Student Dropout Prediction Challenge: Data wrangling to Feature Engineering

Hamed

3/24/2020

This R markdown document contains well commented code and output of the whole process of machine learning from Data wrangling to Modelling.

DATA WRANGLING

Financial aid data loading

```
financial_aid<-read.csv("financial_aid.csv",header = T)
summary(financial_aid)
```

```
##    StudentID       cohort_term      Marital.Status  Adjusted.Gross.Income
##  Min.   : 20932   Min.   :1.000            : 2154   Min.   : -24326
##  1st Qu.:305677   1st Qu.:1.000   Divorced :  236   1st Qu.:      0
##  Median :322283   Median :1.000   Married  : 1024   Median :   2637
##  Mean   :317095   Mean   :1.451   Separated:  200   Mean   :  13125
##  3rd Qu.:344790   3rd Qu.:1.000   Single   :10155   3rd Qu.:  16323
##  Max.   :364184   Max.   :3.000                     Max.   :2576425
##                                                     NA's   :2154
##  Parent.Adjusted.Gross.Income Father.s.Highest.Grade.Level
##  Min.   :-62979                          :2292
##  1st Qu.:     0               College    :3284
##  Median : 12372               High School  :5092
##  Mean   : 28102               Middle School:1330
##  3rd Qu.: 38587               Unknown    :1771
##  Max.   :657631
##  NA's   :2154
##  Mother.s.Highest.Grade.Level              Housing       X2012.Loan
##                :2520                           :2164   Min.   :  337
##  College       :3215           Off Campus       :5373   1st Qu.: 3500
##  High School   :5024           On Campus Housing:1624   Median : 5500
##  Middle School:1296            With Parent      :4608   Mean   : 7169
##  Unknown       :1714                                    3rd Qu.: 9500
##                                                         Max.   :55626
##                                                         NA's   :12532
##  X2012.Scholarship X2012.Work.Study  X2012.Grant         X2013.Loan
##  Min.   :  283     Min.   : 200     Min.   :   79.09   Min.   :  103
##  1st Qu.: 2000     1st Qu.:1700     1st Qu.: 3368.25   1st Qu.: 3500
##  Median : 4000     Median :2000     Median : 5794.00   Median : 5500
##  Mean   : 5225     Mean   :1873     Mean   : 6660.93   Mean   : 7156
##  3rd Qu.: 6000     3rd Qu.:2121     3rd Qu.:10714.00   3rd Qu.: 9500
```

```
##  Max.   :27632     Max.   :3000     Max.   :13263.00   Max.   :50555
##  NA's   :13598     NA's   :13666    NA's   :12415      NA's   :11582
##  X2013.Scholarship X2013.Work.Study  X2013.Grant      X2014.Loan
##  Min.   :   23     Min.   :  25     Min.   :  162   Min.   :  128
##  1st Qu.: 2000     1st Qu.:2000     1st Qu.: 3683   1st Qu.: 3783
##  Median : 3549     Median :2000     Median : 6089   Median : 6250
##  Mean   : 4793     Mean   :2084     Mean   : 7094   Mean   : 7280
##  3rd Qu.: 6409     3rd Qu.:2200     3rd Qu.:11040   3rd Qu.:10500
##  Max.   :28737     Max.   :4000     Max.   :13790   Max.   :49845
##  NA's   :13459     NA's   :13590    NA's   :11450   NA's   :11028
##  X2014.Scholarship X2014.Work.Study  X2014.Grant         X2015.Loan
##  Min.   :  100     Min.   :  70     Min.   :   97.24   Min.   :   25
##  1st Qu.: 2000     1st Qu.:2000     1st Qu.: 3528.00   1st Qu.: 4162
##  Median : 4000     Median :2000     Median : 6245.00   Median : 6250
##  Mean   : 4999     Mean   :1933     Mean   : 7208.11   Mean   : 7241
##  3rd Qu.: 6000     3rd Qu.:2000     3rd Qu.:11725.89   3rd Qu.:10500
##  Max.   :38851     Max.   :3300     Max.   :14001.00   Max.   :47824
##  NA's   :13353     NA's   :13526    NA's   :10840      NA's   :10718
##  X2015.Scholarship X2015.Work.Study  X2015.Grant      X2016.Loan
##  Min.   :  200     Min.   :  10     Min.   :  209   Min.   :  103
##  1st Qu.: 2000     1st Qu.:2000     1st Qu.: 3880   1st Qu.: 4500
##  Median : 4000     Median :2000     Median : 6358   Median : 6420
##  Mean   : 4755     Mean   :2127     Mean   : 7370   Mean   : 7625
##  3rd Qu.: 5730     3rd Qu.:2800     3rd Qu.:11592   3rd Qu.:10500
##  Max.   :30478     Max.   :4600     Max.   :19038   Max.   :52880
##  NA's   :13174     NA's   :13520    NA's   :10365   NA's   :10594
##  X2016.Scholarship X2016.Work.Study  X2016.Grant         X2017.Loan
##  Min.   :   28.3   Min.   :  75     Min.   :    9.69   Min.   :  103
##  1st Qu.: 2000.0   1st Qu.:2000     1st Qu.: 3963.25   1st Qu.: 5354
##  Median : 4000.0   Median :2000     Median : 6428.00   Median : 6500
##  Mean   : 4897.3   Mean   :2036     Mean   : 7458.96   Mean   : 8256
##  3rd Qu.: 6000.0   3rd Qu.:2000     3rd Qu.:11717.50   3rd Qu.:11812
##  Max.   :31265.5   Max.   :4000     Max.   :18505.00   Max.   :60118
##  NA's   :13084     NA's   :13497    NA's   :10075      NA's   :10445
##  X2017.Scholarship X2017.Work.Study  X2017.Grant
##  Min.   :  100     Min.   :  45     Min.   :    0.1
##  1st Qu.: 2000     1st Qu.:1500     1st Qu.: 4261.0
##  Median : 4000     Median :2000     Median : 7305.0
##  Mean   : 5024     Mean   :1929     Mean   : 7794.2
##  3rd Qu.: 6906     3rd Qu.:2000     3rd Qu.:12173.0
##  Max.   :33848     Max.   :3000     Max.   :19823.0
##  NA's   :12784     NA's   :13402    NA's   :9732
```

- The data loaded has periodic capture financial information whereby variables; loan, workstudy, grant and scholarship are recorded multiple times.
- Therefore to get independent variables out of these variables; Combine by summing the loan, scholarship, grant and work/study data from 2011 to 2017 as follows;

```
library(tidyverse)
```

```
## -- Attaching packages ------------------------------ tidyverse 1.3.0 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ---------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
attach(financial_aid)
financial_aid=financial_aid %>%
  mutate(Total_loan = select(.,c(X2012.Loan,X2013.Loan,X2014.Loan,
           X2015.Loan,X2016.Loan,X2017.Loan)) %>%
           rowSums(na.rm = TRUE))

financial_aid<-financial_aid%>%
  mutate(Total_grant=select(.,c(X2012.Grant,X2013.Grant,X2014.Grant,
          X2015.Grant,X2016.Grant,X2017.Grant))%>%
           rowSums(na.rm = TRUE))

financial_aid<-financial_aid%>%

mutate(Total_scholarship=select(.,c(X2012.Scholarship,X2013.Scholarship,X2014
.Scholarship,
            X2015.Scholarship,X2016.Scholarship,X2017.Scholarship))%>%
           rowSums(na.rm = TRUE))

financial_aid<-financial_aid%>%
  mutate(Total_WorkStudy=select(.,c(X2012.Work.Study,X2013.Work.Study,
    X2014.Work.Study,X2015.Work.Study,X2016.Work.Study,X2017.Work.Study))%>%
           rowSums(na.rm = TRUE))

# Variables
colnames(financial_aid)
```

```
##  [1] "StudentID"                   "cohort_term"
##  [3] "Marital.Status"              "Adjusted.Gross.Income"
##  [5] "Parent.Adjusted.Gross.Income" "Father.s.Highest.Grade.Level"
##  [7] "Mother.s.Highest.Grade.Level" "Housing"
##  [9] "X2012.Loan"                  "X2012.Scholarship"
## [11] "X2012.Work.Study"            "X2012.Grant"
## [13] "X2013.Loan"                  "X2013.Scholarship"
## [15] "X2013.Work.Study"            "X2013.Grant"
## [17] "X2014.Loan"                  "X2014.Scholarship"
## [19] "X2014.Work.Study"            "X2014.Grant"
## [21] "X2015.Loan"                  "X2015.Scholarship"
## [23] "X2015.Work.Study"            "X2015.Grant"
## [25] "X2016.Loan"                  "X2016.Scholarship"
## [27] "X2016.Work.Study"            "X2016.Grant"
## [29] "X2017.Loan"                  "X2017.Scholarship"
```

```
## [31] "X2017.Work.Study"              "X2017.Grant"
## [33] "Total_loan"                    "Total_grant"
## [35] "Total_scholarship"             "Total_WorkStudy"
```

Data cleaning : drop the periodic columns after creating the independent variables in order to have relevant columns only.

```
# drop the extra periodical financial data
financial_aid<-financial_aid[,-c(9:32)]

# Load the train labels and the TestIDs
train_labels=read.csv("DropoutTrainLabels.csv",header = T)
testIDs=read.csv("TestIDs.csv",header = T)
```

JOIN the train labels and the financial aid data

```
library(dplyr)
financial_aid_train=left_join(train_labels,financial_aid,by="StudentID")
dim(financial_aid_train)

## [1] 12261    13

# JOIN the test IDs and the financial aid data
financial_aid_test=left_join(testIDs,financial_aid,by="StudentID")
dim(financial_aid_test)

## [1] 1000   12
```

Check for duplicated StudentID (TRUE=No duplicates, FALSE=duplicates present)

```
length(unique(financial_aid_train$StudentID)) == nrow(financial_aid_train)

## [1] TRUE
```

STATIC Data

STATIC data files loaded and merged using a function that iterates through a folder and picks up all files loads and merges them into one data frame.

```
#load all files and merge them
myETL=function(mypath){
  filenames = list.files(path=mypath, full.names=TRUE)
  file_load = function(x){read.csv(file=x,header=T)}
  datalist = lapply(filenames, file_load)
  data2 = do.call(rbind, lapply(datalist, as.data.frame))
  return(data2)
}
# call the function
mergedStatic<-myETL("D:/Hamed/KAGGLE COMPETITION/Student Retention Challenge
Data/Student Static Data")
```

```r
# Variable names in static dataset
colnames(mergedStatic)
```

```
##  [1] "StudentID"                 "Cohort"
##  [3] "CohortTerm"                "Campus"
##  [5] "Address1"                  "Address2"
##  [7] "City"                      "State"
##  [9] "Zip"                       "RegistrationDate"
## [11] "Gender"                    "BirthYear"
## [13] "BirthMonth"                "Hispanic"
## [15] "AmericanIndian"            "Asian"
## [17] "Black"                     "NativeHawaiian"
## [19] "White"                     "TwoOrMoreRace"
## [21] "HSDip"                     "HSDipYr"
## [23] "HSGPAUnwtd"                "HSGPAWtd"
## [25] "FirstGen"                  "DualHSSummerEnroll"
## [27] "EnrollmentStatus"          "NumColCredAttemptTransfer"
## [29] "NumColCredAcceptTransfer"  "CumLoanAtEntry"
## [31] "HighDeg"                   "MathPlacement"
## [33] "EngPlacement"              "GatewayMathStatus"
## [35] "GatewayEnglishStatus"
```

Check for duplicated StudentID (TRUE=No duplicates, FALSE=duplicates present)

```r
length(unique(mergedStatic$StudentID)) == nrow(mergedStatic)
```

```
## [1] TRUE
```

Since there were no duplicated student IDs, JOIN the train labels & test IDs and the static data

```r
library(dplyr)
static_train=left_join(train_labels,mergedStatic,by="StudentID")
summary(static_train)
```

```
##    StudentID         Dropout           Cohort       CohortTerm
##  Min.   : 20932   Min.   :0.0000   2011-12:2131   Min.   :1.000
##  1st Qu.:305164   1st Qu.:0.0000   2012-13:2059   1st Qu.:1.000
##  Median :321580   Median :0.0000   2013-14:1936   Median :1.000
##  Mean   :316079   Mean   :0.3861   2014-15:2080   Mean   :1.393
##  3rd Qu.:343608   3rd Qu.:1.0000   2015-16:2184   3rd Qu.:1.000
##  Max.   :359783   Max.   :1.0000   2016-17:1871   Max.   :3.000
##
##   Campus                            Address1
##  Mode:logical                           :  103
##  NA's:12261     NJCU-Registrar's Office:    6
##                 Summit Apts             :    5
##                 Jackson Garden Apt      :    4
##                 Westview Towers         :    4
##                 John F                  :    4
##                 (Other)                 :12135
```

```
##                       Address2                City          State
##                           :11906   Jersey City  :3285   NJ     :11869
##  1                        :   14   Bayonne      :1138   NY     :  120
##  2                        :   11   Newark       : 683          :  103
##  Apt 2                    :   10   North Bergen : 557   FL     :   29
##  2039 John F Kennedy Blvd:    6   Union City   : 549   CA     :   16
##  2nd Floor                :    5   West New York: 418   MD     :   15
##  (Other)                  :  309   (Other)      :5631   (Other):  109
##       Zip        RegistrationDate       Gender        BirthYear
##  Min.   :  747   Min.   :20110111   Min.   :1.000   Min.   :1945
##  1st Qu.: 7060   1st Qu.:20120710   1st Qu.:1.000   1st Qu.:1986
##  Median : 7304   Median :20140122   Median :2.000   Median :1991
##  Mean   : 7800   Mean   :20136172   Mean   :1.597   Mean   :1989
##  3rd Qu.: 7307   3rd Qu.:20150624   3rd Qu.:2.000   3rd Qu.:1994
##  Max.   :98118   Max.   :20160912   Max.   :2.000   Max.   :2000
##  NA's   :121                                        NA's   :1
##    BirthMonth        Hispanic       AmericanIndian        Asian
##  Min.   : 1.000   Min.   :-1.0000   Min.   :-1.0000   Min.   :-1.00000
##  1st Qu.: 4.000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.00000
##  Median : 7.000   Median : 0.0000   Median : 0.0000   Median : 0.00000
##  Mean   : 6.585   Mean   : 0.2567   Mean   :-0.0668   Mean   : 0.01974
##  3rd Qu.:10.000   3rd Qu.: 1.0000   3rd Qu.: 0.0000   3rd Qu.: 0.00000
##  Max.   :12.000   Max.   : 1.0000   Max.   : 1.0000   Max.   : 1.00000
##
##      Black         NativeHawaiian         White         TwoOrMoreRace
##  Min.   :-1.0000   Min.   :-1.00000   Min.   :-1.0000   Min.   :-1.00000
##  1st Qu.: 0.0000   1st Qu.: 0.00000   1st Qu.: 0.0000   1st Qu.: 0.00000
##  Median : 0.0000   Median : 0.00000   Median : 0.0000   Median : 0.00000
##  Mean   : 0.1467   Mean   :-0.06696   Mean   : 0.1824   Mean   :-0.05122
##  3rd Qu.: 0.0000   3rd Qu.: 0.00000   3rd Qu.: 1.0000   3rd Qu.: 0.00000
##  Max.   : 1.0000   Max.   : 1.00000   Max.   : 1.0000   Max.   : 1.00000
##
##      HSDip            HSDipYr         HSGPAUnwtd         HSGPAWtd
FirstGen
##  Min.   :-1.0000   Min.   :  -1.0   Min.   :-1.0000   Min.   :-1   Min.
:-1
##  1st Qu.: 1.0000   1st Qu.:  -1.0   1st Qu.:-1.0000   1st Qu.:-1   1st
Qu.:-1
##  Median : 1.0000   Median :  -1.0   Median :-1.0000   Median :-1   Median
:-1
##  Mean   : 0.9647   Mean   : 547.6   Mean   : 0.1395   Mean   :-1   Mean
:-1
##  3rd Qu.: 1.0000   3rd Qu.:2010.0   3rd Qu.: 2.3800   3rd Qu.:-1   3rd
Qu.:-1
##  Max.   : 4.0000   Max.   :2016.0   Max.   : 4.0000   Max.   :-1   Max.
:-1
##
##  DualHSSummerEnroll EnrollmentStatus NumColCredAttemptTransfer
##  Min.   :0          Min.   :1.000    Min.   : -2.00
##  1st Qu.:0          1st Qu.:1.000    1st Qu.: -2.00
```

```
##  Median :0          Median :2.000    Median : 16.00
##  Mean   :0          Mean   :1.596    Mean   : 37.46
##  3rd Qu.:0          3rd Qu.:2.000    3rd Qu.: 73.00
##  Max.   :0          Max.   :2.000    Max.   :150.00
##
##  NumColCredAcceptTransfer CumLoanAtEntry      HighDeg       MathPlacement
##  Min.   :-2.00            Min.   :-2.000   Min.   :0.0000   Min.   :-
1.0000
##  1st Qu.:-2.00            1st Qu.:-2.000   1st Qu.:0.0000   1st Qu.:
0.0000
##  Median :24.00            Median :-1.000   Median :0.0000   Median :
0.0000
##  Mean   :32.14            Mean   :-1.404   Mean   :0.5912   Mean   :
0.2742
##  3rd Qu.:66.00            3rd Qu.:-1.000   3rd Qu.:2.0000   3rd Qu.:
1.0000
##  Max.   :96.00            Max.   :-1.000   Max.   :4.0000   Max.   :
1.0000
##
##   EngPlacement     GatewayMathStatus GatewayEnglishStatus
##  Min.   :-1.0000   Min.   :0.0000    Min.   :0.0000
##  1st Qu.: 0.0000   1st Qu.:0.0000    1st Qu.:0.0000
##  Median : 0.0000   Median :0.0000    Median :0.0000
##  Mean   : 0.1839   Mean   :0.1196    Mean   :0.1871
##  3rd Qu.: 0.0000   3rd Qu.:0.0000    3rd Qu.:0.0000
##  Max.   : 1.0000   Max.   :1.0000    Max.   :1.0000
##
```

```
# JOIN the test IDs and the static data
static_test=left_join(testIDs,mergedStatic,by="StudentID")
```

PROGRESS Data

Loaded and merged using a function that iterates through a folder and picks up all files loads and merges them into one data frame.

```
myETL=function(mypath){
filenames = list.files(path=mypath, full.names=TRUE)
file_load = function(x){read.csv(file=x,header=T)}
datalist = lapply(filenames, file_load)
data2 = do.call(rbind, lapply(datalist, as.data.frame))
return(data2)
}

# call the function
mergedProgress<-myETL("D:/Hamed/KAGGLE COMPETITION/Student Retention
Challenge Data/Student Progress Data")

# variable names in the progress data
colnames(mergedProgress)
```

```
##  [1] "StudentID"          "Cohort"             "CohortTerm"
##  [4] "Term"               "AcademicYear"       "CompleteDevMath"
##  [7] "CompleteDevEnglish" "Major1"             "Major2"
## [10] "Complete1"          "Complete2"          "CompleteCIP1"
## [13] "CompleteCIP2"       "TransferIntent"     "DegreeTypeSought"
## [16] "TermGPA"            "CumGPA"
```

```
dim(mergedProgress)
```

```
## [1] 57945    17
```

Check for duplicated StudentID (TRUE=No duplicates, FALSE=duplicates present)

```
length(unique(mergedProgress$StudentID)) == nrow(mergedProgress)
```

```
## [1] FALSE
```

- The data had alot of duplicated student IDs thus to get only unique student ID the data frame was summarised and grouped by student ID.
- Thereafter a left join with the train labels to capter the students appearing only in the train labels data frame.

```
library(plyr)
```

```
## ------------------------------------------------------------------------
----

## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first,
then dplyr:
## library(plyr); library(dplyr)

## ------------------------------------------------------------------------
----

##
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

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

```
prog1=ddply(mergedProgress,.(StudentID),summarize,

CompleteDevMath=mean(CompleteDevMath),CompleteDevEnglish=mean(CompleteDevEngl
ish),

Major1=mean(Major1),Major2=mean(Major2),Complete1=mean(Complete1),
```

```
            Complete2=mean(Complete2),CompleteCIP1=mean(CompleteCIP1),

CompleteCIP2=mean(CompleteCIP2),TransferIntent=mean(TransferIntent),
            DegreeTypeSought=mean(DegreeTypeSought),TermGPA=mean(CumGPA),
            CumGPA=mean(CumGPA),number=length(StudentID))
dim(prog1)

## [1] 13767    14

# drop the irrelevant frequency column
prog1<-prog1[,-14]
dim(prog1)

## [1] 13767    13

# JOIN the train labels and the progress data
library(dplyr)
progress_train=left_join(train_labels,prog1,by="StudentID")
# JOIN the test IDs and the progress data
progress_test=left_join(testIDs,prog1,by="StudentID")

dim(progress_train)

## [1] 12261    14

# variables names for the progress data
summary(progress_train)

##    StudentID         Dropout       CompleteDevMath  CompleteDevEnglish
##  Min.   : 20932   Min.   :0.0000   Min.   :-2.000   Min.   :-2.000
##  1st Qu.:305164   1st Qu.:0.0000   1st Qu.:-2.000   1st Qu.:-2.000
##  Median :321580   Median :0.0000   Median :-2.000   Median :-2.000
##  Mean   :316079   Mean   :0.3861   Mean   :-1.258   Mean   :-1.426
##  3rd Qu.:343608   3rd Qu.:1.0000   3rd Qu.: 0.000   3rd Qu.:-1.000
##  Max.   :359783   Max.   :1.0000   Max.   : 1.000   Max.   : 1.000
##      Major1          Major2         Complete1        Complete2
##  Min.   :-1.00   Min.   :-1.000   Min.   :0.0000   Min.   :0
##  1st Qu.:26.01   1st Qu.:-1.000   1st Qu.:0.0000   1st Qu.:0
##  Median :43.02   Median :-1.000   Median :0.0000   Median :0
##  Mean   :36.62   Mean   :-0.136   Mean   :0.4482   Mean   :0
##  3rd Qu.:51.38   3rd Qu.:-1.000   3rd Qu.:0.7778   3rd Qu.:0
##  Max.   :54.01   Max.   :52.140   Max.   :4.0000   Max.   :0
##   CompleteCIP1      CompleteCIP2 TransferIntent DegreeTypeSought
##  Min.   :-2.0000   Min.   :-2    Min.   :-1     Min.   :6
##  1st Qu.:-2.0000   1st Qu.:-2    1st Qu.:-1     1st Qu.:6
##  Median :-2.0000   Median :-2    Median :-1     Median :6
##  Mean   : 0.7489   Mean   :-2    Mean   :-1     Mean   :6
##  3rd Qu.: 2.0927   3rd Qu.:-2    3rd Qu.:-1     3rd Qu.:6
##  Max.   :26.0051   Max.   :-2    Max.   :-1     Max.   :6
##     TermGPA          CumGPA
##  Min.   :0.000   Min.   :0.000
```

```
##  1st Qu.:2.395   1st Qu.:2.395
##  Median :3.075   Median :3.075
##  Mean   :2.817   Mean   :2.817
##  3rd Qu.:3.578   3rd Qu.:3.578
##  Max.   :4.000   Max.   :4.000
```

Check to ensure that all the datasets are of the same row and column sizes before merging them

```
dim(financial_aid_train)

## [1] 12261    13

dim(financial_aid_test)

## [1] 1000   12

dim(static_train)

## [1] 12261    36

dim(static_test)

## [1] 1000   35

dim(progress_train)

## [1] 12261    14

dim(progress_test)

## [1] 1000   13
```

TRAINING DATASET

- The train data set was created by use of a inner join function between financial , static and progress. The same was done for test data.
- The data was cleaned by removing irrelevant(duplicated) columns and and renaming the Dropout variable.
- The missing values were clearly assigned NA's to make sure all of them are identified.
- The categorical variables with wrong data types were corrected and assigned factor data types.

```
library(dplyr)
join1<-inner_join(financial_aid_train,static_train,by="StudentID")
TRAIN_DATA=inner_join(join1,progress_train,by="StudentID")
colnames(TRAIN_DATA)

##  [1] "StudentID"                    "Dropout.x"
##  [3] "cohort_term"                  "Marital.Status"
##  [5] "Adjusted.Gross.Income"        "Parent.Adjusted.Gross.Income"
```

```
##  [7] "Father.s.Highest.Grade.Level" "Mother.s.Highest.Grade.Level"
##  [9] "Housing"                      "Total_loan"
## [11] "Total_grant"                  "Total_scholarship"
## [13] "Total_WorkStudy"              "Dropout.y"
## [15] "Cohort"                       "CohortTerm"
## [17] "Campus"                       "Address1"
## [19] "Address2"                     "City"
## [21] "State"                        "Zip"
## [23] "RegistrationDate"             "Gender"
## [25] "BirthYear"                    "BirthMonth"
## [27] "Hispanic"                     "AmericanIndian"
## [29] "Asian"                        "Black"
## [31] "NativeHawaiian"               "White"
## [33] "TwoOrMoreRace"                "HSDip"
## [35] "HSDipYr"                      "HSGPAUnwtd"
## [37] "HSGPAWtd"                     "FirstGen"
## [39] "DualHSSummerEnroll"           "EnrollmentStatus"
## [41] "NumColCredAttemptTransfer"    "NumColCredAcceptTransfer"
## [43] "CumLoanAtEntry"               "HighDeg"
## [45] "MathPlacement"                "EngPlacement"
## [47] "GatewayMathStatus"            "GatewayEnglishStatus"
## [49] "Dropout"                      "CompleteDevMath"
## [51] "CompleteDevEnglish"           "Major1"
## [53] "Major2"                       "Complete1"
## [55] "Complete2"                    "CompleteCIP1"
## [57] "CompleteCIP2"                 "TransferIntent"
## [59] "DegreeTypeSought"             "TermGPA"
## [61] "CumGPA"

dim(TRAIN_DATA)

## [1] 12261    61

# TESTING DATASET
join2<-inner_join(financial_aid_test,static_test,by="StudentID")
TEST_DATA<-inner_join(join2,progress_test,by="StudentID")
dim(TEST_DATA)

## [1] 1000   58

# Clean the train data by removing extra dropout variables
(Dropout.y,Dropout)
TRAIN_DATA<-select(TRAIN_DATA,-c("Dropout.y"))
TRAIN_DATA<-select(TRAIN_DATA,-c("Dropout"))

# Rename the droput.x to just dropout for consistency
library(dplyr)
colnames(TRAIN_DATA)[colnames(TRAIN_DATA)=="Dropout.x"] <- "Dropout"

dim(TRAIN_DATA)
```

```
## [1] 12261    59
```

```
dim(TEST_DATA)
```

```
## [1] 1000   58
```

```
# The response variable
TRAIN_DATA$Dropout[1:5]
```

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

```
# The features of our project that will be analysed further
colnames(TEST_DATA)
```

```
##  [1] "StudentID"                    "cohort_term"
##  [3] "Marital.Status"               "Adjusted.Gross.Income"
##  [5] "Parent.Adjusted.Gross.Income" "Father.s.Highest.Grade.Level"
##  [7] "Mother.s.Highest.Grade.Level" "Housing"
##  [9] "Total_loan"                   "Total_grant"
## [11] "Total_scholarship"            "Total_WorkStudy"
## [13] "Cohort"                       "CohortTerm"
## [15] "Campus"                       "Address1"
## [17] "Address2"                     "City"
## [19] "State"                        "Zip"
## [21] "RegistrationDate"             "Gender"
## [23] "BirthYear"                    "BirthMonth"
## [25] "Hispanic"                     "AmericanIndian"
## [27] "Asian"                        "Black"
## [29] "NativeHawaiian"               "White"
## [31] "TwoOrMoreRace"                "HSDip"
## [33] "HSDipYr"                      "HSGPAUnwtd"
## [35] "HSGPAWtd"                     "FirstGen"
## [37] "DualHSSummerEnroll"           "EnrollmentStatus"
## [39] "NumColCredAttemptTransfer"    "NumColCredAcceptTransfer"
## [41] "CumLoanAtEntry"               "HighDeg"
## [43] "MathPlacement"                "EngPlacement"
## [45] "GatewayMathStatus"            "GatewayEnglishStatus"
## [47] "CompleteDevMath"              "CompleteDevEnglish"
## [49] "Major1"                       "Major2"
## [51] "Complete1"                    "Complete2"
## [53] "CompleteCIP1"                 "CompleteCIP2"
## [55] "TransferIntent"               "DegreeTypeSought"
## [57] "TermGPA"                      "CumGPA"
```

```
# Convert empty spaces and (-1) to NULL in order to facilitate imputation
later on
TRAIN_DATA[TRAIN_DATA=="-1"]<-NA
TRAIN_DATA[TRAIN_DATA==""]<-NA


# Convert the theorically categorical variables to  factors
```

```
names=c('DegreeTypeSought','TransferIntent','CompleteDevEnglish','CompleteDev
Math',
  'GatewayEnglishStatus','GatewayMathStatus','EngPlacement','MathPlacement',
  'HighDeg','EnrollmentStatus','FirstGen','HSDip','BirthMonth','BirthYear',
  'Gender','Campus','CohortTerm','cohort_term','Dropout')

TRAIN_DATA[,names] <- lapply(TRAIN_DATA[,names] , factor)
str(TRAIN_DATA)

## 'data.frame':    12261 obs. of  59 variables:
##  $ StudentID                   : int  285848 302176 301803 302756 301067
297371 273211 302772 280023 300412 ...
##  $ Dropout                     : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 2
2 1 2 ...
##  $ cohort_term                 : Factor w/ 2 levels "1","3": 1 1 1 1 1 1 1
1 1 1 ...
##  $ Marital.Status              : Factor w/ 5 levels "","Divorced",..: 3 NA
5 NA 3 5 5 5 5 5 ...
##  $ Adjusted.Gross.Income       : int  116846 NA 1528 NA 69036 0 0 2069
10033 3602 ...
##  $ Parent.Adjusted.Gross.Income: int  0 NA 0 NA 0 0 0 73993 19467 65801
...
##  $ Father.s.Highest.Grade.Level: Factor w/ 5 levels "","College","High
School",..: 3 NA 2 NA 4 5 2 2 5 3 ...
##  $ Mother.s.Highest.Grade.Level: Factor w/ 5 levels "","College","High
School",..: 2 NA 3 NA 3 2 3 2 3 3 ...
##  $ Housing                     : Factor w/ 4 levels "","Off Campus",..: 2
NA 2 NA 2 4 3 4 4 3 ...
##  $ Total_loan                  : num  35000 28896 54057 0 0 ...
##  $ Total_grant                 : num  0 0 0 0 0 ...
##  $ Total_scholarship           : num  0 0 0 0 21643 ...
##  $ Total_WorkStudy             : num  0 0 0 0 0 0 0 0 0 745 ...
##  $ Cohort                      : Factor w/ 6 levels "2011-12","2012-
13",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ CohortTerm                  : Factor w/ 2 levels "1","3": 1 1 1 1 1 1 1
1 1 1 ...
##  $ Campus                      : Factor w/ 0 levels: NA NA NA NA NA NA NA
NA NA NA ...
##  $ Address1                    : Factor w/ 12704 levels "","1 Brookside
Ave",..: 916 279 168 954 673 370 769 1347 750 1275 ...
##  $ Address2                    : Factor w/ 291 levels "","#5","1 St.
Floor",..: NA NA NA NA NA NA NA NA NA NA ...
##  $ City                        : Factor w/ 677 levels "","Allentown",..:
102 110 194 110 110 110 39 207 110 111 ...
##  $ State                       : Factor w/ 40 levels "","AZ","CA","CO",..:
15 15 11 15 15 15 15 15 15 20 ...
##  $ Zip                         : int  7030 7305 4769 7302 7302 7305 7306
8872 7307 2919 ...
##  $ RegistrationDate            : int  20110808 20110804 20110809 20110823
20110420 20110628 20110810 20110908 20110714 20110607 ...
```

