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:

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

# 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

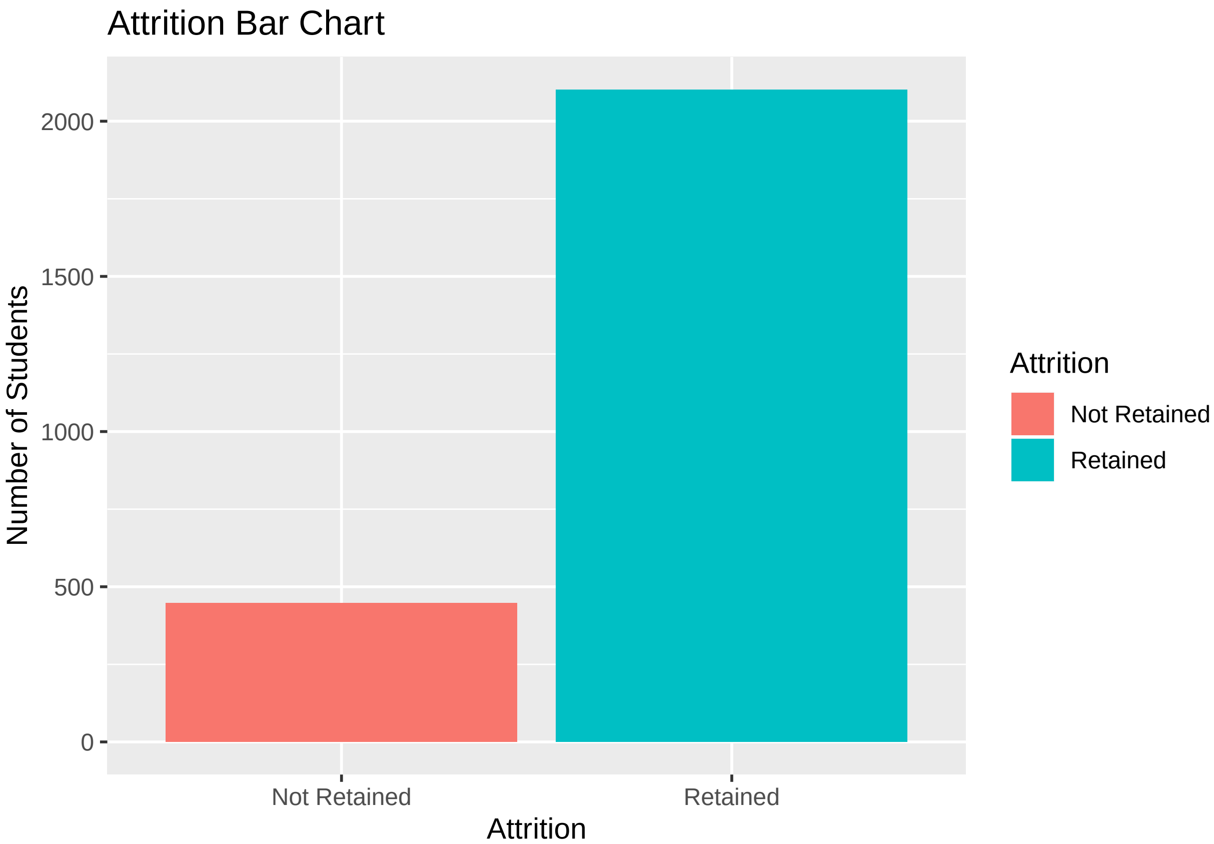
Summary each variables using a table

```r
summary(uniData)
```

