**Student Academic Dropout Prediction**

**GROUP-2**

SWATI SANJEEV DESHPANDE (SXD210145)

VAISHNAVI GAWALI (VXG220023)

MODASIYA, MOHMADSAFIN MOHMADYUSUF (MXM210198)

RATAN VIKRAM(VXR220042)

BHUMIREDDY, DINAKAR REDDY(DXB220049)

****

**MIS 6356.001 BUSINESS ANALYTICS WITH R SPRING 2023**

**PROF. JIANQING CHEN**

**Table of Contents**

1. [**Executive Summary**](#BK1) 4
2. [**Objective**](#BK2)4
3. [**Data**](#BK3)4
4. [**Data Description**](#BK4) 5
5. [**Data Analysis**](#BK5)8
6. [**Data Preprocessing and Evaluating plots**](#BK6)9

6.a Importing the Data9

6.b Names in Data9

6.c Class of Variables10

6.d Missing Values10

6.e Summary Statistics10

6.f Plots to understand distribution11

6.g Plotting Distribution of various categories for key variables13

1. [**Correlation Matrix**](#BK7) 17
2. [**BI models**](#BK8)18
3. [**Model Evaluation**](#BK9)22
4. [**Results**](#BK10)23
5. [**Conclusion**](#BK11)23
6. [**References**](#BK11) 23

**List of Figures:**

1. [Values for attribute Marital Status](#BK1)
2. [Values for attribute Gender](#F2)
3. [Values for attribute Nationality](#F3)
4. [Values for attribute Application mode](#F4)
5. [Values for attribute Course](#F5)
6. [Values for attribute Previous Qualification](#F6)
7. [Values for attribute qualification of parents](#F7)
8. [Values for attribute attendance](#F8)
9. [Values for Miscellaneous attributes](#F9)
10. [Values for attribute occupation of parents](#F10)
11. [Conversion of numerical categories to nominal categories](#F11)
12. [Names of columns in dataset](#F12)
13. [Class of variables](#F13)
14. [Summary Statistics](#F14)
15. [Histogram and boxplot for variable ‘Age at Enrollment’](#F15)
16. [Histogram and boxplot for variable ‘GDP’](#F16)
17. [Histogram and boxplot for variable ‘Inflation Rate’](#F17)
18. [Summary of target variable](#F18)
19. [Distribution of Gender](#F19)
20. [Gender wise distribution of target variable](#F20)
21. [Distribution based on Course Category](#F21)
22. [Distribution of attendance](#F22)
23. [Age wise distribution](#F23)
24. [Correlation Matrix](#F24)
25. [Summary of Logistic Regression](#F25)
26. [Confusion Matrix and Statistics for Training Data](#F26)
27. [Confusion Matrix and Statistics for Validation Data](#F27)
28. [ROC Curve](#F28)

**1.Executive Summary:**

Higher education in technical institutes across the Globe is key to better socio-economic status, better job opportunities and overall progress in any country. Countries where individuals have completed higher education seem to have better stability, low crime rates and equality. Hence, we want to understand how and why students drop out from various academic programs and help academic institutions to retain and encourage students to perform better in their studies.

The dataset provides information on numerous demographics, social-economic, and academic performance variables related to students enrolled in higher education. The dataset offers insightful information about the factors that influence student achievement and may be used to direct interventions and regulations affecting student retention.

**2.Objectives:**

Our main objective in this project is to understand factors which are linked to success and dropout of students and how these factors contribute to their academic performance. We will be studying the historical data and identity factors that have caused students to drop out in the past. After studying these factors, we can predict which students are likely to drop out. It is also necessary to identify interdependencies and interactions between variables affecting students’ overall performance, areas of improvements, so that timely suggestions can be given to educational institutions to implement precautionary measures required to reduce attrition rate of students. Further, these educational institutes can take initiatives in certain groups of students based on parent occupation, debt, GDP, inflation rate etc.

**3.Data:**

We are using a second-hand dataset from Kaggle. Data consists of 4424 records of 35 variables such as debt, gender, unemployment rate and so on. We anticipate that the data might be skewed as most of the students typically enroll for university between the age of 20-30. It also covers a wide variety of academic disciplines, parents’ education, and occupation.

<https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention>

**4.Data Description:**

This dataset contains demographic data, socioeconomic and macroeconomic data, data at the time of student enrollment, and data at the end of the first and second semesters. These include data from 17 undergraduate degrees from different fields of knowledge, such as agronomy, design, education, nursing, journalism, management, social service, and technologies.

The dataset contains 35 variables which are coded as numeric. For example, Marital status is coded as numbers from 1 to 6, each with different categories as below:

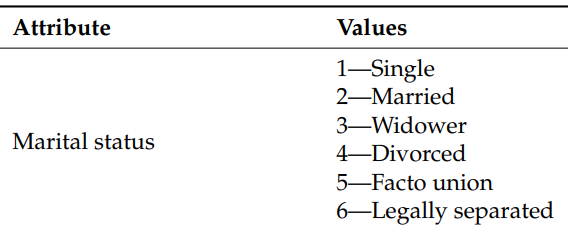


Figure 1: Values for attribute Marital Status

Similarly, gender variable is coded as 1 and 2, with categories as below:

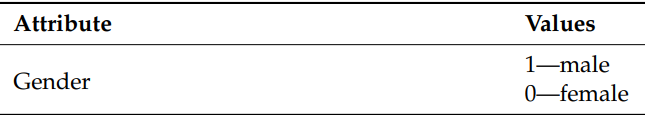


Figure 2: Values for attribute Gender

The main outcome variable is named “Target” in the dataset with categories “Dropout”, “Enrolled” and “Graduate”.

The Description of all the other coded categories is given below:

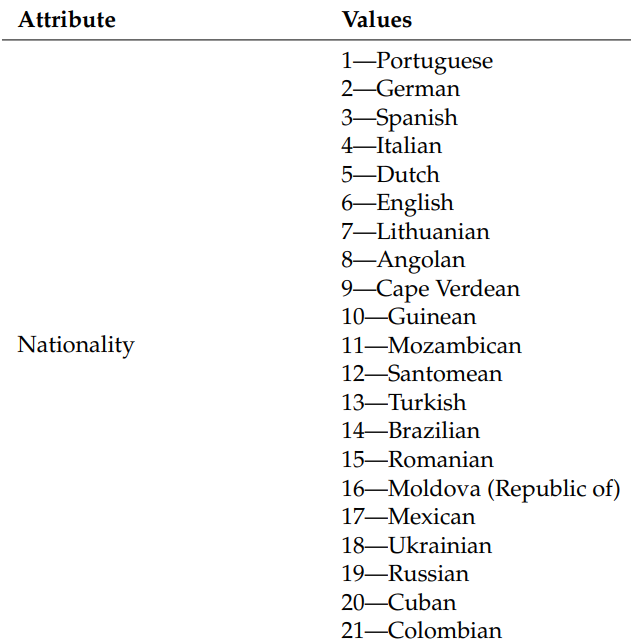
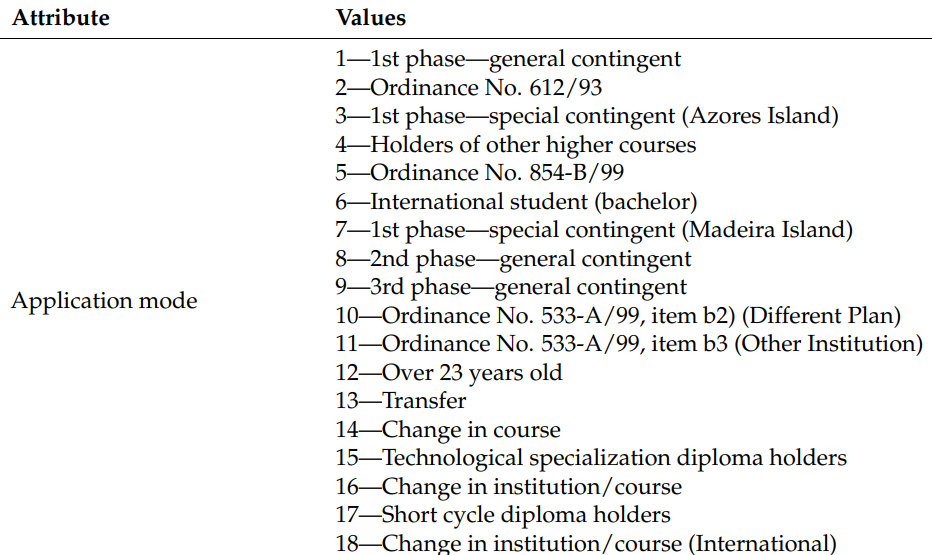