```
##  $ Gender                        : Factor w/ 2 levels "1","2": 2 1 2 2 1 2 2
1 2 2 ...
##  $ BirthYear                     : Factor w/ 55 levels "1945","1946",..: 33
25 39 41 24 47 41 45 42 48 ...
##  $ BirthMonth                    : Factor w/ 12 levels "1","2","3","4",..: 9
4 4 1 4 8 8 6 12 2 ...
##  $ Hispanic                      : int  0 0 0 0 0 0 NA NA 1 0 ...
##  $ AmericanIndian                : int  0 0 0 0 0 0 NA NA 0 0 ...
##  $ Asian                         : int  0 0 0 0 0 0 NA NA 0 0 ...
##  $ Black                         : int  0 0 0 0 0 1 NA NA 0 1 ...
##  $ NativeHawaiian                : int  0 0 0 0 0 0 NA NA 0 0 ...
##  $ White                         : int  1 1 1 1 1 0 NA NA 0 0 ...
##  $ TwoOrMoreRace                 : int  0 0 0 0 0 0 NA NA 0 0 ...
##  $ HSDip                         : Factor w/ 3 levels "1","2","4": 1 1 1 NA
1 1 1 1 1 1 ...
##  $ HSDipYr                       : int  NA NA NA NA NA 2010 NA NA NA 2011
...
##  $ HSGPAUnwtd                    : num  NA NA NA NA NA 3.5 NA NA NA 2.5 ...
##  $ HSGPAWtd                      : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ FirstGen                      : Factor w/ 0 levels: NA NA NA NA NA NA NA
NA NA NA ...
##  $ DualHSSummerEnroll            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ EnrollmentStatus              : Factor w/ 2 levels "1","2": 2 2 2 2 2 1 2
2 2 1 ...
##  $ NumColCredAttemptTransfer     : num  0 96 0 54 70 -2 62 53 52 -2 ...
##  $ NumColCredAcceptTransfer      : num  0 45 0 87.5 66 -2 66 45 66 -2 ...
##  $ CumLoanAtEntry                : int  NA NA NA NA NA -2 NA NA NA -2 ...
##  $ HighDeg                       : Factor w/ 4 levels "0","2","3","4": 1 1 1
1 2 1 2 1 1 1 ...
##  $ MathPlacement                 : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1
1 1 2 ...
##  $ EngPlacement                  : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 1
1 1 1 ...
##  $ GatewayMathStatus             : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 1
1 1 1 ...
##  $ GatewayEnglishStatus          : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1
1 1 1 ...
##  $ CompleteDevMath               : Factor w/ 36 levels "-2","-0.5","0",..: 1
1 1 1 1 1 1 1 1 3 ...
##  $ CompleteDevEnglish            : Factor w/ 40 levels "-2","-1.5","-
1.25",..: 1 1 1 1 1 23 1 1 1 1 ...
##  $ Major1                        : num  51.2 51.4 51.2 45.1 23 ...
##  $ Major2                        : num  NA NA NA NA 13.1 ...
##  $ Complete1                     : num  2.667 1.333 2.667 1.75 0.875 ...
##  $ Complete2                     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ CompleteCIP1                  : num  15.8 6.9 15.8 9.77 1.13 ...
##  $ CompleteCIP2                  : num  -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 ...
##  $ TransferIntent                : Factor w/ 0 levels: NA NA NA NA NA NA NA
NA NA NA ...
##  $ DegreeTypeSought              : Factor w/ 1 level "6": 1 1 1 1 1 1 1 1 1
```

```
1 ...
##  $ TermGPA                     : num  3.48 3.54 3.94 3.79 4 ...
##  $ CumGPA                      : num  3.48 3.54 3.94 3.79 4 ...
```

EXPLORATORY DATA ANALYSIS

We use ggplot for our boxplots, bar plots, scatter plots to display insightful analysis

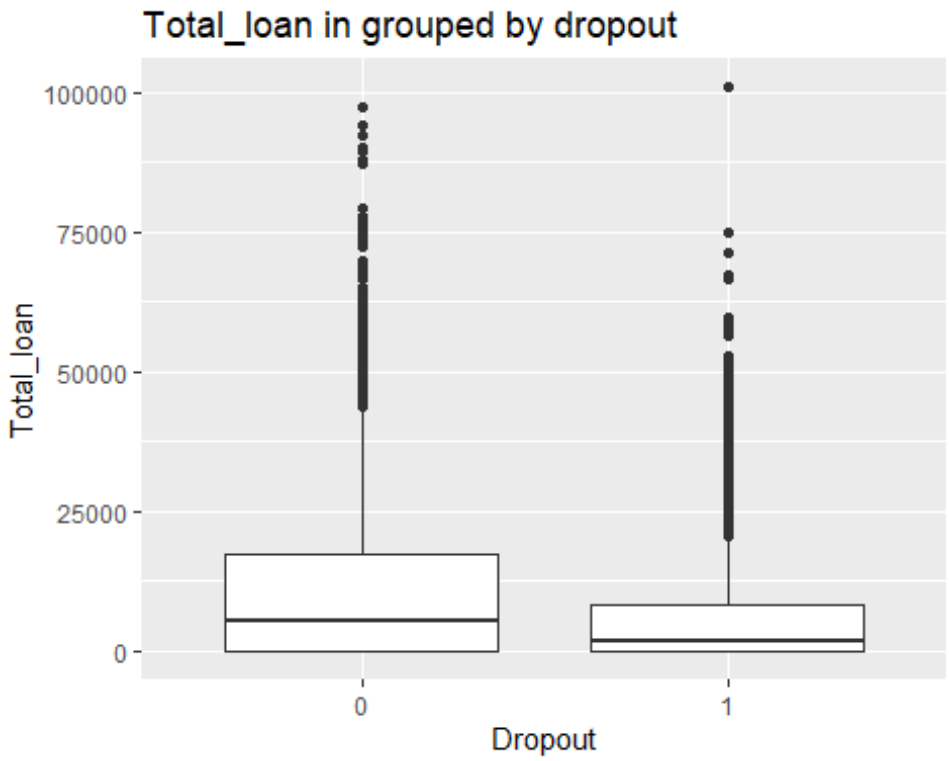
```
library(ggplot2)
# Summarize the data
summary(TRAIN_DATA)

##    StudentID      Dropout  cohort_term   Marital.Status
Adjusted.Gross.Income
##  Min.   : 20932   0:7527   1:9851              :   0   Min.   : -24326
##  1st Qu.:305164   1:4734   3:2410      Divorced : 208   1st Qu.:      0
##  Median :321580                       Married  : 924   Median :   2768
##  Mean   :316079                       Separated: 185   Mean   :  13263
##  3rd Qu.:343608                       Single   :9103   3rd Qu.:  16491
##  Max.   :359783                       NA's     :1841   Max.   :2576425
##                                                        NA's   :1841
##  Parent.Adjusted.Gross.Income Father.s.Highest.Grade.Level
##  Min.   :-49406                           :   0
##  1st Qu.:     0               College      :2916
##  Median : 12373               High School  :4578
##  Mean   : 28318               Middle School:1201
##  3rd Qu.: 38805               Unknown      :1598
##  Max.   :657631               NA's         :1968
##  NA's   :1841
##  Mother.s.Highest.Grade.Level              Housing       Total_loan
##                :   0                           :   0   Min.   :     0
##  College       :2896          Off Campus       :4846   1st Qu.:     0
##  High School   :4516          On Campus Housing:1430   Median :  3745
##  Middle School:1153           With Parent      :4120   Mean   :  8834
##  Unknown       :1535          NA's             :1865   3rd Qu.: 13429
##  NA's          :2161                                   Max.   :100960
##
##   Total_grant    Total_scholarship Total_WorkStudy       Cohort
CohortTerm
##  Min.   :    0   Min.   :     0    Min.   :    0.0   2011-12:2131   1:9851
##  1st Qu.:    0   1st Qu.:     0    1st Qu.:    0.0   2012-13:2059   3:2410
##  Median : 5265   Median :     0    Median :    0.0   2013-14:1936
##  Mean   : 9690   Mean   :  1170    Mean   :  208.5   2014-15:2080
##  3rd Qu.:14100   3rd Qu.:     0    3rd Qu.:    0.0   2015-16:2184
##  Max.   :80873   Max.   :125497    Max.   :14820.0   2016-17:1871
##
##   Campus                          Address1                          Address2
##  NA's:12261   NJCU-Registrar's Office:    6   1                          :
14
```

```
##               Summit Apts            :    5   2                        :
11
##               Jackson Garden Apt     :    4   Apt 2                    :
10
##               Westview Towers        :    4   2039 John F Kennedy Blvd:
6
##               John F                 :    4   2nd Floor                :
5
##               (Other)                :12135   (Other)                  :
309
##               NA's                   :  103   NA's
:11906
##            City          State            Zip        RegistrationDate
##  Jersey City :3285   NJ     :11869   Min.   :  747   Min.   :20110111
##  Bayonne     :1138   NY     :  120   1st Qu.: 7060   1st Qu.:20120710
##  Newark      : 683   FL     :   29   Median : 7304   Median :20140122
##  North Bergen: 557   CA     :   16   Mean   : 7800   Mean   :20136172
##  Union City  : 549   MD     :   15   3rd Qu.: 7307   3rd Qu.:20150624
##  (Other)     :5945   (Other):  109   Max.   :98118   Max.   :20160912
##  NA's        : 104   NA's   :  103   NA's   :121
##  Gender     BirthYear      BirthMonth      Hispanic      AmericanIndian
##  1:4947   1993   :1173   9      :1119   Min.   :0.0000   Min.   :0.000
##  2:7314   1994   :1051   7      :1098   1st Qu.:0.0000   1st Qu.:0.000
##           1995   : 864   8      :1093   Median :0.0000   Median :0.000
##           1992   : 832   1      :1058   Mean   :0.3494   Mean   :0.002
##           1996   : 811   10     :1029   3rd Qu.:1.0000   3rd Qu.:0.000
##           (Other):7529   12     :1028   Max.   :1.0000   Max.   :1.000
##           NA's   :   1   (Other):5836   NA's   :842      NA's   :842
##      Asian            Black        NativeHawaiian       White
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.0949   Mean   :0.2313   Mean   :0.0018   Mean   :0.2696
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  NA's   :842      NA's   :842      NA's   :842      NA's   :842
##  TwoOrMoreRace     HSDip          HSDipYr       HSGPAUnwtd       HSGPAWtd
##  Min.   :0.0000   1   :11916   Min.   :1963   Min.   :0.900   Min.   : NA
##  1st Qu.:0.0000   2   :   69   1st Qu.:2011   1st Qu.:2.500   1st Qu.: NA
##  Median :0.0000   4   :   10   Median :2013   Median :2.880   Median : NA
##  Mean   :0.0187   NA's:  266   Mean   :2013   Mean   :2.909   Mean   :NaN
##  3rd Qu.:0.0000                3rd Qu.:2015   3rd Qu.:3.300   3rd Qu.: NA
##  Max.   :1.0000                Max.   :2016   Max.   :4.000   Max.   : NA
##  NA's   :842                   NA's   :8921   NA's   :8687    NA's
:12261
##  FirstGen     DualHSSummerEnroll EnrollmentStatus
NumColCredAttemptTransfer
##  NA's:12261   Min.   :0          1:4952           Min.   : -2.00
##               1st Qu.:0          2:7309           1st Qu.: -2.00
##               Median :0                           Median : 24.00
```

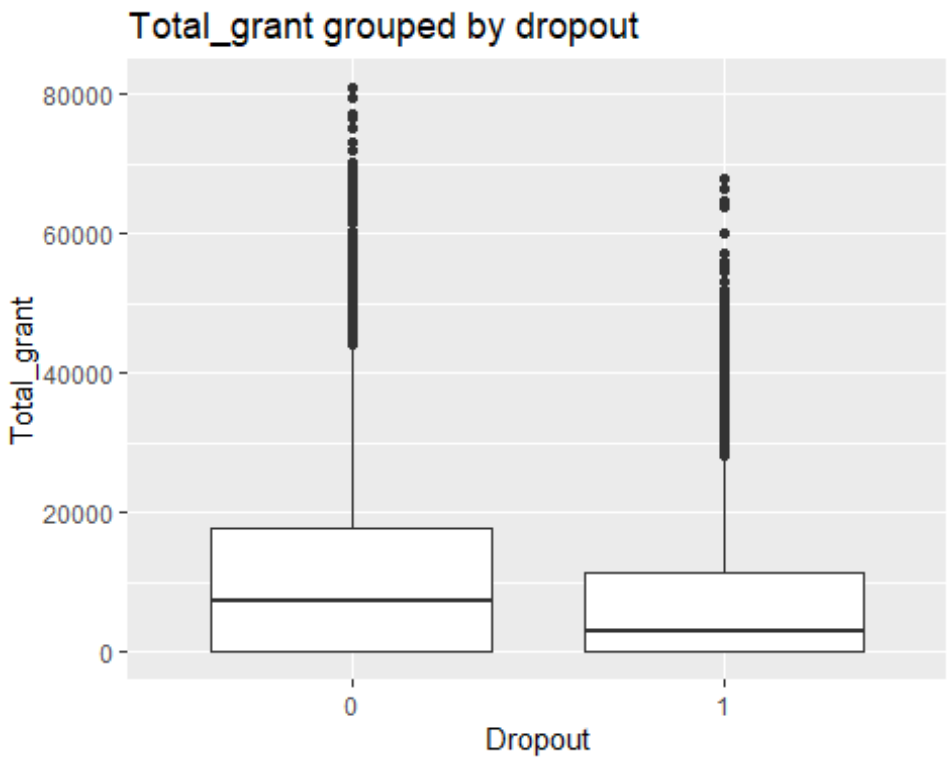
```
##                Mean   :0                              Mean   : 38.66
##                3rd Qu.:0                              3rd Qu.: 74.00
##                Max.   :0                              Max.   :150.00
##                                                       NA's   :370
##  NumColCredAcceptTransfer CumLoanAtEntry HighDeg  MathPlacement
EngPlacement
##  Min.   :-2.00            Min.   :-2     0:8710   0   :7859     0   :8966
##  1st Qu.:-2.00            1st Qu.:-2     2:3406   1   :3882     1   :2775
##  Median :24.00            Median :-2     3: 143   NA's: 520     NA's: 520
##  Mean   :32.14            Mean   :-2     4:   2
##  3rd Qu.:66.00            3rd Qu.:-2
##  Max.   :96.00            Max.   :-2
##  NA's   :1                NA's   :7309
##  GatewayMathStatus GatewayEnglishStatus CompleteDevMath CompleteDevEnglish
##  0:10794           0:9967               -2     :7854    -2     :8860
##  1: 1467           1:2294               0      :1478    0      : 773
##                                         0.5    : 443    0.5    : 319
##                                         0.25   : 379    1      : 311
##                                         1      : 213    0.25   : 197
##                                         (Other):1371    (Other):1274
##                                         NA's   : 523    NA's   : 527
##      Major1          Major2         Complete1        Complete2
##  Min.   :-0.50   Min.   : 0.003   Min.   :0.0000   Min.   :0
##  1st Qu.:26.01   1st Qu.: 6.060   1st Qu.:0.0000   1st Qu.:0
##  Median :43.02   Median : 9.575   Median :0.0000   Median :0
##  Mean   :37.02   Mean   :12.564   Mean   :0.4482   Mean   :0
##  3rd Qu.:51.38   3rd Qu.:13.121   3rd Qu.:0.7778   3rd Qu.:0
##  Max.   :54.01   Max.   :52.140   Max.   :4.0000   Max.   :0
##  NA's   :129     NA's   :11480
##   CompleteCIP1      CompleteCIP2 TransferIntent DegreeTypeSought
##  Min.   :-2.0000   Min.   :-2    NA's:12261     6:12261
##  1st Qu.:-2.0000   1st Qu.:-2
##  Median :-2.0000   Median :-2
##  Mean   : 0.7489   Mean   :-2
##  3rd Qu.: 2.0927   3rd Qu.:-2
##  Max.   :26.0051   Max.   :-2
##
##     TermGPA          CumGPA
##  Min.   :0.000   Min.   :0.000
##  1st Qu.:2.395   1st Qu.:2.395
##  Median :3.075   Median :3.075
##  Mean   :2.817   Mean   :2.817
##  3rd Qu.:3.578   3rd Qu.:3.578
##  Max.   :4.000   Max.   :4.000
##
```

```
# Visualize the financial information in relation to the dropout variable
whereby we look at the distribution (mean, outliers and spread) of each
variable grouped by the dropout variable
```

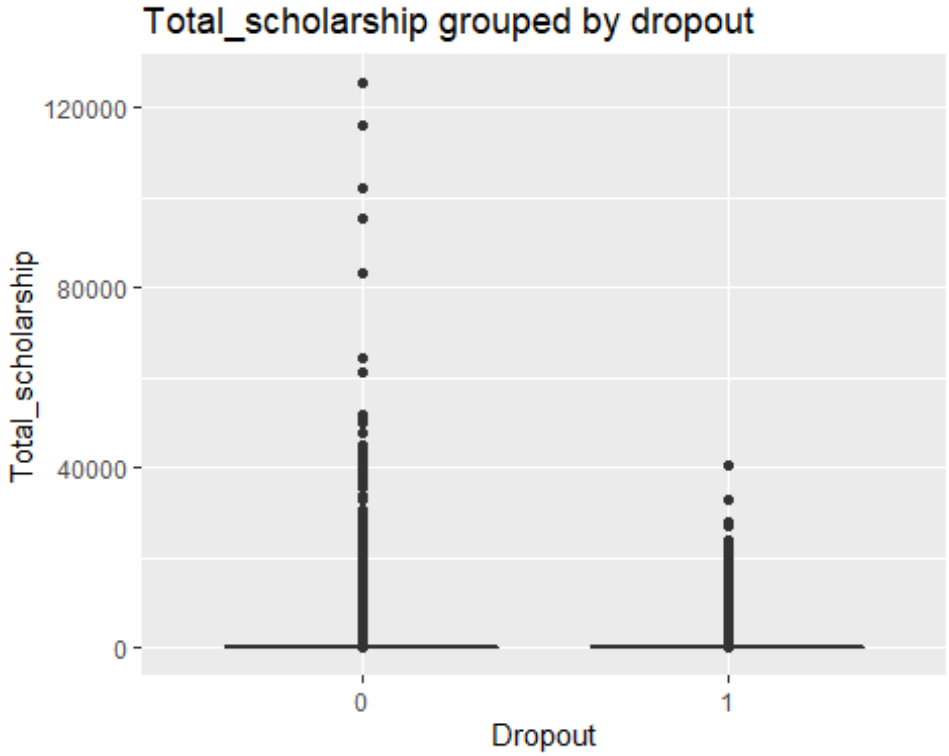
```
library(dplyr)
ggplot(data = TRAIN_DATA, mapping = aes(x =Dropout,y=Total_loan))+
  ggtitle("Total_loan in grouped by dropout") +
  geom_boxplot()
```



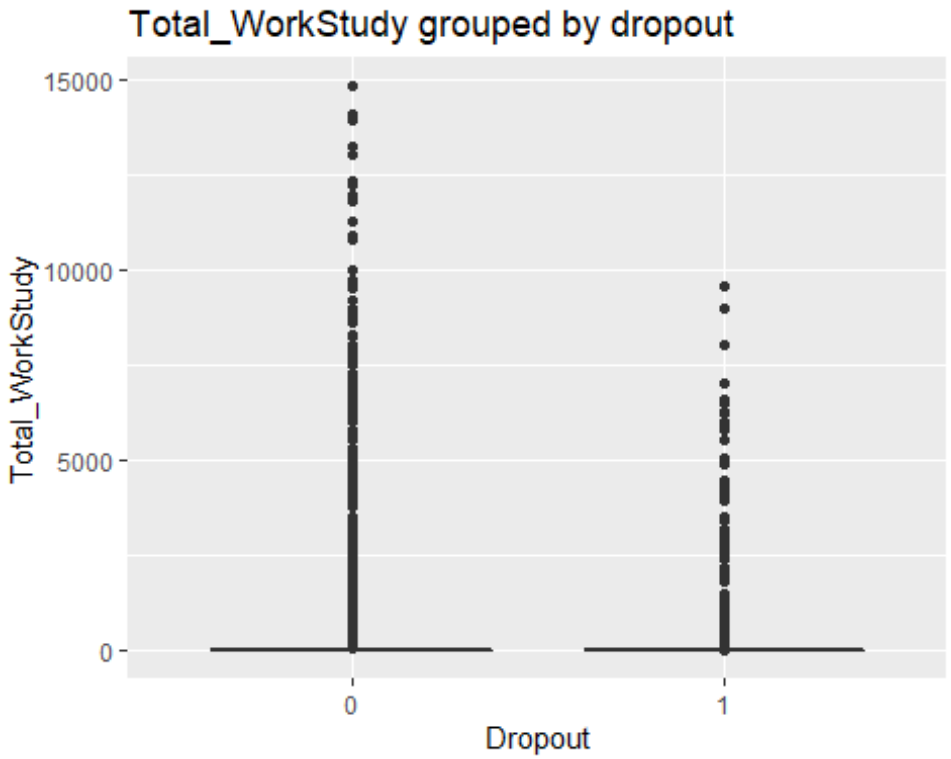

```
ggplot(data = TRAIN_DATA, mapping = aes(x =Dropout,y=Total_grant))+
  ggtitle("Total_grant grouped by dropout") +
  geom_boxplot()
```




```
ggplot(data = TRAIN_DATA, mapping = aes(x =Dropout,y=Total_scholarship))+
  ggtitle("Total_scholarship grouped by dropout") +
  geom_boxplot()
```

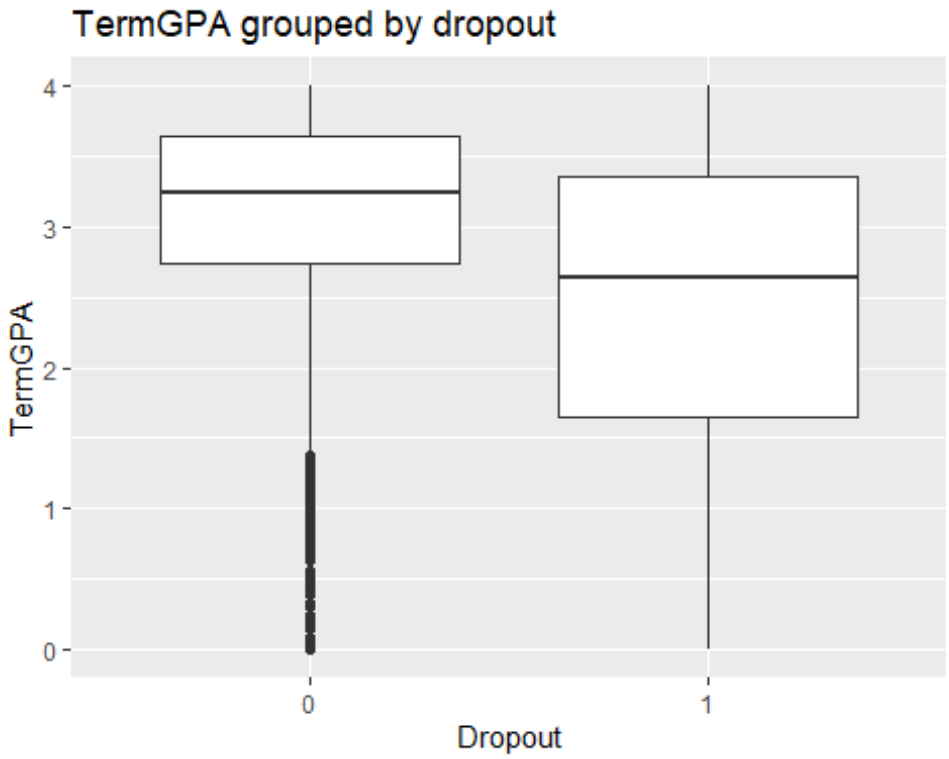


```
ggplot(data = TRAIN_DATA, mapping = aes(x =Dropout,y=Total_WorkStudy))+
ggtitle("Total_WorkStudy grouped by dropout") +
    geom_boxplot()
```

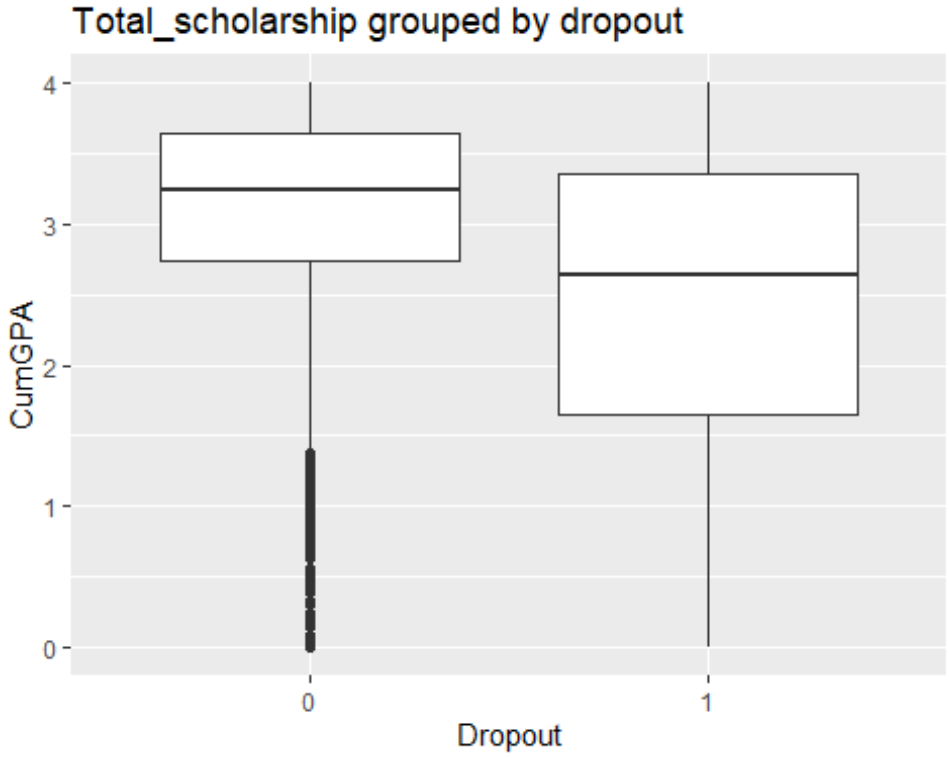


Visualize the student performance information in relation to the dropout variable whereby we look at the distribution (mean, outliers and spread) of each variable grouped by the dropout variable

```
ggplot(data = TRAIN_DATA, mapping = aes(x =Dropout,y=TermGPA))+
  ggtitle("TermGPA grouped by dropout") +
  geom_boxplot()
```

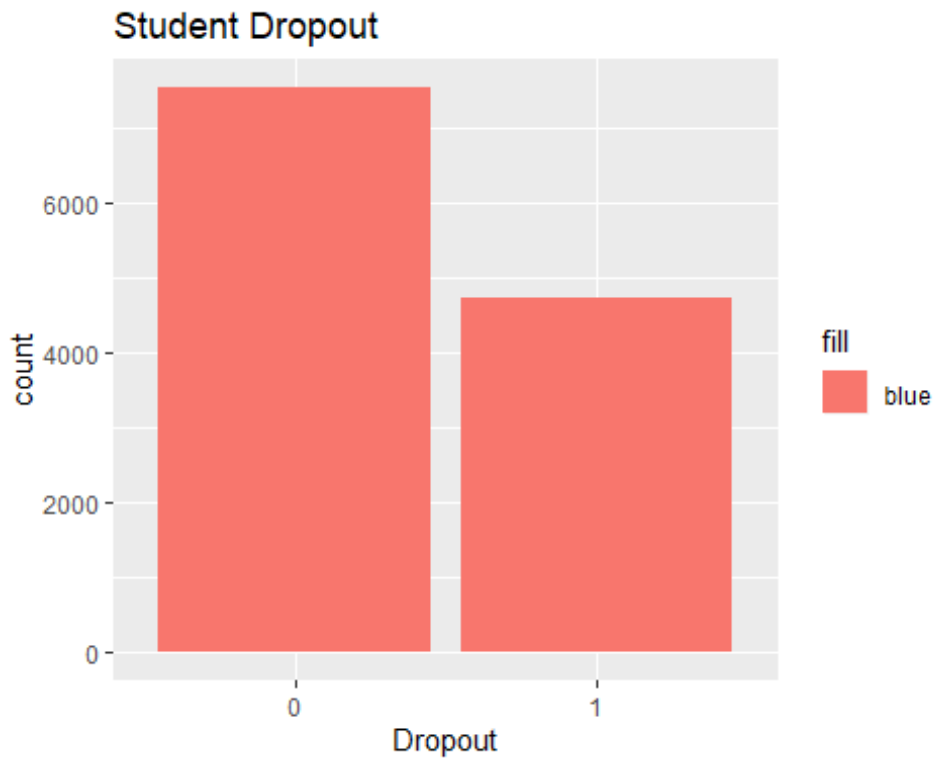