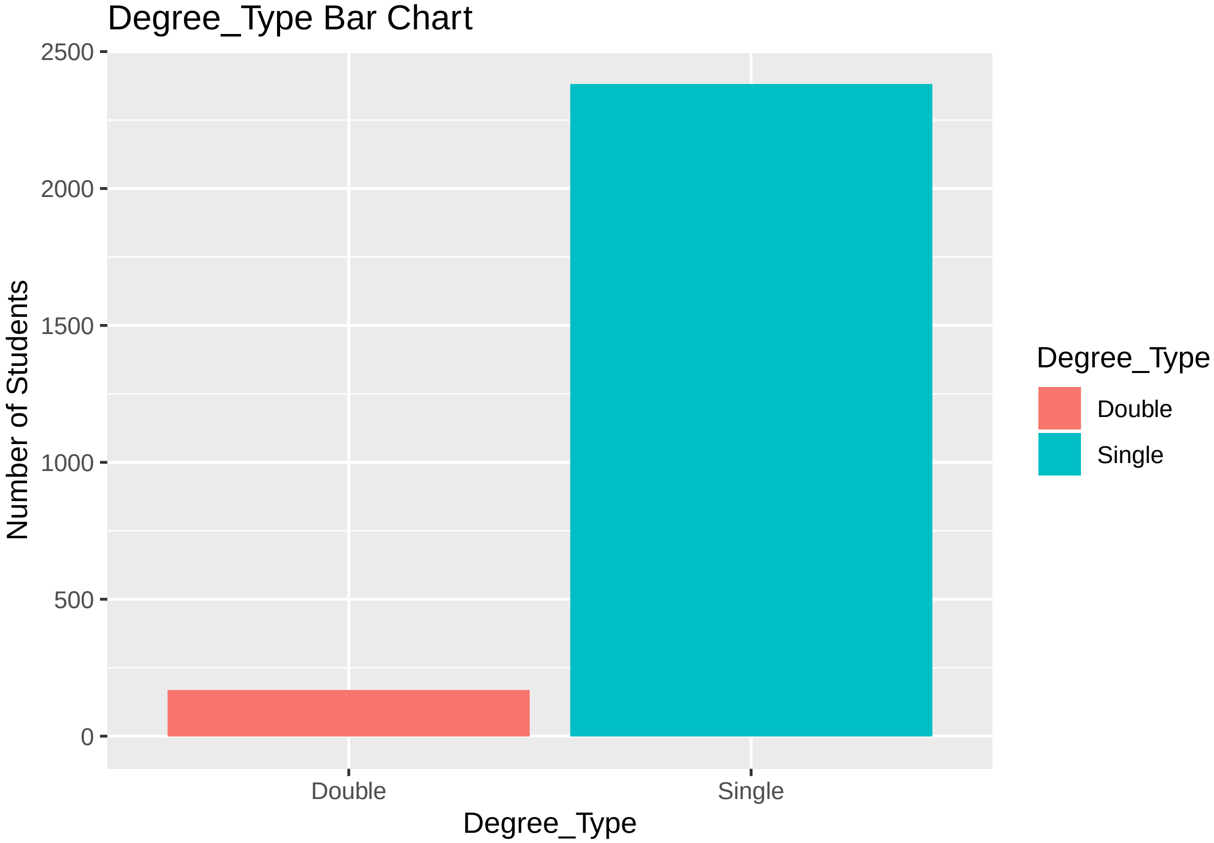
```
##         Attrition    Degree_Type   Achieved_Credit_Points  Attendance_Type
##  Not Retained: 448   Double: 169   Min.   :  0.00         Full Time:2308
##  Retained    :2102   Single:2381   1st Qu.: 60.00         Part Time: 242
##                                    Median : 96.00
##                                    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

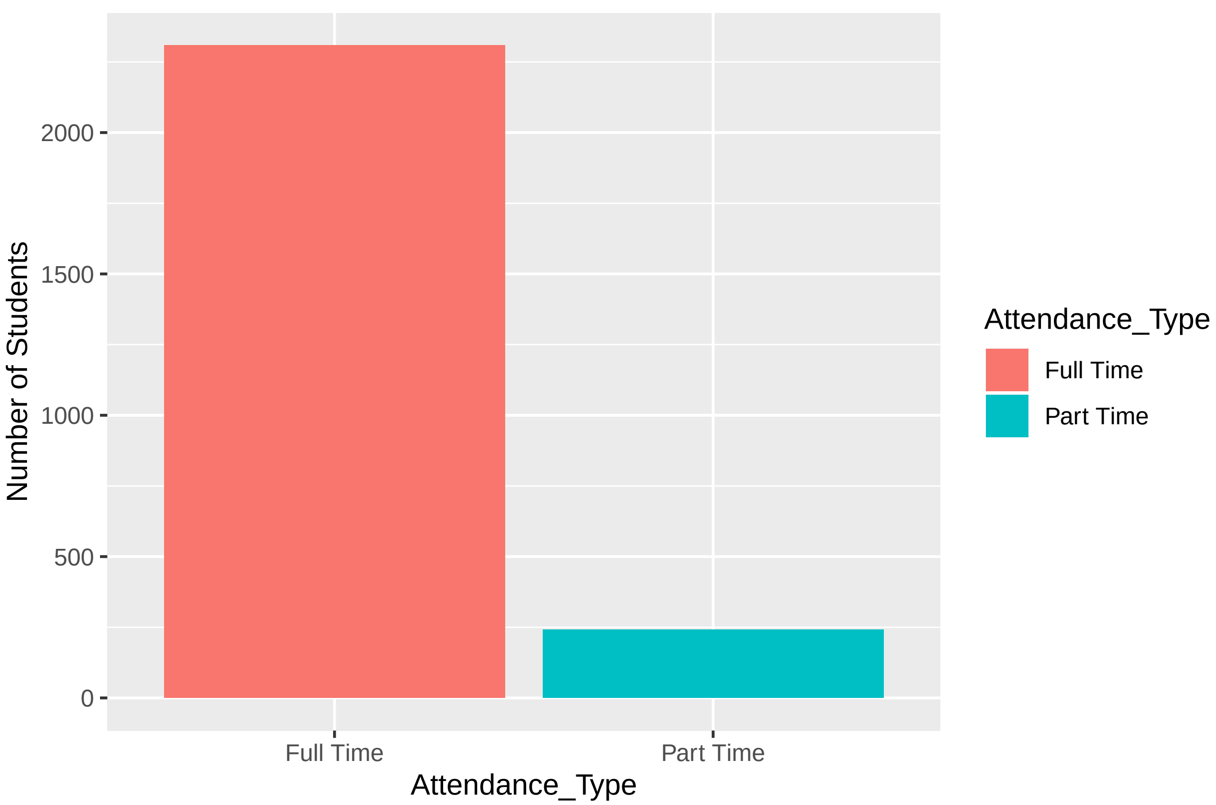
```
visualize_categorical_data(uniData)
```



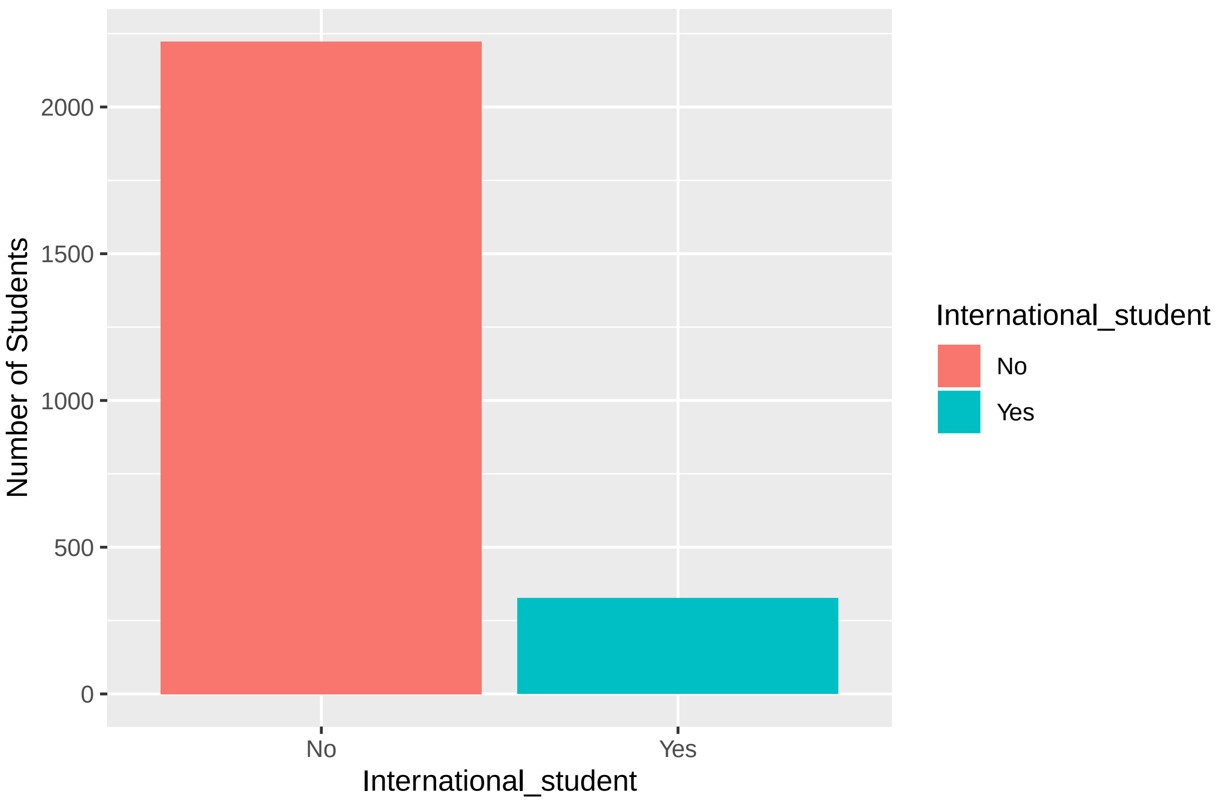
Attrition Bar Chart



Degree_Type Bar Chart

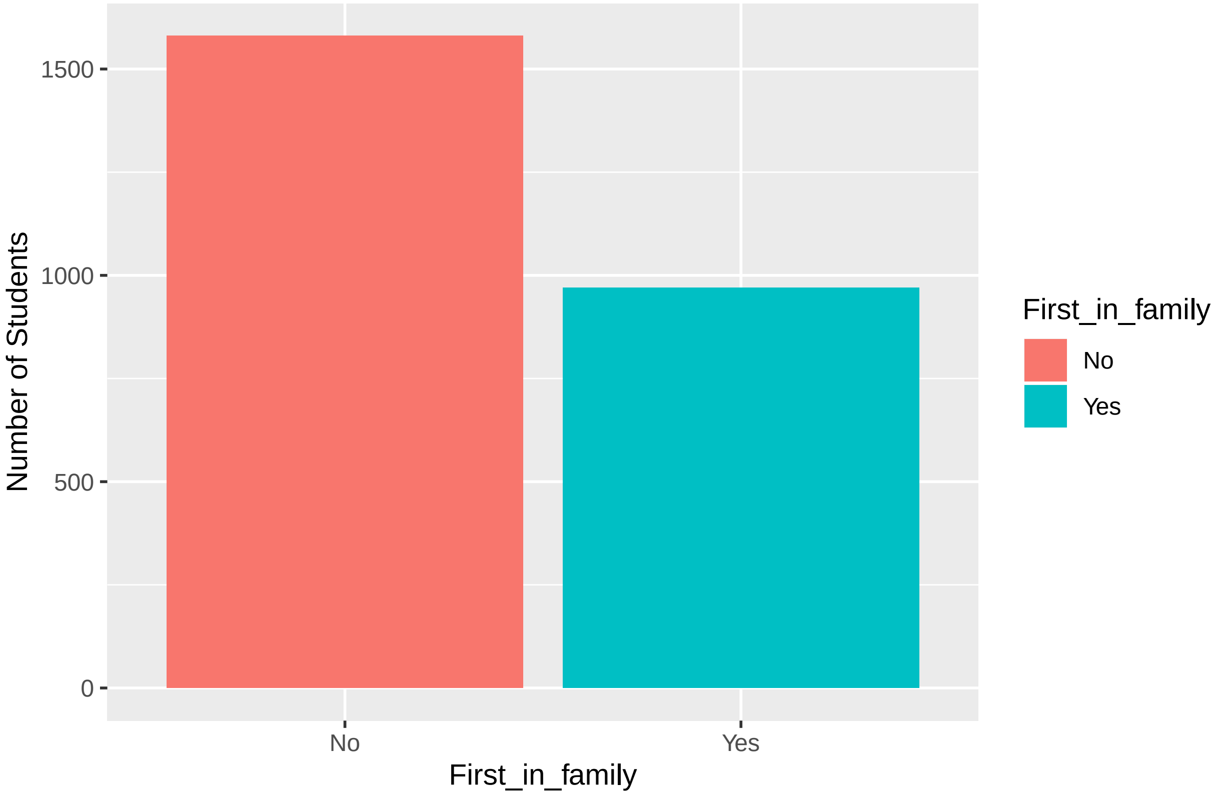


Attendance_Type Bar Chart

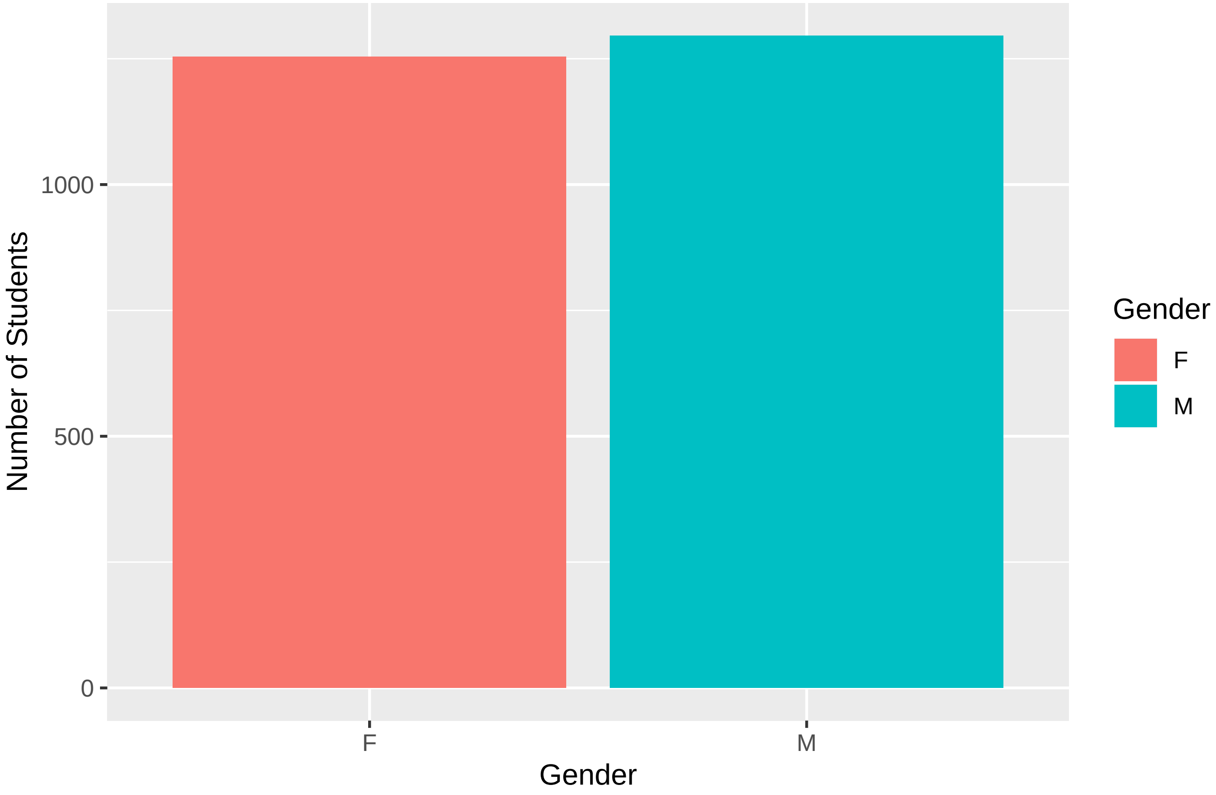


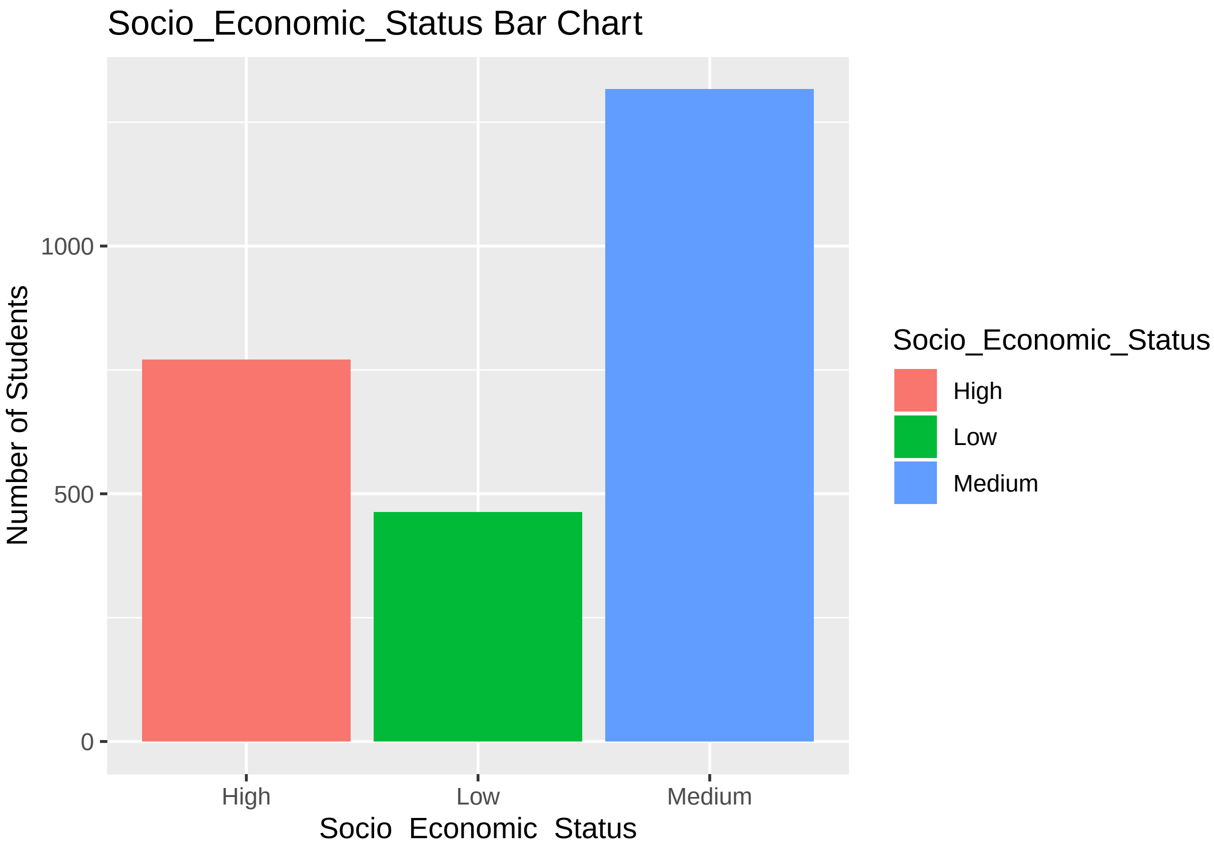
International_student Bar Chart

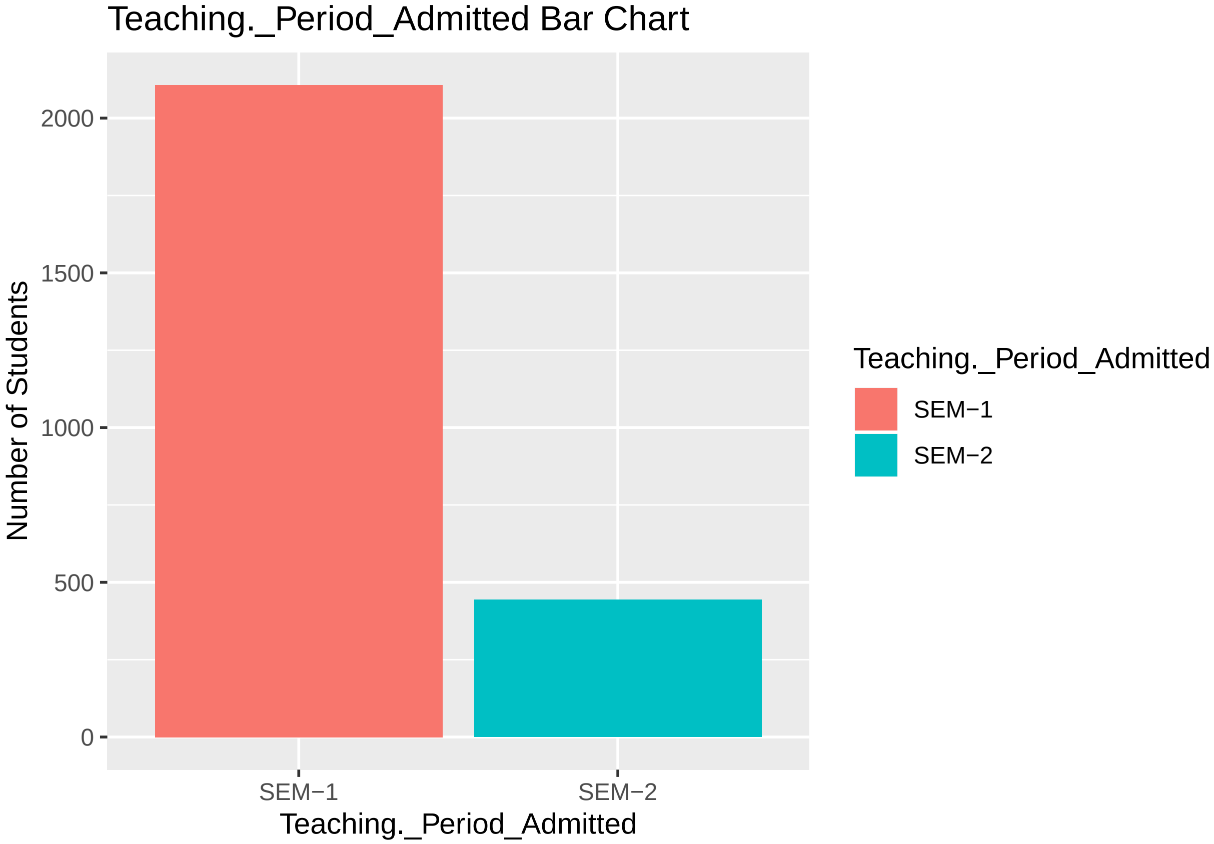
First_in_family Bar Chart

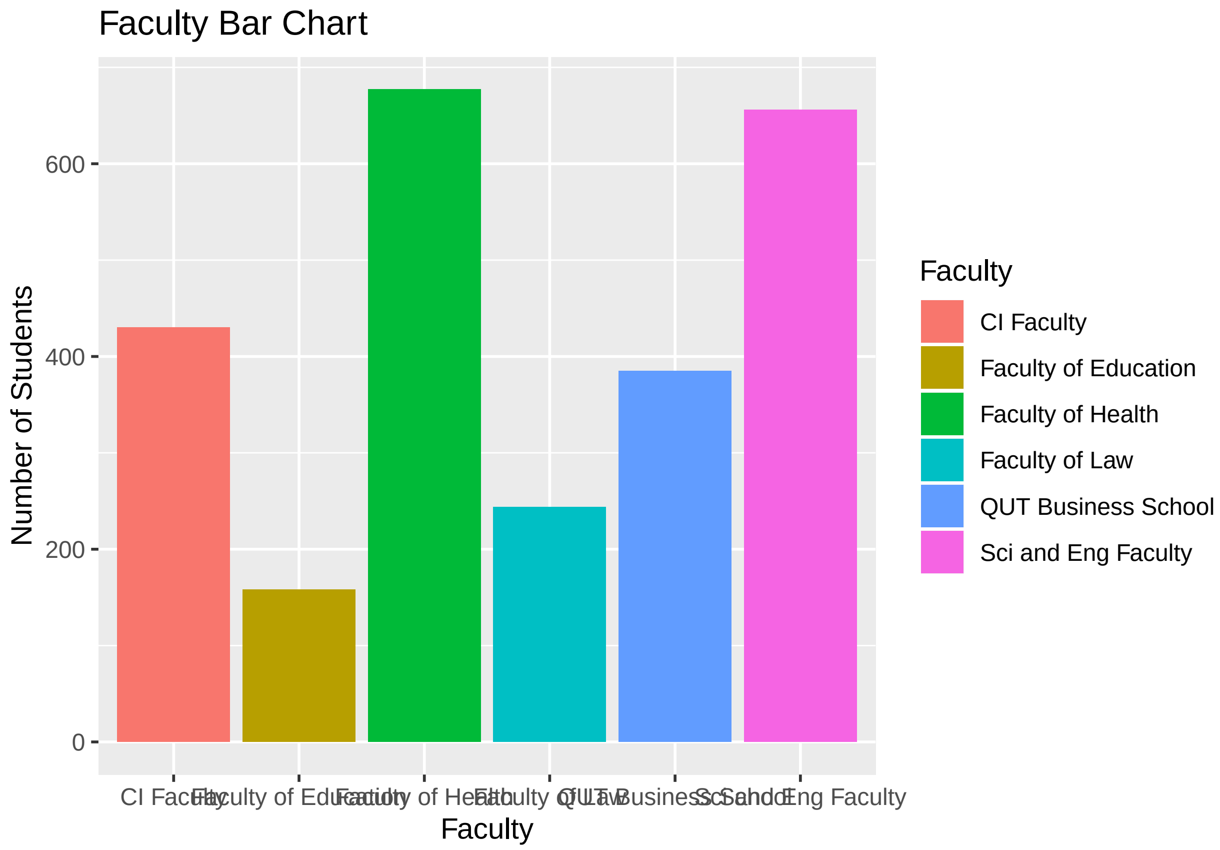


Gender Bar Chart




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 data frame. The summaries of each chart 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% of students are doing a single degree, while the rest are doing a double degree.

3. The attendance type distribution among students