Figure 4: Values for attribute Application mode

Figure 3: Values for attribute Nationality

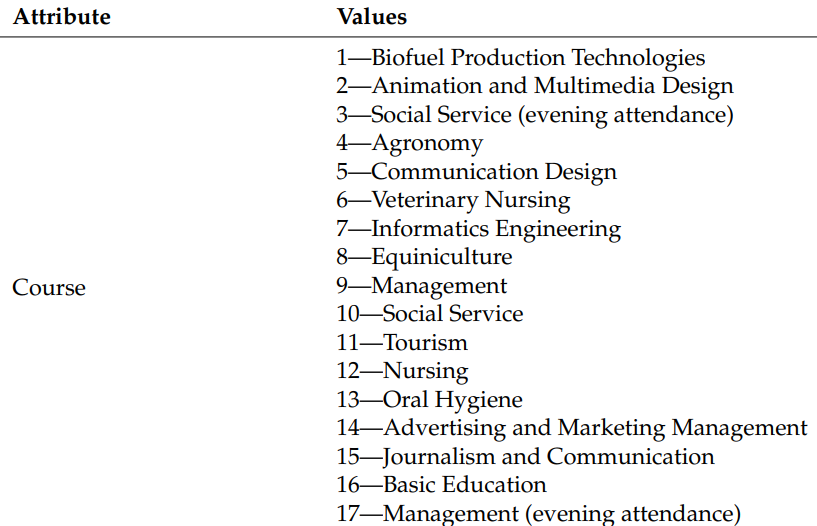


Figure 5: Values for attribute Course

Text

Description automatically generated

Text

Description automatically generated

Figure 6: Values for attribute Previous Qualification

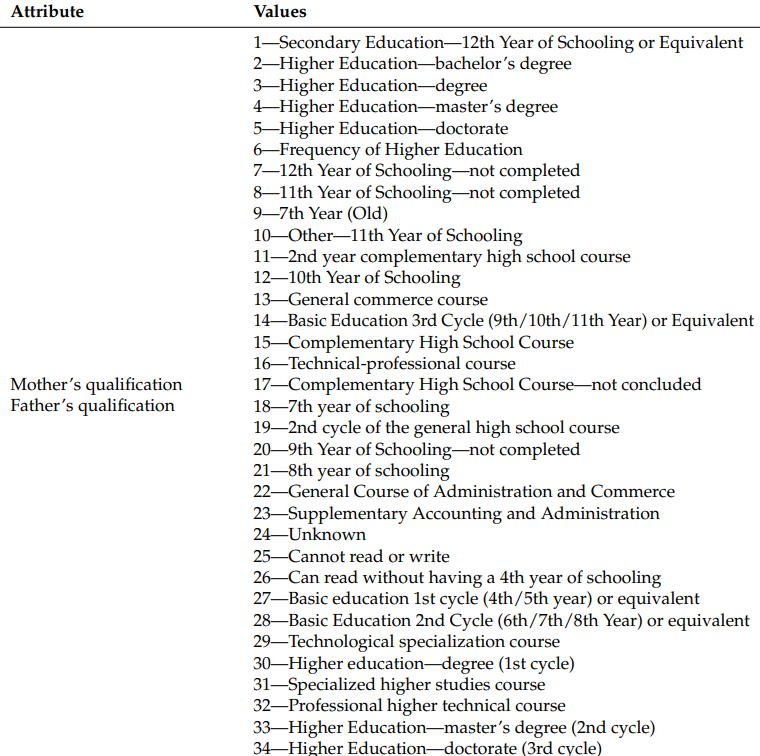


Figure 7: Values for attribute qualification of parents

Shape

Description automatically generated with medium confidence

Figure 8: Values for attribute attendance

Text

Description automatically generated

Figure 9: Values for Miscellaneous attributes

A picture containing text

Description automatically generated

Text

Description automatically generated

Figure 10: Values for attribute occupation of parents

**5.Data Analysis:**

We analyze the complete data set to find if there are any missing values, irregularities, unusuality, null values in the data. We can see that the data is skewed as most of the students enroll in the age group 20-30, so we have more records with age 20-30.

**6.Data Pre-processing and evaluating plots:**

Since most of the variables are coded as numbers representing categories, the first step in data cleaning is converting the numerical categories into nominal categories. We converted the categories in excel to make the data easily readable and meaningful in RStudio. The example of updated data is below:

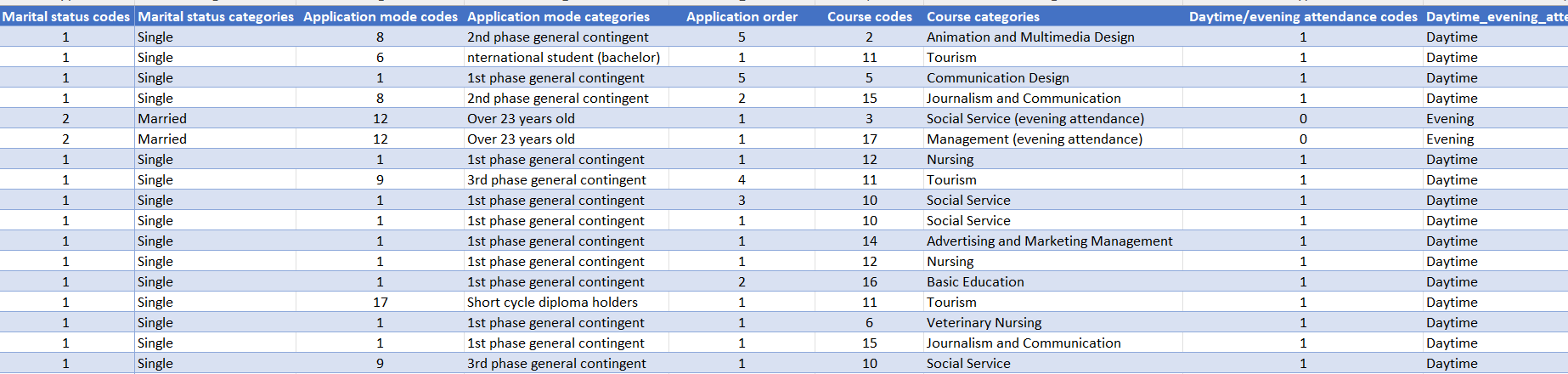


Figure 11: Conversion of numerical categories to nominal categories

1. **Importing the data in R:**

*#loading dataset into R*

*library(readxl)*

*School<-read\_excel("Dataset\_schooling\_with\_categories\_and\_codes.xlsx")*

*View(School)*

1. **Checking names of columns in the dataset:**

*names(School)*

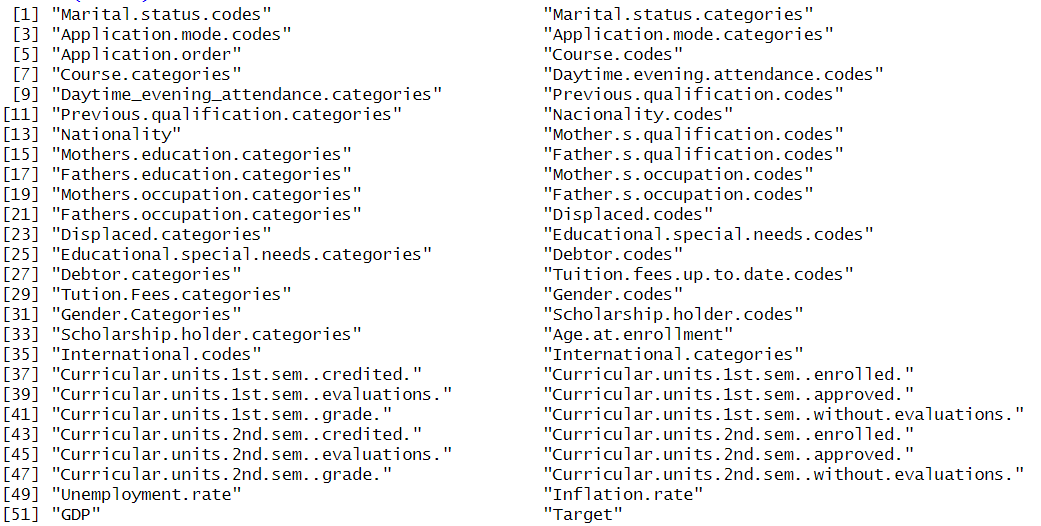
****

Figure 12: Names of columns in dataset

1. **Checking class of variables:**

*sapply(School, class)*

**Graphical user interface, text

Description automatically generated**

**Graphical user interface, text

Description automatically generated with medium confidence**

Figure 13: Class of variables

1. **Checking if there are any missing values in the dataset:**

*sum(is.na(School))*

**> sum(is.na(School))**

**[1] 0**

There are no missing values in the dataset.

1. **Running basic summary statistics:**

#remove non numeric columns

School\_1<- School[,-c(2,4,7,9,11,13,15,17,19,21,23,25,27,29,31,33,36,52)]

View(School\_1)

#getting basic statistics of the numerical variables

data.frame(mean=sapply(School\_1, mean),

sd= sapply(School\_1, sd),

median= sapply(School\_1, median),

max= sapply(School\_1, max),

min= sapply(School\_1, min))

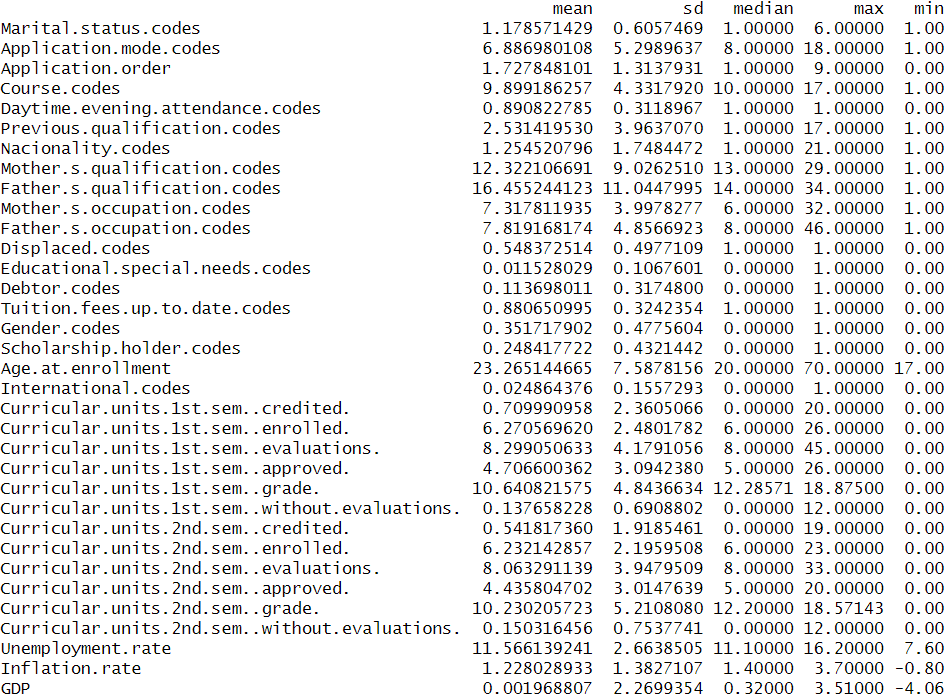


Figure 14: Summary Statistics

1. **Basic plots to understand distribution of important variables:**

Age at Enrollment:

*hist(School$Age.at.enrollment, main = "Age at Enrollment", xlab = "Age at Enrollment", ylab = "Frequency")*

*boxplot(School$Age.at.enrollment)*

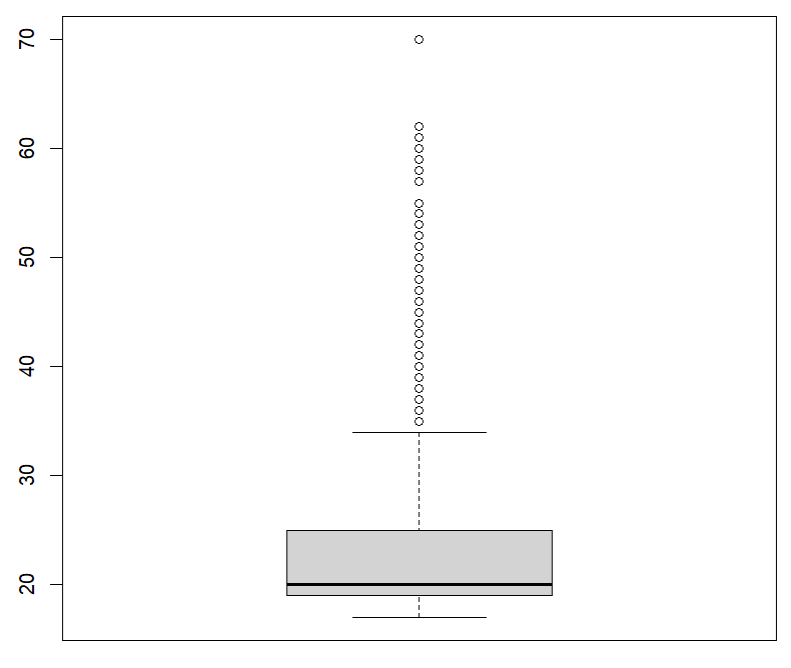
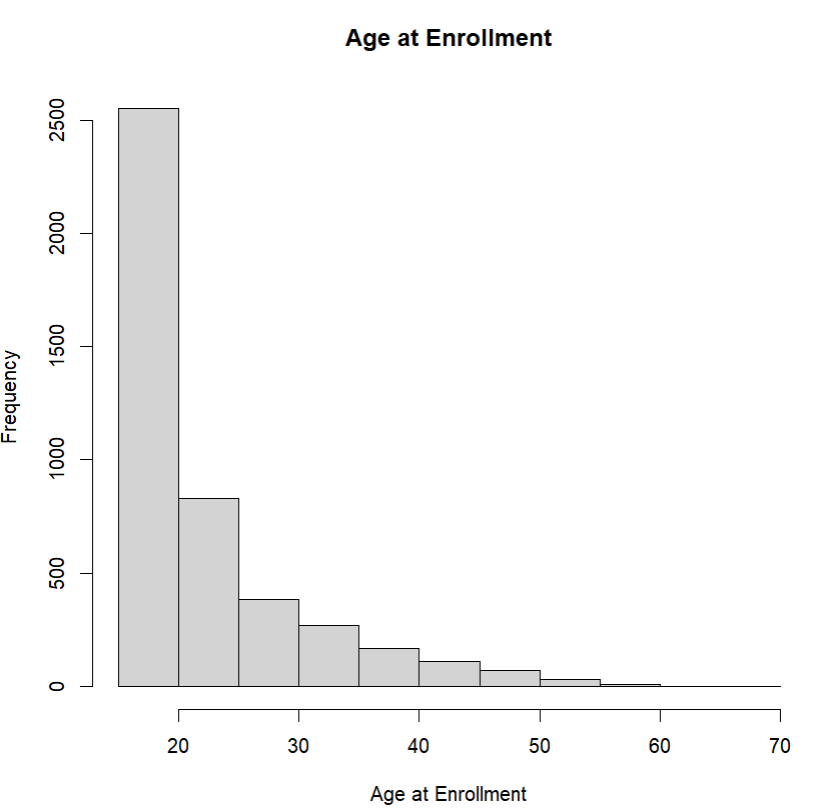