```
ggplot(data = TRAIN_DATA, mapping = aes(x =Dropout,y=CumGPA))+
  ggtitle("Total_scholarship grouped by dropout") +
  geom_boxplot()
```

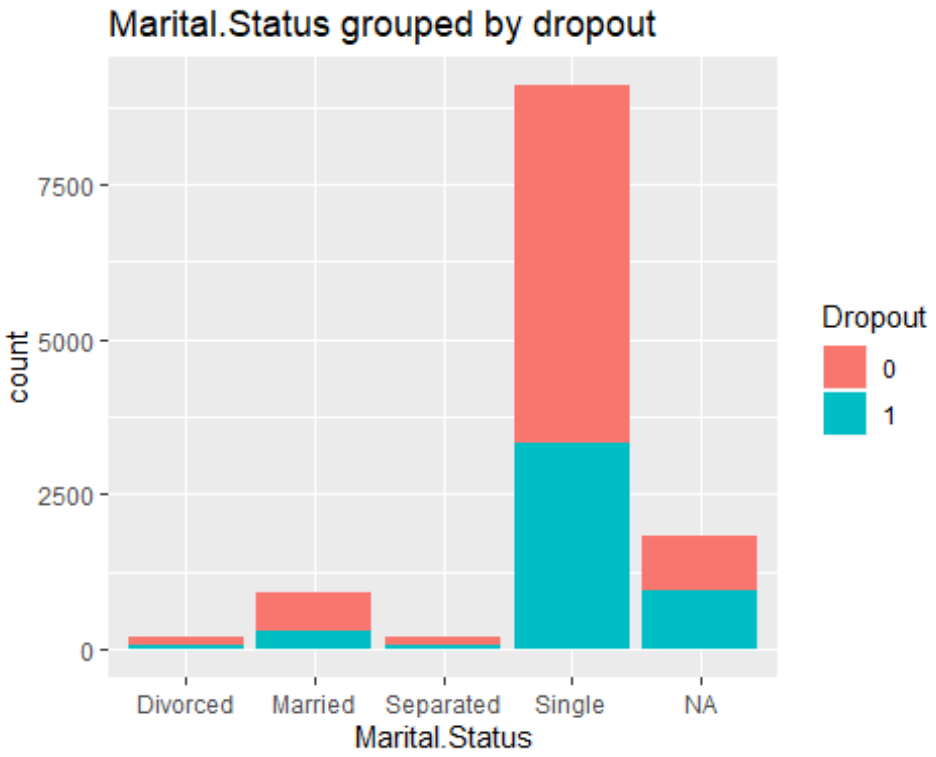


Frequency analysis for categorical variables in relation to the Dropout variable

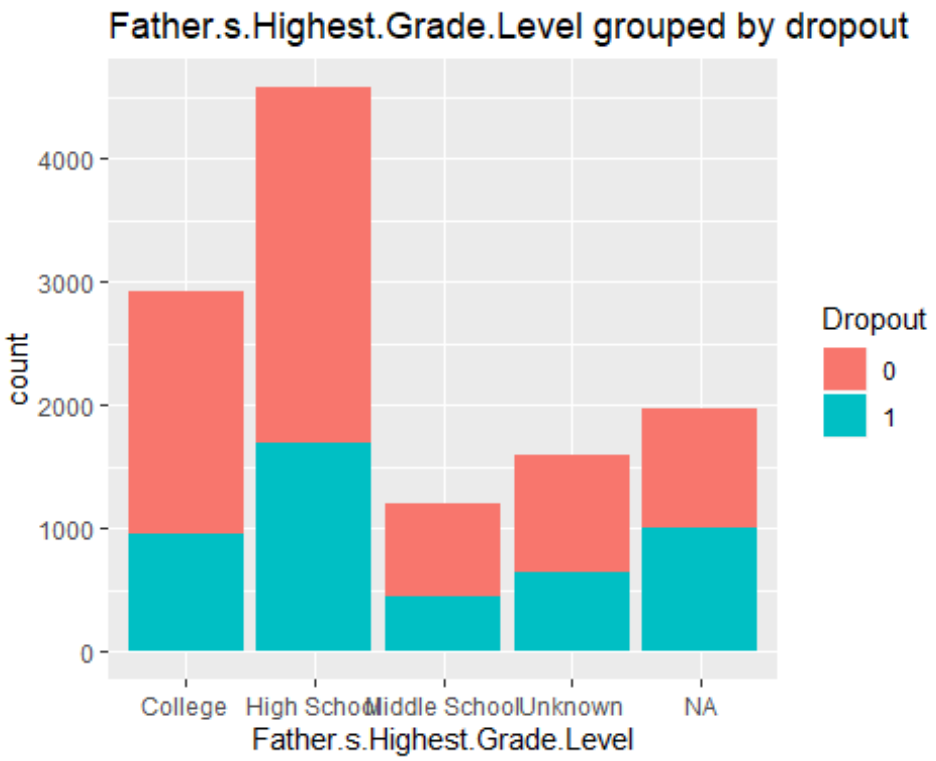
```
# Dropout summary statistics
table(TRAIN_DATA$Dropout,exclude = NULL)

##
##    0    1
## 7527 4734

ggplot(data = TRAIN_DATA) +
  ggtitle("Student Dropout") +
  geom_bar(mapping = aes(x = Dropout,fill='blue'))
```



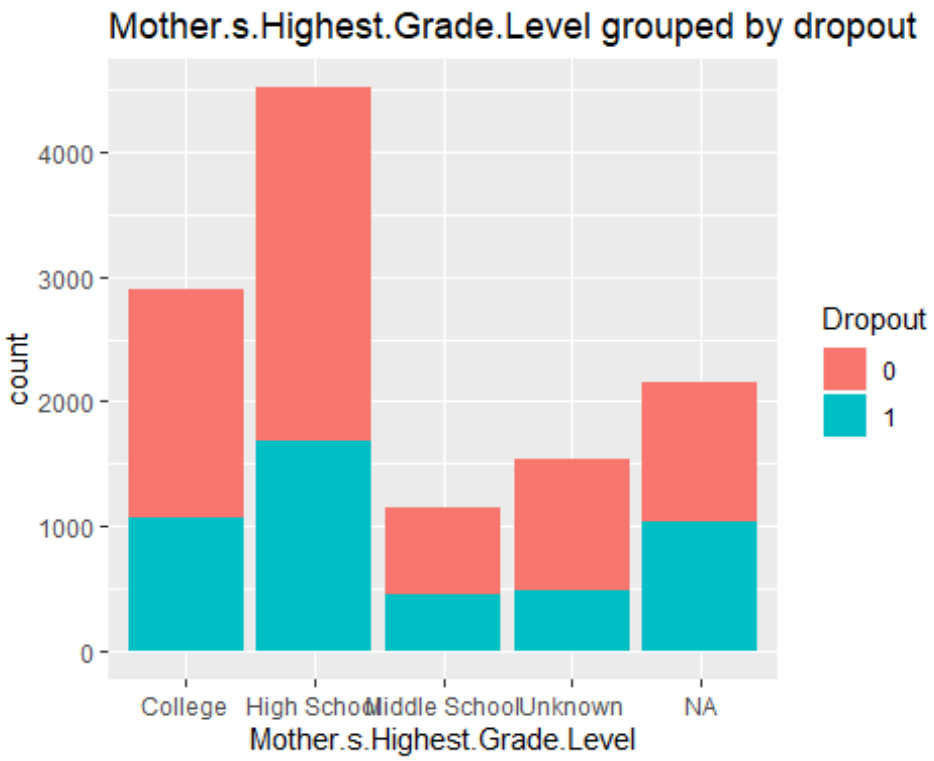

```
# Marital.Status
table(TRAIN_DATA$Marital.Status,TRAIN_DATA$Dropout,exclude = NULL)

##
##                0    1
##                0    0
##   Divorced   136   72
##   Married    616  308
##   Separated  108   77
##   Single    5777 3326
##   <NA>       890  951

ggplot(data = TRAIN_DATA) +
  ggtitle("Marital.Status grouped by dropout")+
  geom_bar(mapping = aes(x = Marital.Status,fill=Dropout))
```



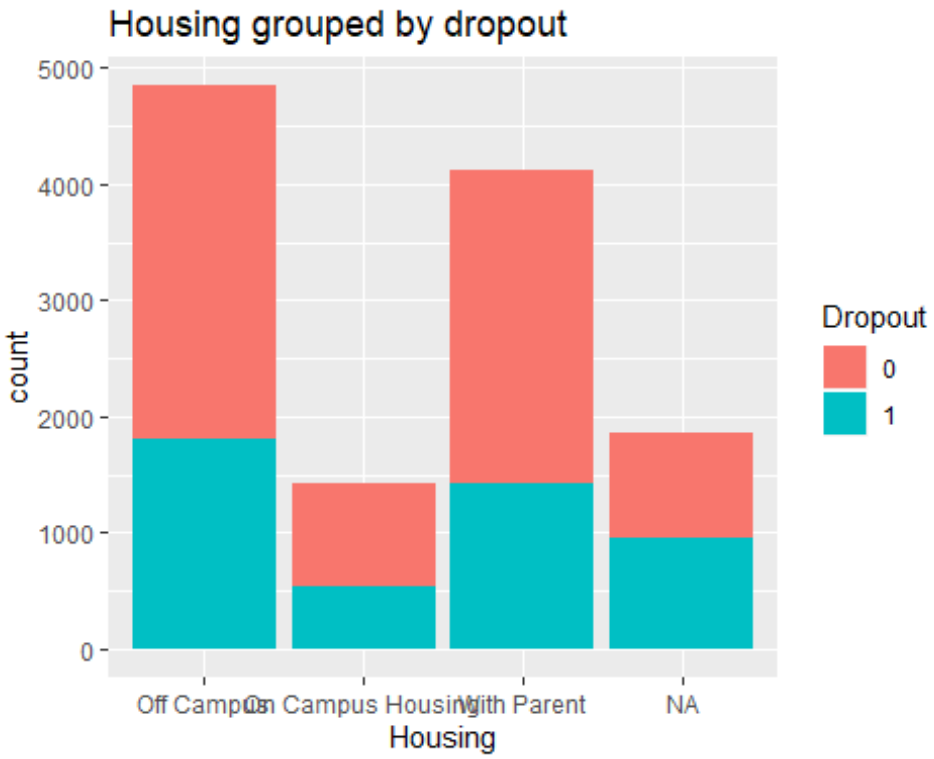

```
# Father.s.Highest.Grade.Level
table(TRAIN_DATA$Father.s.Highest.Grade.Level,TRAIN_DATA$Dropout,exclude =
NULL)

##
##                     0    1
##                     0    0
##   College        1959  957
##   High School    2893 1685
##   Middle School   755  446
##   Unknown         961  637
##   <NA>            959 1009

ggplot(data = TRAIN_DATA) +
  ggtitle("Father.s.Highest.Grade.Level grouped by dropout")+
    geom_bar(mapping = aes(x = Father.s.Highest.Grade.Level,fill=Dropout))
```



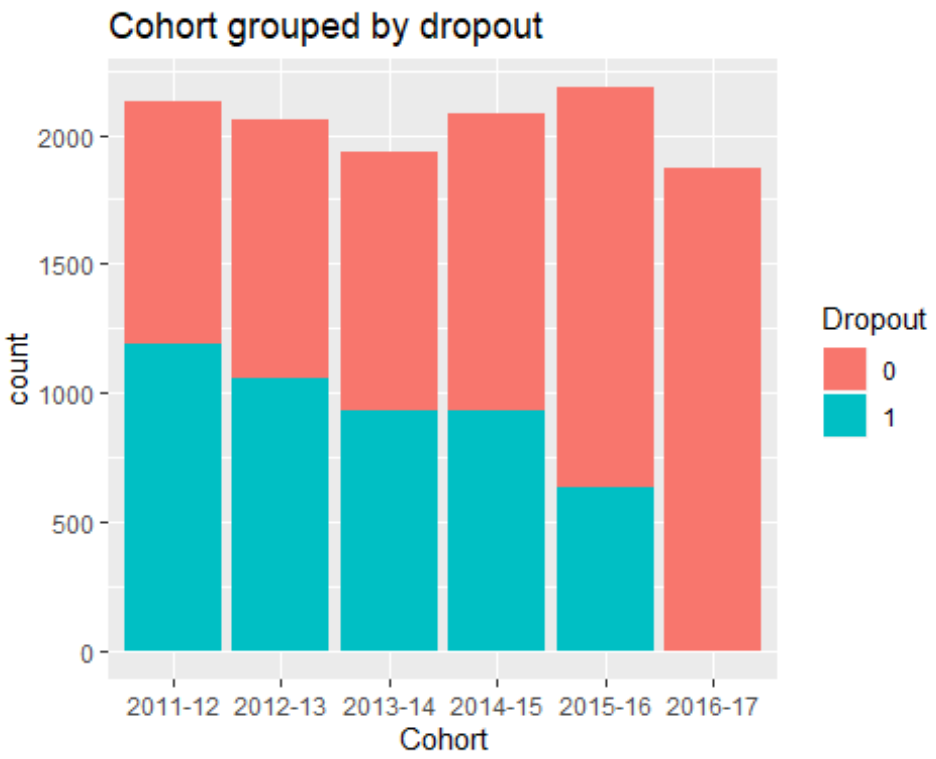

```
# Mother.s.Highest.Grade.Level
table(TRAIN_DATA$Mother.s.Highest.Grade.Level,TRAIN_DATA$Dropout,exclude =
NULL)

##
##                     0    1
##                     0    0
##   College        1824 1072
##   High School    2827 1689
##   Middle School   698  455
##   Unknown        1055  480
##   <NA>           1123 1038

ggplot(data = TRAIN_DATA) +
  ggtitle("Mother.s.Highest.Grade.Level grouped by dropout")+
  geom_bar(mapping = aes(x = Mother.s.Highest.Grade.Level,fill=Dropout))
```



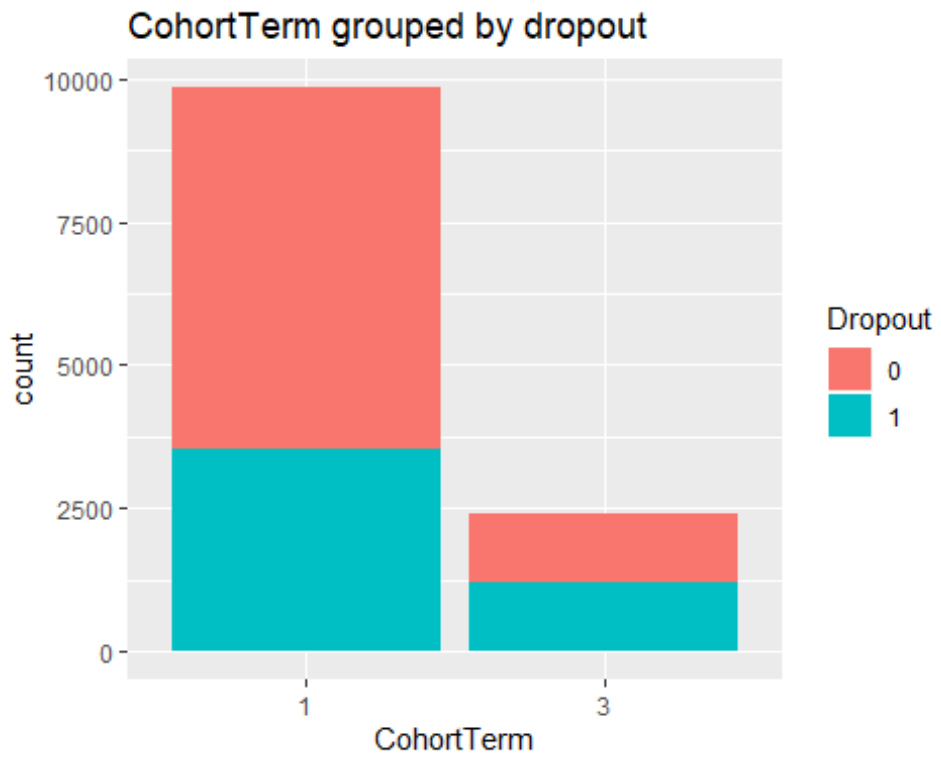

```
# Housing
table(TRAIN_DATA$Housing,TRAIN_DATA$Dropout,exclude = NULL)

##
##                         0    1
##                         0    0
##   Off Campus         3044 1802
##   On Campus Housing   884  546
##   With Parent        2697 1423
##   <NA>                902  963

ggplot(data = TRAIN_DATA) +
 ggtitle("Housing grouped by dropout")+
    geom_bar(mapping = aes(x = Housing,fill=Dropout))
```



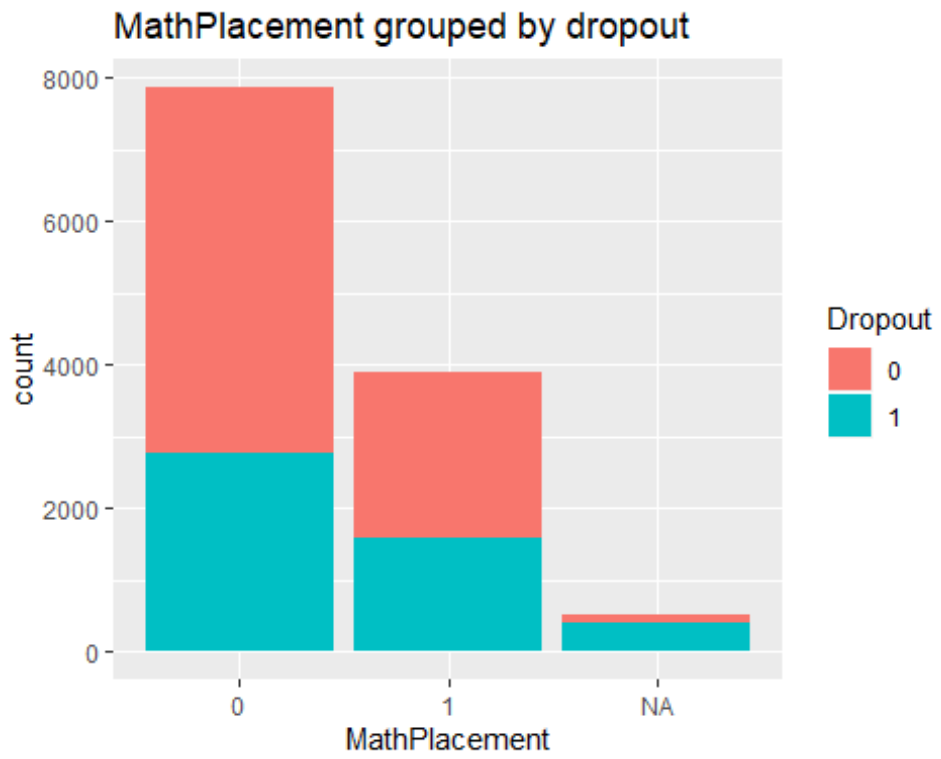

```
# Cohort
table(TRAIN_DATA$Cohort,TRAIN_DATA$Dropout,exclude = NULL)

##
##                0    1
##   2011-12  943 1188
##   2012-13 1005 1054
##   2013-14 1008  928
##   2014-15 1148  932
##   2015-16 1552  632
##   2016-17 1871    0

ggplot(data = TRAIN_DATA) +
 ggtitle("Cohort grouped by dropout")+
   geom_bar(mapping = aes(x = Cohort,fill=Dropout))
```




```
# CohortTerm
table(TRAIN_DATA$CohortTerm,TRAIN_DATA$Dropout,exclude = NULL)

##
##        0    1
##   1 6309 3542
##   3 1218 1192

ggplot(data = TRAIN_DATA) +
ggtitle("CohortTerm grouped by dropout")+
    geom_bar(mapping = aes(x = CohortTerm,fill=Dropout))
```



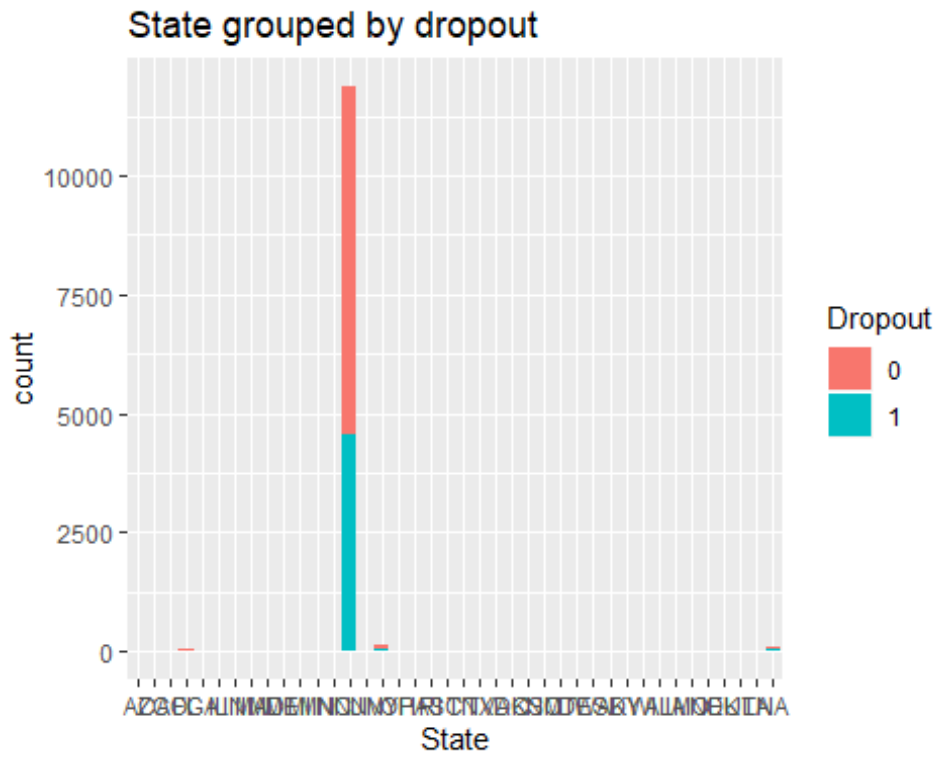

```
# MathPlacement
table(TRAIN_DATA$MathPlacement,TRAIN_DATA$Dropout,exclude = NULL)

##
##            0    1
##   0     5104 2755
##   1     2305 1577
##   <NA>   118  402

ggplot(data = TRAIN_DATA) +
ggtitle("MathPlacement grouped by dropout")+
    geom_bar(mapping = aes(x = MathPlacement,fill=Dropout))
```




```
# State
table(TRAIN_DATA$State,TRAIN_DATA$Dropout,exclude = NULL)

##
##           0    1
##           0    0
##    AZ     3    1
##    CA     7    9
##    CO     1    1
##    FL    17   12
##    GA     7    1
##    IL     2    1
##    IN     1    2
##    MA     2    5
##    MD    10    5
##    ME     1    0
##    MI     1    1
##    MN     1    2
##    NC     2    2
##    NJ  7316 4553
##    NV     2    0
##    NY    66   54
##    OH     2    2
##    PA     8    6
##    RI     1    1
##    SC     2    1
##    TN     1    0
```

```
##    TX       6    4
##    VA       2    3
##    DC       1    1
##    KS       1    1
##    NM       1    1
##    CT       2    1
##    DE       1    0
##    WA       2    0
##    SD       0    1
##    KY       0    1
##    WI       3    3
##    AL       2    1
##    IA       2    1
##    MO       1    0
##    NE       1    0
##    OK       1    0
##    UT       1    0
##    LA       0    1
##    <NA>    47   56
```

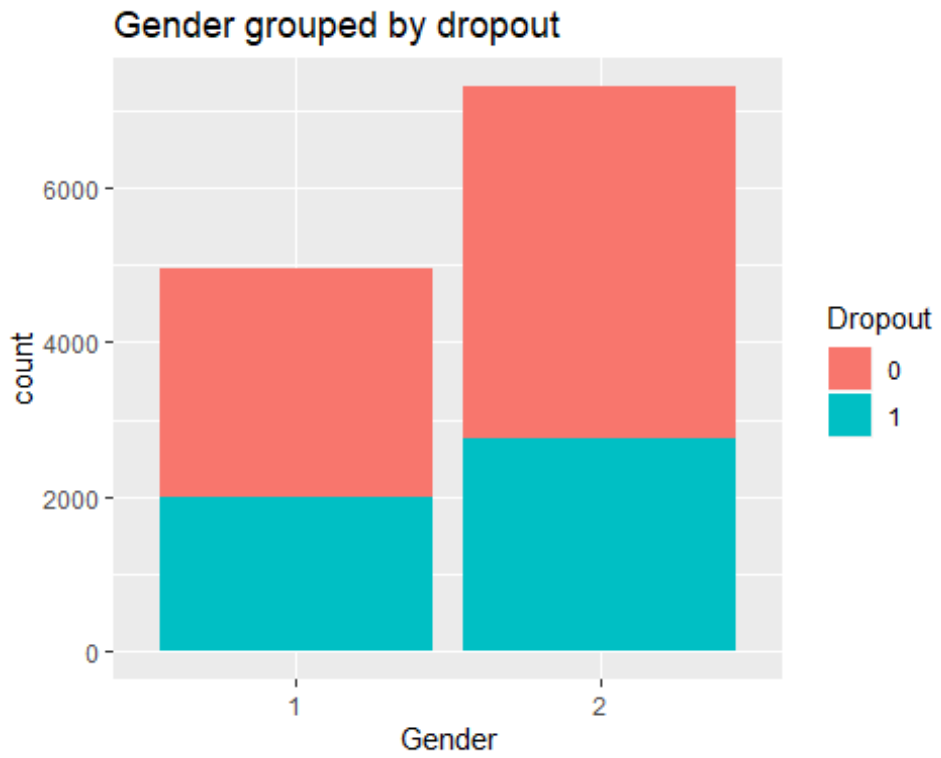
```
ggplot(data = TRAIN_DATA) +
ggtitle("State grouped by dropout")+
    geom_bar(mapping = aes(x = State,fill=Dropout))
```




```
# Gender
table(TRAIN_DATA$Gender,TRAIN_DATA$Dropout,exclude = NULL)
```

```
##
##        0    1
##   1 2958 1989
##   2 4569 2745
```

```
ggplot(data = TRAIN_DATA) +
ggtitle("Gender grouped by dropout")+
    geom_bar(mapping = aes(x = Gender,fill=Dropout))
```

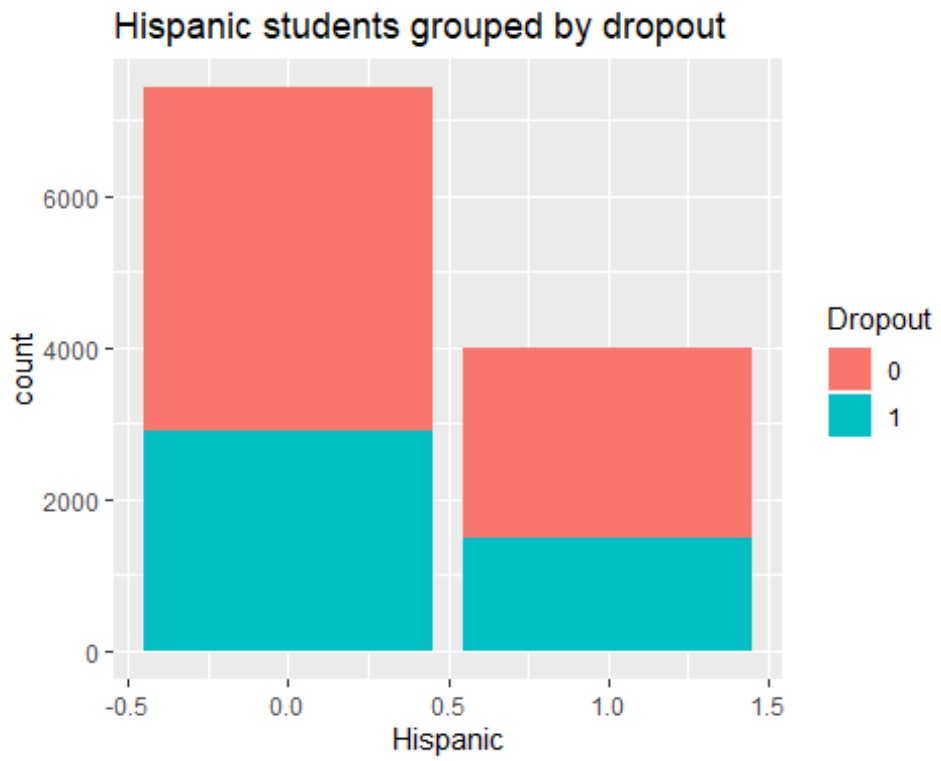



```
# Hispanic
table(TRAIN_DATA$Hispanic,TRAIN_DATA$Dropout,exclude = NULL)
```

```
##
##           0    1
##   0    4528 2901
##   1    2499 1491
##   <NA>  500  342
```

```
ggplot(data = TRAIN_DATA) +
ggtitle("Hispanic students grouped by dropout")+
    geom_bar(mapping = aes(x = Hispanic,fill=Dropout))
```

```
## Warning: Removed 842 rows containing non-finite values (stat_count).
```

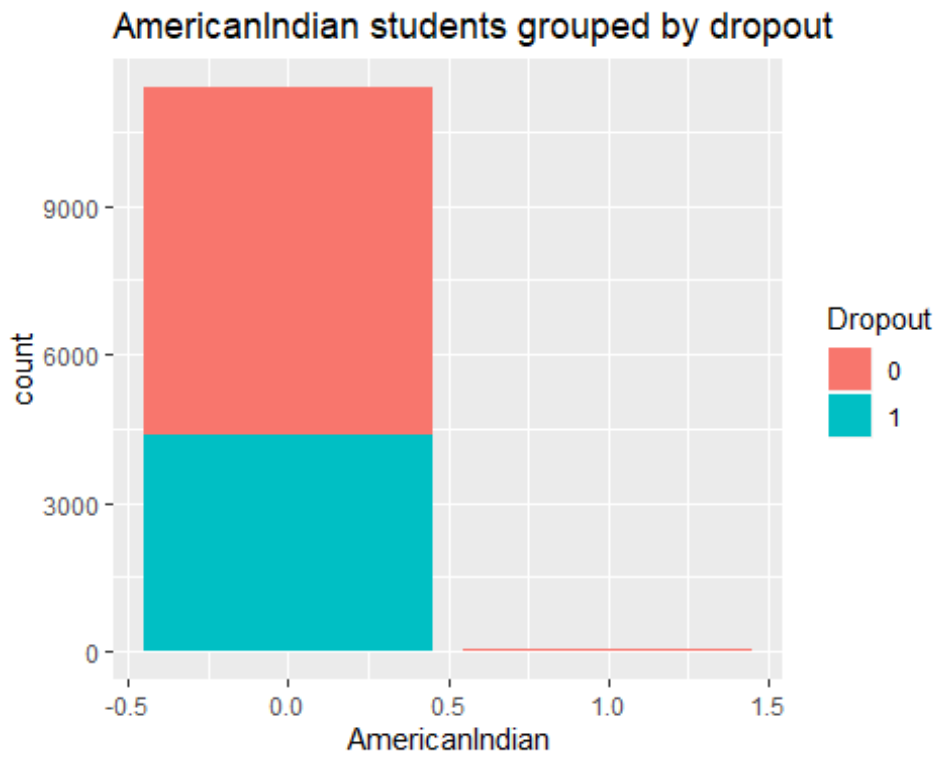