Not surprisingly, most of the students are studying full-time at university. On the other hand, around ten per cent of students is a part-time student.

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

There are approximately 95% of students are local students in the university, while the remainder is international 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 families. Approximately 30 per cent of students are in high-income families, while around 18% of students were heavily concerning their economic status.

7. The distribution of the period students admitted to university

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

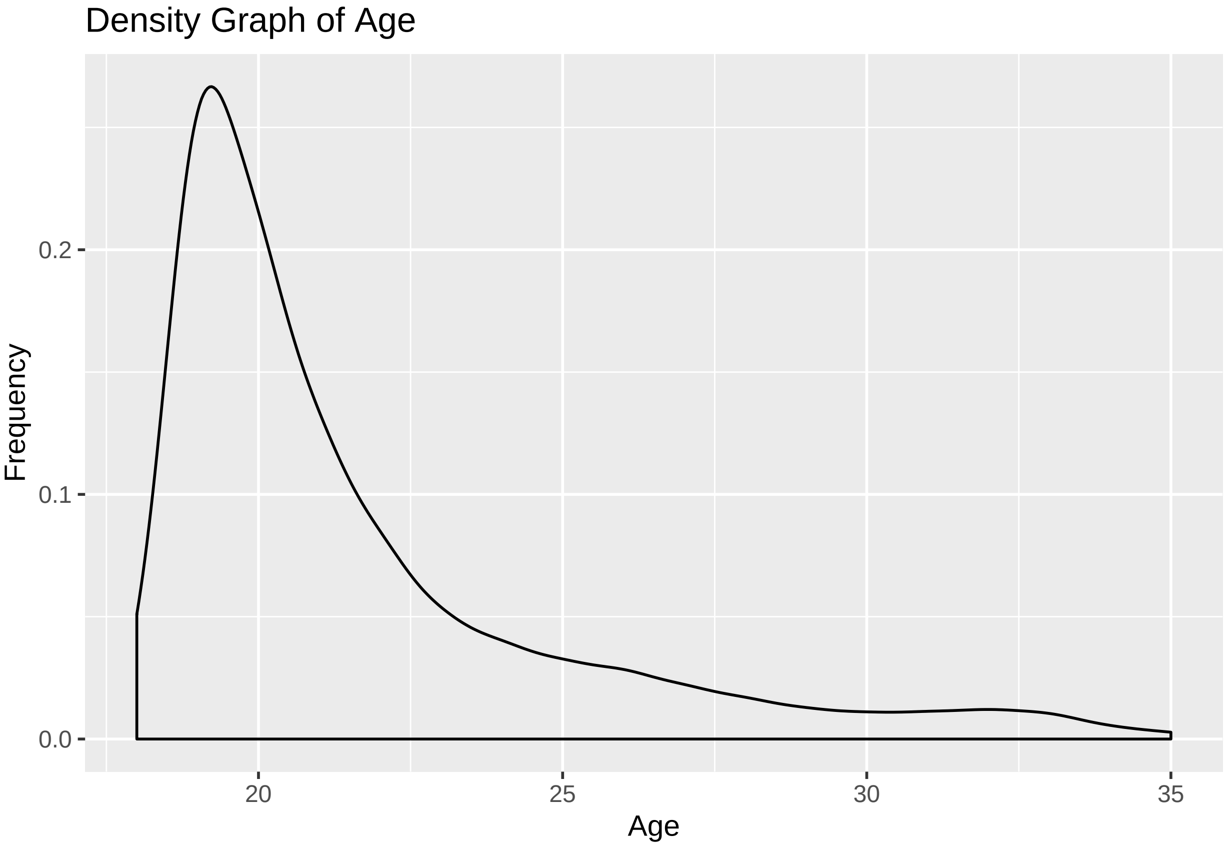
8. The distribution of students in each faculty

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

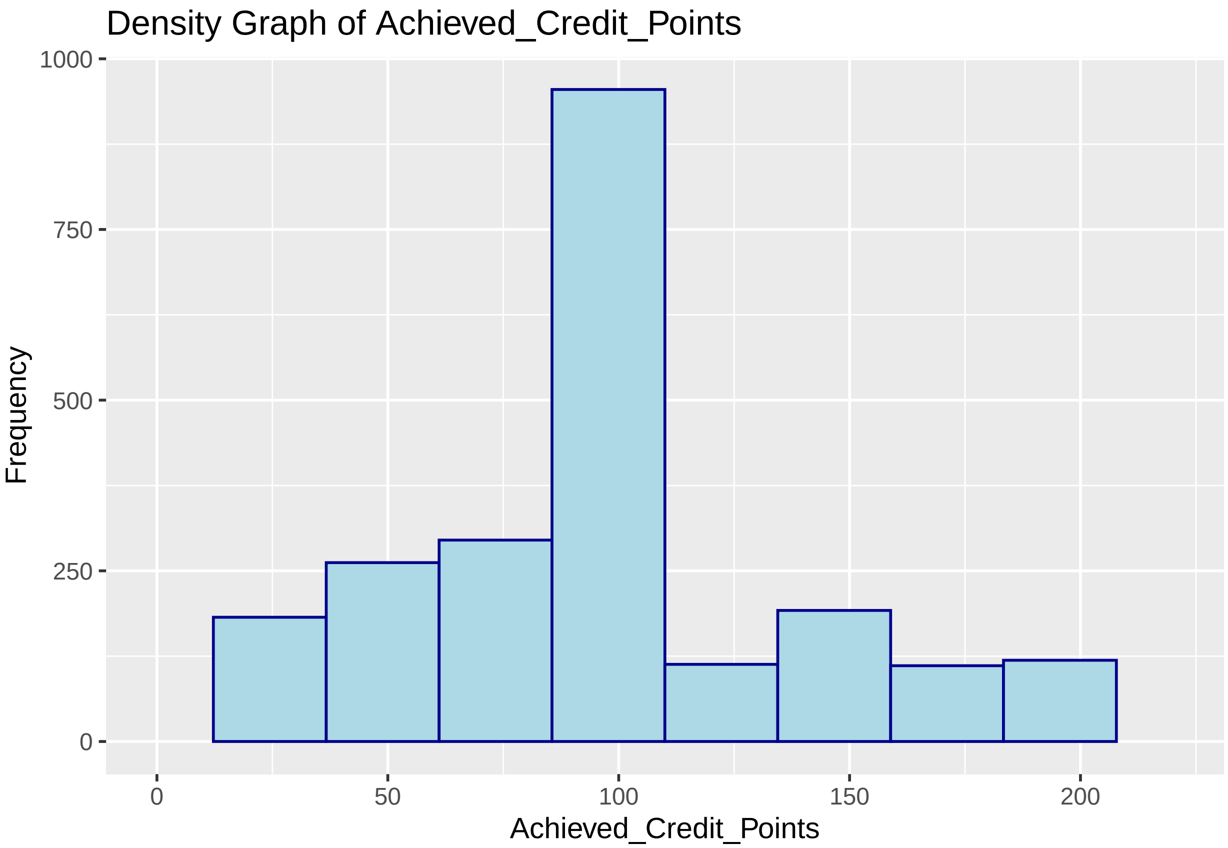
Summary each numerical data using appropriate graphs

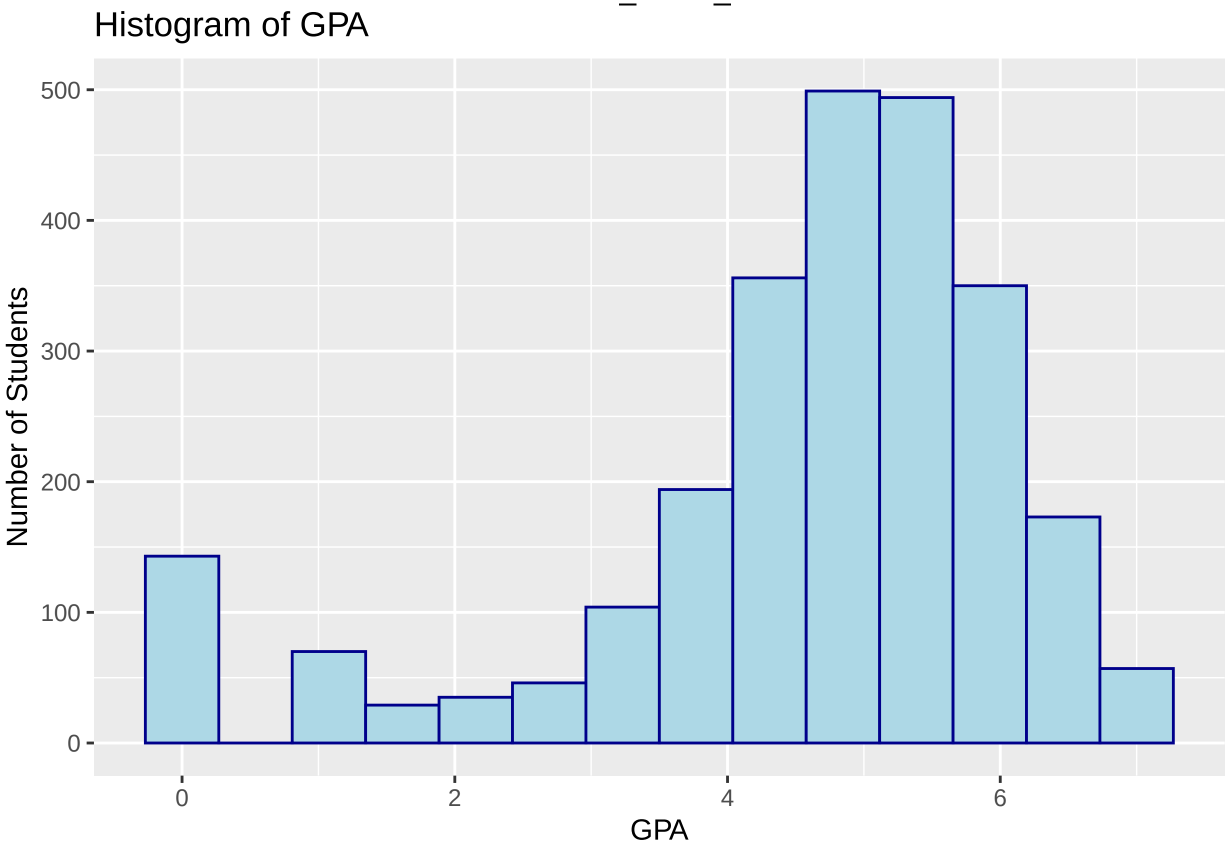
```r
# A function print out each appropriate graphs
visualize_numerical_data(uniData)
```

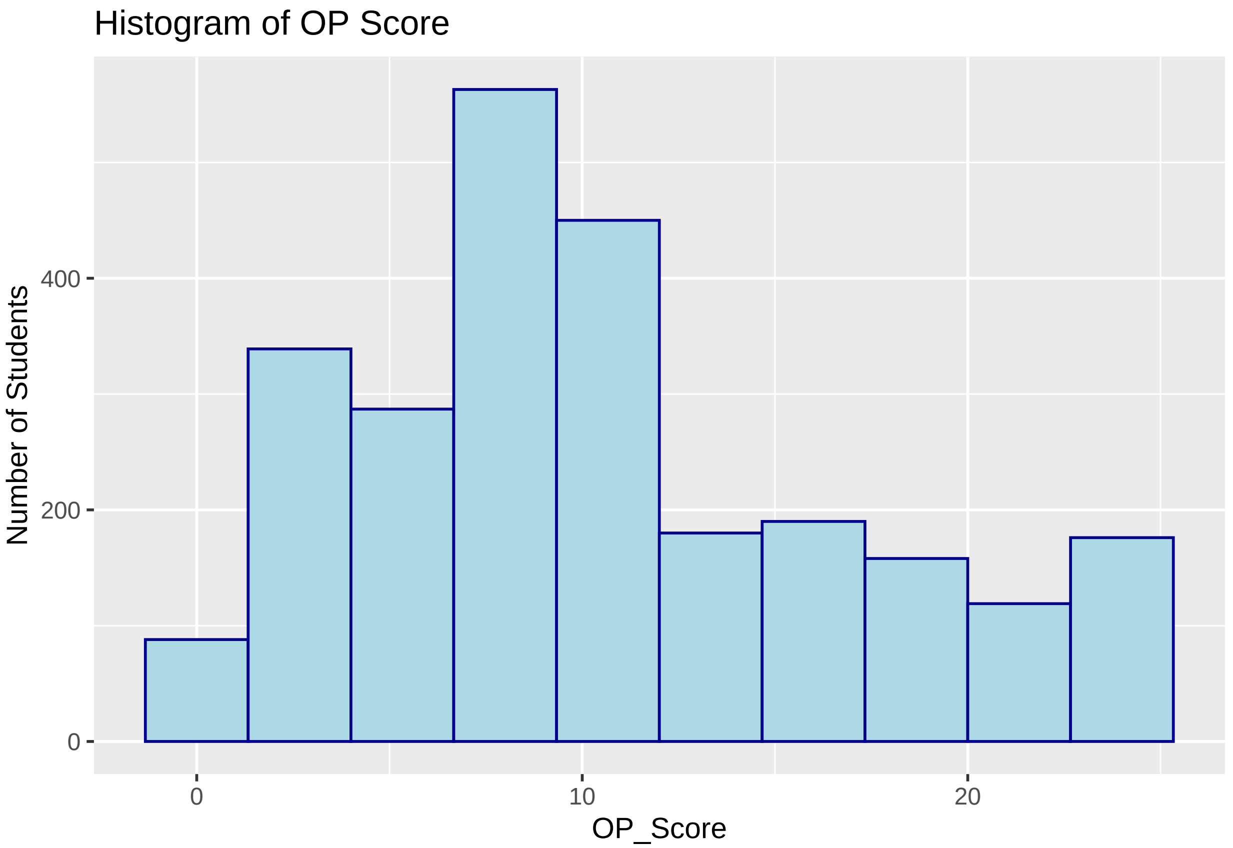
```
## Warning: Removed 132 rows containing non-finite values (stat_density).
```

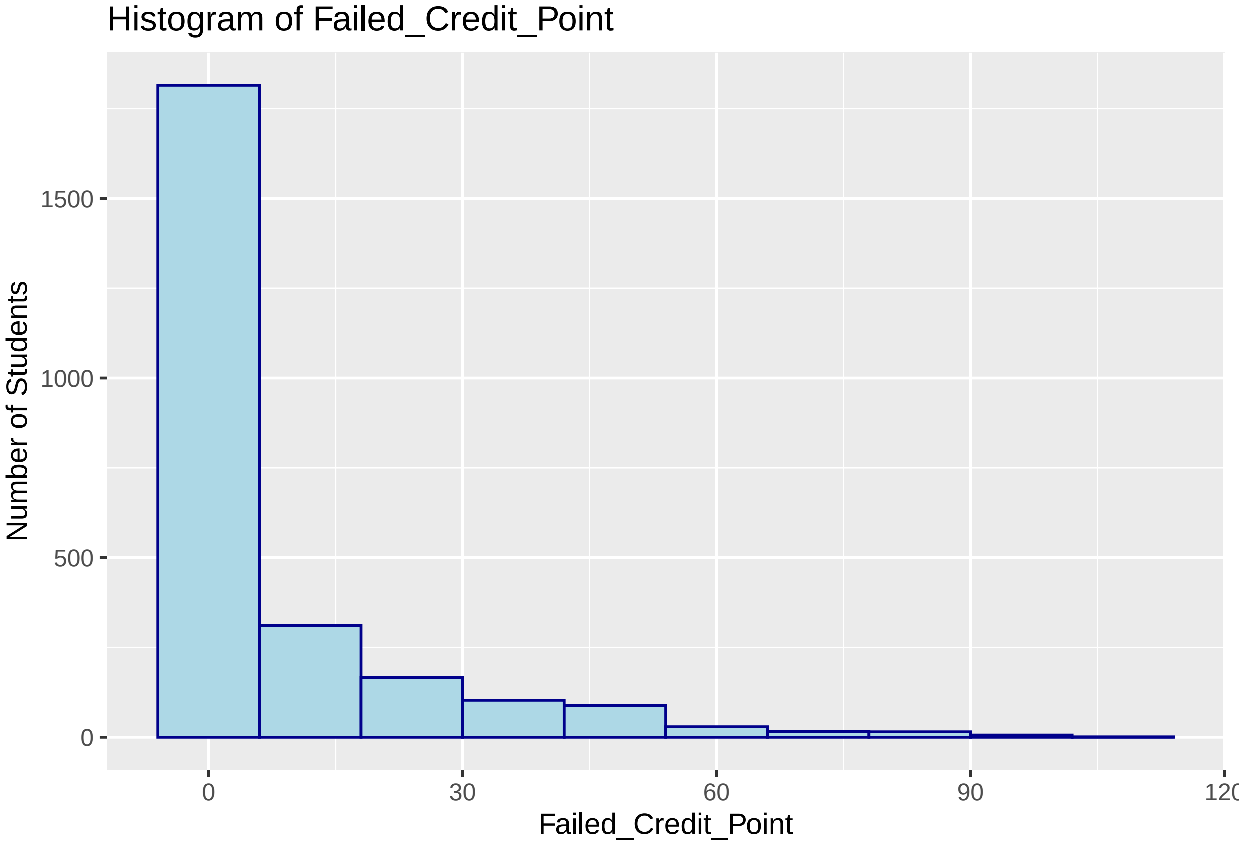