Figure 15: Histogram and boxplot for variable ‘Age at Enrollment’

The variable Age at Enrollment looks skewed to the right, with most students enrolled below the age of 20. This observation is supported by the following skewness measure.

> skewness(School$Age.at.enrollment)

[1] 2.054292

GDP:

*hist(School$GDP, main = "Histogram of GDP", xlab = "GDP", ylab = "Frequency", breaks = 10)*

*boxplot(School$GDP)*

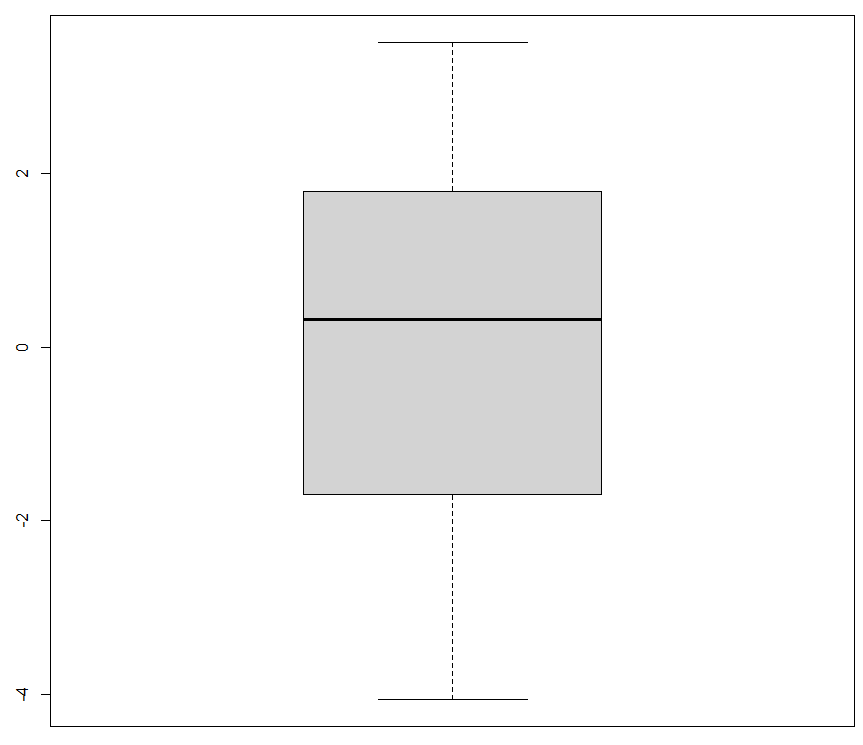
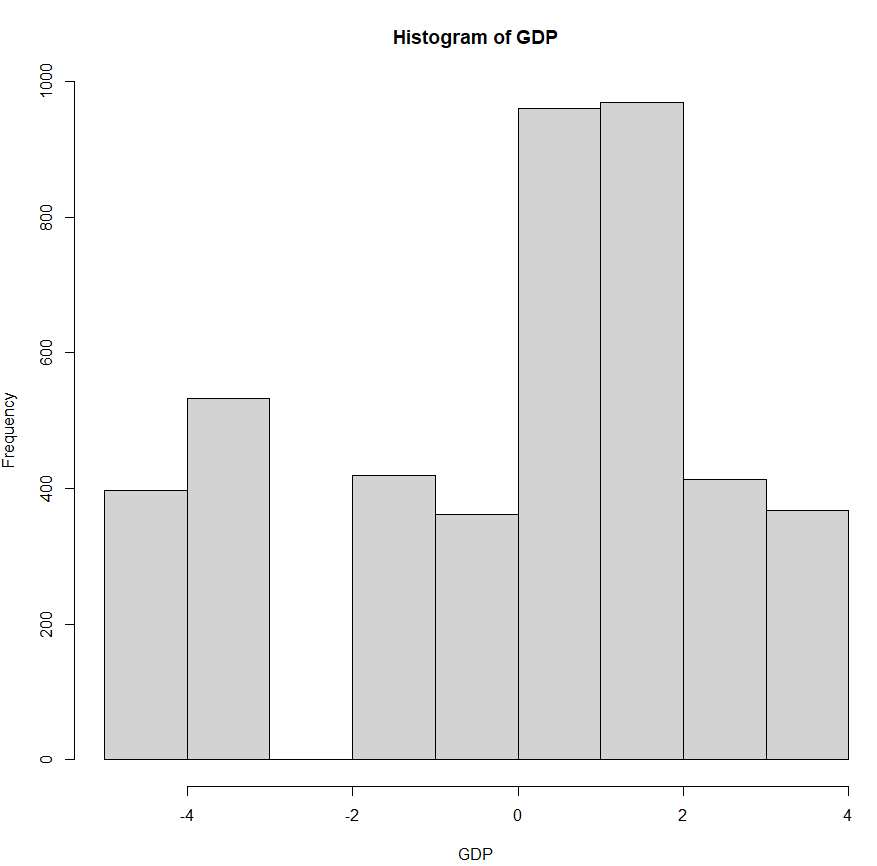


Figure 16: Histogram and boxplot for variable ‘GDP

> skewness(School$GDP)

[1] -0.3939346

Inflation Rate:

*hist(School$Inflation.rate, main = "Histogram of Inflation Rate", xlab = "Inflation Rate", ylab = "Frequency", breaks = 5)*