```
# AmericanIndian
table(TRAIN_DATA$AmericanIndian,TRAIN_DATA$Dropout,exclude = NULL)
```

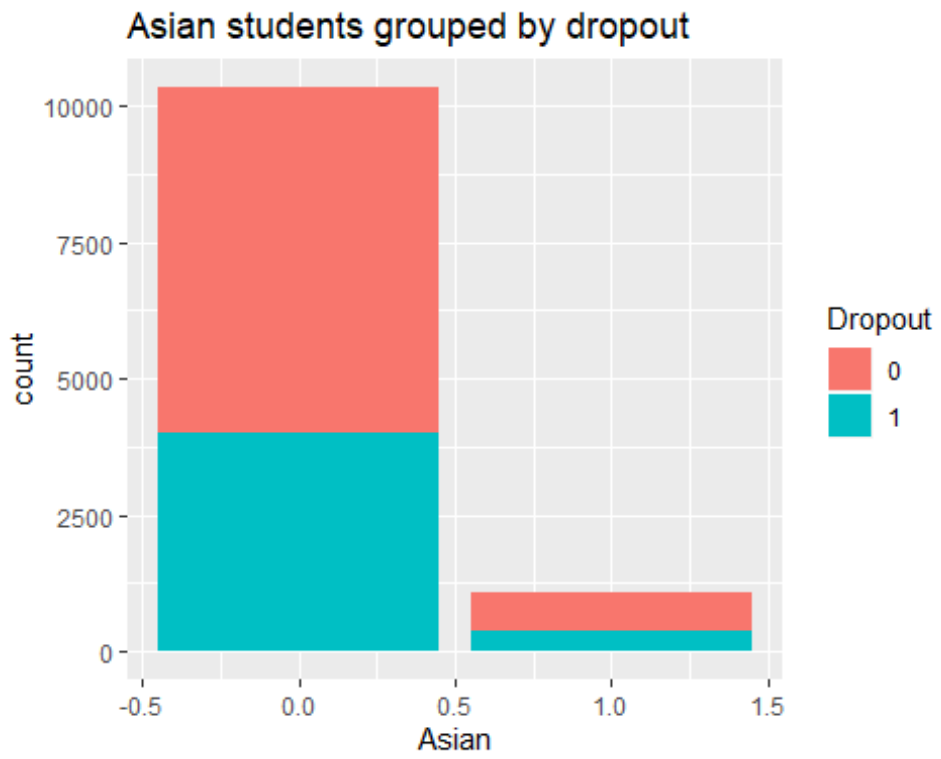
```
##
##            0    1
##   0    7010 4386
##   1      17    6
##   <NA>  500  342
```

```
ggplot(data = TRAIN_DATA) +
ggtitle("AmericanIndian students grouped by dropout")+
    geom_bar(mapping = aes(x = AmericanIndian,fill=Dropout))
```

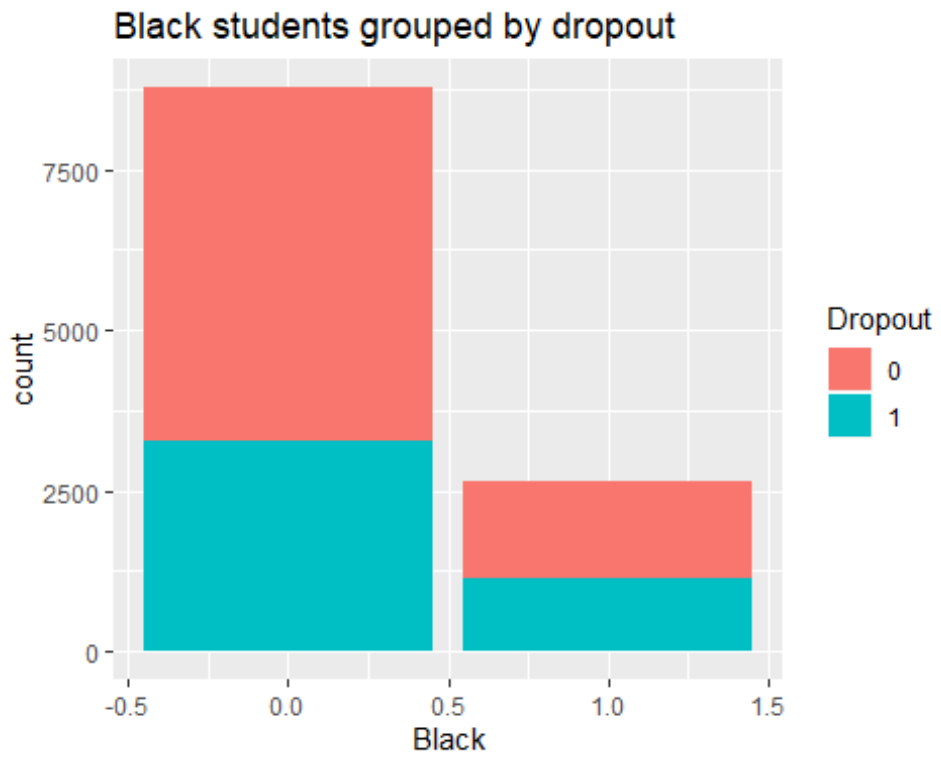
```
## Warning: Removed 842 rows containing non-finite values (stat_count).
```



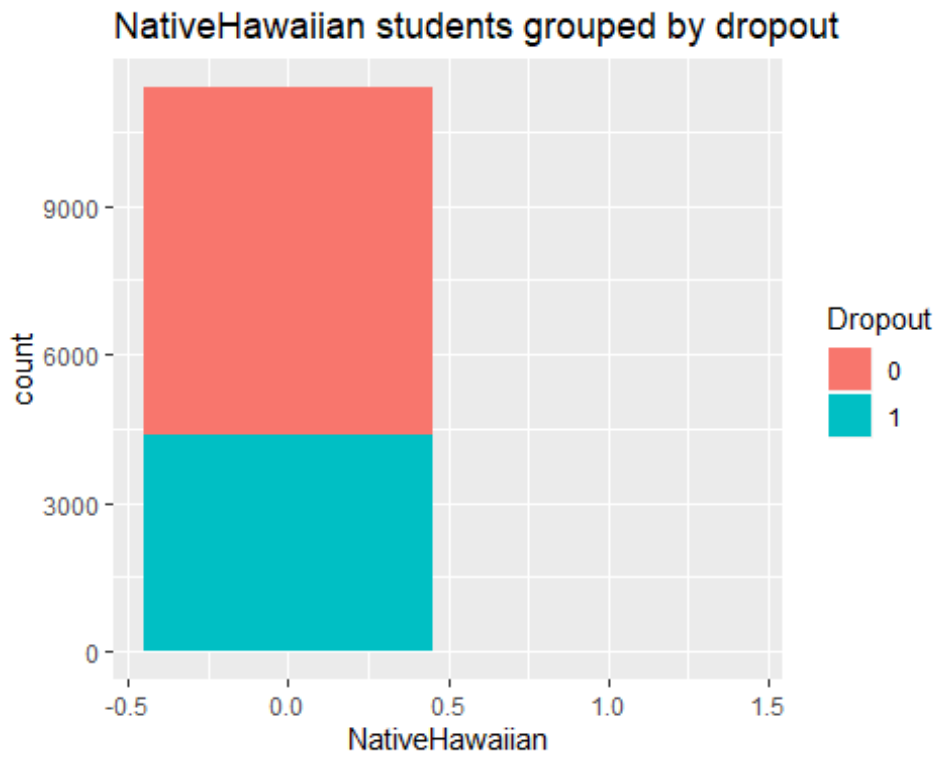

```
# Asian
table(TRAIN_DATA$Asian,TRAIN_DATA$Dropout,exclude = NULL)

##
##           0    1
##   0    6318 4017
##   1     709  375
##   <NA>  500  342

ggplot(data = TRAIN_DATA) +
ggtitle("Asian students grouped by dropout")+
    geom_bar(mapping = aes(x = Asian,fill=Dropout))

## Warning: Removed 842 rows containing non-finite values (stat_count).
```



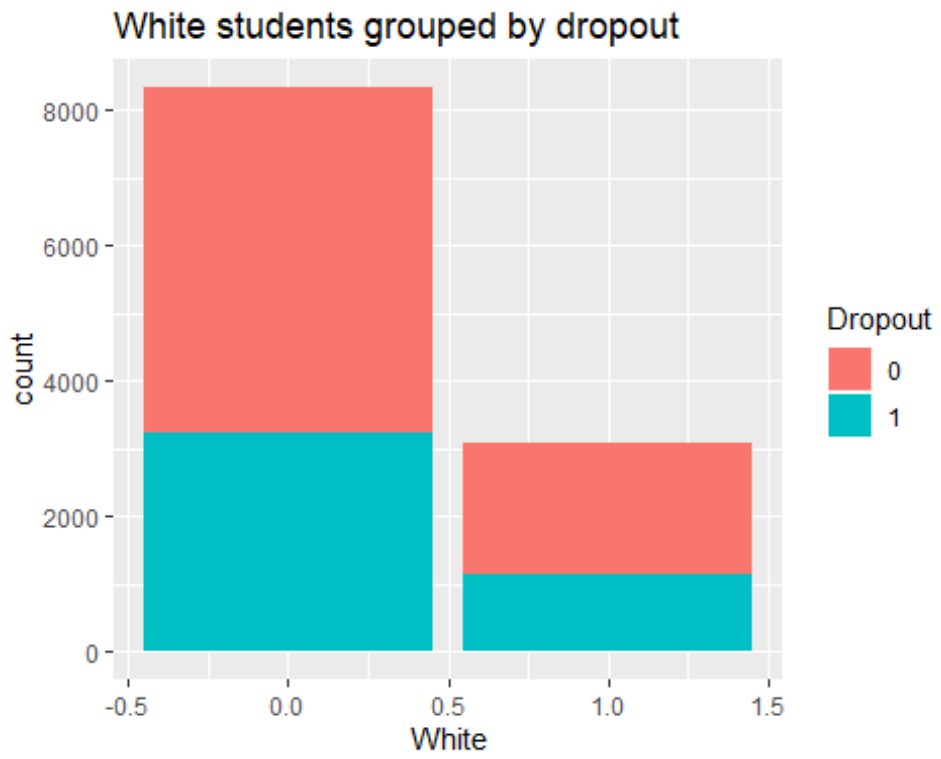

```
# Black
table(TRAIN_DATA$Black,TRAIN_DATA$Dropout,exclude = NULL)

##
##           0    1
##   0    5514 3264
##   1    1513 1128
##   <NA>  500  342

ggplot(data = TRAIN_DATA) +
ggtitle("Black students grouped by dropout")+
    geom_bar(mapping = aes(x = Black,fill=Dropout))

## Warning: Removed 842 rows containing non-finite values (stat_count).
```



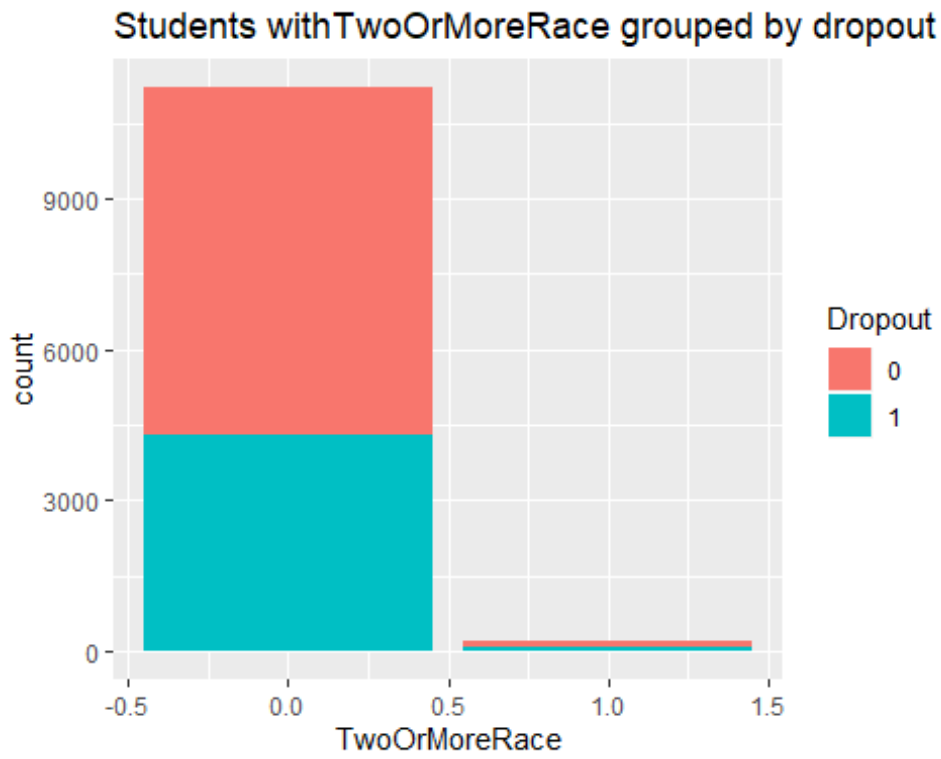

```
# NativeHawaiian
table(TRAIN_DATA$NativeHawaiian,TRAIN_DATA$Dropout,exclude = NULL)

##
##            0    1
##   0    7007 4391
##   1      20    1
##   <NA>  500  342

ggplot(data = TRAIN_DATA) +
  ggtitle("NativeHawaiian students grouped by dropout")+
  geom_bar(mapping = aes(x = NativeHawaiian,fill=Dropout))

## Warning: Removed 842 rows containing non-finite values (stat_count).
```



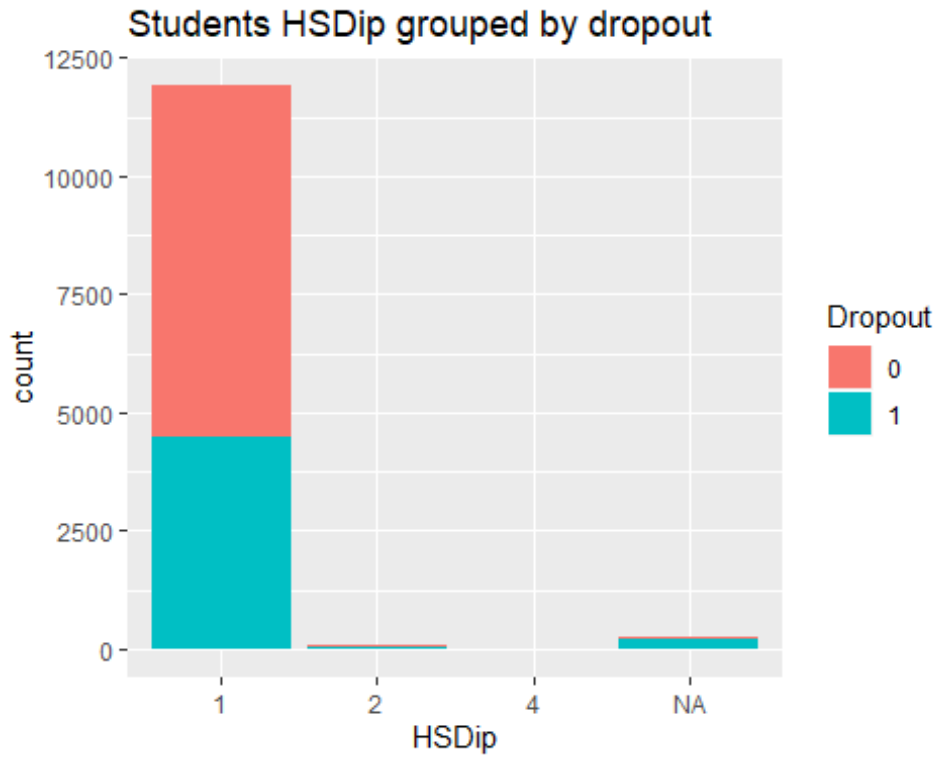

```
# White
table(TRAIN_DATA$White,TRAIN_DATA$Dropout,exclude = NULL)

##
##           0    1
##   0    5099 3242
##   1    1928 1150
##   <NA>  500  342

ggplot(data = TRAIN_DATA) +
ggtitle("White students grouped by dropout")+
    geom_bar(mapping = aes(x = White,fill=Dropout))

## Warning: Removed 842 rows containing non-finite values (stat_count).
```




```
# TwoOrMoreRace
table(TRAIN_DATA$TwoOrMoreRace,TRAIN_DATA$Dropout,exclude = NULL)

##
##            0    1
##   0    6900 4305
##   1     127   87
##   <NA>  500  342

ggplot(data = TRAIN_DATA) +
ggtitle("Students withTwoOrMoreRace grouped by dropout")+
    geom_bar(mapping = aes(x = TwoOrMoreRace,fill=Dropout))

## Warning: Removed 842 rows containing non-finite values (stat_count).
```



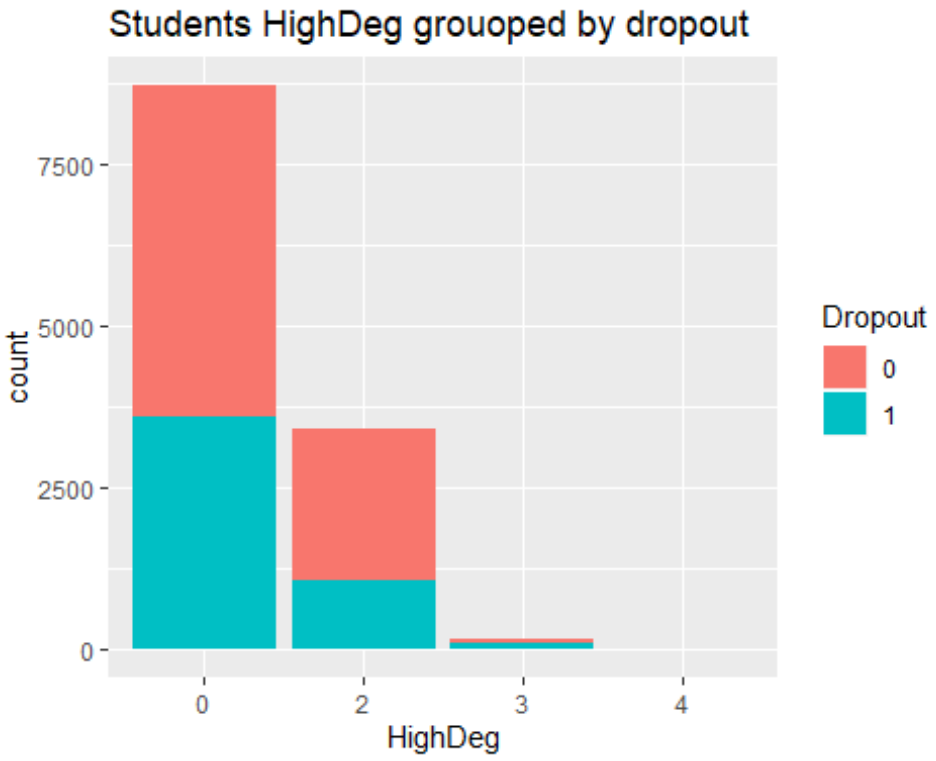

```
# HSDip
table(TRAIN_DATA$HSDip,TRAIN_DATA$Dropout,exclude = NULL)

##
##            0    1
##   1     7426 4490
##   2       26   43
##   4        6    4
##   <NA>    69  197

ggplot(data = TRAIN_DATA) +
ggtitle("Students HSDip grouped by dropout")+
    geom_bar(mapping = aes(x = HSDip,fill=Dropout))
```



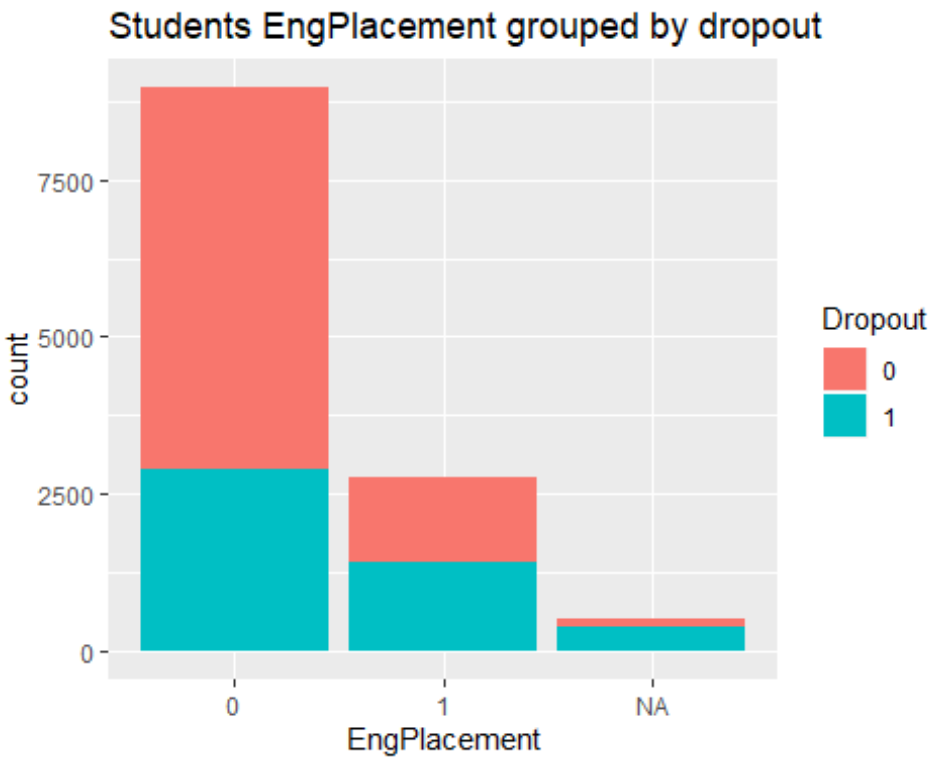

```
# HighDeg
table(TRAIN_DATA$HighDeg,TRAIN_DATA$Dropout,exclude = NULL)
```

```
##
##        0    1
##   0 5132 3578
##   2 2335 1071
##   3   59   84
##   4    1    1
```

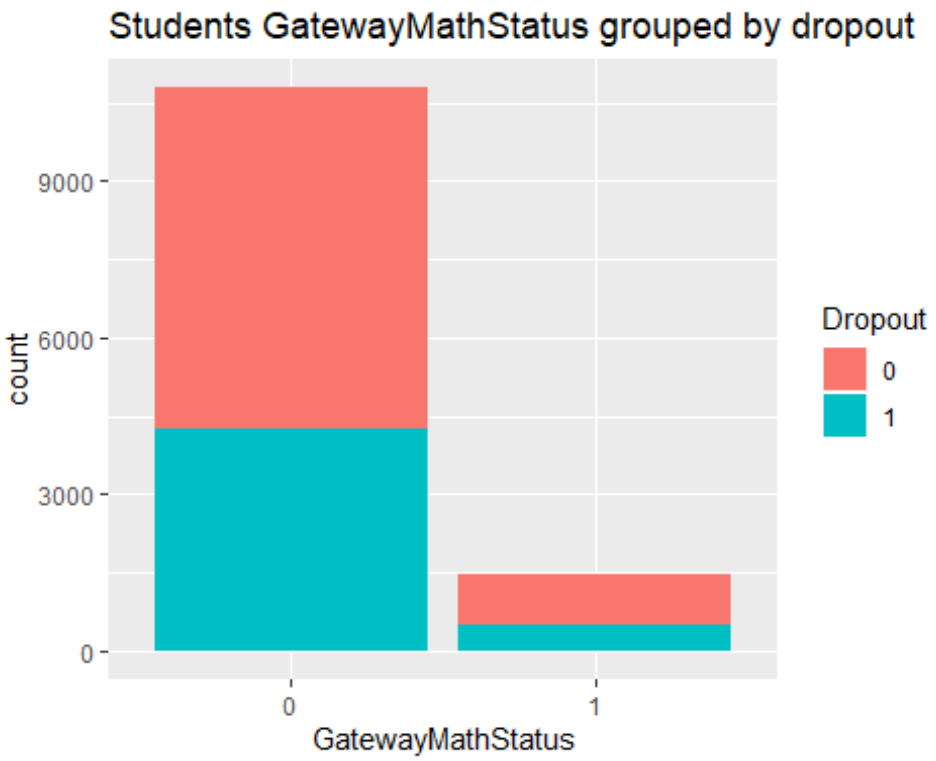
```
ggplot(data = TRAIN_DATA) +
  ggtitle('Students HighDeg grouoped by dropout')+
  geom_bar(mapping = aes(x = HighDeg,fill=Dropout))
```




```
# EngPlacement
table(TRAIN_DATA$EngPlacement,TRAIN_DATA$Dropout,exclude = NULL)

##
##           0    1
##   0    6063 2903
##   1    1346 1429
##   <NA>  118  402

ggplot(data = TRAIN_DATA) +
  ggtitle('Students EngPlacement grouped by dropout')+
  geom_bar(mapping = aes(x = EngPlacement,fill=Dropout))
```

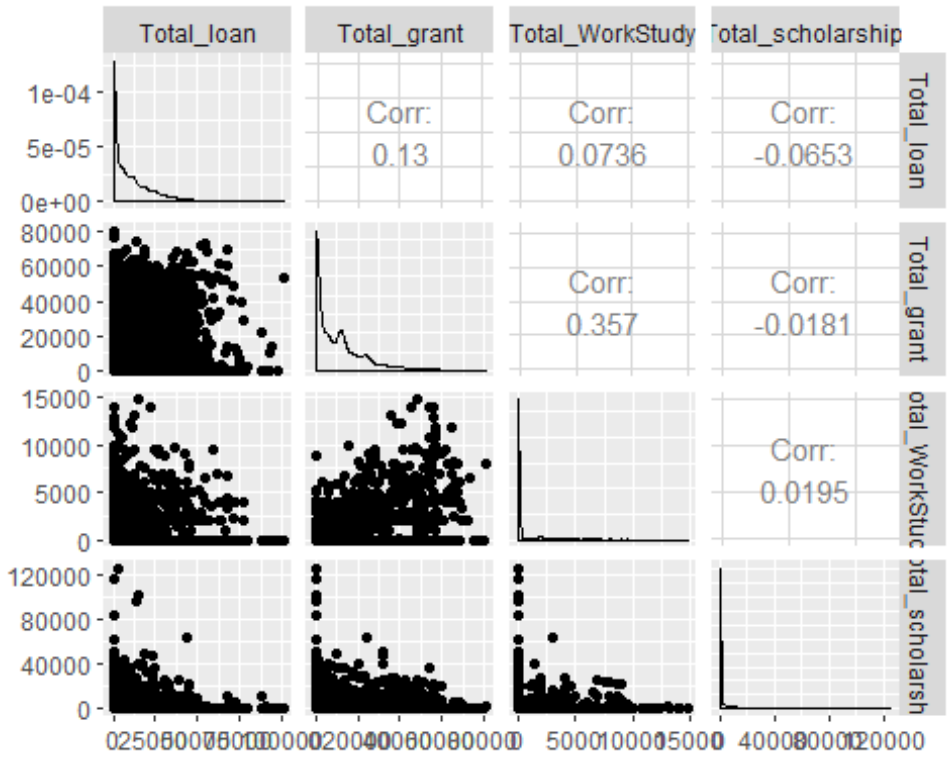



```
# GatewayMathStatus
table(TRAIN_DATA$GatewayMathStatus,TRAIN_DATA$Dropout)

##
##        0    1
##   0 6551 4243
##   1  976  491

ggplot(data = TRAIN_DATA) +
ggtitle('Students GatewayMathStatus grouped by dropout')+
    geom_bar(mapping = aes(x = GatewayMathStatus,fill=Dropout))
```

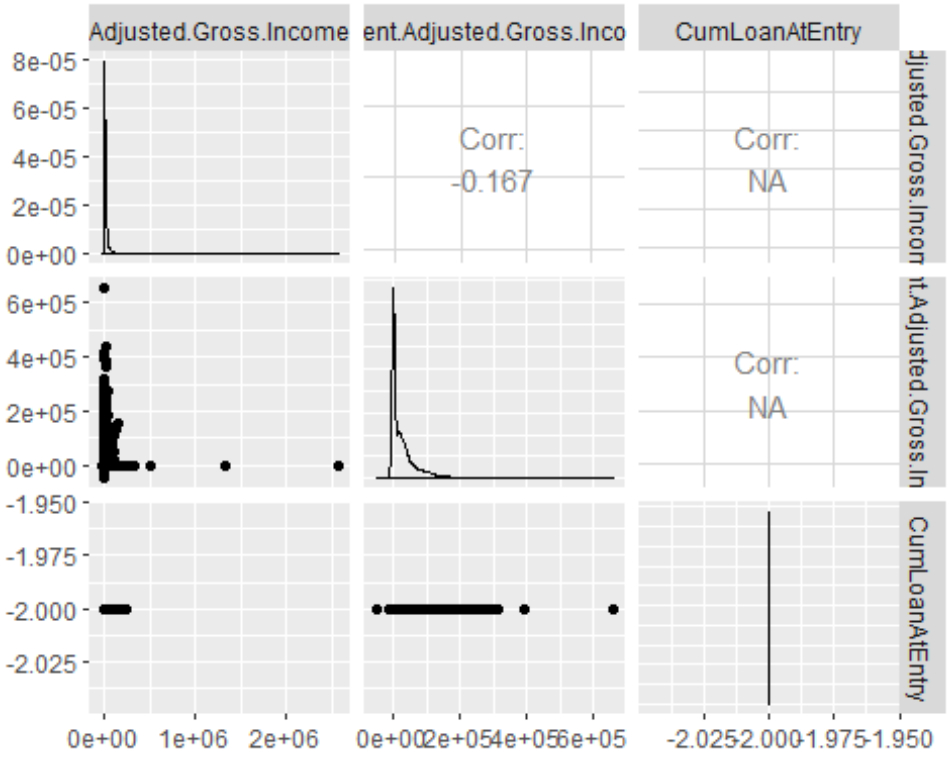


Correlation matrix of the financial aid data to check for multicollinearity among the continuous independent variables (The correlation is expected to be less than 0.5 so as to facilitate regression with multicollinearity)

```
# Load the GGally package
library(GGally)

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2

##
## Attaching package: 'GGally'

## The following object is masked from 'package:dplyr':
##
##     nasa

# Create a scatter plot matrix
vars <- c("Total_loan", "Total_grant", "Total_WorkStudy",
"Total_scholarship")
ggpairs(TRAIN_DATA[vars])
```




```r
# Load the GGally package
library(GGally)
# Create a scatter plot matrix
vars <-
c("Adjusted.Gross.Income","Parent.Adjusted.Gross.Income","CumLoanAtEntry")
ggpairs(TRAIN_DATA[vars])
```



Hypotheses Tests

Hypotheses tests to test the relation between individual variables and the dependent variable (Dropout) - The Null hypothesis: These variables have no relationship with Dropout. - The Alternative hypothesis: These variables have a relationshop with Dropout. - Decision rule: We reject Null hypothesis if the P-value is less than 0.05 and conclude that the variables have a relationship with Dropout and that Dropout of students is dependent on these variables.

