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3.1854527  0.0688092   46.29   <2e-16 ***
## Achieved_Credit_Points 0.0149224  0.0006455   23.12   <2e-16 ***
## ---
```

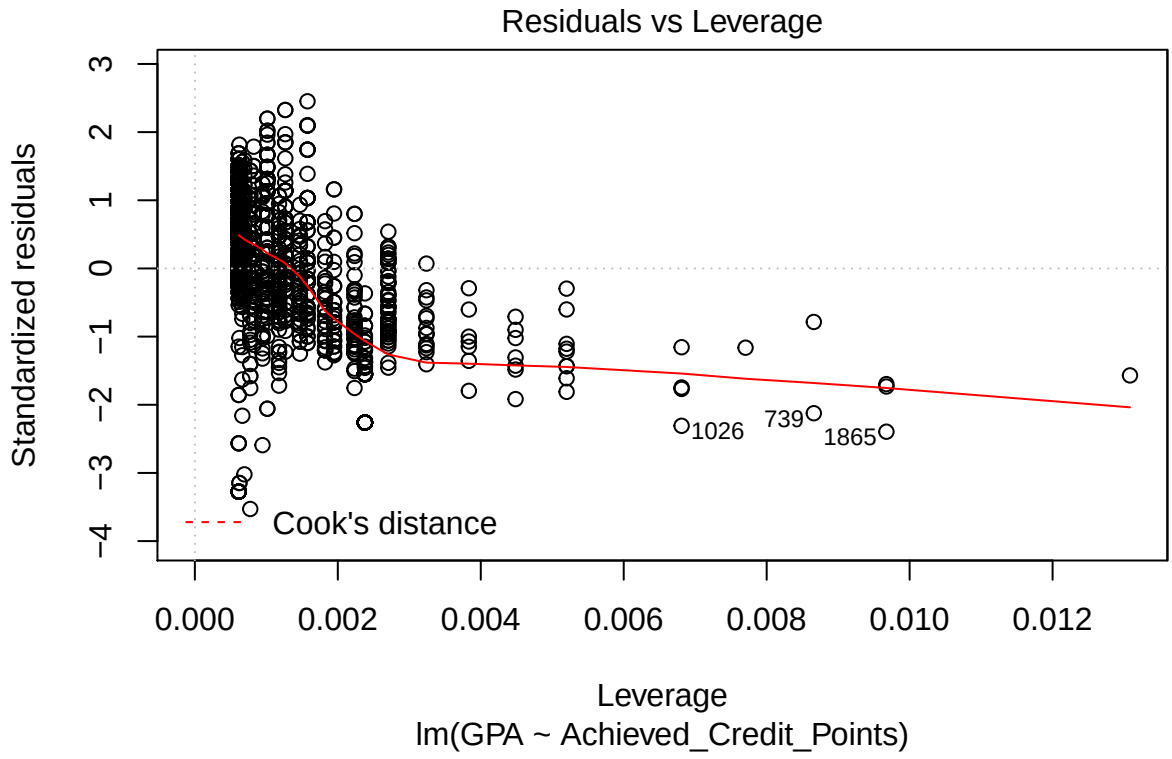
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.411 on 1637 degrees of freedom
## Multiple R-squared:  0.2461, Adjusted R-squared:  0.2457
## F-statistic: 534.4 on 1 and 1637 DF,  p-value: < 2.2e-16
```

Plot the Linear Regression Prediction Line

```
plot(linear_model)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(GPA ~ Achieved_Credit_Points)
Scale−Location
√|Standardized residuals|
Fitted values
lm(GPA ~ Achieved_Credit_Points)



Task 5 Logistic Regression

```
# Logistic Regression Library
library(DAAG)
```

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

```
library(cowplot)
```

```
##
## ************************************************************
## Note: As of version 1.0.0, cowplot does not change the
##   default ggplot2 theme anymore. To recover the previous
##   behavior, execute:
##   theme_set(theme_cowplot())
## ************************************************************
```

```
library(InformationValue)
```