*boxplot(School$Inflation.rate)*

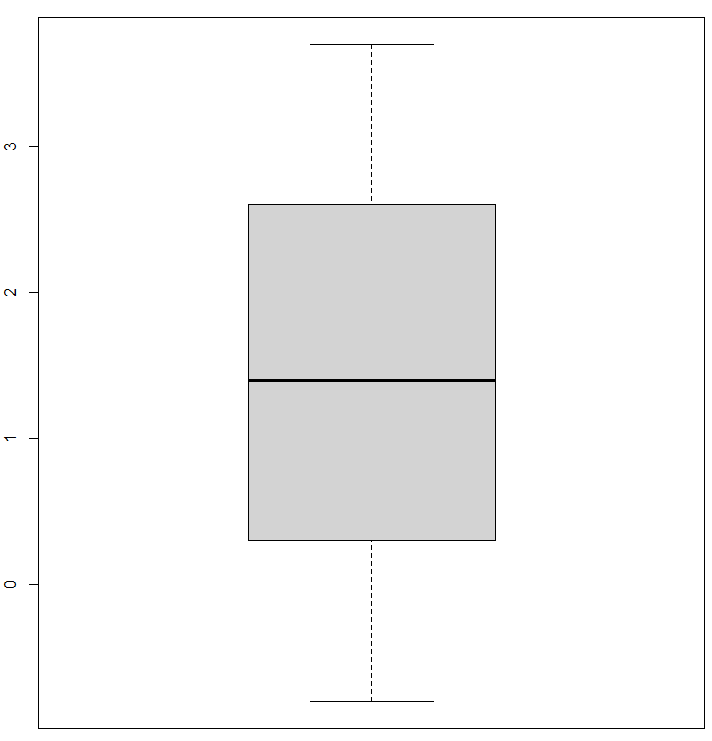
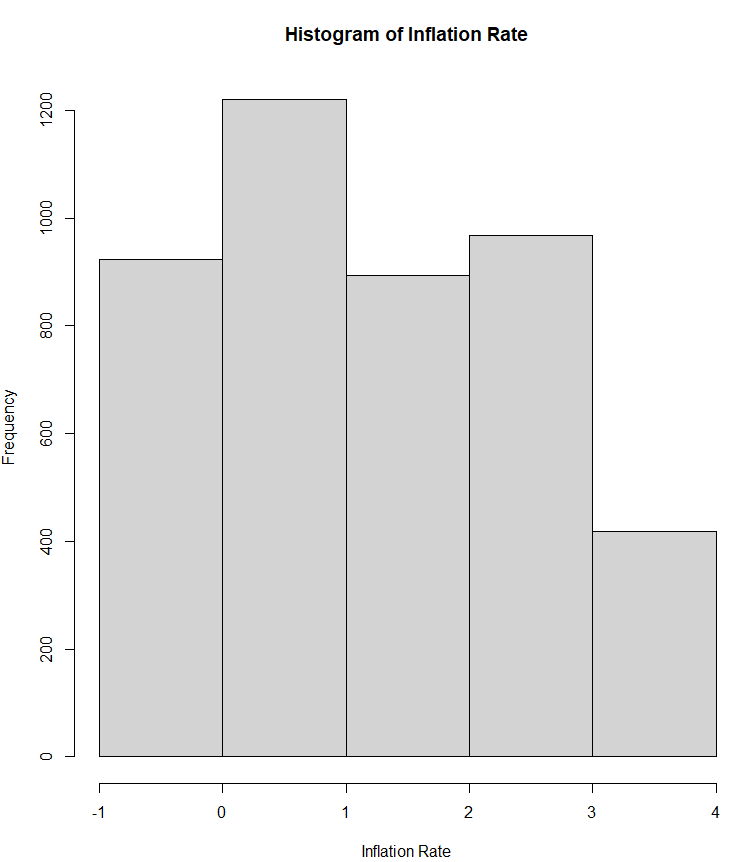


Figure 17: Histogram and boxplot for variable ‘Inflation Rate’

> skewness(School$Inflation.rate)

[1] 0.2522898

1. **Plotting Distribution of various categories for key variables:**

Target:

The outcome variable is Target with categories as Dropout, Enrolled and Graduate.

*#plotting distribution of various categories*

*# count of students by dropout vs enrolled vs graduate*

*table(School$Target)*

*library(ggplot2)*

*ggplot(School, aes(x=School$Target, fill=School$Target)) +*

*geom\_bar()*

Dropout Enrolled Graduate

1421 794 2209

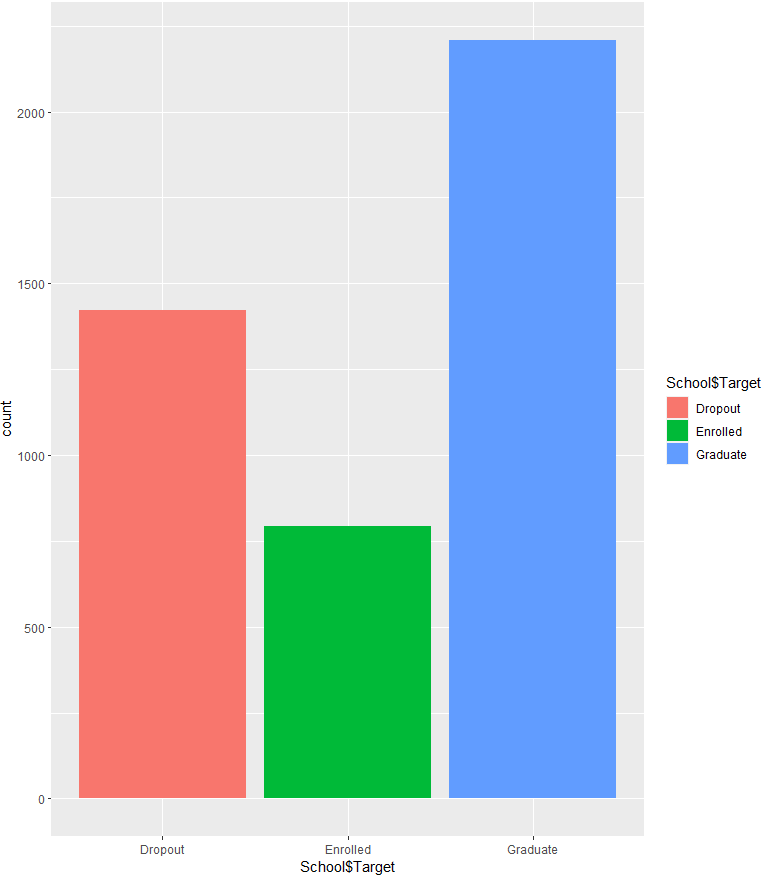


Figure 18: Count of students by Drop out vs Enrolled vs Graduate

So, in the dataset we have a greater number of graduate and dropout students.

Gender:

*ggplot(School, aes(x=School$Gender.Categories, fill= School$Gender.Categories)) +*

*geom\_bar()*

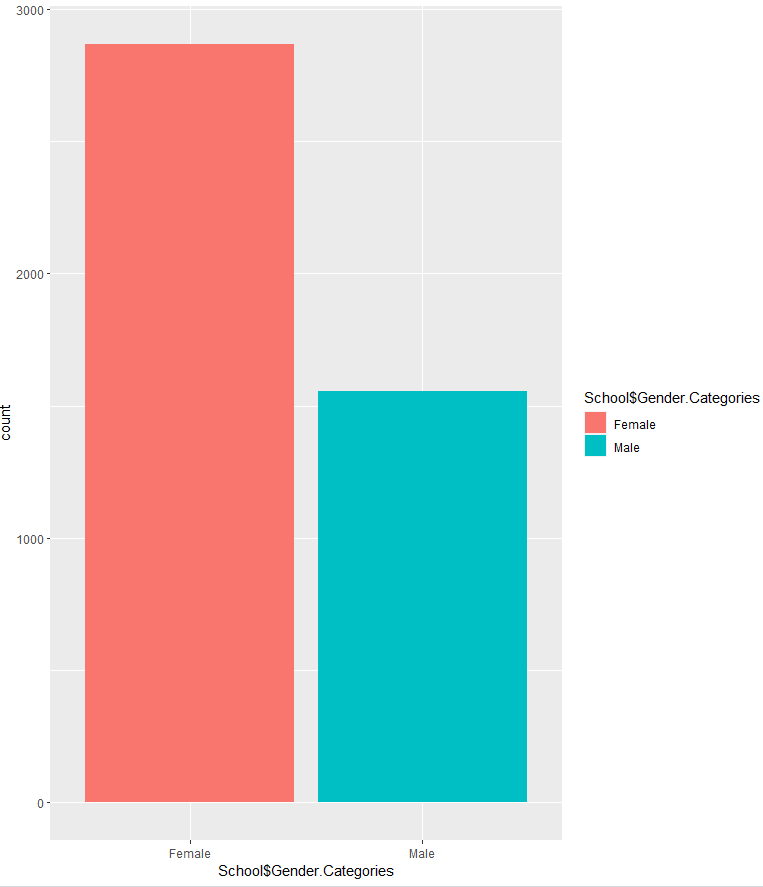


Figure 19: Distribution of Gender

To check the distribution of different gender categories in target variable we plotted a stacked bar chart with percentage representing for each category.