```
# Test for association
# Null hypothesis: The dropout of students is not related to gender of a
student
attach(TRAIN_DATA)

## The following objects are masked from financial_aid:
##
##     Adjusted.Gross.Income, cohort_term, Father.s.Highest.Grade.Level,
##     Housing, Marital.Status, Mother.s.Highest.Grade.Level,
##     Parent.Adjusted.Gross.Income, StudentID

chisq.test(Dropout,Gender)

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  Dropout and Gender
## X-squared = 8.7991, df = 1, p-value = 0.003014
```

```r
# Null hypothesis: The dropout of students is not related to Marital.Status
of a student
chisq.test(Marital.Status,Dropout)
```

```
##
##  Pearson's Chi-squared test
##
## data:  Marital.Status and Dropout
## X-squared = 6.2593, df = 3, p-value = 0.09965
```

```r
# Null hypothesis: The dropout of students is not related to the housing of a
student
chisq.test(Housing,Dropout)
```

```
##
##  Pearson's Chi-squared test
##
## data:  Housing and Dropout
## X-squared = 9.3589, df = 2, p-value = 0.009284
```

```r
 #Null hypothesis: The dropout of students is not related to the
Father.s.Highest.Grade.Level
chisq.test(Father.s.Highest.Grade.Level,Dropout)
```

```
##
##  Pearson's Chi-squared test
##
## data:  Father.s.Highest.Grade.Level and Dropout
## X-squared = 24.901, df = 3, p-value = 1.62e-05
```

```r
# Null hypothesis: The dropout of students is not related to the
Mother.s.Highest.Grade.Level
chisq.test(Mother.s.Highest.Grade.Level,Dropout)
```

```
##
##  Pearson's Chi-squared test
##
## data:  Mother.s.Highest.Grade.Level and Dropout
## X-squared = 24.326, df = 3, p-value = 2.136e-05
```

```r
# Null hypothesis: The dropout of students is not related to
GatewayMathStatus of the student
chisq.test(GatewayEnglishStatus,Dropout)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  GatewayEnglishStatus and Dropout
## X-squared = 99.031, df = 1, p-value < 2.2e-16
```

NOTE: Most variables had a relationship with dropout except marital status with p-value: 0.09 (p-value>0.05) where we do not reject null hypothesis and conclude that there is no

enough statistical evidence to support the claim that dropout of students is related to mariatal status of students.

Imputation of Missing Values

Summarize and look for the count of NA's in Individual variables

```
summary(TRAIN_DATA)
```

```
##    StudentID      Dropout  cohort_term   Marital.Status
Adjusted.Gross.Income
##  Min.   : 20932   0:7527   1:9851              :   0   Min.   : -24326
##  1st Qu.:305164   1:4734   3:2410      Divorced : 208   1st Qu.:      0
##  Median :321580                       Married  : 924   Median :   2768
##  Mean   :316079                       Separated: 185   Mean   :  13263
##  3rd Qu.:343608                       Single   :9103   3rd Qu.:  16491
##  Max.   :359783                       NA's     :1841   Max.   :2576425
##                                                        NA's   :1841
##  Parent.Adjusted.Gross.Income Father.s.Highest.Grade.Level
##  Min.   :-49406                          :   0
##  1st Qu.:     0               College    :2916
##  Median : 12373               High School  :4578
##  Mean   : 28318               Middle School:1201
##  3rd Qu.: 38805               Unknown    :1598
##  Max.   :657631               NA's       :1968
##  NA's   :1841
##  Mother.s.Highest.Grade.Level             Housing       Total_loan
##                :   0                           :   0   Min.   :     0
##  College       :2896          Off Campus       :4846   1st Qu.:     0
##  High School   :4516          On Campus Housing:1430   Median :  3745
##  Middle School:1153           With Parent      :4120   Mean   :  8834
##  Unknown       :1535          NA's             :1865   3rd Qu.: 13429
##  NA's          :2161                                   Max.   :100960
##
##   Total_grant    Total_scholarship Total_WorkStudy       Cohort
CohortTerm
##  Min.   :    0   Min.   :     0    Min.   :    0.0   2011-12:2131   1:9851
##  1st Qu.:    0   1st Qu.:     0    1st Qu.:    0.0   2012-13:2059   3:2410
##  Median : 5265   Median :     0    Median :    0.0   2013-14:1936
##  Mean   : 9690   Mean   :  1170    Mean   :  208.5   2014-15:2080
##  3rd Qu.:14100   3rd Qu.:     0    3rd Qu.:    0.0   2015-16:2184
##  Max.   :80873   Max.   :125497    Max.   :14820.0   2016-17:1871
##
##   Campus                          Address1                          Address2
##  NA's:12261   NJCU-Registrar's Office:    6   1                         :
14
##               Summit Apts            :    5   2                         :
11
##               Jackson Garden Apt     :    4   Apt 2                     :
10
```

```
##               Westview Towers        :    4   2039 John F Kennedy Blvd:
6
##               John F                 :    4   2nd Floor               :
5
##               (Other)                :12135   (Other)                 :
309
##               NA's                   :  103   NA's
:11906
##            City          State            Zip        RegistrationDate
##  Jersey City :3285   NJ     :11869   Min.   :  747   Min.   :20110111
##  Bayonne     :1138   NY     :  120   1st Qu.: 7060   1st Qu.:20120710
##  Newark      : 683   FL     :   29   Median : 7304   Median :20140122
##  North Bergen: 557   CA     :   16   Mean   : 7800   Mean   :20136172
##  Union City  : 549   MD     :   15   3rd Qu.: 7307   3rd Qu.:20150624
##  (Other)     :5945   (Other):  109   Max.   :98118   Max.   :20160912
##  NA's        : 104   NA's   :  103   NA's   :121
##  Gender     BirthYear      BirthMonth      Hispanic      AmericanIndian
##  1:4947   1993   :1173   9      :1119   Min.   :0.0000   Min.   :0.000
##  2:7314   1994   :1051   7      :1098   1st Qu.:0.0000   1st Qu.:0.000
##           1995   : 864   8      :1093   Median :0.0000   Median :0.000
##           1992   : 832   1      :1058   Mean   :0.3494   Mean   :0.002
##           1996   : 811   10     :1029   3rd Qu.:1.0000   3rd Qu.:0.000
##           (Other):7529   12     :1028   Max.   :1.0000   Max.   :1.000
##           NA's   :   1   (Other):5836   NA's   :842      NA's   :842
##      Asian            Black        NativeHawaiian       White
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.0949   Mean   :0.2313   Mean   :0.0018   Mean   :0.2696
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  NA's   :842      NA's   :842      NA's   :842      NA's   :842
##  TwoOrMoreRace     HSDip          HSDipYr       HSGPAUnwtd       HSGPAWtd
##  Min.   :0.0000   1   :11916   Min.   :1963   Min.   :0.900   Min.   : NA
##  1st Qu.:0.0000   2   :   69   1st Qu.:2011   1st Qu.:2.500   1st Qu.: NA
##  Median :0.0000   4   :   10   Median :2013   Median :2.880   Median : NA
##  Mean   :0.0187   NA's:  266   Mean   :2013   Mean   :2.909   Mean   :NaN
##  3rd Qu.:0.0000                3rd Qu.:2015   3rd Qu.:3.300   3rd Qu.: NA
##  Max.   :1.0000                Max.   :2016   Max.   :4.000   Max.   : NA
##  NA's   :842                   NA's   :8921   NA's   :8687    NA's
:12261
##  FirstGen     DualHSSummerEnroll EnrollmentStatus
NumColCredAttemptTransfer
##  NA's:12261   Min.   :0          1:4952           Min.   : -2.00
##               1st Qu.:0          2:7309           1st Qu.: -2.00
##               Median :0                           Median : 24.00
##               Mean   :0                           Mean   : 38.66
##               3rd Qu.:0                           3rd Qu.: 74.00
##               Max.   :0                           Max.   :150.00
##                                                   NA's   :370
```

```
##  NumColCredAcceptTransfer CumLoanAtEntry HighDeg  MathPlacement
EngPlacement
##  Min.   :-2.00            Min.   :-2     0:8710   0   :7859     0   :8966
##  1st Qu.:-2.00            1st Qu.:-2     2:3406   1   :3882     1   :2775
##  Median :24.00            Median :-2     3: 143   NA's: 520     NA's: 520
##  Mean   :32.14            Mean   :-2     4:   2
##  3rd Qu.:66.00            3rd Qu.:-2
##  Max.   :96.00            Max.   :-2
##  NA's   :1                NA's   :7309
##  GatewayMathStatus GatewayEnglishStatus CompleteDevMath CompleteDevEnglish
##  0:10794           0:9967               -2     :7854    -2     :8860
##  1: 1467           1:2294               0      :1478    0      : 773
##                                         0.5    : 443    0.5    : 319
##                                         0.25   : 379    1      : 311
##                                         1      : 213    0.25   : 197
##                                         (Other):1371    (Other):1274
##                                         NA's   : 523    NA's   : 527
##      Major1          Major2         Complete1        Complete2
##  Min.   :-0.50   Min.   : 0.003   Min.   :0.0000   Min.   :0
##  1st Qu.:26.01   1st Qu.: 6.060   1st Qu.:0.0000   1st Qu.:0
##  Median :43.02   Median : 9.575   Median :0.0000   Median :0
##  Mean   :37.02   Mean   :12.564   Mean   :0.4482   Mean   :0
##  3rd Qu.:51.38   3rd Qu.:13.121   3rd Qu.:0.7778   3rd Qu.:0
##  Max.   :54.01   Max.   :52.140   Max.   :4.0000   Max.   :0
##  NA's   :129     NA's   :11480
##   CompleteCIP1      CompleteCIP2 TransferIntent DegreeTypeSought
##  Min.   :-2.0000   Min.   :-2    NA's:12261     6:12261
##  1st Qu.:-2.0000   1st Qu.:-2
##  Median :-2.0000   Median :-2
##  Mean   : 0.7489   Mean   :-2
##  3rd Qu.: 2.0927   3rd Qu.:-2
##  Max.   :26.0051   Max.   :-2
##
##     TermGPA          CumGPA
##  Min.   :0.000   Min.   :0.000
##  1st Qu.:2.395   1st Qu.:2.395
##  Median :3.075   Median :3.075
##  Mean   :2.817   Mean   :2.817
##  3rd Qu.:3.578   3rd Qu.:3.578
##  Max.   :4.000   Max.   :4.000
##
```

List the columns with missing values

```
colnames(TRAIN_DATA)[colSums(is.na(TRAIN_DATA)) > 0]
```

```
##  [1] "Marital.Status"               "Adjusted.Gross.Income"
##  [3] "Parent.Adjusted.Gross.Income" "Father.s.Highest.Grade.Level"
##  [5] "Mother.s.Highest.Grade.Level" "Housing"
##  [7] "Campus"                       "Address1"
```

```
##  [9] "Address2"                      "City"
## [11] "State"                         "Zip"
## [13] "BirthYear"                     "Hispanic"
## [15] "AmericanIndian"                "Asian"
## [17] "Black"                         "NativeHawaiian"
## [19] "White"                         "TwoOrMoreRace"
## [21] "HSDip"                         "HSDipYr"
## [23] "HSGPAUnwtd"                    "HSGPAWtd"
## [25] "FirstGen"                      "NumColCredAttemptTransfer"
## [27] "NumColCredAcceptTransfer"      "CumLoanAtEntry"
## [29] "MathPlacement"                 "EngPlacement"
## [31] "CompleteDevMath"               "CompleteDevEnglish"
## [33] "Major1"                        "Major2"
## [35] "TransferIntent"
```

Looking at the summary statistics of the TRAIN data frame there are variables:('HSGPAWtd','FirstGen','TransferIntent','Campus',,'Major2') that have missing values close to 90% of the whole column (>11,000), these variables if imputed may introduce bias, hence they were dropped.

- Also drop other variables that are string/nominal in nature: (state, address1, zip, Birthyear, BirthMonth and city) which even if they are imputed they have too many levels hence would increase number predictors unnecessarily after 'OneHot' encoding of categorical variables is done later hence too much bias.

```
library(tidyverse)

drop.cols <- c('HSGPAWtd', 'FirstGen','TransferIntent','Campus','Address1',
'Address2','Major2','State','Zip','BirthYear','BirthMonth','City')

TRAIN_DATA<-TRAIN_DATA %>% select(-drop.cols)
dim(TRAIN_DATA)

## [1] 12261    47

# which columns have missing data
colnames(TRAIN_DATA)[colSums(is.na(TRAIN_DATA)) > 0]

##  [1] "Marital.Status"               "Adjusted.Gross.Income"
##  [3] "Parent.Adjusted.Gross.Income" "Father.s.Highest.Grade.Level"
##  [5] "Mother.s.Highest.Grade.Level" "Housing"
##  [7] "Hispanic"                     "AmericanIndian"
##  [9] "Asian"                        "Black"
## [11] "NativeHawaiian"               "White"
## [13] "TwoOrMoreRace"                "HSDip"
## [15] "HSDipYr"                      "HSGPAUnwtd"
## [17] "NumColCredAttemptTransfer"    "NumColCredAcceptTransfer"
## [19] "CumLoanAtEntry"               "MathPlacement"
## [21] "EngPlacement"                 "CompleteDevMath"
## [23] "CompleteDevEnglish"           "Major1"
```

Impute categorical variables with mode (most common category level in the variable)

```
val<-unique(TRAIN_DATA$Marital.Status[!is.na(TRAIN_DATA$Marital.Status)]) #
Values in vec_miss
mode <- val[which.max(tabulate(match(TRAIN_DATA$Marital.Status, val)))] #
Mode of vec_miss
TRAIN_DATA$Marital.Status[is.na(TRAIN_DATA$Marital.Status)]<-mode # Impute by
mode

val<-
unique(TRAIN_DATA$Father.s.Highest.Grade.Level[!is.na(TRAIN_DATA$Father.s.Hig
hest.Grade.Level)])
mode<-val[which.max(tabulate(match(TRAIN_DATA$Father.s.Highest.Grade.Level,
val)))]
TRAIN_DATA$Father.s.Highest.Grade.Level[is.na(TRAIN_DATA$Father.s.Highest.Gra
de.Level)]<-mode

val<-
unique(TRAIN_DATA$Mother.s.Highest.Grade.Level[!is.na(TRAIN_DATA$Mother.s.Hig
hest.Grade.Level)])
mode<-val[which.max(tabulate(match(TRAIN_DATA$Mother.s.Highest.Grade.Level,
val)))]
TRAIN_DATA$Mother.s.Highest.Grade.Level[is.na(TRAIN_DATA$Mother.s.Highest.Gra
de.Level)]<-mode

val<-unique(TRAIN_DATA$Housing[!is.na(TRAIN_DATA$Housing)])
mode<-val[which.max(tabulate(match(TRAIN_DATA$Housing, val)))]
TRAIN_DATA$Housing[is.na(TRAIN_DATA$Housing)]<-mode


val<-unique(TRAIN_DATA$EngPlacement[!is.na(TRAIN_DATA$EngPlacement)])
mode<-val[which.max(tabulate(match(TRAIN_DATA$EngPlacement, val)))]
TRAIN_DATA$EngPlacement[is.na(TRAIN_DATA$EngPlacement)]<-mode

val<-unique(TRAIN_DATA$MathPlacement[!is.na(TRAIN_DATA$MathPlacement)])
mode<-val[which.max(tabulate(match(TRAIN_DATA$MathPlacement, val)))]
TRAIN_DATA$MathPlacement[is.na(TRAIN_DATA$MathPlacement)]<-mode

val<-unique(TRAIN_DATA$CompleteDevEnglish[!is.na(TRAIN_DATA$MathPlacement)])
mode<-val[which.max(tabulate(match(TRAIN_DATA$CompleteDevEnglish, val)))]
TRAIN_DATA$CompleteDevEnglish[is.na(TRAIN_DATA$CompleteDevEnglish)]<-mode

val<-unique(TRAIN_DATA$CompleteDevMath[!is.na(TRAIN_DATA$CompleteDevMath)])
mode<-val[which.max(tabulate(match(TRAIN_DATA$CompleteDevMath, val)))]
TRAIN_DATA$CompleteDevMath[is.na(TRAIN_DATA$CompleteDevMath)]<-mode


val<-unique(TRAIN_DATA$Hispanic[!is.na(TRAIN_DATA$Hispanic)])
mode<-val[which.max(tabulate(match(TRAIN_DATA$Hispanic, val)))]
```

```
TRAIN_DATA$Hispanic[is.na(TRAIN_DATA$Hispanic)]<-mode

val<-unique(TRAIN_DATA$AmericanIndian[!is.na(TRAIN_DATA$AmericanIndian)])
mode<-val[which.max(tabulate(match(TRAIN_DATA$AmericanIndian, val)))]
TRAIN_DATA$AmericanIndian[is.na(TRAIN_DATA$AmericanIndian)]<-mode

val<-unique(TRAIN_DATA$Asian[!is.na(TRAIN_DATA$Asian)])
mode<-val[which.max(tabulate(match(TRAIN_DATA$Asian, val)))]
TRAIN_DATA$Asian[is.na(TRAIN_DATA$Asian)]<-mode

val<-unique(TRAIN_DATA$Black[!is.na(TRAIN_DATA$Black)])
mode<-val[which.max(tabulate(match(TRAIN_DATA$Black, val)))]
TRAIN_DATA$Black[is.na(TRAIN_DATA$Black)]<-mode

val<-unique(TRAIN_DATA$NativeHawaiian[!is.na(TRAIN_DATA$NativeHawaiian)])
mode<-val[which.max(tabulate(match(TRAIN_DATA$NativeHawaiian, val)))]
TRAIN_DATA$NativeHawaiian[is.na(TRAIN_DATA$NativeHawaiian)]<-mode

val<-unique(TRAIN_DATA$White[!is.na(TRAIN_DATA$White)])
mode<-val[which.max(tabulate(match(TRAIN_DATA$White, val)))]
TRAIN_DATA$White[is.na(TRAIN_DATA$White)]<-mode

val<-unique(TRAIN_DATA$TwoOrMoreRace[!is.na(TRAIN_DATA$TwoOrMoreRace)])
mode<-val[which.max(tabulate(match(TRAIN_DATA$TwoOrMoreRace, val)))]
TRAIN_DATA$TwoOrMoreRace[is.na(TRAIN_DATA$TwoOrMoreRace)]<-mode

val<-unique(TRAIN_DATA$HSDip[!is.na(TRAIN_DATA$HSDip)])
mode<-val[which.max(tabulate(match(TRAIN_DATA$HSDip, val)))]
TRAIN_DATA$HSDip[is.na(TRAIN_DATA$HSDip)]<-mode

val<-unique(TRAIN_DATA$HSDipYr[!is.na(TRAIN_DATA$HSDipYr)])
mode<-val[which.max(tabulate(match(TRAIN_DATA$HSDipYr, val)))]
TRAIN_DATA$HSDipYr[is.na(TRAIN_DATA$HSDipYr)]<-mode

#Check which columns have missing data to ensure the imputed columns are not
there
colnames(TRAIN_DATA)[colSums(is.na(TRAIN_DATA)) > 0]

## [1] "Adjusted.Gross.Income"        "Parent.Adjusted.Gross.Income"
## [3] "HSGPAUnwtd"                   "NumColCredAttemptTransfer"
## [5] "NumColCredAcceptTransfer"     "CumLoanAtEntry"
## [7] "Major1"
```

Imputing continuous variables

- First check for normal distribution in order to impute with mean or median if they are not normally distributed (bell-shaped histogram and density plot shows normality)
- Adjusted.Gross.Income

```r
#Density plot
library(ggpubr)
```

```
## Loading required package: magrittr
```

```
##
## Attaching package: 'magrittr'
```

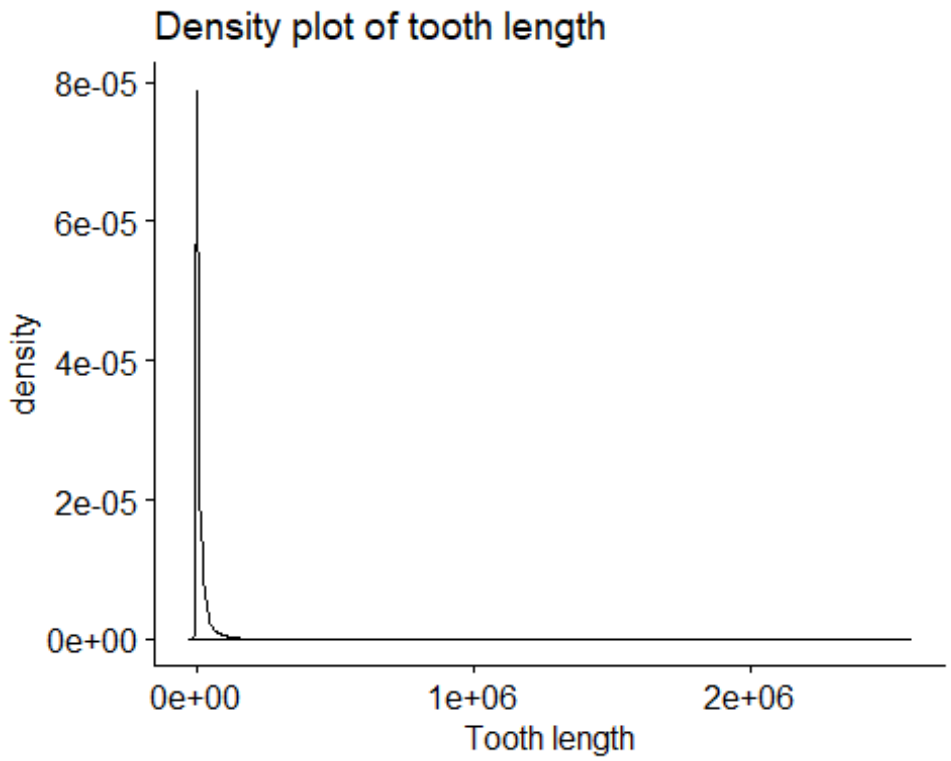
```
## The following object is masked from 'package:purrr':
##
##     set_names
```

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

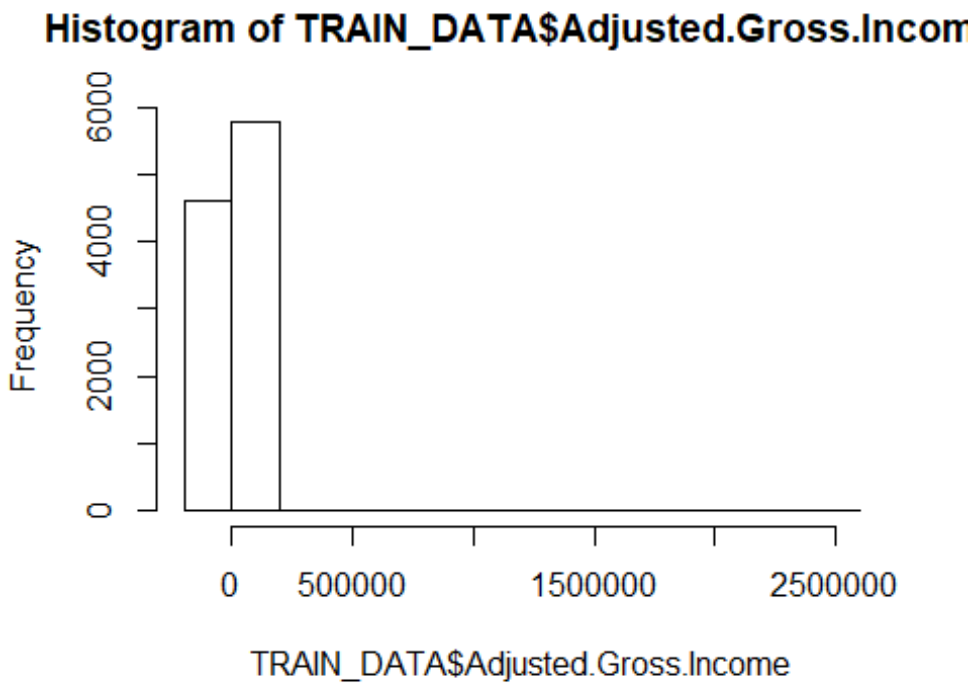
```
##
## Attaching package: 'ggpubr'
```

```
## The following object is masked from 'package:plyr':
##
##     mutate
```

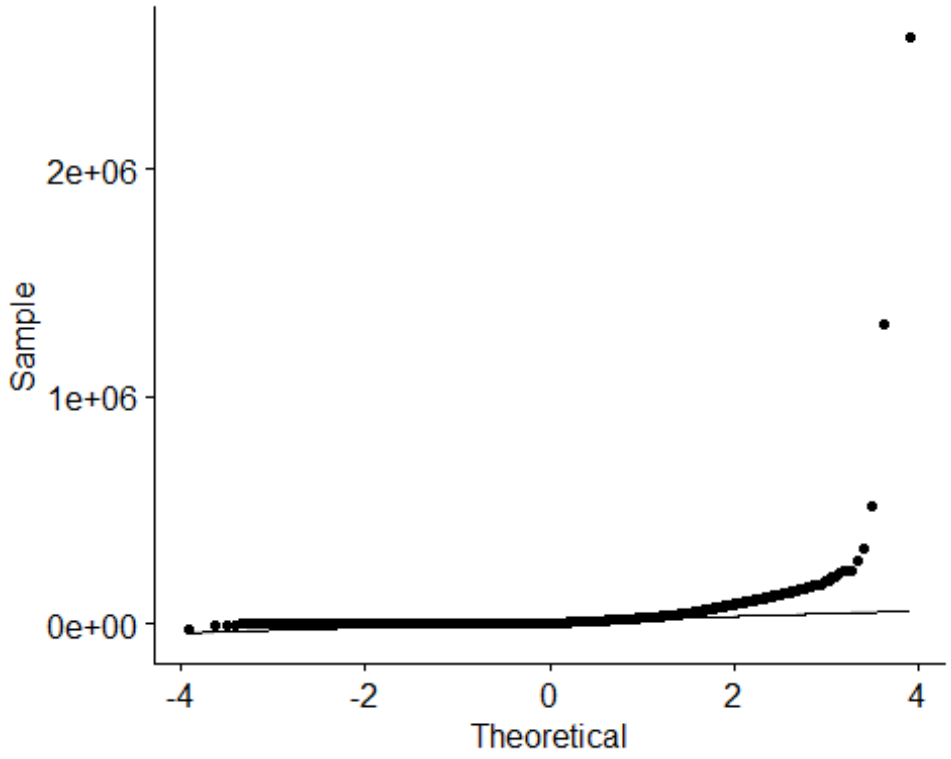
```r
ggdensity(TRAIN_DATA$Adjusted.Gross.Income,
          main = "Density plot of tooth length",
          xlab = "Tooth length")
```



```r
#Histogram
hist(TRAIN_DATA$Adjusted.Gross.Income) #histogram
```

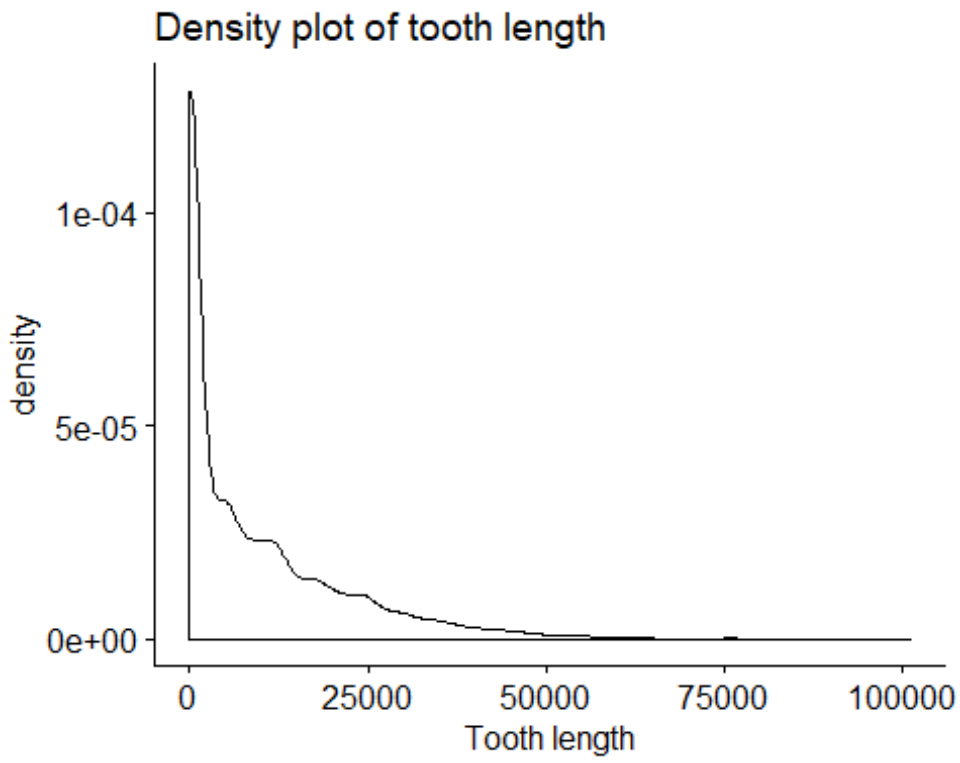