```
## Warning: Removed 46 rows containing non-finite values (stat_bin).

## Warning: Removed 2 rows containing missing values (geom_bar).
```


Density Graph of Achieved_Credit_Points
Frequency
1000
750
500
250
0
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





The function operated above generates one density graphs and four histogram illustrating the distribution of each numerical variables in the data frame. The summaries of each chart are listed below.

1. The distribution of the age of the students

Not surprisingly, most of the students are around 17 and 19 years old. They normally enrolled in the university after graduated from high school. However, also some students are over 20 years old. It could be some students enrolled in the university after finishing a lower education, such as a diploma or certificate IV.

2. The distribution of achieved credit points among students

The histogram does not show any interesting fact to be noted. Most of the students in the record are in second year of their study.

3. The GPA distribution among students

Most of the student average around a GPA of 4 to 6. It can be summarised that there are approximately 80% of student with a GPA higher than 3.5, while the remainders are with a lower GPA less than 3.5.

4. The OP Score distribution among students

Clearly, most of the students get OP score around 5 to 10. The diagram occurs right-skewed.

5.The distribution of failed credit points among students

Not surprisingly, most of the students are highly possible that never fail any unit (0 point) or one single unit (12 point). It makes the diagrams tend to be frequent on the left-hand side.

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

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

summary(male_data$GPA)
```