*School\_gender\_target <- table(School$Gender.Categories, School$Target)*

> School\_gender\_target

Dropout Enrolled Graduate

Female 720 487 1661

Male 701 307 548

*barplot(School\_gender\_target, legend.text = TRUE)*

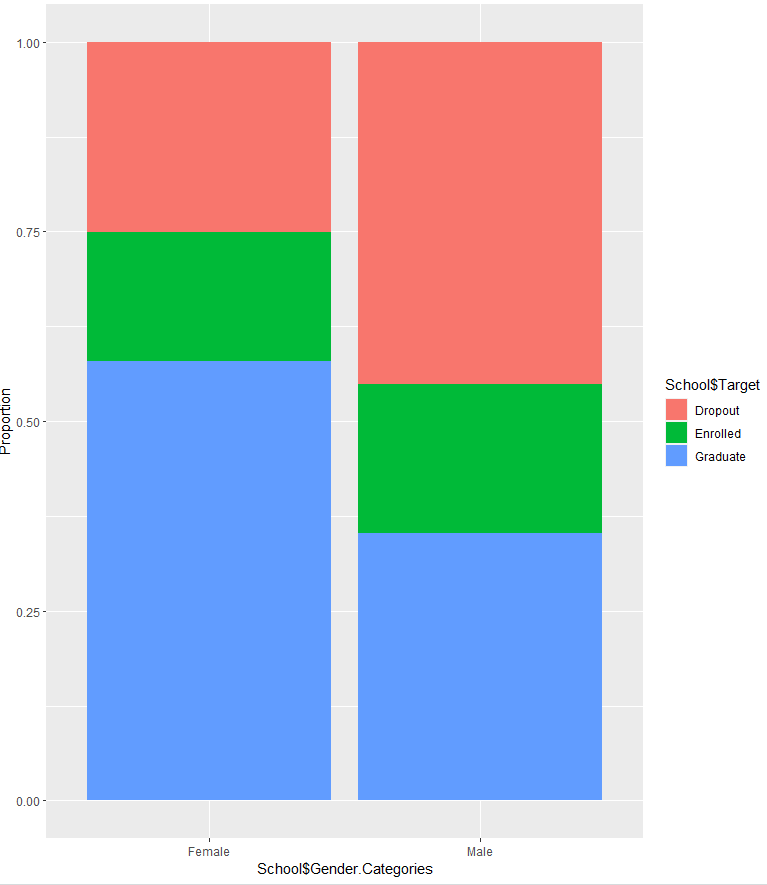


Figure 20: Gender wise distribution of target variable

We can see that almost the same percentage of female and male dropout students is present in the dataset, while the number of graduate female students is more than graduate male students. In the chart below we can see the number of each category by gender category.

Course Categories:

To check the proportion of various target categories in each course category, we plotted the stacked bar chart.

*ggplot(School,aes(x = School$Course.categories,*

*fill = School$Target)) +*

*geom\_bar(position = "fill") +*

*labs(y = "Proportion")+*

*theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1))*

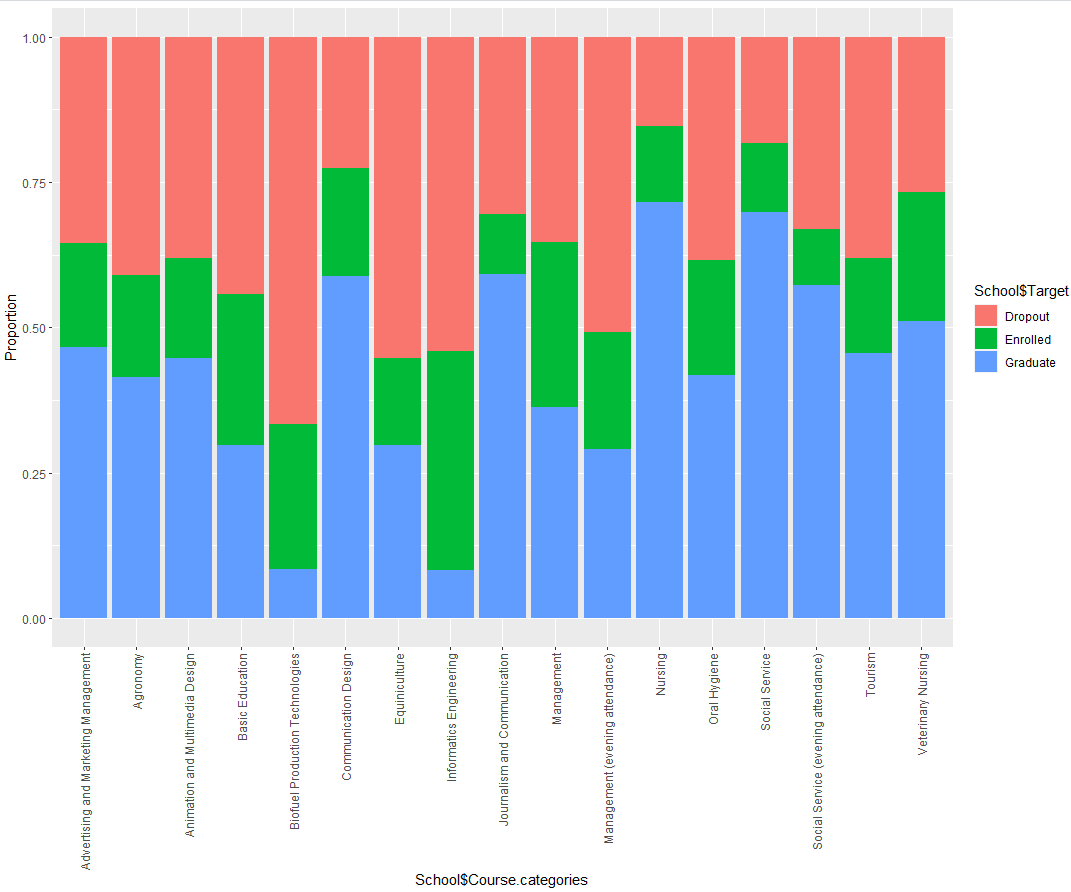


Figure 21: Distribution based on Course Category

We can see most of the dropout students are from the Biofuel Production Technologies course and the least are from the Communication Design course.

Daytime and evening attendance:

The stacked bar chart shows that most dropout students are in the evening attendance category.