```r
# QQplot
ggqqplot(TRAIN_DATA$Adjusted.Gross.Income) #qqplot
```

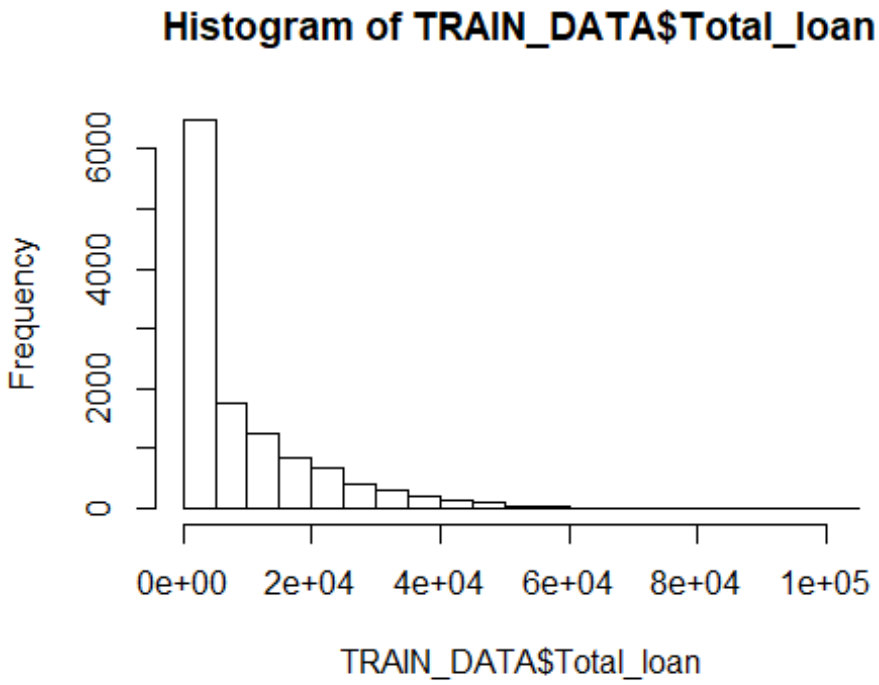


- Total_loan

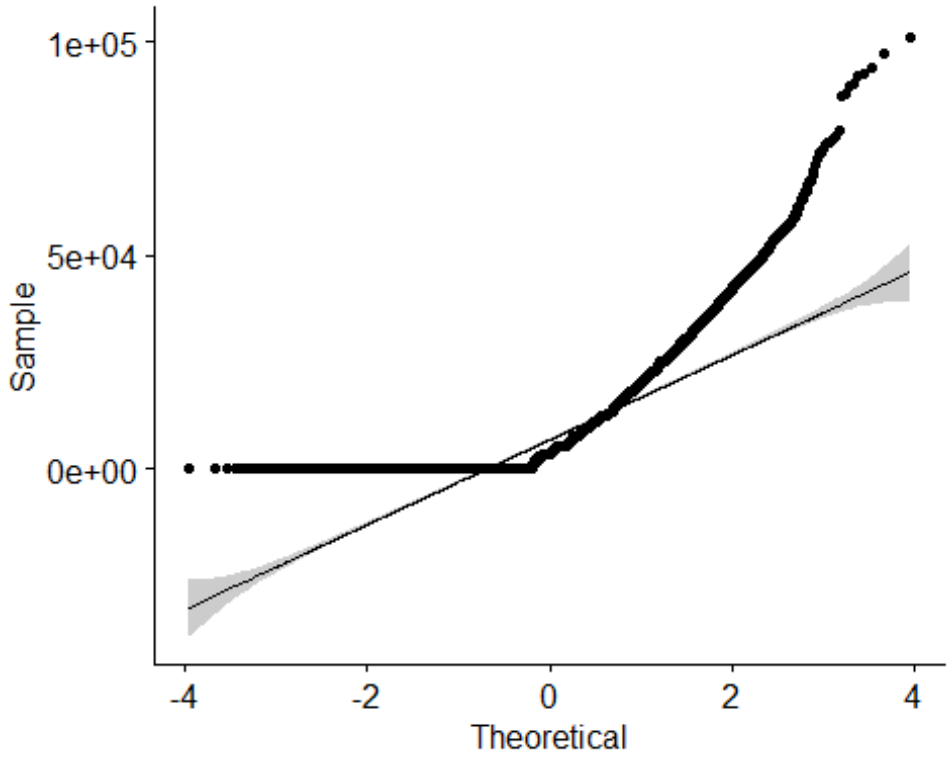
```
#Density plot
library(ggpubr)
ggdensity(TRAIN_DATA$Total_loan,
          main = "Density plot of tooth length",
          xlab = "Tooth length")
```




```r
#Histogram
hist(TRAIN_DATA$Total_loan) #histogram
```

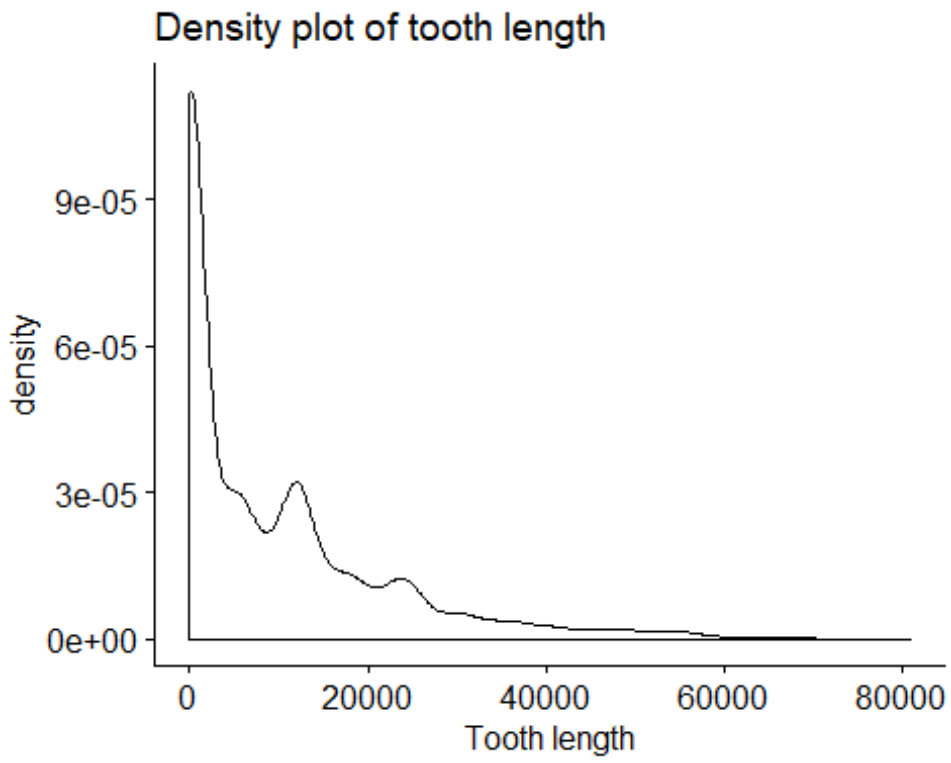


```
# QQplot
ggqqplot(TRAIN_DATA$Total_loan) #qqplot
```

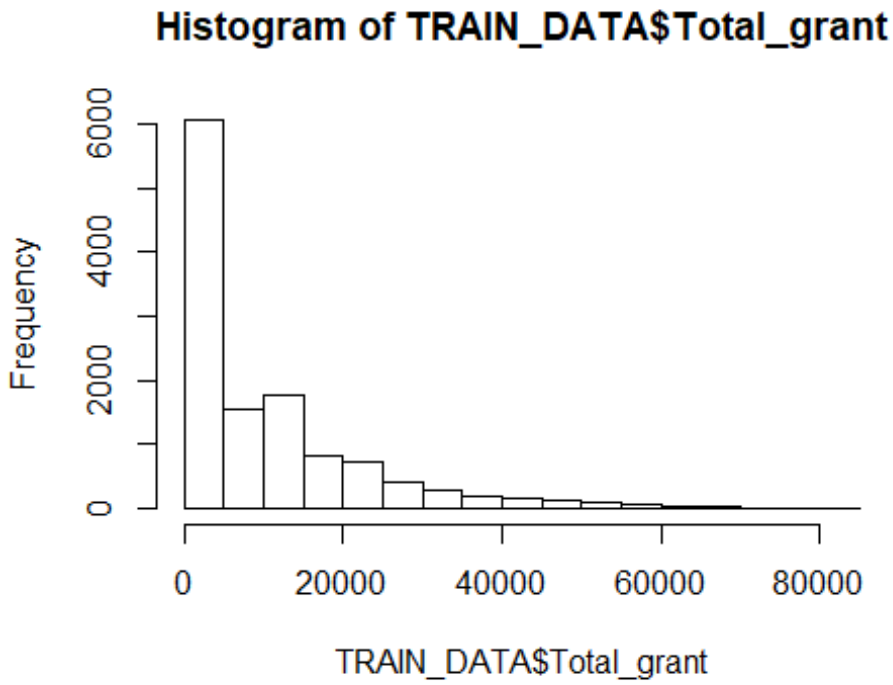


- Total_grant

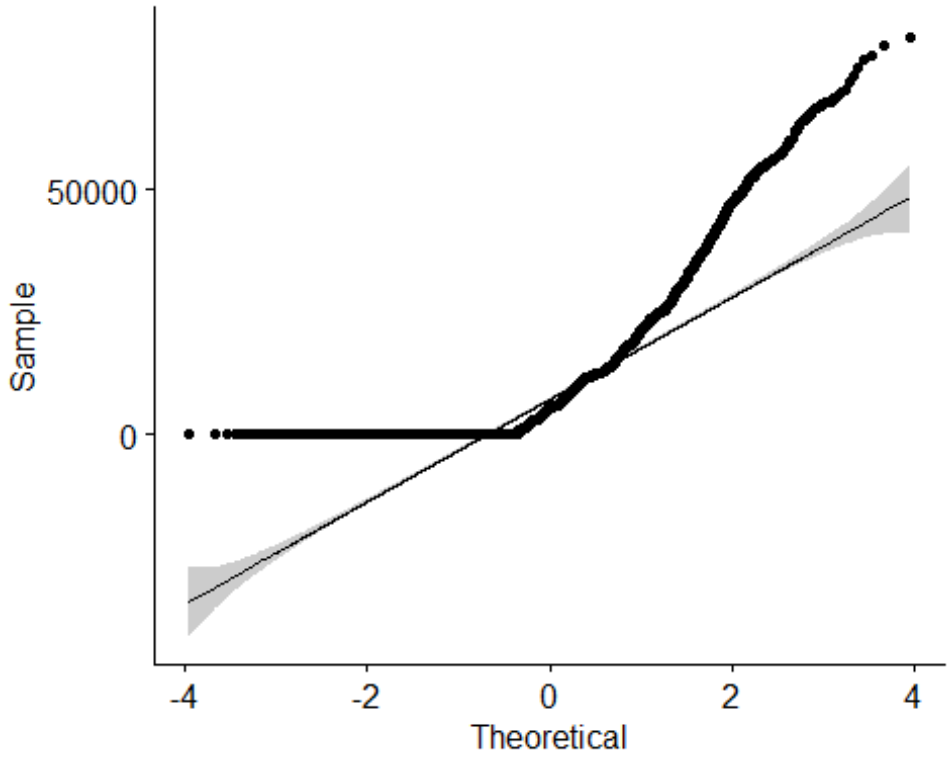
```
#Density plot
library(ggpubr)
ggdensity(TRAIN_DATA$Total_grant,
          main = "Density plot of tooth length",
          xlab = "Tooth length")
```




```r
#Histogram
hist(TRAIN_DATA$Total_grant) #histogram
```

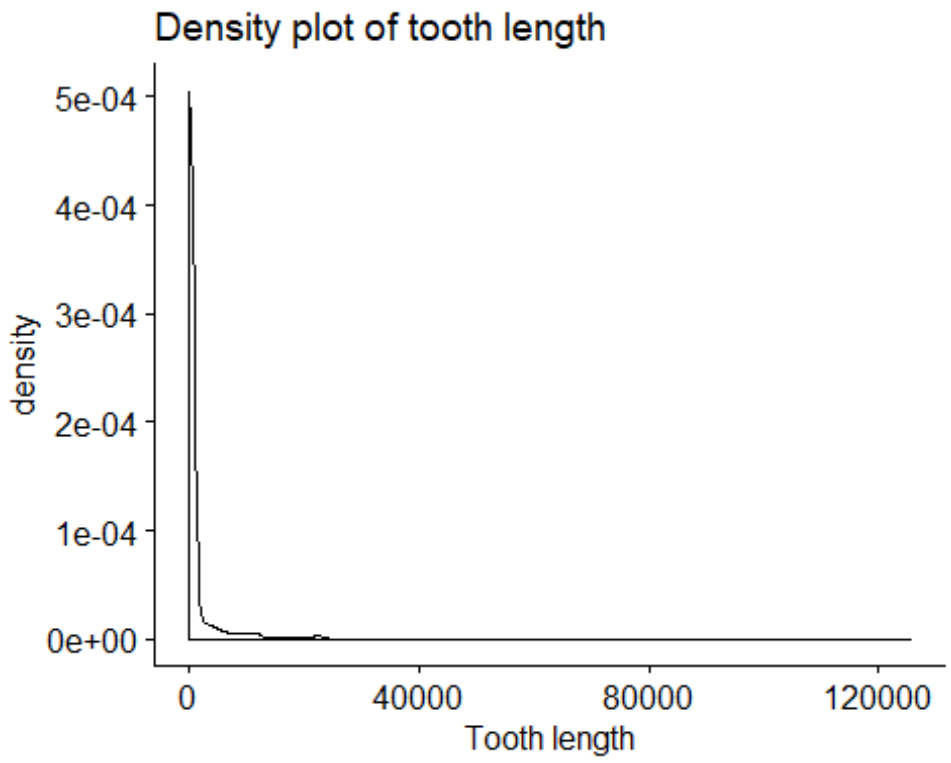


```
# QQplot
ggqqplot(TRAIN_DATA$Total_grant) #qqplot
```

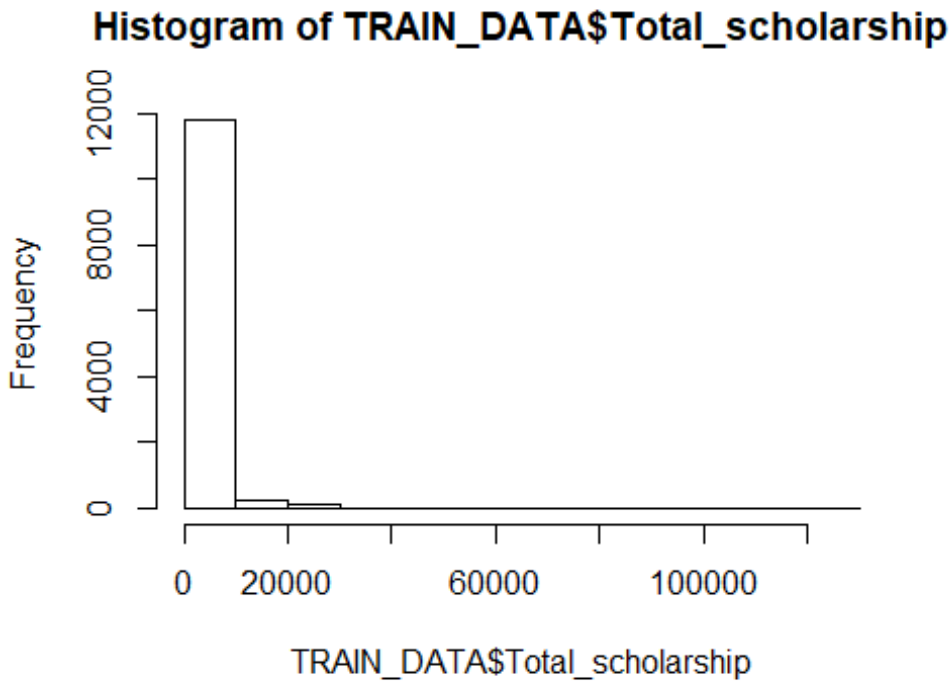


- Total_scholarship

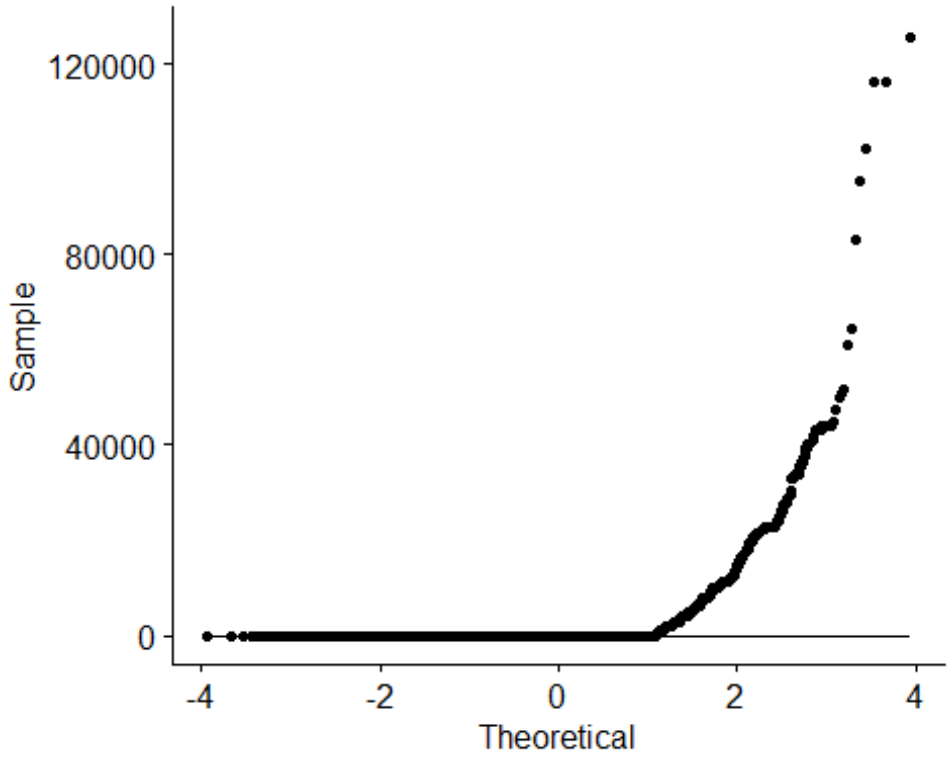
```
# Density plot
ggdensity(TRAIN_DATA$Total_scholarship,
          main = "Density plot of tooth length",
          xlab = "Tooth length")
```




```r
# Histogram
hist(TRAIN_DATA$Total_scholarship) #histogram
```

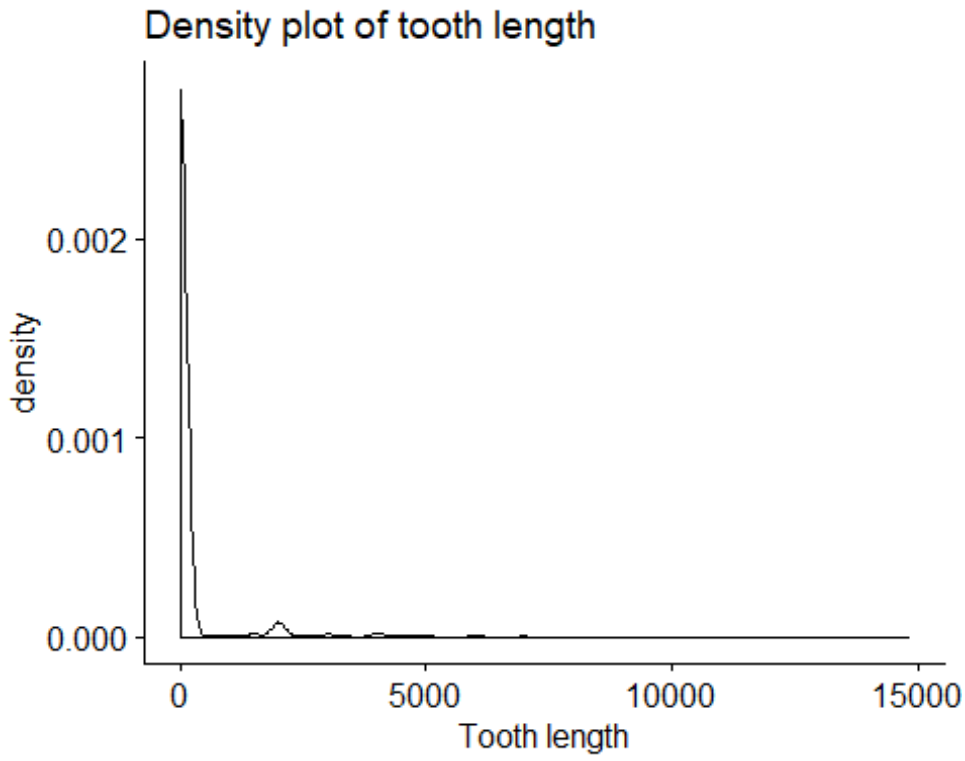


```
# QQplot
ggqqplot(TRAIN_DATA$Total_scholarship) #qqplot
```

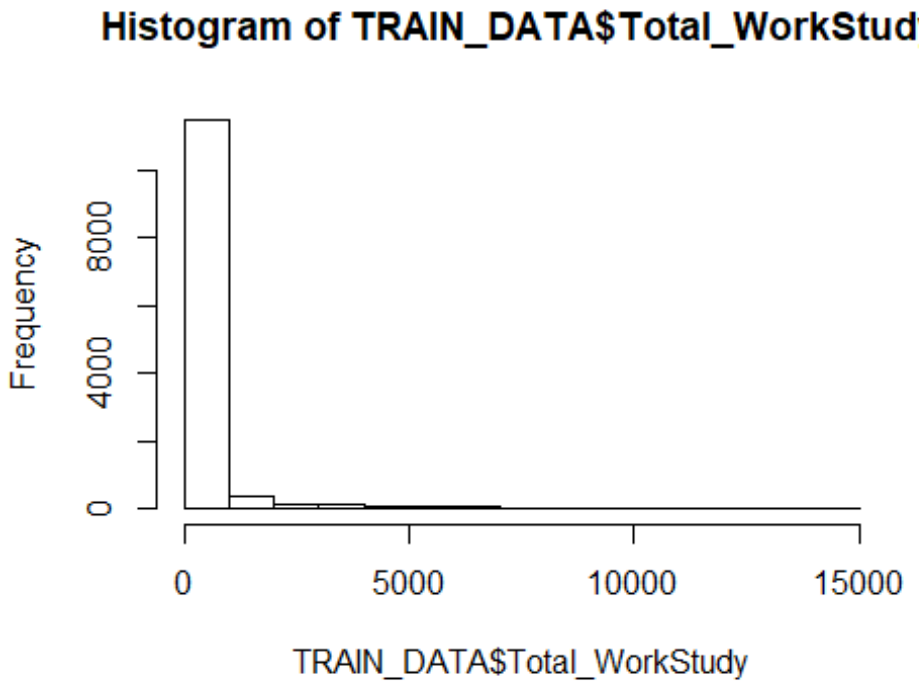


- Total_WorkStudy

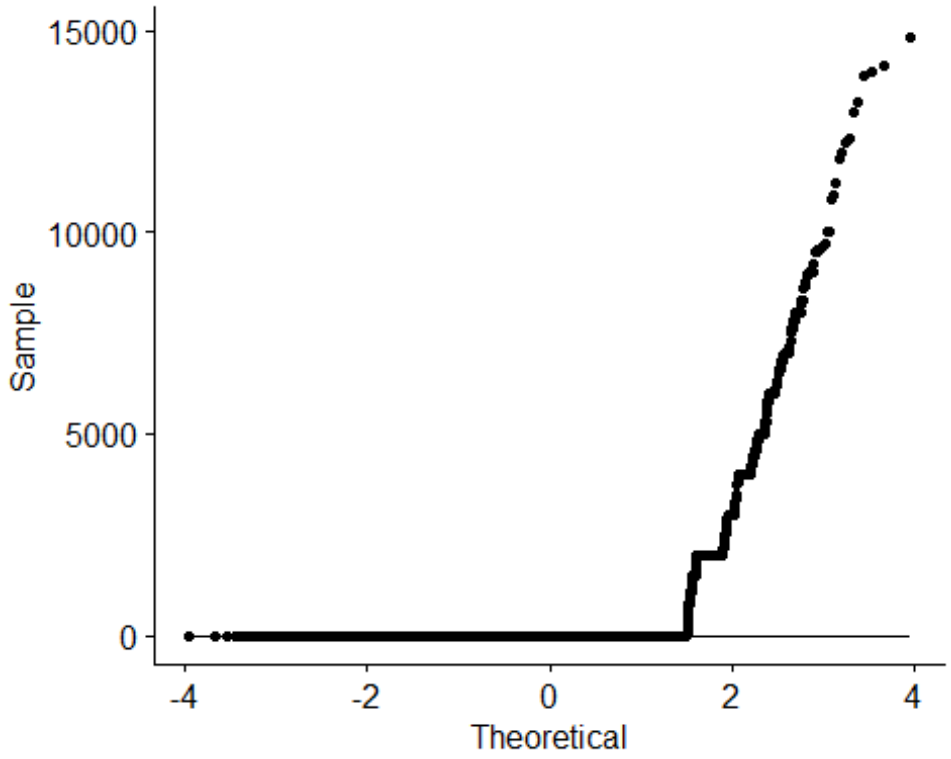
```
# Density plot
ggdensity(TRAIN_DATA$Total_WorkStudy,
          main = "Density plot of tooth length",
          xlab = "Tooth length")
```




```r
# Histogram
hist(TRAIN_DATA$Total_WorkStudy) #histogram
```

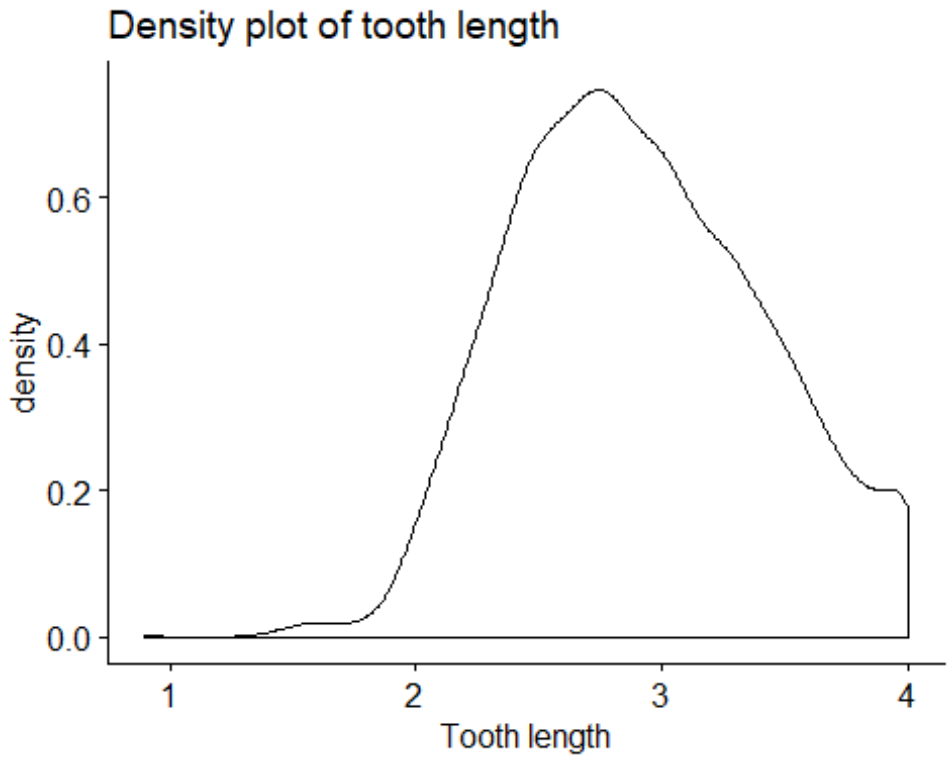


```r
# QQplot
ggqqplot(TRAIN_DATA$Total_WorkStudy) #qqplot
```

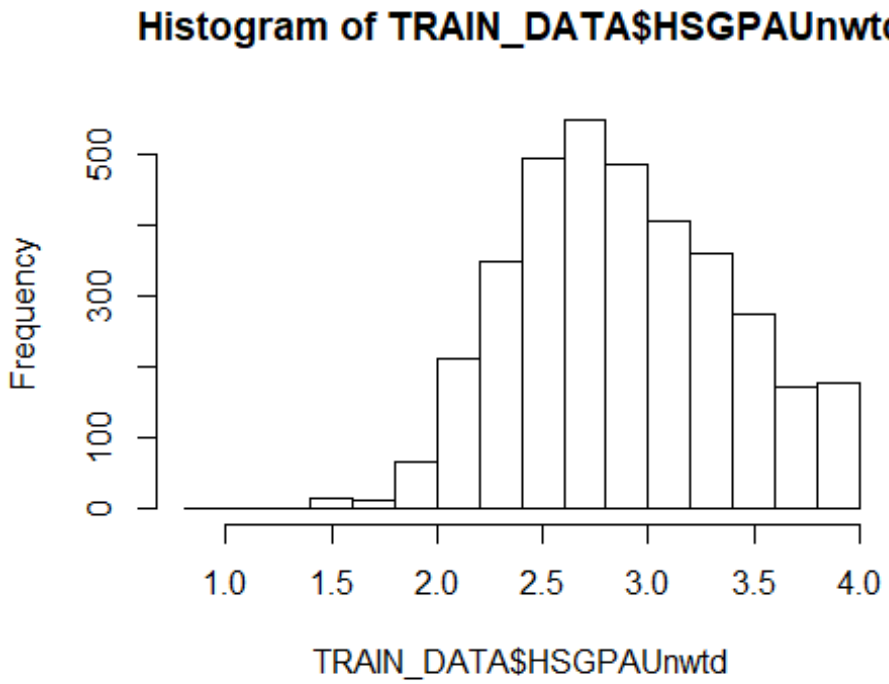


- HSGPAUnwtd

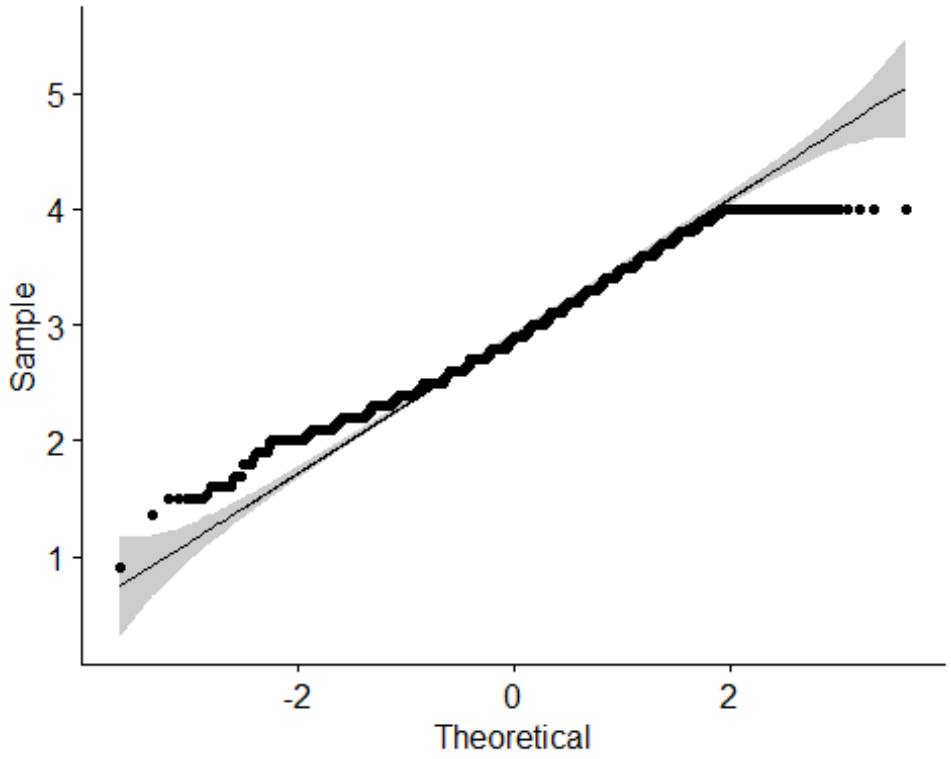
```r
# Density plot
ggdensity(TRAIN_DATA$HSGPAUnwtd,
          main = "Density plot of tooth length",
          xlab = "Tooth length")
```