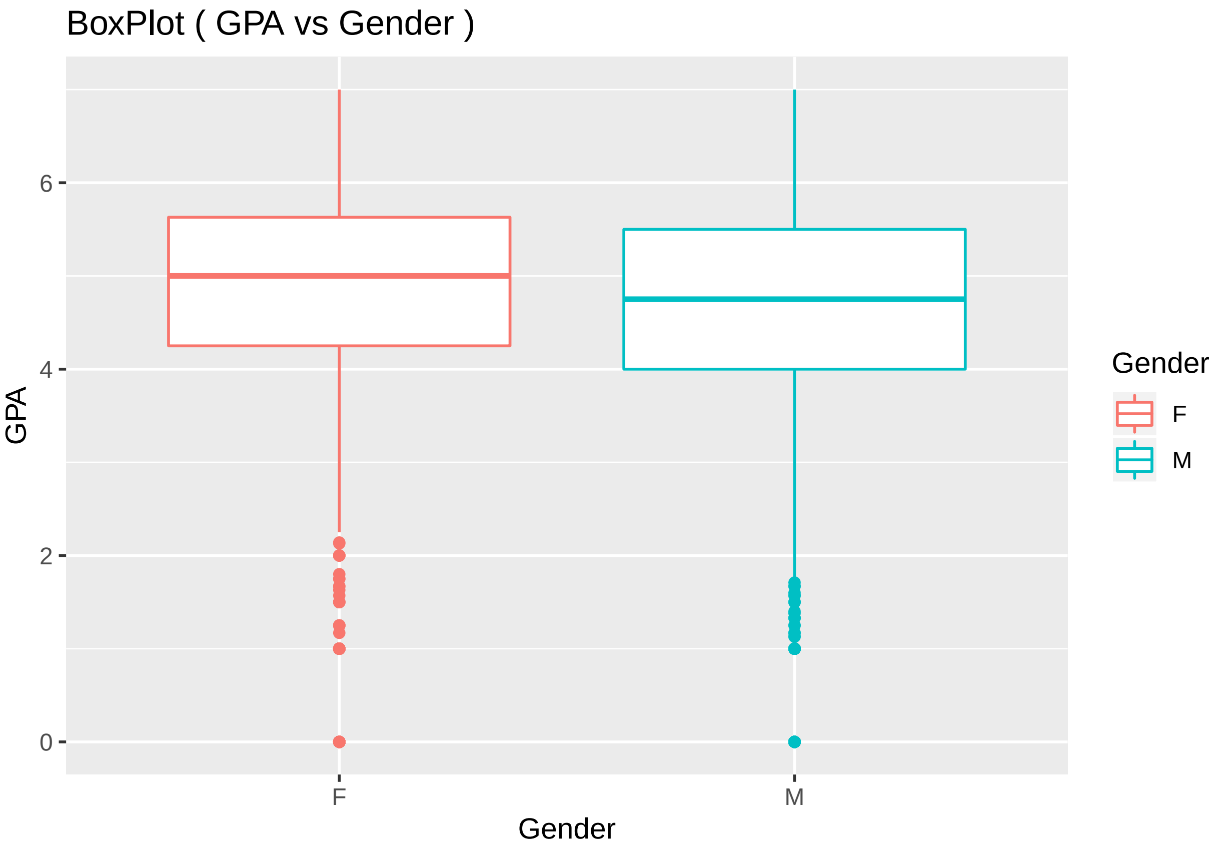
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   4.000   4.750   4.472   5.500   7.000
```

Summary GPA for Female

```r
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
```

```r
# 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)
```

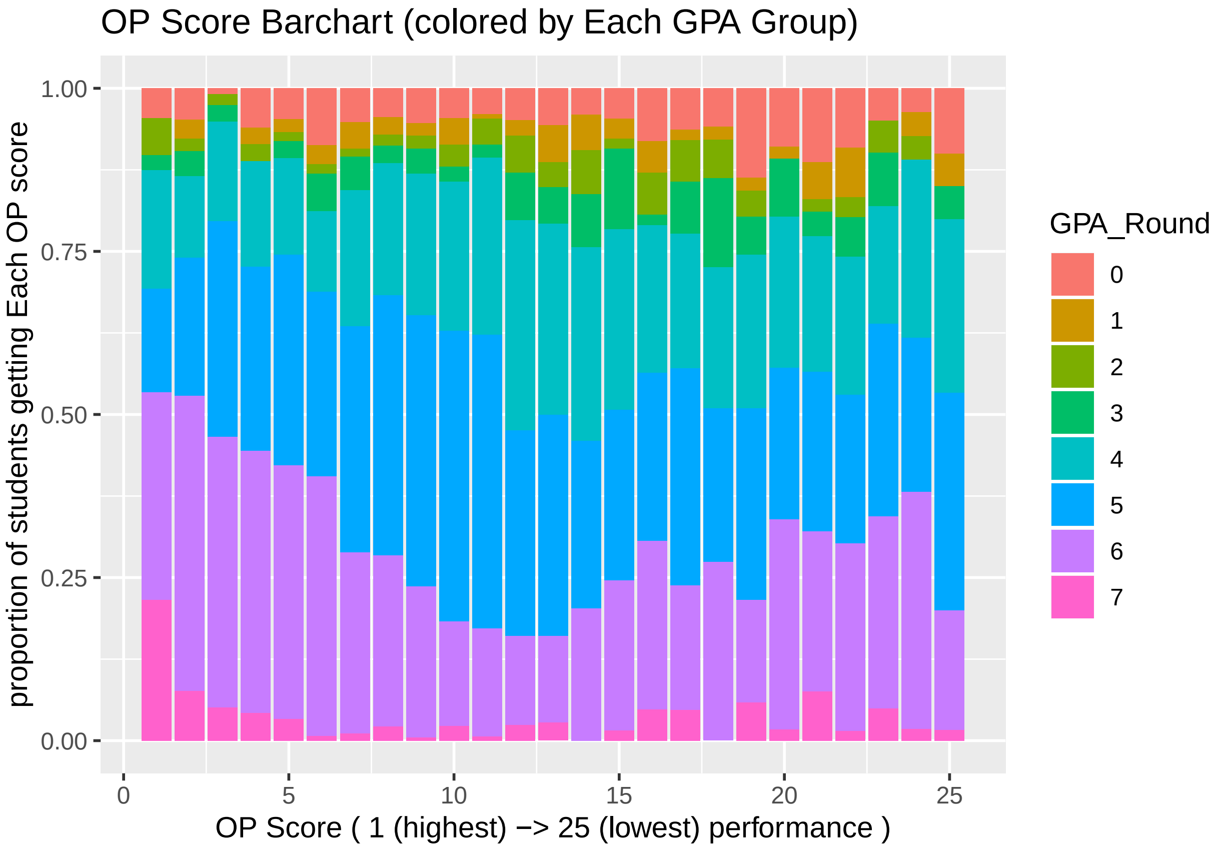
```
##
##  Welch Two Sample t-test
##
## data:  uniData$GPA by uniData$Gender
## t = 2.4454, df = 2539.7, p-value = 0.01453
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.03111718 0.28297210
## sample estimates:
## mean in group F mean in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

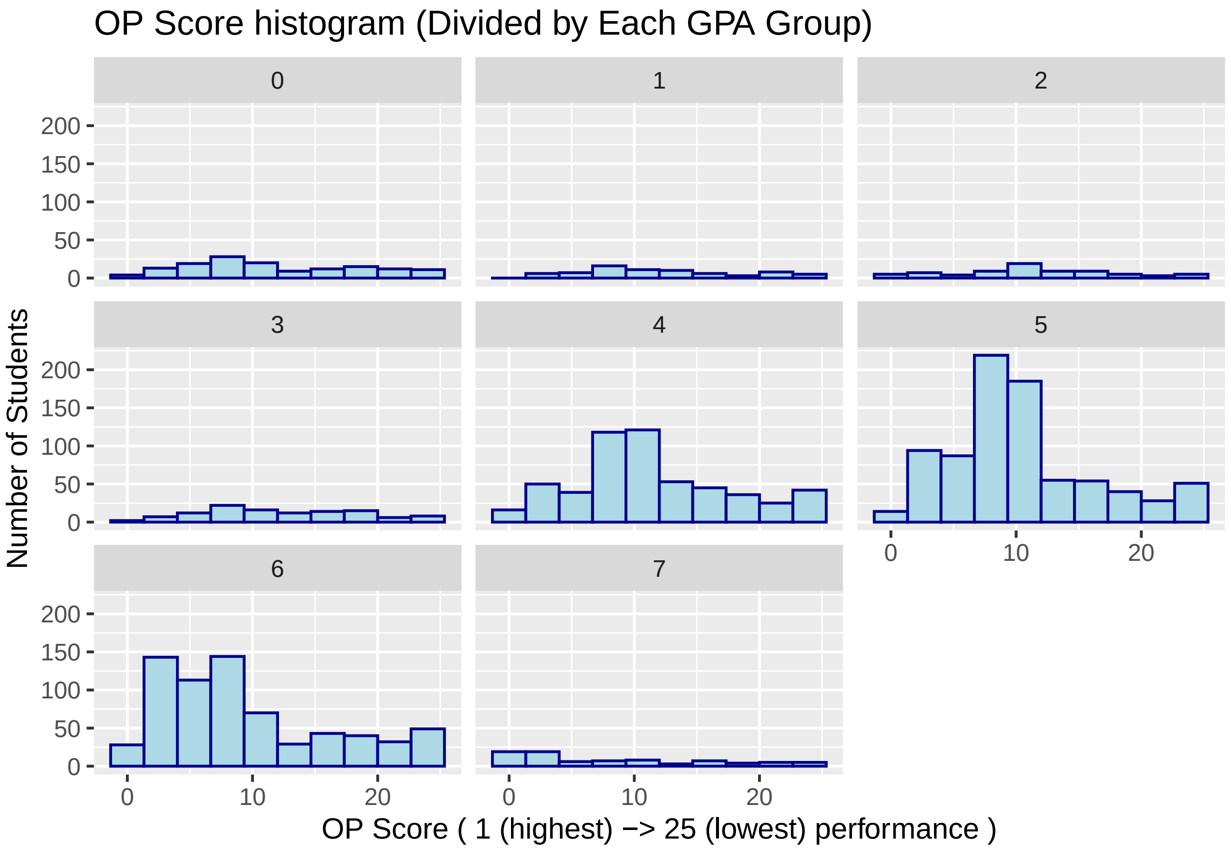
```
# Variance
var.test(uniData$GPA ~ uniData$Gender)
```

```
##
##  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)
```





Bar chart (OP Score VS GPA)

The first bar chart displayed the relationship between OP score and GPA. Each bar indicates every student achieves in the OP exam, while each bar is filled with 8 different colours which indicate how these students perform 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, tend to perform better in the university. Approximately 50% of students, who get 1 OP score, archived above GPA 6 when they are studying in university. In contrast, about 40% of students, who get 25 OP score, archived 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

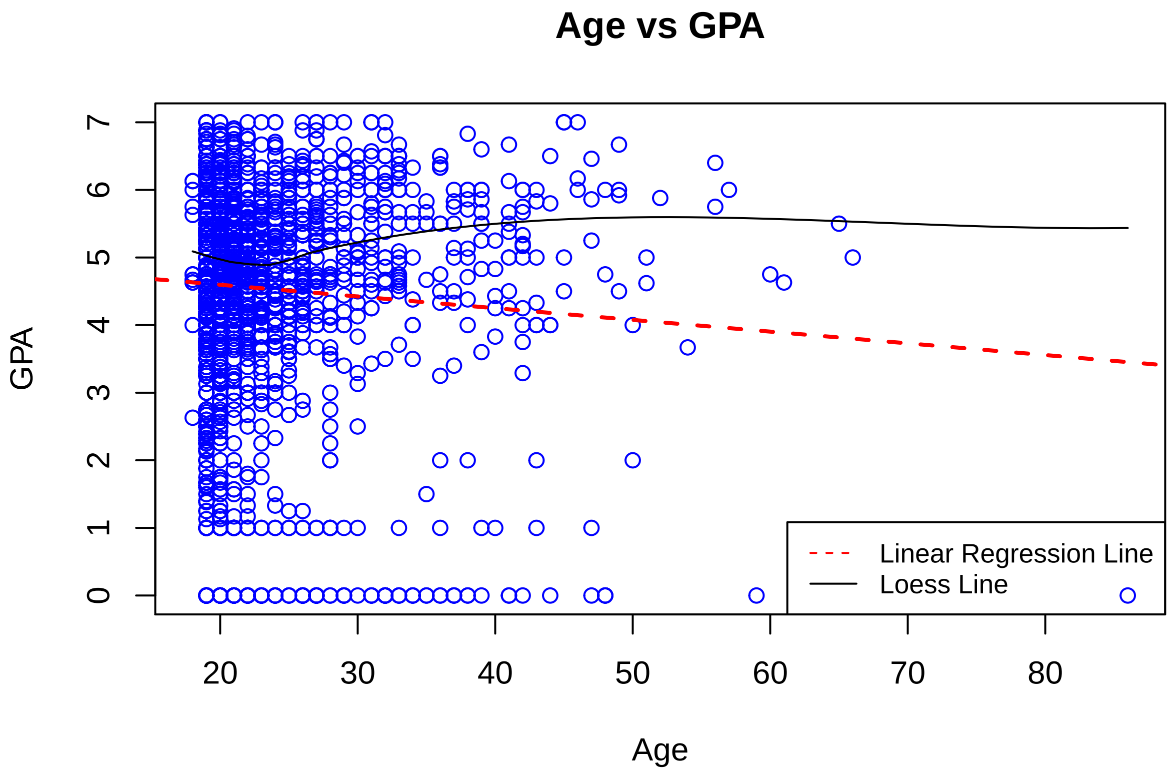
Develop a linear regression model of GPA using the given data. You need to describe your choice of predictors, examine your model's assumptions, assess model fit, and interpret the final model's regression coefficients.