*ggplot(School,*

*aes(x = School$Daytime\_evening\_attendance.categories,*

*fill = School$Target)) +*

*geom\_bar(position = "fill") +*

*labs(y = "Proportion")+*

*theme(axis.text.x = element\_text(vjust = 0.5, hjust=1))*

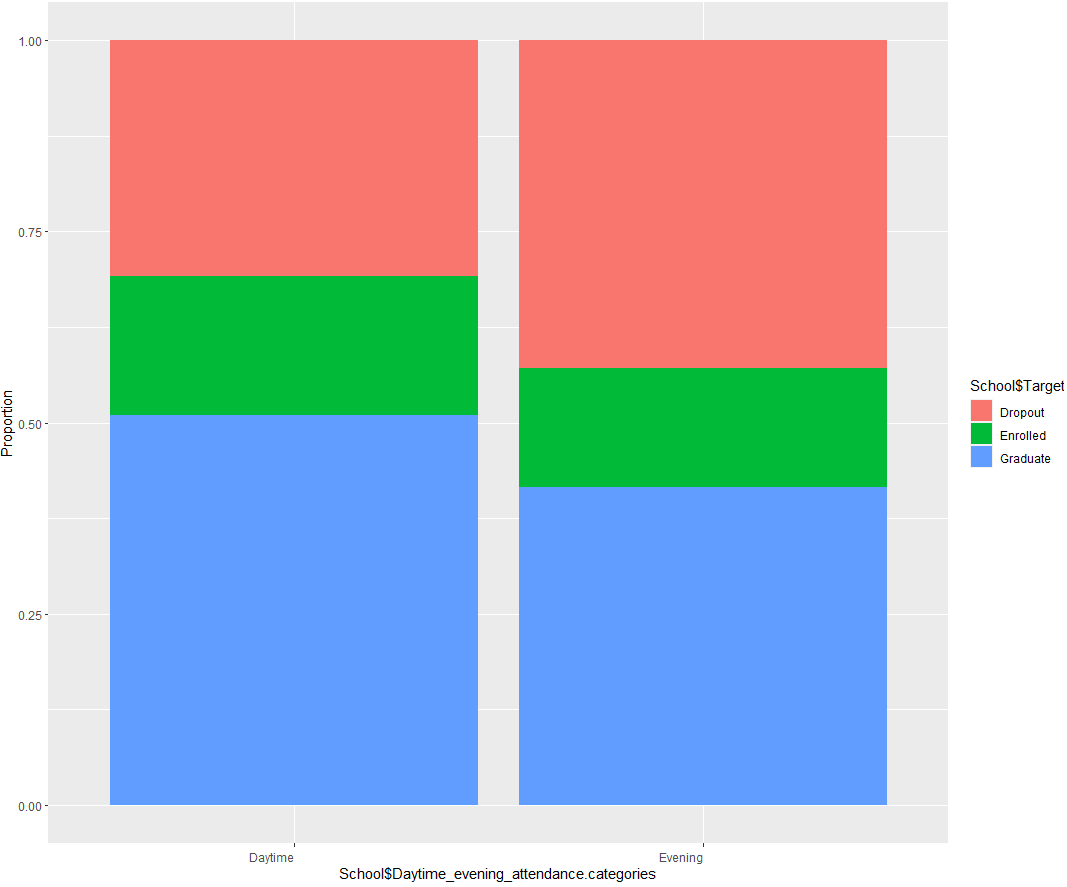


Figure 22: Distribution of attendance

Age at Enrollment:

The stacked bar chart below shows that the proportion of dropout students increases from age 20-30 years, with the least proportion happening below the age of 20 years.

*ggplot(School,*

*aes(x = School$Age.at.enrollment,*

*fill = School$Target)) +*

*geom\_bar(position = "fill") +*

*labs(y = "Proportion")+*

*theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1))*

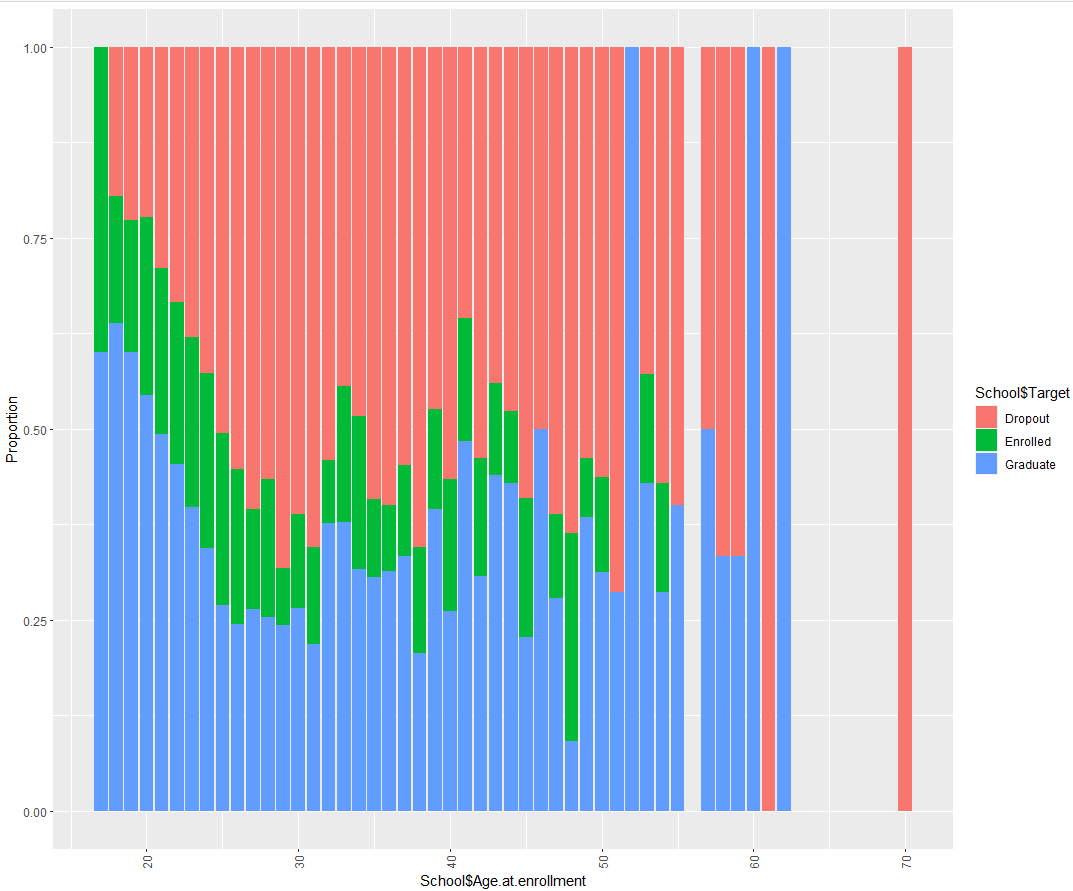


Figure 23: Age wise distribution

**7.** **Correlation Matrix:**

Correlation Matrix tells us how different variables in data set are related to each other. We plot

We removed categorical variable columns and plotted the correlation matrix/ heatmap using the ggplot library.

*#need to remove categorical codes columns*

*School\_2<- School\_1[,-c(1,2,4:17,19)]*

*View(School\_2)*

*cor(School\_2)*

*install.packages("ggcorrplot")*

*library(ggcorrplot)*

*ggcorrplot(cor(School\_2), hc.order = TRUE, type = "upper", lab = TRUE)*

*# correlation matrix shows that following variables are strongly correlated*

*# curricular units 2nd sem approved and curricular units 1st sem approved*

*# curricular units 1st sem grade and curricular units 2nd sem grade*

*# curricular units 1st sem evaluations and curricular units 2nd sem evaluations*