```r
# Histogram
hist(TRAIN_DATA$HSGPAUnwtd) #histogram
```

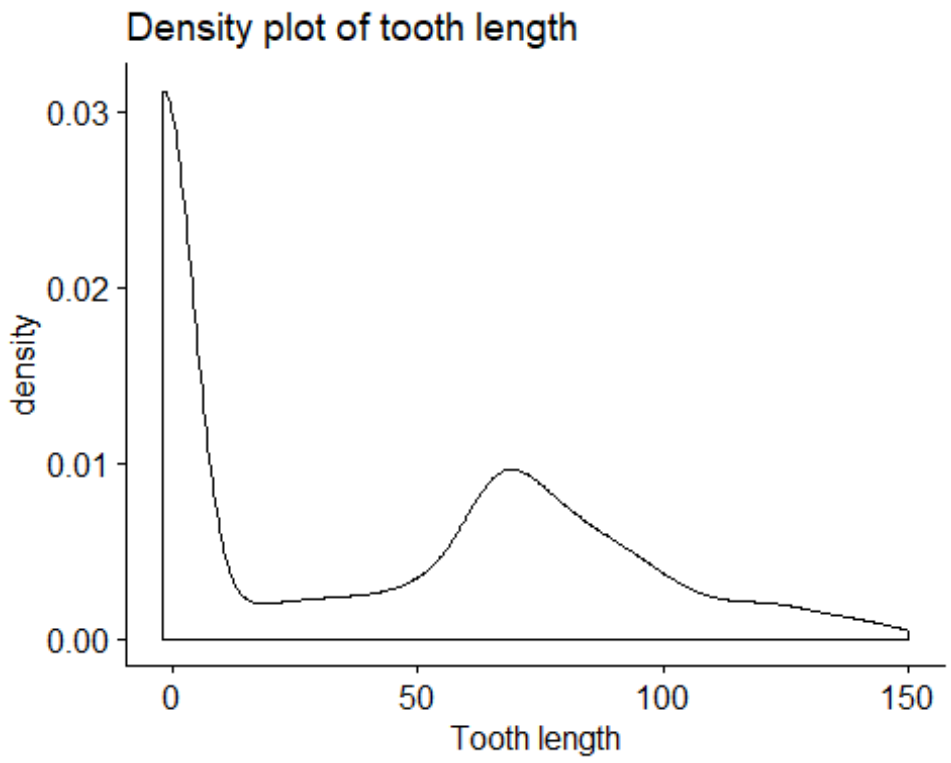


```
# QQplot
ggqqplot(TRAIN_DATA$HSGPAUnwtd) #qqplot
```

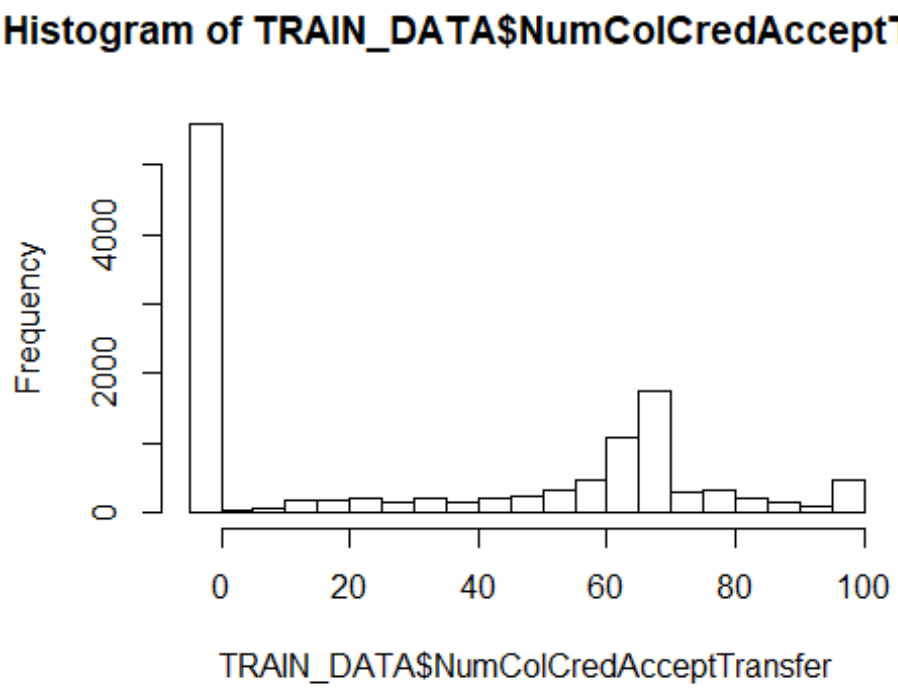


- NumColCredAttemptTransfer

```
# Density plot
ggdensity(TRAIN_DATA$NumColCredAttemptTransfer,
          main = "Density plot of tooth length",
          xlab = "Tooth length")

## Warning: Removed 370 rows containing non-finite values (stat_density).
```




```r
# Histogram
hist(TRAIN_DATA$NumColCredAcceptTransfer) #histogram
```

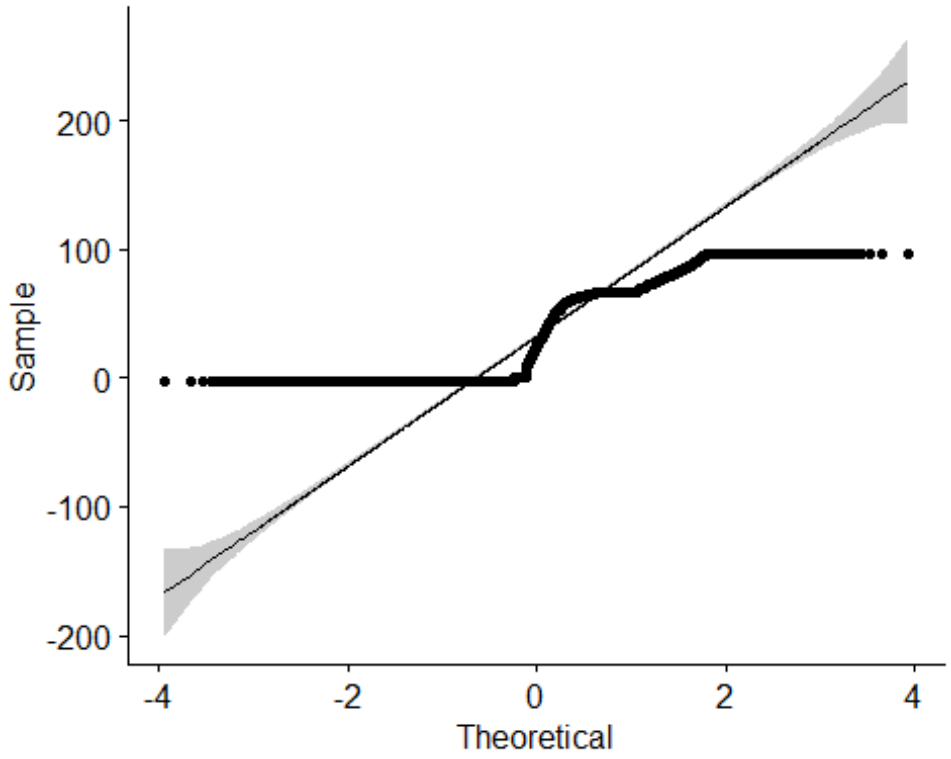


```
# QQplot
ggqqplot(TRAIN_DATA$NumColCredAcceptTransfer) #qqplot
```

```
## Warning: Removed 1 rows containing non-finite values (stat_qq).
```

```
## Warning: Removed 1 rows containing non-finite values (stat_qq_line).
```

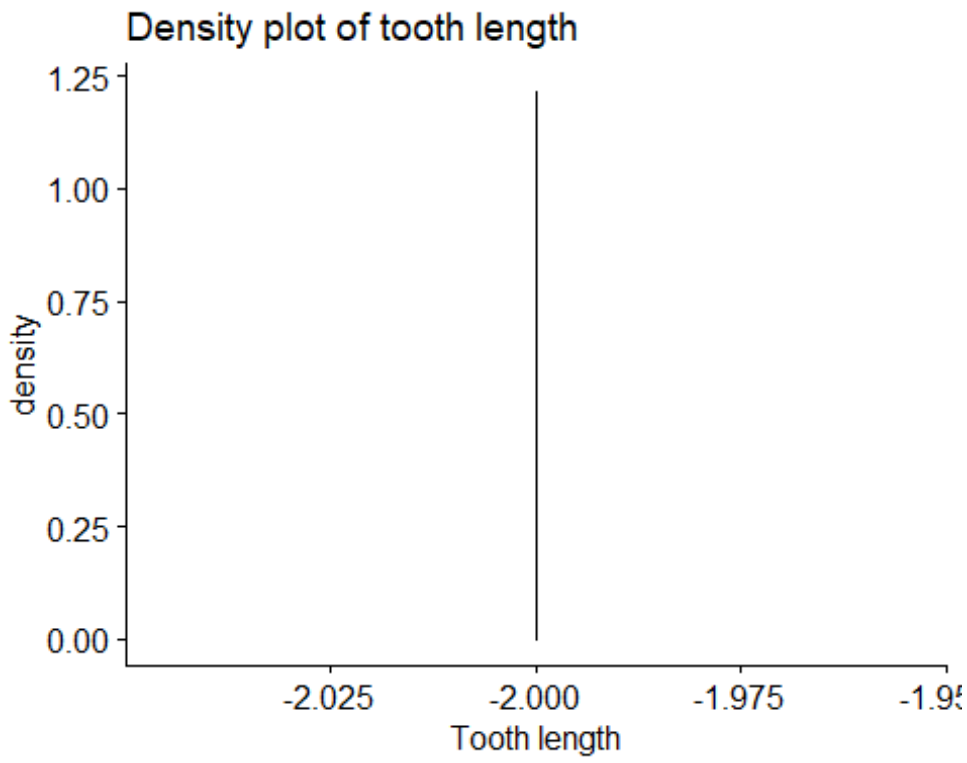
```
## Warning: Removed 1 rows containing non-finite values (stat_qq_line).
```



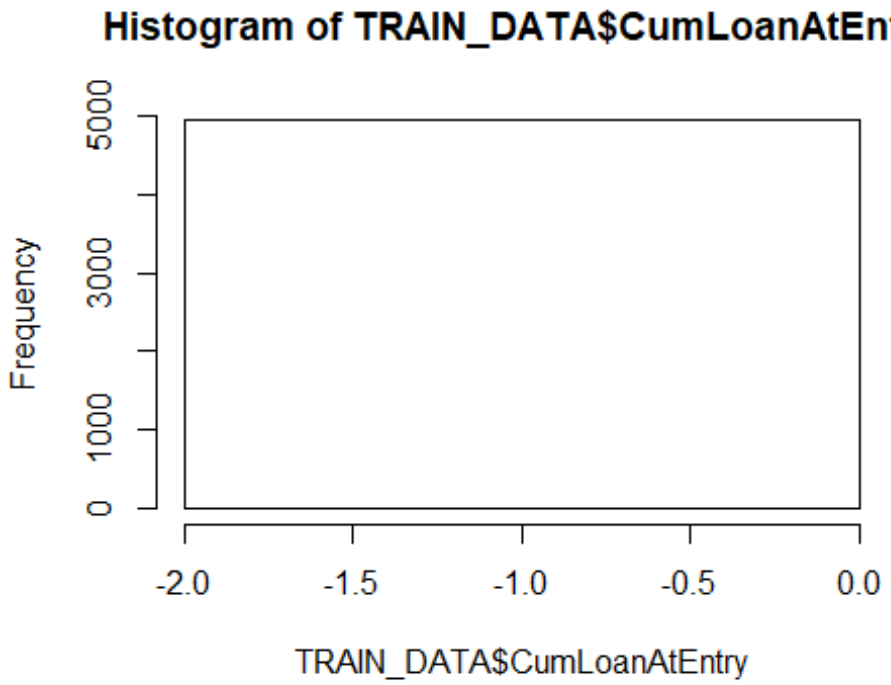
- CumLoanAtEntry

```
# Density plot
ggdensity(TRAIN_DATA$CumLoanAtEntry,
          main = "Density plot of tooth length",
          xlab = "Tooth length")
```

```
## Warning: Removed 7309 rows containing non-finite values (stat_density).
```




```r
# Histogram
hist(TRAIN_DATA$CumLoanAtEntry) #histogram
```

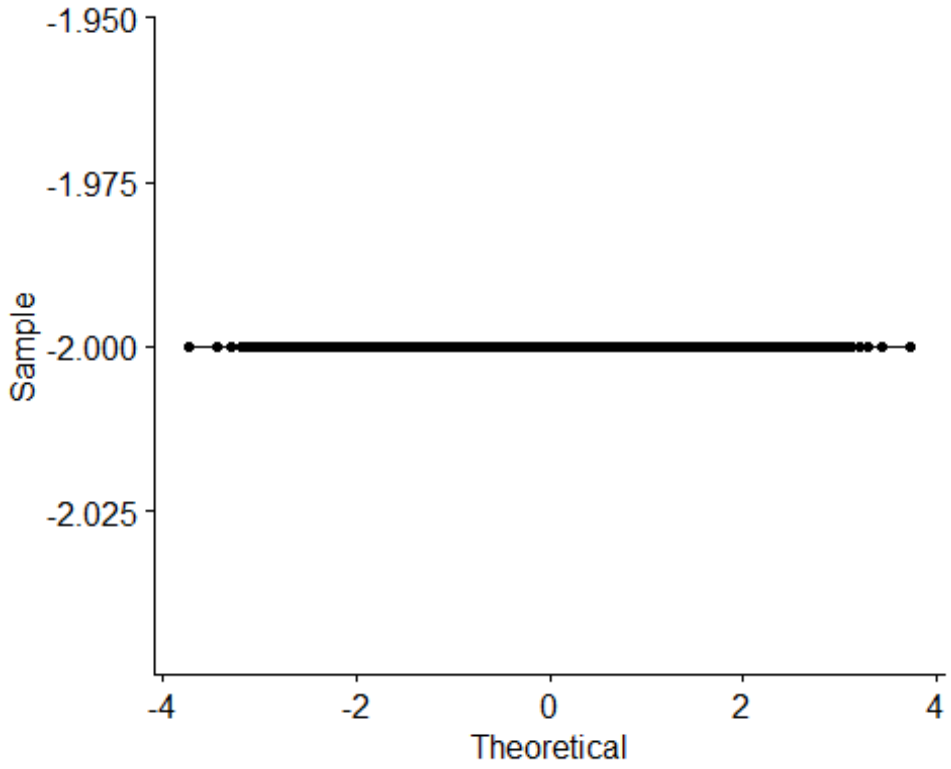


```
# QQplot
ggqqplot(TRAIN_DATA$CumLoanAtEntry) #qqplot
```

```
## Warning: Removed 7309 rows containing non-finite values (stat_qq).
```

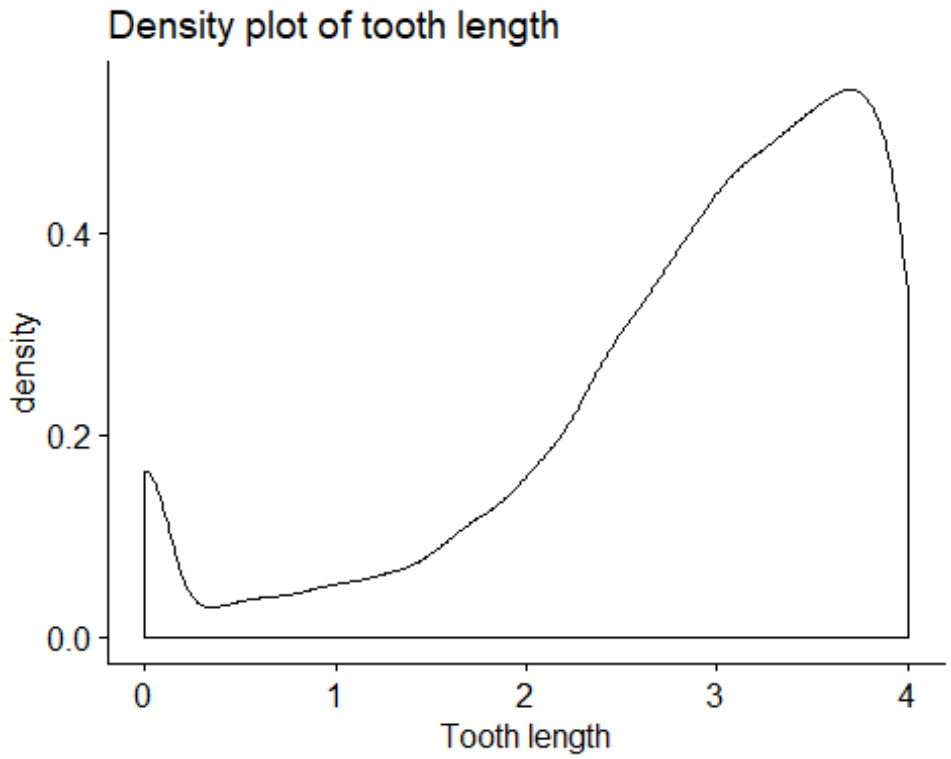
```
## Warning: Removed 7309 rows containing non-finite values (stat_qq_line).
```

```
## Warning: Removed 7309 rows containing non-finite values (stat_qq_line).
```

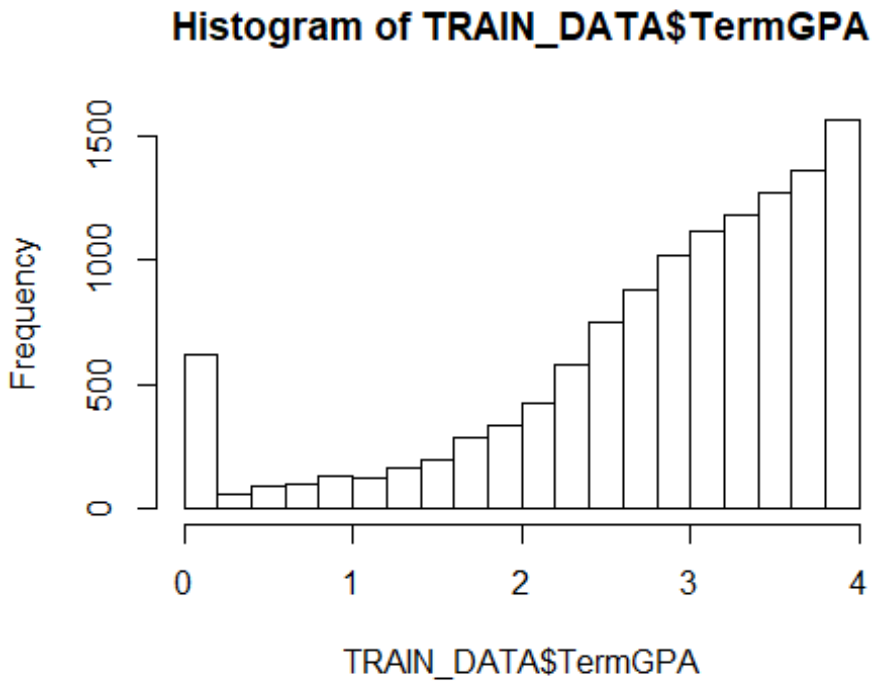


- TermGPA

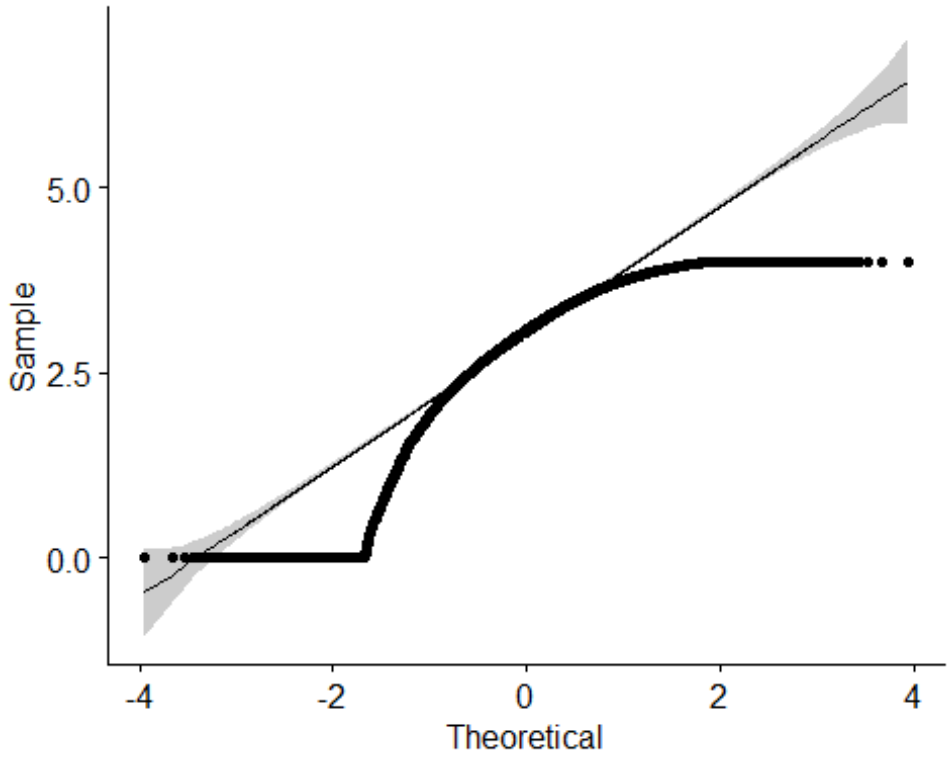
```
# Density plot
ggdensity(TRAIN_DATA$TermGPA,
          main = "Density plot of tooth length",
          xlab = "Tooth length")
```




```
# Histogram
hist(TRAIN_DATA$TermGPA) #histogram
```



```
# QQplot
ggqqplot(TRAIN_DATA$TermGPA) #qqplot
```



- NOTE: the continous variables did not display normal distribution properties, thus imputation will be done using median intead of mean

Impute continuous variables using median

```
for(i in 1:nrow(TRAIN_DATA)) {
  TRAIN_DATA[i, ][is.na(TRAIN_DATA[i,])]<-median(as.numeric(TRAIN_DATA[i,]),
na.rm = TRUE)
}

# Count any missing data
sum(is.na(TRAIN_DATA))

## [1] 0

# which columns have missing data
colnames(TRAIN_DATA)[colSums(is.na(TRAIN_DATA)) > 0]

## character(0)

TRAIN_Imputed=TRAIN_DATA
```

TEST Data Imputation

- Now lets clean and impute missing data in the TEST_DATA as well

- Looking at the dataframe there are variables that have missing values close to 90% of the whole column (>11,000), these variables if imputed may introduce bias, hence they are dropped. Also drop the nominal variables with too many levels or irrelevant qualitative information like BirthMonth, BirthYear, etc.

```
library(tidyverse)

drop.cols <- c('HSGPAWtd', 'FirstGen','TransferIntent','Campus','Address1',
'Address2','Major2','State','Zip','BirthYear','BirthMonth','City')


TEST_DATA<-TEST_DATA %>% select(-drop.cols)
colnames(TEST_DATA)
```

```
##  [1] "StudentID"                    "cohort_term"
##  [3] "Marital.Status"               "Adjusted.Gross.Income"
##  [5] "Parent.Adjusted.Gross.Income" "Father.s.Highest.Grade.Level"
##  [7] "Mother.s.Highest.Grade.Level" "Housing"
##  [9] "Total_loan"                   "Total_grant"
## [11] "Total_scholarship"            "Total_WorkStudy"
## [13] "Cohort"                       "CohortTerm"
## [15] "RegistrationDate"             "Gender"
## [17] "Hispanic"                     "AmericanIndian"
## [19] "Asian"                        "Black"
## [21] "NativeHawaiian"               "White"
## [23] "TwoOrMoreRace"                "HSDip"
## [25] "HSDipYr"                      "HSGPAUnwtd"
## [27] "DualHSSummerEnroll"           "EnrollmentStatus"
## [29] "NumColCredAttemptTransfer"    "NumColCredAcceptTransfer"
## [31] "CumLoanAtEntry"               "HighDeg"
## [33] "MathPlacement"                "EngPlacement"
## [35] "GatewayMathStatus"            "GatewayEnglishStatus"
## [37] "CompleteDevMath"              "CompleteDevEnglish"
## [39] "Major1"                       "Complete1"
## [41] "Complete2"                    "CompleteCIP1"
## [43] "CompleteCIP2"                 "DegreeTypeSought"
## [45] "TermGPA"                      "CumGPA"
```

Ensure that all missing values and empty values are captured

```
# Convert empty spaces and (-1) to NULL in order to facilitate imputation
later on
TEST_DATA[TEST_DATA=="-1"]<-NA
TEST_DATA[TEST_DATA==""]<-NA

# Check for missing columns
colnames(TEST_DATA)[colSums(is.na(TEST_DATA)) > 0]
```

```
##  [1] "Marital.Status"               "Adjusted.Gross.Income"
##  [3] "Parent.Adjusted.Gross.Income" "Father.s.Highest.Grade.Level"
##  [5] "Mother.s.Highest.Grade.Level" "Housing"
```

```
##  [7] "Hispanic"                      "AmericanIndian"
##  [9] "Asian"                         "Black"
## [11] "NativeHawaiian"                "White"
## [13] "TwoOrMoreRace"                 "HSDip"
## [15] "HSDipYr"                       "HSGPAUnwtd"
## [17] "NumColCredAttemptTransfer"     "CumLoanAtEntry"
## [19] "MathPlacement"                 "EngPlacement"
## [21] "CompleteDevMath"               "CompleteDevEnglish"
## [23] "Major1"

# Impute categorical variables using mode

val<-unique(TEST_DATA$Marital.Status[!is.na(TEST_DATA$Marital.Status)]) #
Values in vec_miss
mode <- val[which.max(tabulate(match(TEST_DATA$Marital.Status, val)))] # Mode
of vec_miss
TEST_DATA$Marital.Status[is.na(TEST_DATA$Marital.Status)]<-mode # Impute by
mode



val<-
unique(TEST_DATA$Father.s.Highest.Grade.Level[!is.na(TEST_DATA$Father.s.Highe
st.Grade.Level)])
mode<-val[which.max(tabulate(match(TEST_DATA$Father.s.Highest.Grade.Level,
val)))]
TEST_DATA$Father.s.Highest.Grade.Level[is.na(TEST_DATA$Father.s.Highest.Grade
.Level)]<-mode

val<-
unique(TEST_DATA$Mother.s.Highest.Grade.Level[!is.na(TEST_DATA$Mother.s.Highe
st.Grade.Level)])
mode<-val[which.max(tabulate(match(TEST_DATA$Mother.s.Highest.Grade.Level,
val)))]
TEST_DATA$Mother.s.Highest.Grade.Level[is.na(TEST_DATA$Mother.s.Highest.Grade
.Level)]<-mode

val<-unique(TEST_DATA$Housing[!is.na(TEST_DATA$Housing)])
mode<-val[which.max(tabulate(match(TEST_DATA$Housing, val)))]
TEST_DATA$Housing[is.na(TEST_DATA$Housing)]<-mode



val<-unique(TEST_DATA$EngPlacement[!is.na(TEST_DATA$EngPlacement)])
mode<-val[which.max(tabulate(match(TEST_DATA$EngPlacement, val)))]
TEST_DATA$EngPlacement[is.na(TEST_DATA$EngPlacement)]<-mode



val<-unique(TEST_DATA$MathPlacement[!is.na(TEST_DATA$MathPlacement)])
mode<-val[which.max(tabulate(match(TEST_DATA$MathPlacement, val)))]
TEST_DATA$MathPlacement[is.na(TEST_DATA$MathPlacement)]<-mode
```

```r
val<-unique(TEST_DATA$CompleteDevEnglish[!is.na(TEST_DATA$MathPlacement)])
mode<-val[which.max(tabulate(match(TEST_DATA$CompleteDevEnglish, val)))]
TEST_DATA$CompleteDevEnglish[is.na(TEST_DATA$CompleteDevEnglish)]<-mode


val<-unique(TEST_DATA$CompleteDevMath[!is.na(TEST_DATA$CompleteDevMath)])
mode<-val[which.max(tabulate(match(TEST_DATA$CompleteDevMath, val)))]
TEST_DATA$CompleteDevMath[is.na(TEST_DATA$CompleteDevMath)]<-mode

val<-unique(TEST_DATA$Hispanic[!is.na(TEST_DATA$Hispanic)])
mode<-val[which.max(tabulate(match(TEST_DATA$Hispanic, val)))]
TEST_DATA$Hispanic[is.na(TEST_DATA$Hispanic)]<-mode

val<-unique(TEST_DATA$AmericanIndian[!is.na(TEST_DATA$AmericanIndian)])
mode<-val[which.max(tabulate(match(TEST_DATA$AmericanIndian, val)))]
TEST_DATA$AmericanIndian[is.na(TEST_DATA$AmericanIndian)]<-mode

val<-unique(TEST_DATA$Asian[!is.na(TEST_DATA$Asian)])
mode<-val[which.max(tabulate(match(TEST_DATA$Asian, val)))]
TEST_DATA$Asian[is.na(TEST_DATA$Asian)]<-mode

val<-unique(TEST_DATA$Black[!is.na(TEST_DATA$Black)])
mode<-val[which.max(tabulate(match(TEST_DATA$Black, val)))]
TEST_DATA$Black[is.na(TEST_DATA$Black)]<-mode

val<-unique(TEST_DATA$NativeHawaiian[!is.na(TEST_DATA$NativeHawaiian)])
mode<-val[which.max(tabulate(match(TEST_DATA$NativeHawaiian, val)))]
TEST_DATA$NativeHawaiian[is.na(TEST_DATA$NativeHawaiian)]<-mode

val<-unique(TEST_DATA$White[!is.na(TEST_DATA$White)])
mode<-val[which.max(tabulate(match(TEST_DATA$White, val)))]
TEST_DATA$White[is.na(TEST_DATA$White)]<-mode

val<-unique(TEST_DATA$TwoOrMoreRace[!is.na(TEST_DATA$TwoOrMoreRace)])
mode<-val[which.max(tabulate(match(TEST_DATA$TwoOrMoreRace, val)))]
TEST_DATA$TwoOrMoreRace[is.na(TEST_DATA$TwoOrMoreRace)]<-mode

val<-unique(TEST_DATA$HSDip[!is.na(TEST_DATA$HSDip)])
mode<-val[which.max(tabulate(match(TEST_DATA$HSDip, val)))]
TEST_DATA$HSDip[is.na(TEST_DATA$HSDip)]<-mode

val<-unique(TEST_DATA$HSDipYr[!is.na(TEST_DATA$HSDipYr)])
mode<-val[which.max(tabulate(match(TEST_DATA$HSDipYr, val)))]
TEST_DATA$HSDipYr[is.na(TEST_DATA$HSDipYr)]<-mode

#Check which columns have missing data to ensure the imputed columns are not
```

```
there
colnames(TEST_DATA)[colSums(is.na(TEST_DATA)) > 0]
```

```
## [1] "Adjusted.Gross.Income"        "Parent.Adjusted.Gross.Income"
## [3] "HSGPAUnwtd"                   "NumColCredAttemptTransfer"
## [5] "CumLoanAtEntry"               "Major1"
```

```
# Impute using median the other numeric variables
for(i in 1:nrow(TEST_DATA)) {
  TEST_DATA[i, ][is.na(TEST_DATA[i,])]<-median(as.numeric(TEST_DATA[i,]),
na.rm = TRUE)
}

# Ensure no missing columns left (we expect zero)
colnames(TEST_DATA)[colSums(is.na(TEST_DATA)) > 0]
```

```
## character(0)
```

```
TEST_Imputed=TEST_DATA
```

FEATURE ENGINEERING

Implement feature selection using Boruta package

```
library(Boruta)
```

```
## Loading required package: ranger
```

```
# Now is the time to implement and check the performance of boruta package.
The syntax of boruta is almost similar to regression (lm) method.

set.seed(123)
boruta.train <- Boruta(Dropout~.-StudentID, data = TRAIN_Imputed, doTrace =
2)
```

```
##  1. run of importance source...