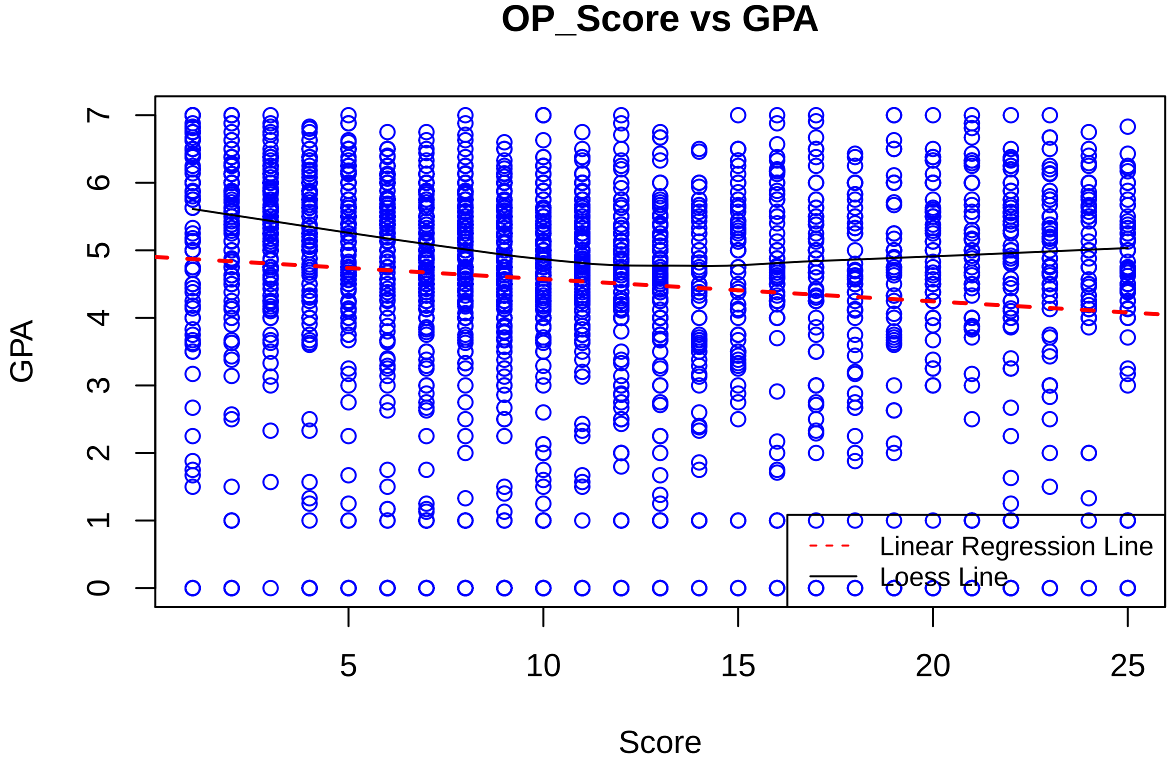
Analyse Each numerical data its relation related to GPA

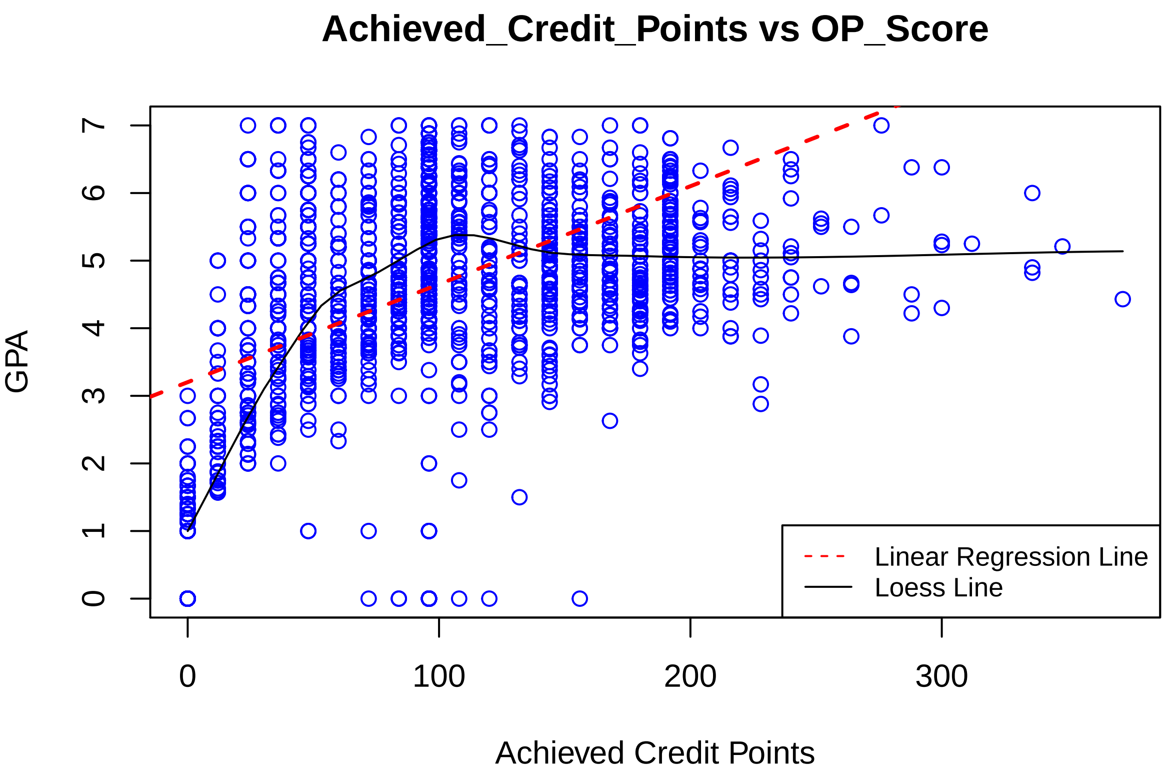
Correlation between each numerical data and GPA

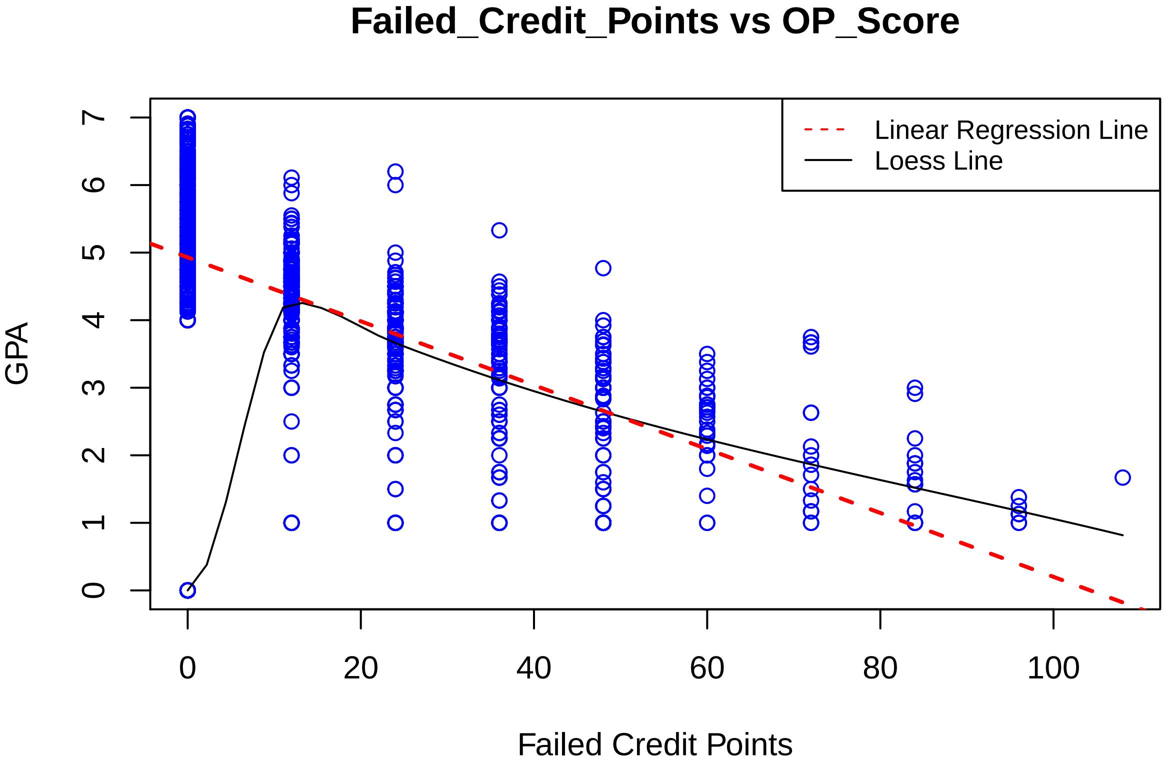
1. Age : -0.0641342
2. OP_Score : -0.129619
3. Achieved_Credit_Points : 0.4920035
4. Failed_Credit_Points : -0.473419

```
visualize_scatterplots_Vs_GPA(uniData)
```









The chosen predictors

It is evident that GPA will always be selected to be Y-axis which is classified as quantitative value. The predictor will need to be a strong data value that could have a significant impact on the analysis. From the correlation score and scatterplot we get above, Achieved Credit Point seems to be a biased and homoscedastic graph. It provides a good fit for the linear regression model as it has a positive linear relationship and a positive correlation value. It also achieves the highest correlation score which indicates that it has the strongest relationship with GPA comparing to the other three. Therefore, Achieved Credit Point is selected to train and test the simple linear regression model.