*# curricular units 1st sem enrolled and curricular units 2nd sem enrolled*

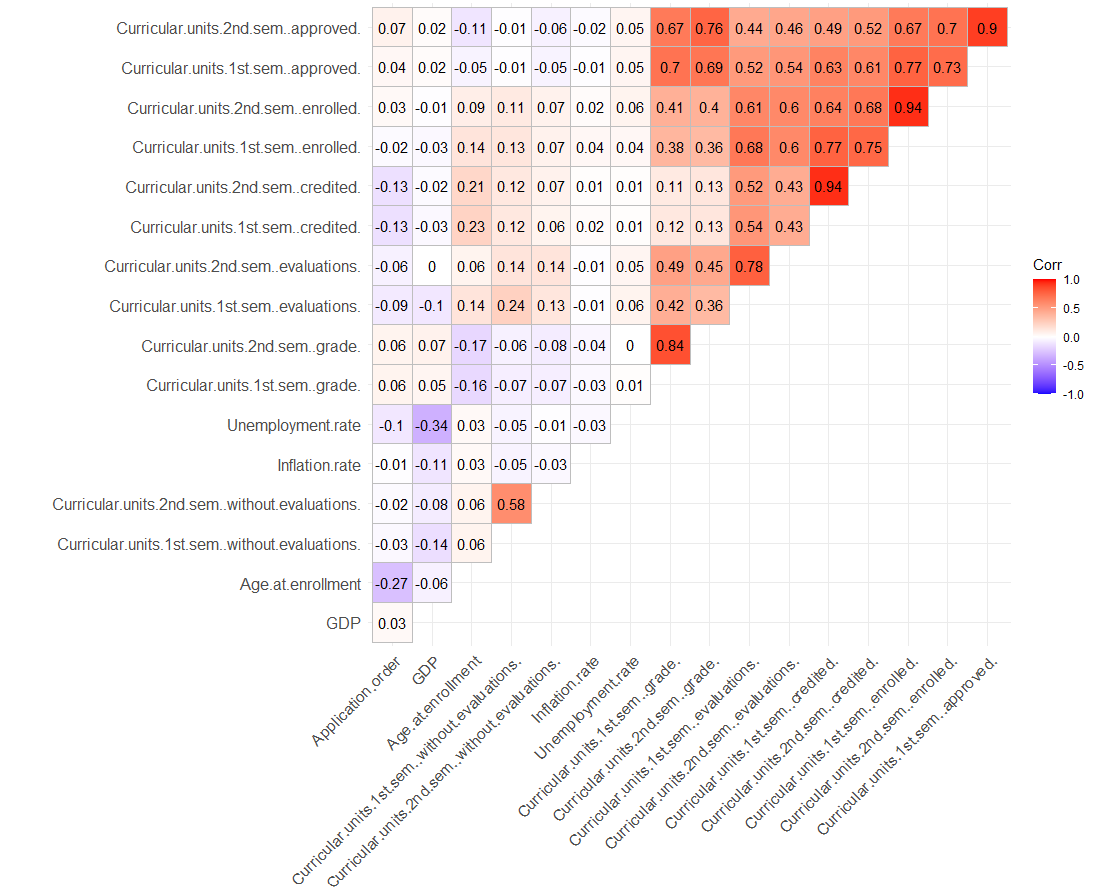


Figure 24: Correlation Matrix

**8.** **BI Models:**

**a) Logistic Regression:**

Removing the categorical variables to run the Logistic Regression model.

*School\_3<- School[,-c(2,4,7,9,11,13,15,17,19,21,23,25,27,29,31,33,36)]*

Partition the data as 70-30 % :

*Sch = School\_3*

*Sch$Target <- factor(School\_3$Target, levels = c("Graduate", "Dropout", "Enrolled"))*

*Sch$Target <- as.numeric(Sch$Target)-1*

*Sch\_filtered <- Sch[Sch$Target != 2, ]*

Run Logistic Regression:

*set.seed(10)*

*train\_index\_2 <- sample(nrow(Model2), (nrow(Model2) \* 0.7), replace = FALSE)*

*train\_S2 <- Model2[train\_index\_2, ]*

*test\_S2 <- Model2[-train\_index\_2, ]*

*install.packages("pROC")*

*library(pROC)*

*install.packages("caret")*

*library(caret)*

*logit.reg2 <- glm(train\_S2$Target ~ . , data = train\_S2, family = "binomial")*

*summary(logit.reg2)*

*predictions\_logit.reg\_train2 <- predict(logit.reg2, train\_S2, type = "response")*

*confusionMatrix(as.factor(ifelse(predictions\_logit.reg\_train2>0.5,1,0)), as.factor(train\_S2$Target))*

Graphical user interface, text

Description automatically generated

Figure 25: Summary of Logistic Regression

**Training Data:**

*predictions\_logit.reg\_train2 <- predict(logit.reg2, train\_S2, type = "response")*

*confusionMatrix(as.factor(ifelse(predictions\_logit.reg\_train2>0.5,1,0)), as.factor(train\_S2$Target))*

Text

Description automatically generated

Figure 26: Confusion Matrix and Statistics for Training Data

**Validation Data:**

*predictions\_logit.reg\_val2 <- predict(logit.reg2, test\_S2, type = "response")*

*confusionMatrix(as.factor(ifelse(predictions\_logit.reg\_val2>0.5,1,0)), as.factor(test\_S2$Target))*

Text

Description automatically generated with medium confidence

Figure 27: Confusion Matrix and Statistics for Validation Data

**9. Model Evaluation using ROC:**

*r\_logi2 <- roc(test\_S2$Target,predictions\_logit.reg\_val2, auc = TRUE)*

*plot.roc(r\_logi2)*

*auc(r\_logi2)*

Area under the curve: 0.9459

We plotted ROC (Receiver Operating Characteristic) curve to determine classification performance at each threshold. The area under the Curve (AUC) is 0.94 showing that the logistic regression model built is performing equally well on training data as well as test data set.

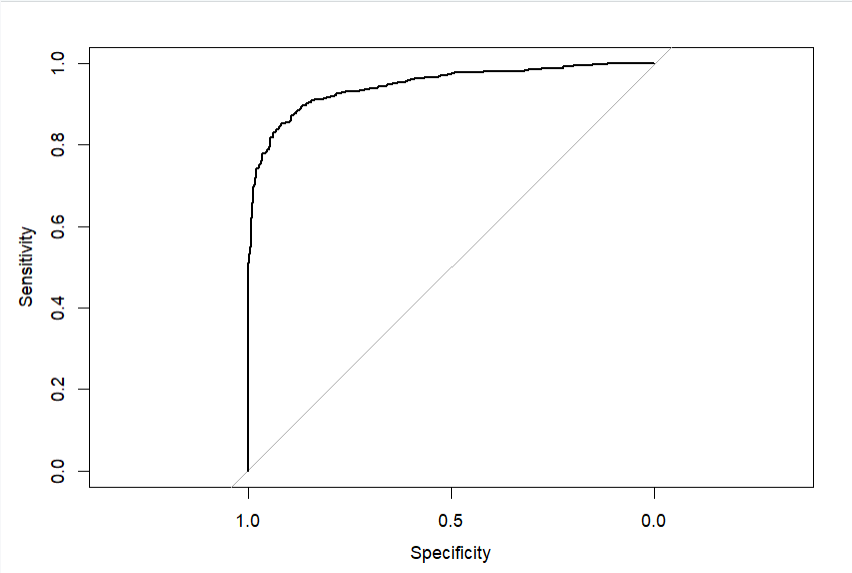


Figure 28: ROC Curve

**10. Results:**

From the above summary of Logistic Regression Model, we can say that “Course Codes” , “Debtor Codes”, “Tuition fees up to date codes”, “Gender codes”, “Scholarship holder codes” , “Curricular units 1st sem (credited)” , “Curricular units 1st sem (enrolled)”, “Curricular units 1st sem (approved)” have a major impact on predicting the drop out probability.

**11.Conclusion:**

Odds of students dropping out is highly dependent on the course they are enrolled in, if they have any debts, age, gender and curricular units and grades in 1st semester.

Based on analysis, we can provide suggestions to academic institutions to reduce the dropout rate among students. Since, drop out rate depends on curricular units, institutions can provide higher flexibility in course options. Students with debt may receive additional student assistance. Many students in the age group above 30 have obligations outside school such as looking after families and employment. To avoid drop out of these students, institutions can avail the facility of online classes.

**12.References:**

<https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention>

<https://cran.r-project.org/>

<https://utdallas.primo.exlibrisgroup.com/permalink/01UT_DALLAS/2hgl0t/alma9927850104601421>