Spiting dataframe into training set and test set

```
# Data Preprocessing Library
library(caTools)
# Set Random seed
set.seed(2)
# Splitting Training and test dataset
split <- sample.split(uniData, SplitRatio = 0.7)
train <- subset(uniData, split==TRUE)
test <- subset(uniData, split==FALSE)
```

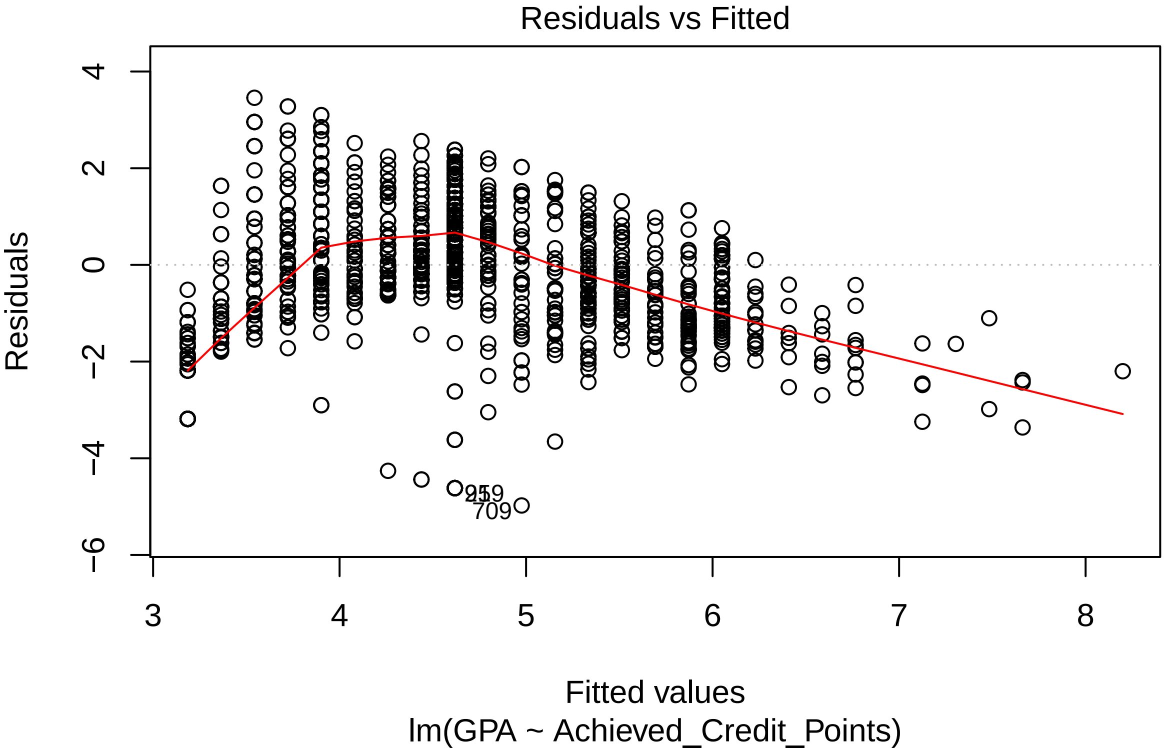
Training Linear Regression Model & Review diagnostic measures.

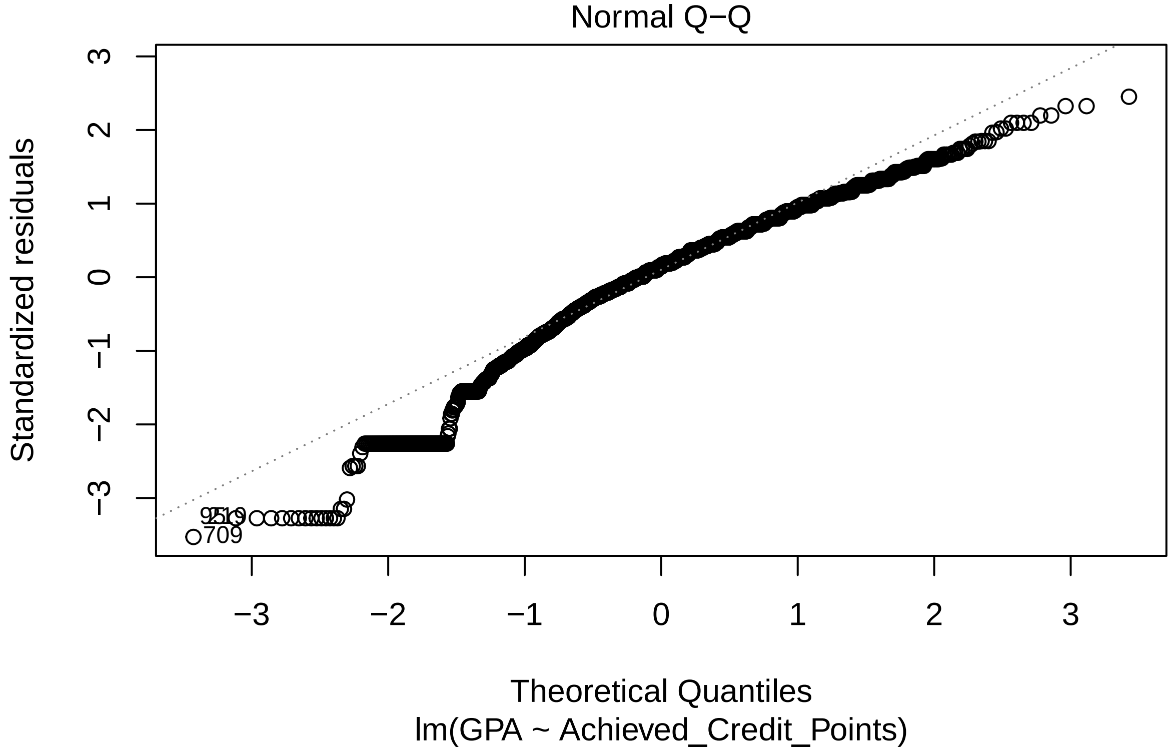
```
linear_model <- lm(GPA ~ Achieved_Credit_Points, data=train)
summary(linear_model)
```

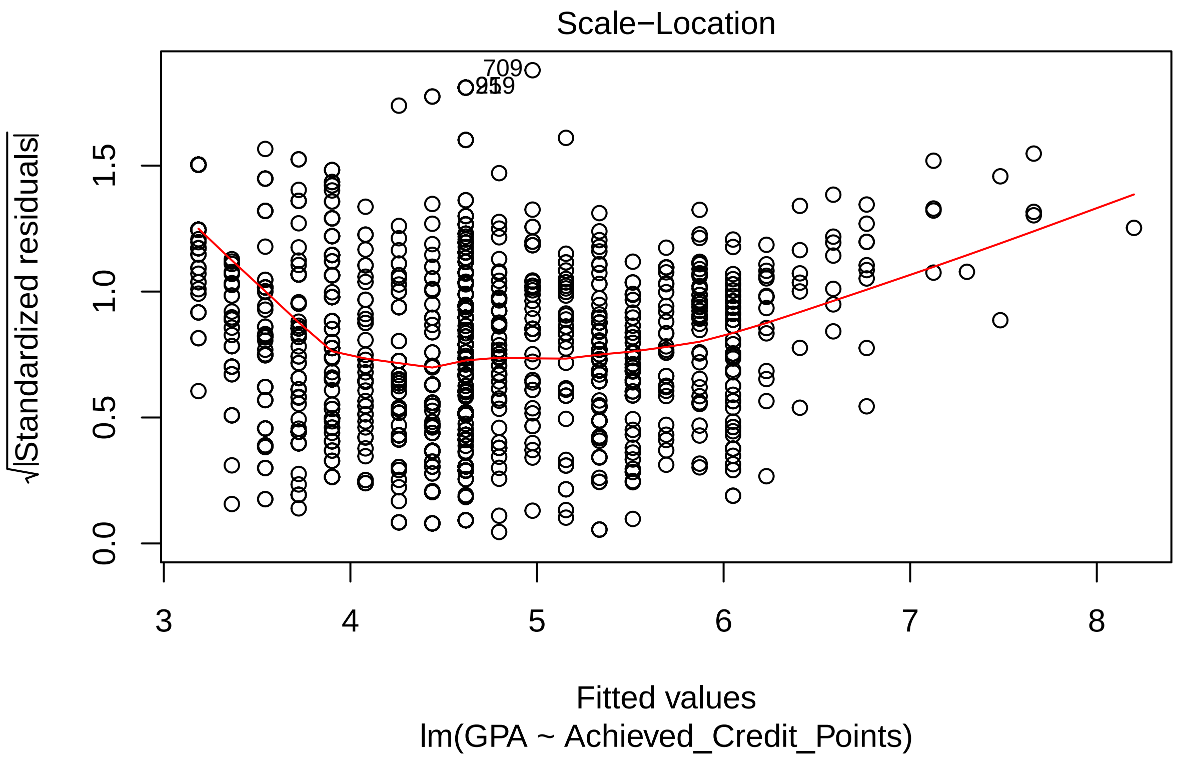
```
##
## Call:
## lm(formula = GPA ~ Achieved_Credit_Points, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9761 -0.7239  0.2029  1.0120  3.4564
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3.1854527  0.0688092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

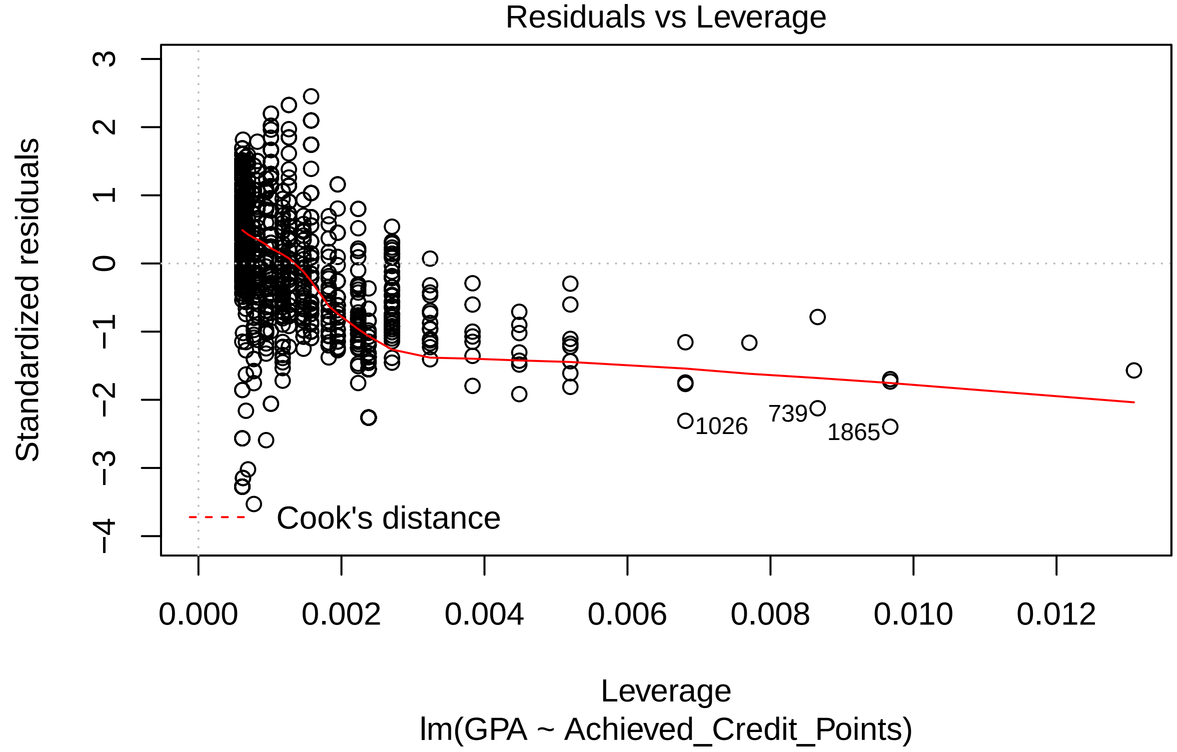
Plot the Linear Regression Prediction Line

```
plot(linear_model)
```


Residuals vs Fitted
Residuals
4
2
0
−2
−4
−6
3
4
5
6
7
8
919
251
709
Fitted values
lm(GPA ~ Achieved_Credit_Points)


Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−3
919
251
709
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(GPA ~ Achieved_Credit_Points)





Task 5 Logistic Regression

```
# Logistic Regression Library
library(DAAG)
source("regression_helper.R")
```

Bivariate exploration

```r
xtabs(~ uniData$Attrition + uniData$Socio_Economic_Status)
```

```
##                    uniData$Socio_Economic_Status
## uniData$Attrition High  Low Medium
##      Not Retained  135   98    215
##      Retained      636  365   1101
```

A simple model with only one predictor

```r
simple_log_model <- glm(Attrition ~ Socio_Economic_Status, data=uniData, family = "binomial")
summary(simple_log_model)
```

```
##
## Call:
## glm(formula = Attrition ~ Socio_Economic_Status, family = "binomial",
##     data = uniData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9035   0.5973   0.5973   0.6205   0.6897
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  1.54992    0.09476  16.356   <2e-16 ***
## Socio_Economic_StatusLow    -0.23499    0.14807  -1.587    0.112
## Socio_Economic_StatusMedium  0.08341    0.12058   0.692    0.489
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2370.4  on 2549  degrees of freedom
## Residual deviance: 2365.1  on 2547  degrees of freedom
## AIC: 2371.1
##
## Number of Fisher Scoring iterations: 4
```

simple model P-value & Pseudo R^2

```r
print_R2_and_pvalue(simple_log_model$null.deviance, simple_log_model$deviance)
```

```
## [1] "R^2 :  0.00225402720346674"
## [1] "P-value :  0.0208056460092459"
```

Summarise the predicted probailities

```r
simple.predicted.data <- data.frame(
  probability.of.Attrition = simple_log_model$fitted.values,
  Socio_Economic_Status = uniData$Socio_Economic_Status
)

xtabs(~ probability.of.Attrition + Socio_Economic_Status ,data=simple.predicted.data)
```

```
##                         Socio_Economic_Status
## probability.of.Attrition High  Low Medium
##        0.788336933045247    0  463      0
##        0.824902723735415  771    0      0
##        0.836626139817635    0    0   1316
```

Logistic Regression model with all predictors

```r
# Logistic Regression with all predictors
log_model <- glm(Attrition ~ ., data=uniData, family = "binomial")
summary(log_model)
```

```
## 
## Call:
## glm(formula = Attrition ~ ., family = "binomial", data = uniData)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -3.5129   0.2020   0.4329   0.5718   1.8857  
## 
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)    
## (Intercept)                     0.961335   0.450604   2.133  0.03289 *  
## Degree_TypeSingle              -0.741475   0.309181  -2.398  0.01648 *  
## Achieved_Credit_Points          0.016922   0.001841   9.192  < 2e-16 ***
## Attendance_TypePart Time        1.425308   0.262767   5.424 5.82e-08 ***
## Age                            -0.026512   0.009730  -2.725  0.00643 ** 
## Failed_Credit_Points           -0.017050   0.003407  -5.005 5.60e-07 ***
## International_studentYes        0.775607   0.247844   3.129  0.00175 ** 
## First_in_familyYes              0.017776   0.121154   0.147  0.88335    
## GenderM                         0.105876   0.125705   0.842  0.39965    
## GPA                             0.075554   0.043064   1.754  0.07935 .  
## OP_Score                       -0.007203   0.009468  -0.761  0.44683    
## Socio_Economic_StatusLow       -0.145242   0.169840  -0.855  0.39246    
## Socio_Economic_StatusMedium     0.049286   0.135809   0.363  0.71667    
## Teaching._Period_AdmittedSEM-2  0.390834   0.171037   2.285  0.02231 *  
## FacultyFaculty of Education     0.596908   0.282377   2.114  0.03453 *  
## FacultyFaculty of Health        0.228437   0.175987   1.298  0.19428    
## FacultyFaculty of Law          -0.256680   0.230529  -1.113  0.26552    
## FacultyQUT Business School      0.524731   0.215330   2.437  0.01482 *  
## FacultySci and Eng Faculty      0.415301   0.184758   2.248  0.02459 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 2370.4  on 2549  degrees of freedom
## Residual deviance: 1914.3  on 2531  degrees of freedom
## AIC: 1952.3
## 
## Number of Fisher Scoring iterations: 6
```

Logistic Regression model with all predictors P-value & Pseudo R^2

```r
print_R2_and_pvalue(log_model$null.deviance, log_model$deviance)
```

```
## [1] "R^2 :  0.192423133796394"
## [1] "P-value :  0"
```

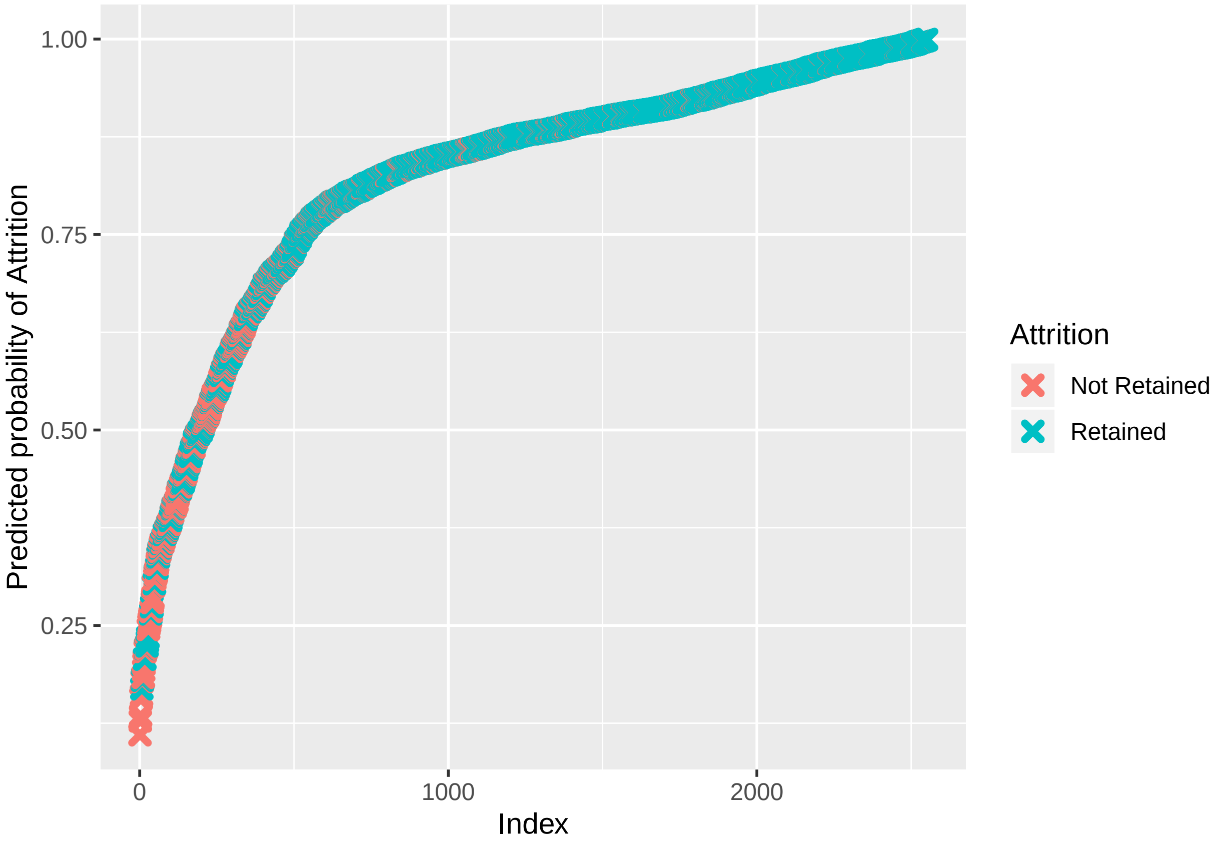
Multicollinearity using VIF

```r
vif(log_model)
```

```
##              Degree_TypeSingle         Achieved_Credit_Points
##                         1.0421                         1.9565
##       Attendance_TypePart Time                            Age
##                         1.0777                         1.1650
##           Failed_Credit_Points       International_studentYes
##                         1.2924                         1.0595
##             First_in_familyYes                        GenderM
##                         1.0537                         1.1668
##                            GPA                       OP_Score
##                         1.9817                         1.0562
##       Socio_Economic_StatusLow    Socio_Economic_StatusMedium
##                         1.3670                         1.3618
## Teaching._Period_AdmittedSEM-2    FacultyFaculty of Education
##                         1.1346                         1.2050
##       FacultyFaculty of Health          FacultyFaculty of Law
##                         1.7609                         1.4829
##     FacultyQUT Business School     FacultySci and Eng Faculty
##                         1.4804                         1.8834
```

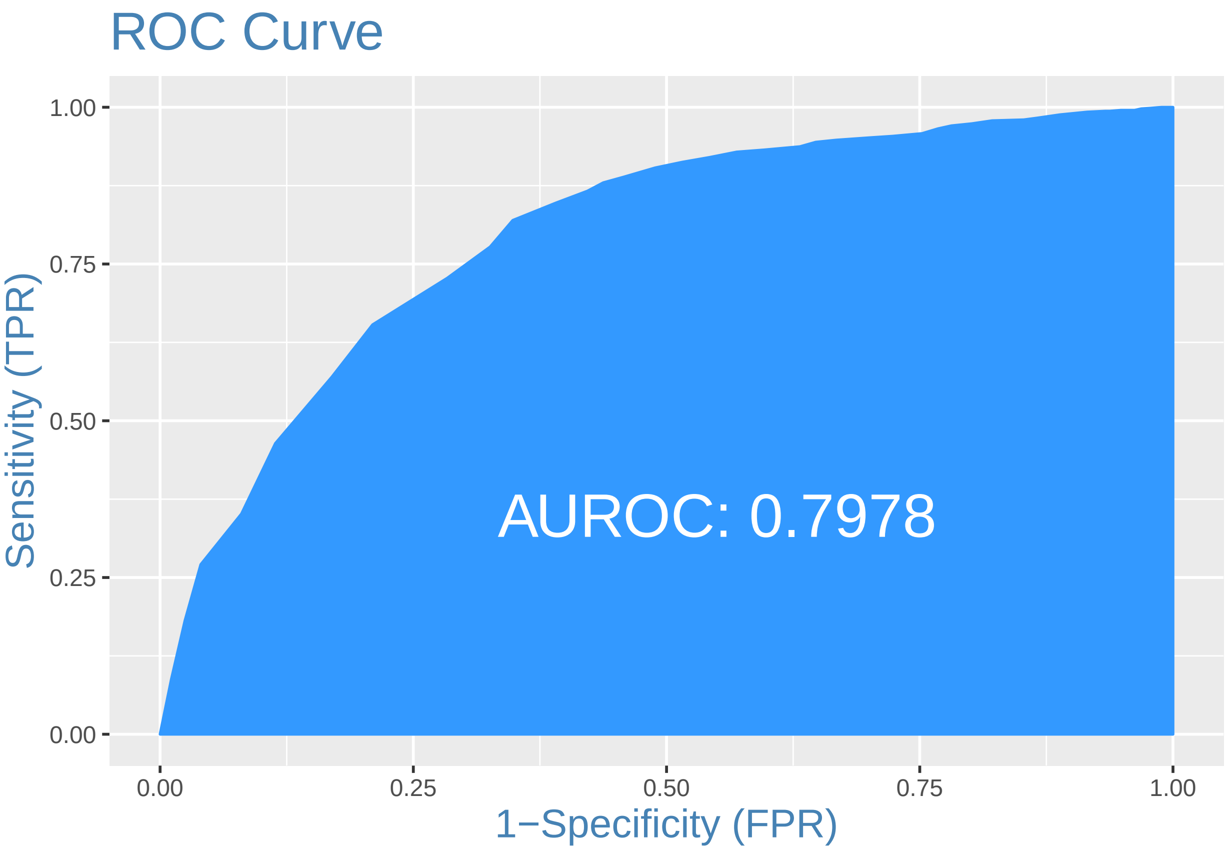
Plot Predicted Probabilities

```r
predicted.data <- data.frame(
  probability.of.Attrition = log_model$fitted.values,
  Attrition = uniData$Attrition
)

# Sort predicted data by Probabilities
predicted.data <- predicted.data[order(predicted.data$probability.of.Attrition, decreasing = FALSE),]
predicted.data$rank <- 1:nrow(predicted.data)

library(ggplot2)
library(cowplot)

ggplot(data=predicted.data, aes(x=rank, y=probability.of.Attrition) ) +
  geom_point(aes(color=Attrition), alpha = 1, shape = 4, stroke = 2) +
  xlab("Index") +
  ylab("Predicted probability of Attrition")
```



ROC Curve

```
#
predicted.data$actuals <- factor(predicted.data$Attrition, labels = c(0,1))
# Shows ROC Curve
visualize_ROC_Curve(predicted.data$actuals, predicted.data$probability.of.Attrition)
```



MisClassification Error, Sensitivity, Specificity

```
print_MCE_Sens_Spec( predicted.data$actuals, predicted.data$probability.of.Attrition)
```

```
## [1] "Optimal Cut off : 0.569423153772749"
## [1] "MisClassification Error :  0.158"
## [1] "sensitivity:  0.9476688867745"
## [1] "specificity :  0.345982142857143"
```

```
print_ConfusionMatrix( predicted.data$actuals, predicted.data$probability.of.Attrition)
```

```
##     0    1
## 0 155  110
## 1 293 1992
```