##  2. run of importance source...

##  3. run of importance source...

##  4. run of importance source...

##  5. run of importance source...

##  6. run of importance source...

##  7. run of importance source...

##  8. run of importance source...

##  9. run of importance source...
```

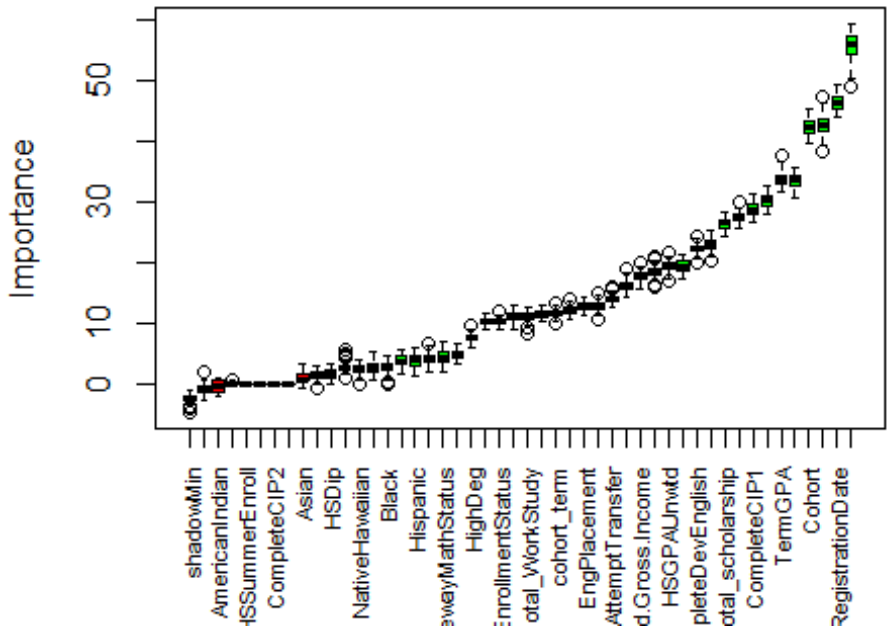
```
##  10. run of importance source...

##  11. run of importance source...

##  12. run of importance source...

##  13. run of importance source...

## After 13 iterations, +3.9 mins:

##  confirmed 29 attributes: Adjusted.Gross.Income, Cohort, cohort_term,
CohortTerm, Complete1 and 24 more;

##  rejected 6 attributes: AmericanIndian, Complete2, CompleteCIP2,
DegreeTypeSought, DualHSSummerEnroll and 1 more;

##  still have 10 attributes left.

##  14. run of importance source...

##  15. run of importance source...

##  16. run of importance source...

##  17. run of importance source...

## After 17 iterations, +5 mins:

##  confirmed 1 attribute: Marital.Status;

##  rejected 1 attribute: Asian;

##  still have 8 attributes left.

##  18. run of importance source...

##  19. run of importance source...

##  20. run of importance source...

## After 20 iterations, +5.7 mins:

##  confirmed 1 attribute: GatewayMathStatus;

##  rejected 1 attribute: Gender;

##  still have 6 attributes left.

##  21. run of importance source...

##  22. run of importance source...

##  23. run of importance source...

##  24. run of importance source...
```

```
## After 24 iterations, +6.7 mins:
##  confirmed 1 attribute: Father.s.Highest.Grade.Level;
##  rejected 1 attribute: HSDip;
##  still have 4 attributes left.
##  25. run of importance source...
##  26. run of importance source...
##  27. run of importance source...
##  28. run of importance source...
##  29. run of importance source...
##  30. run of importance source...
## After 30 iterations, +8 mins:
##  confirmed 1 attribute: Hispanic;
##  still have 3 attributes left.
##  31. run of importance source...
##  32. run of importance source...
##  33. run of importance source...
##  34. run of importance source...
##  35. run of importance source...
##  36. run of importance source...
##  37. run of importance source...
##  38. run of importance source...
##  39. run of importance source...
##  40. run of importance source...
##  41. run of importance source...
##  42. run of importance source...
##  43. run of importance source...
##  44. run of importance source...
##  45. run of importance source...
```

```
##  46. run of importance source...
##  47. run of importance source...
##  48. run of importance source...
##  49. run of importance source...
##  50. run of importance source...
##  51. run of importance source...
##  52. run of importance source...
##  53. run of importance source...
##  54. run of importance source...
##  55. run of importance source...
##  56. run of importance source...
##  57. run of importance source...
##  58. run of importance source...
##  59. run of importance source...
##  60. run of importance source...
##  61. run of importance source...
##  62. run of importance source...
##  63. run of importance source...
##  64. run of importance source...
##  65. run of importance source...
##  66. run of importance source...
##  67. run of importance source...
##  68. run of importance source...
##  69. run of importance source...
##  70. run of importance source...
##  71. run of importance source...
##  72. run of importance source...
##  73. run of importance source...
```

```
##  74. run of importance source...
##  75. run of importance source...
##  76. run of importance source...
##  77. run of importance source...
##  78. run of importance source...
##  79. run of importance source...
##  80. run of importance source...
##  81. run of importance source...
##  82. run of importance source...
##  83. run of importance source...
##  84. run of importance source...
##  85. run of importance source...
##  86. run of importance source...
##  87. run of importance source...
##  88. run of importance source...
##  89. run of importance source...
##  90. run of importance source...
##  91. run of importance source...
##  92. run of importance source...
##  93. run of importance source...
##  94. run of importance source...
##  95. run of importance source...
##  96. run of importance source...
##  97. run of importance source...
##  98. run of importance source...
##  99. run of importance source...
print(boruta.train)

## Boruta performed 99 iterations in 23.05626 mins.
##  33 attributes confirmed important: Adjusted.Gross.Income, Cohort,
```

```
## cohort_term, CohortTerm, Complete1 and 28 more;
##  9 attributes confirmed unimportant: AmericanIndian, Asian, Complete2,
## CompleteCIP2, DegreeTypeSought and 4 more;
##  3 tentative attributes left: Black, NativeHawaiian, White;

# Now, we'll plot the boruta variable importance chart.
plot(boruta.train, xlab = "", xaxt = "n")
 lz<-lapply(1:ncol(boruta.train$ImpHistory),function(i)
  boruta.train$ImpHistory[is.finite(boruta.train$ImpHistory[,i]),i])
 names(lz) <- colnames(boruta.train$ImpHistory)
 Labels <- sort(sapply(lz,median))
 axis(side = 1,las=2,labels = names(Labels),
       at = 1:ncol(boruta.train$ImpHistory), cex.axis = 0.7)
```



- Blue boxplots correspond to minimal, average and maximum Z score of a shadow attribute. Red, yellow and green boxplots represent Z scores of rejected, tentative and confirmed attributes respectively.
- Now is the time to take decision on tentative attributes. The tentative attributes will be classified as confirmed or rejected by comparing the median Z score of the attributes with the median Z score of the best shadow attribute. Let's do it.

```
final.boruta <- TentativeRoughFix(boruta.train)
print(final.boruta)

## Boruta performed 99 iterations in 23.05626 mins.
## Tentatives roughfixed over the last 99 iterations.
##  36 attributes confirmed important: Adjusted.Gross.Income, Black,
```

```
## Cohort, cohort_term, CohortTerm and 31 more;
##  9 attributes confirmed unimportant: AmericanIndian, Asian, Complete2,
## CompleteCIP2, DegreeTypeSought and 4 more;
```

We'll create a data frame of the final result derived from Boruta.

```
boruta.df <- attStats(final.boruta)
class(boruta.df)

## [1] "data.frame"

print(boruta.df)

##                                   meanImp  medianImp       minImp     maxImp
## cohort_term                    11.4644063 11.4890541  9.748180445 13.0654723
## Marital.Status                  4.8001656  4.8011813  3.253116694  6.6341844
## Adjusted.Gross.Income          17.5834106 17.5396540 15.558539531 19.8420250
## Parent.Adjusted.Gross.Income   19.3928709 19.2948620 17.277263752 21.2453945
## Father.s.Highest.Grade.Level    3.7422215  3.7021947  1.540689513  5.6151279
## Mother.s.Highest.Grade.Level    4.0561694  3.9986348  1.771297864  6.6144269
## Housing                        10.8900635 10.8735631  8.997804971 12.7478507
## Total_loan                     42.5251197 42.4642185 38.298559210 47.2951999
## Total_grant                    55.4045865 55.9224708 48.747025701 59.2426057
## Total_scholarship              26.2770581 26.4027287 24.225882143 28.2172163
## Total_WorkStudy                11.0508074 11.0451278  8.198797381 12.6695747
## Cohort                         42.2167406 42.2184180 39.396587241 45.1987074
## CohortTerm                     11.4519135 11.4464315 10.194097204 12.9837520
## RegistrationDate               46.3384203 46.2726345 43.959364846 49.2282809
## Gender                          1.3115292  1.4183324 -0.808447376  2.7471088
## Hispanic                        3.7528627  3.7986604  1.170683057  5.8318774
## AmericanIndian                 -0.4970859 -0.3365582 -2.049282861  0.9825939
## Asian                           0.8789662  0.6885190 -0.866592666  3.0748625
## Black                           2.7563634  2.8204223 -0.263307986  4.4728354
## NativeHawaiian                  2.4073579  2.5199452  0.002989199  3.8246302
## White                           2.5735889  2.5650832  0.532827248  5.1532921
## TwoOrMoreRace                  -0.7872690 -0.9413356 -2.636313484  1.7166095
## HSDip                           1.5571518  1.5701008 -0.080618946  3.2618418
## HSDipYr                        18.4271222 18.3762186 16.006170796 20.9253378
## HSGPAUnwtd                     19.2657042 19.2174180 17.028717003 21.4887519
## DualHSSummerEnroll              0.0000000  0.0000000  0.000000000  0.0000000
## EnrollmentStatus               10.2968792 10.2425672  8.953024844 12.0079643
## NumColCredAttemptTransfer      13.8599568 13.8144037 12.533440935 15.9759968
## NumColCredAcceptTransfer       16.1377078 16.1212677 14.314978327 18.9158072
## CumLoanAtEntry                 10.1817785 10.1916459  8.955420653 11.3805630
## HighDeg                         7.5435041  7.5699375  5.767598541  9.5547090
## MathPlacement                  12.1082994 11.9884078 10.474289978 14.0035540
## EngPlacement                   12.7356027 12.7507211 11.321221199 14.1868192
## GatewayMathStatus               4.3183216  4.1573323  1.881066736  6.7146035
## GatewayEnglishStatus           12.7957729 12.8514080 10.394398869 14.7856407
## CompleteDevMath                22.9028779 22.7871243 20.162655784 25.1403590
## CompleteDevEnglish             22.1818595 22.2213170 19.872623628 24.1322784
```

```
## Major1                          27.4737358 27.4522212 25.608579480 29.9199472
## Complete1                       30.0916956 30.1738274 27.792953306 32.6064043
## Complete2                        0.0000000  0.0000000  0.000000000  0.0000000
## CompleteCIP1                    28.8298497 28.7031964 26.483003792 31.2678100
## CompleteCIP2                     0.0000000  0.0000000  0.000000000  0.0000000
## DegreeTypeSought                 0.0000000  0.0000000  0.000000000  0.0000000
## TermGPA                         33.4941703 33.3777534 31.604048131 37.5905229
## CumGPA                          33.3570918 33.4047757 30.693412868 35.6499898
##                                   normHits  decision
## cohort_term                     1.00000000 Confirmed
## Marital.Status                  0.96969697 Confirmed
## Adjusted.Gross.Income           1.00000000 Confirmed
## Parent.Adjusted.Gross.Income    1.00000000 Confirmed
## Father.s.Highest.Grade.Level    0.88888889 Confirmed
## Mother.s.Highest.Grade.Level    0.90909091 Confirmed
## Housing                         1.00000000 Confirmed
## Total_loan                      1.00000000 Confirmed
## Total_grant                     1.00000000 Confirmed
## Total_scholarship               1.00000000 Confirmed
## Total_WorkStudy                 1.00000000 Confirmed
## Cohort                          1.00000000 Confirmed
## CohortTerm                      1.00000000 Confirmed
## RegistrationDate                1.00000000 Confirmed
## Gender                          0.02020202  Rejected
## Hispanic                        0.83838384 Confirmed
## AmericanIndian                  0.00000000  Rejected
## Asian                           0.01010101  Rejected
## Black                           0.63636364 Confirmed
## NativeHawaiian                  0.44444444 Confirmed
## White                           0.58585859 Confirmed
## TwoOrMoreRace                   0.00000000  Rejected
## HSDip                           0.03030303  Rejected
## HSDipYr                         1.00000000 Confirmed
## HSGPAUnwtd                      1.00000000 Confirmed
## DualHSSummerEnroll              0.00000000  Rejected
## EnrollmentStatus                1.00000000 Confirmed
## NumColCredAttemptTransfer       1.00000000 Confirmed
## NumColCredAcceptTransfer        1.00000000 Confirmed
## CumLoanAtEntry                  1.00000000 Confirmed
## HighDeg                         1.00000000 Confirmed
## MathPlacement                   1.00000000 Confirmed
## EngPlacement                    1.00000000 Confirmed
## GatewayMathStatus               0.93939394 Confirmed
## GatewayEnglishStatus            1.00000000 Confirmed
## CompleteDevMath                 1.00000000 Confirmed
## CompleteDevEnglish              1.00000000 Confirmed
## Major1                          1.00000000 Confirmed
## Complete1                       1.00000000 Confirmed
## Complete2                       0.00000000  Rejected
## CompleteCIP1                    1.00000000 Confirmed
```

```
## CompleteCIP2                0.00000000  Rejected
## DegreeTypeSought            0.00000000  Rejected
## TermGPA                     1.00000000 Confirmed
## CumGPA                      1.00000000 Confirmed
```

It's time for results now. Let's obtain the list of confirmed attributes

```
features=getSelectedAttributes(final.boruta, withTentative = F)
features #List all selected features
```

```
##  [1] "cohort_term"                  "Marital.Status"
##  [3] "Adjusted.Gross.Income"        "Parent.Adjusted.Gross.Income"
##  [5] "Father.s.Highest.Grade.Level" "Mother.s.Highest.Grade.Level"
##  [7] "Housing"                      "Total_loan"
##  [9] "Total_grant"                  "Total_scholarship"
## [11] "Total_WorkStudy"              "Cohort"
## [13] "CohortTerm"                   "RegistrationDate"
## [15] "Hispanic"                     "Black"
## [17] "NativeHawaiian"               "White"
## [19] "HSDipYr"                      "HSGPAUnwtd"
## [21] "EnrollmentStatus"             "NumColCredAttemptTransfer"
## [23] "NumColCredAcceptTransfer"     "CumLoanAtEntry"
## [25] "HighDeg"                      "MathPlacement"
## [27] "EngPlacement"                 "GatewayMathStatus"
## [29] "GatewayEnglishStatus"         "CompleteDevMath"
## [31] "CompleteDevEnglish"           "Major1"
## [33] "Complete1"                    "CompleteCIP1"
## [35] "TermGPA"                      "CumGPA"
```

```
# Apply the features onto the TRAIN imputed data
TRAIN_Features=TRAIN_Imputed[,features]

# Attach the Dropout variable back to the TRAIN Features data frame
TRAIN_Features$Dropout=TRAIN_Imputed$Dropout
dim(TRAIN_Features)
```

```
## [1] 12261    37
```

```
# Also apply the feature selection on the TEST DATA
TEST_Features=TEST_Imputed[,features]


# Save the TRAIN and TEST features in csv files because we can load them
later to avoid the time consuming process of feature engineering.

#write.csv(TRAIN_Features,"D: /Hamed/KAGGLE
COMPETITION/FEATURES/clean_features/TRAIN_Features.csv")

#write.csv(TEST_Features,"D:/Hamed/KAGGLE
COMPETITION/FEATURES/clean_features/TEST_Features.csv")
```

Let's understand the parameters used in Boruta as follows:

- maxRuns: maximal number of random forest runs. You can consider increasing this parameter if tentative attributes are left.Default is 100.
- doTrace: It refers to verbosity level. 0 means no tracing. 1 means reporting attribute decision as soon as it is cleared. 2 means all of 1 plus additionally reporting each iteration. Default is 0.
- holdHistory: The full history of importance runs is stored if set to TRUE (Default). Gives a plot of Classifier run vs. Importance when the plotImpHistory function is called to run.

Data Standardization:

- Derive Dummy variables out of categorical variables and normalization of continuous variables

```
# Create Dummy varibales
library(dummies)

## dummies-1.5.6 provided by Decision Patterns

TRAIN_Features <- dummy.data.frame(TRAIN_Features, names =
c("cohort_term","Marital.Status" ,
"Father.s.Highest.Grade.Level","Mother.s.Highest.Grade.Level",
                  "Housing","Cohort") , sep = ".")

## Warning in model.matrix.default(~x - 1, model.frame(~x - 1), contrasts =
FALSE):
## non-list contrasts argument ignored

## Warning in model.matrix.default(~x - 1, model.frame(~x - 1), contrasts =
FALSE):
## non-list contrasts argument ignored

## Warning in model.matrix.default(~x - 1, model.frame(~x - 1), contrasts =
FALSE):
## non-list contrasts argument ignored

## Warning in model.matrix.default(~x - 1, model.frame(~x - 1), contrasts =
FALSE):
## non-list contrasts argument ignored

## Warning in model.matrix.default(~x - 1, model.frame(~x - 1), contrasts =
FALSE):
## non-list contrasts argument ignored

## Warning in model.matrix.default(~x - 1, model.frame(~x - 1), contrasts =
FALSE):
## non-list contrasts argument ignored

dim(TRAIN_Features)
```

```
## [1] 12261    54
```

Standardization of continuous variables

TRAIN DATA transformation

```
library(caret)

## Loading required package: lattice

##
## Attaching package: 'caret'

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

# calculate the pre-process parameters from the dataset
preprocessParams <- preProcess(TRAIN_Features[,1:52], method=c("center",
"scale"))
# summarize transform parameters
print(preprocessParams)

## Created from 12261 samples and 52 variables
##
## Pre-processing:
##   - centered (43)
##   - ignored (9)
##   - scaled (43)

# transform the dataset using the parameters
transformed_train <- predict(preprocessParams, TRAIN_Features[,1:52])
#exclude Dropout variable
# summarize the transformed dataset
summary(transformed_train)

##  cohort_term.1     cohort_term.3     Marital.Status.Divorced
##  Min.   :-2.0217   Min.   :-0.4946   Min.   :-0.1314
##  1st Qu.: 0.4946   1st Qu.:-0.4946   1st Qu.:-0.1314
##  Median : 0.4946   Median :-0.4946   Median :-0.1314
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.4946   3rd Qu.:-0.4946   3rd Qu.:-0.1314
##  Max.   : 0.4946   Max.   : 2.0217   Max.   : 7.6120
##
##  Marital.Status.Married Marital.Status.Separated Marital.Status.Single
##  Min.   :-0.2855        Min.   :-0.1238          Min.   :-2.8826
##  1st Qu.:-0.2855        1st Qu.:-0.1238          1st Qu.: 0.3469
##  Median :-0.2855        Median :-0.1238          Median : 0.3469
##  Mean   : 0.0000        Mean   : 0.0000          Mean   : 0.0000
##  3rd Qu.:-0.2855        3rd Qu.:-0.1238          3rd Qu.: 0.3469
##  Max.   : 3.5026        Max.   : 8.0790          Max.   : 0.3469
##
```

```
##  Adjusted.Gross.Income Parent.Adjusted.Gross.Income
##  Min.   :-1.03904      Min.   :-1.7836
##  1st Qu.:-0.32900      1st Qu.:-0.5842
##  Median :-0.32897      Median :-0.5842
##  Mean   : 0.00000      Mean   : 0.0000
##  3rd Qu.: 0.06688      3rd Qu.: 0.2164
##  Max.   :74.87315      Max.   :15.3805
##
##  Father.s.Highest.Grade.Level.College Father.s.Highest.Grade.Level.High
School
##  Min.   :-0.5586                      Min.   :-1.0702
##  1st Qu.:-0.5586                      1st Qu.:-1.0702
##  Median :-0.5586                      Median : 0.9343
##  Mean   : 0.0000                      Mean   : 0.0000
##  3rd Qu.:-0.5586                      3rd Qu.: 0.9343
##  Max.   : 1.7901                      Max.   : 0.9343
##
##  Father.s.Highest.Grade.Level.Middle School
##  Min.   :-0.3295
##  1st Qu.:-0.3295
##  Median :-0.3295
##  Mean   : 0.0000
##  3rd Qu.:-0.3295
##  Max.   : 3.0345
##
##  Father.s.Highest.Grade.Level.Unknown Mother.s.Highest.Grade.Level.College
##  Min.   :-0.3871                      Min.   :-0.5561
##  1st Qu.:-0.3871                      1st Qu.:-0.5561
##  Median :-0.3871                      Median :-0.5561
##  Mean   : 0.0000                      Mean   : 0.0000
##  3rd Qu.:-0.3871                      3rd Qu.:-0.5561
##  Max.   : 2.5831                      Max.   : 1.7982
##
##  Mother.s.Highest.Grade.Level.High School
##  Min.   :-1.0935
##  1st Qu.:-1.0935
##  Median : 0.9145
##  Mean   : 0.0000
##  3rd Qu.: 0.9145
##  Max.   : 0.9145
##
##  Mother.s.Highest.Grade.Level.Middle School
##  Min.   :-0.3222
##  1st Qu.:-0.3222
##  Median :-0.3222
##  Mean   : 0.0000
##  3rd Qu.:-0.3222
##  Max.   : 3.1037
##
##  Mother.s.Highest.Grade.Level.Unknown Housing.Off Campus
```

```
##  Min.   :-0.3783                      Min.   :-1.0996
##  1st Qu.:-0.3783                      1st Qu.:-1.0996
##  Median :-0.3783                      Median : 0.9094
##  Mean   : 0.0000                      Mean   : 0.0000
##  3rd Qu.:-0.3783                      3rd Qu.: 0.9094
##  Max.   : 2.6433                      Max.   : 0.9094
##
##  Housing.On Campus Housing Housing.With Parent   Total_loan
##  Min.   :-0.3633           Min.   :-0.7114     Min.   :-0.7311
##  1st Qu.:-0.3633           1st Qu.:-0.7114     1st Qu.:-0.7311
##  Median :-0.3633           Median :-0.7114     Median :-0.4212
##  Mean   : 0.0000           Mean   : 0.0000     Mean   : 0.0000
##  3rd Qu.:-0.3633           3rd Qu.: 1.4056     3rd Qu.: 0.3803
##  Max.   : 2.7520           Max.   : 1.4056     Max.   : 7.6244
##
##   Total_grant      Total_scholarship Total_WorkStudy   Cohort.2011-12
##  Min.   :-0.7725   Min.   :-0.2404   Min.   :-0.2127   Min.   :-0.4586
##  1st Qu.:-0.7725   1st Qu.:-0.2404   1st Qu.:-0.2127   1st Qu.:-0.4586
##  Median :-0.3528   Median :-0.2404   Median :-0.2127   Median :-0.4586
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.3516   3rd Qu.:-0.2404   3rd Qu.:-0.2127   3rd Qu.:-0.4586
##  Max.   : 5.6750   Max.   :25.5464   Max.   :14.9001   Max.   : 2.1802
##
##  Cohort.2012-13    Cohort.2013-14   Cohort.2014-15   Cohort.2015-16
##  Min.   :-0.4492   Min.   :-0.433   Min.   :-0.452   Min.   :-0.4655
##  1st Qu.:-0.4492   1st Qu.:-0.433   1st Qu.:-0.452   1st Qu.:-0.4655
##  Median :-0.4492   Median :-0.433   Median :-0.452   Median :-0.4655
##  Mean   : 0.0000   Mean   : 0.000   Mean   : 0.000   Mean   : 0.0000
##  3rd Qu.:-0.4492   3rd Qu.:-0.433   3rd Qu.:-0.452   3rd Qu.:-0.4655
##  Max.   : 2.2259   Max.   : 2.309   Max.   : 2.212   Max.   : 2.1479
##
##  Cohort.2016-17    CohortTerm RegistrationDate     Hispanic
##  Min.   :-0.4243   1:9851     Min.   :-1.5339   Min.   :-0.6945
##  1st Qu.:-0.4243   3:2410     1st Qu.:-0.9101   1st Qu.:-0.6945
##  Median :-0.4243              Median : 0.2325   Median :-0.6945
##  Mean   : 0.0000              Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.:-0.4243              3rd Qu.: 0.8506   3rd Qu.: 1.4397
##  Max.   : 2.3564              Max.   : 1.4562   Max.   : 1.4397
##
##      Black          NativeHawaiian          White            HSDipYr
##  Min.   :-0.5239   Min.   :-0.04142   Min.   :-0.5789   Min.   :-21.4281
##  1st Qu.:-0.5239   1st Qu.:-0.04142   1st Qu.:-0.5789   1st Qu.:  0.3636
##  Median :-0.5239   Median :-0.04142   Median :-0.5789   Median :  0.3636
##  Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000   Mean   :  0.0000
##  3rd Qu.:-0.5239   3rd Qu.:-0.04142   3rd Qu.: 1.7272   3rd Qu.:  0.3636
##  Max.   : 1.9085   Max.   :24.14145   Max.   : 1.7272   Max.   :  0.3636
##
##    HSGPAUnwtd      EnrollmentStatus NumColCredAttemptTransfer
##  Min.   :-1.1848   1:4952           Min.   :-0.9100
##  1st Qu.:-0.6341   2:7309           1st Qu.:-0.9100
```

```
##  Median :-0.6341                    Median :-0.4955
##  Mean   : 0.0000                    Mean   : 0.0000
##  3rd Qu.: 0.8860                    3rd Qu.: 0.8170
##  Max.   : 2.6703                    Max.   : 2.5901
##
##  NumColCredAcceptTransfer CumLoanAtEntry    HighDeg  MathPlacement
EngPlacement
##  Min.   :-0.9887          Min.   :-1.2103   0:8710   0:8379        0:9486
##  1st Qu.:-0.9887          1st Qu.:-1.2103   2:3406   1:3882        1:2775
##  Median :-0.2358          Median : 0.7988   3: 143
##  Mean   : 0.0000          Mean   : 0.0000   4:   2
##  3rd Qu.: 0.9805          3rd Qu.: 0.7988
##  Max.   : 1.8493          Max.   : 1.4685
##
##  GatewayMathStatus GatewayEnglishStatus          CompleteDevMath
##  0:10794           0:9967                -2                :8377
##  1: 1467           1:2294                0                 :1478
##                                          0.5               : 443
##                                          0.25              : 379
##                                          1                 : 213
##                                          0.333333333333333: 205
##                                          (Other)           :1166
##          CompleteDevEnglish     Major1          Complete1
##  -2                :9387     Min.   :-2.2302   Min.   :-0.5385
##  0                 : 773     1st Qu.:-0.6385   1st Qu.:-0.5385
##  0.5               : 319     Median : 0.3827   Median :-0.5385
##  1                 : 311     Mean   : 0.0000   Mean   : 0.0000
##  0.25              : 197     3rd Qu.: 0.8847   3rd Qu.: 0.3959
##  0.333333333333333: 170     Max.   : 1.0426   Max.   : 4.2669
##  (Other)           :1104
##   CompleteCIP1        TermGPA
##  Min.   :-0.5149   Min.   :-2.7625
##  1st Qu.:-0.5149   1st Qu.:-0.4139
##  Median :-0.5149   Median : 0.2529
##  Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.2517   3rd Qu.: 0.7465
##  Max.   : 4.7306   Max.   : 1.1600
##
```

```
dim(transformed_train)
```

```
## [1] 12261    52
```

```
transformed_train$Dropout=TRAIN_Features$Dropout #re-include Dropout variable


# save the claen and transformed features into a local folder for future use
#write.csv(transformed_train,"D:/Hamed/KAGGLE
COMPETITION/FEATURES/clean_features/transformed_train.csv")
```

TEST DATA transformation

```
vars=c("cohort_term","Marital.Status" ,"Father.s.Highest.Grade.Level",
       "Mother.s.Highest.Grade.Level","Housing","Cohort")
TEST_Features[,vars] <- lapply(TEST_Features[,vars] , factor)
TEST_Features<-dummy.data.frame(TEST_Features,names = vars,sep=".")
dim(TEST_Features)
```

```
## [1] 1000   53
```

```
# Also standardize the test features
library(caret)
# calculate the pre-process parameters from the dataset
preprocessParams <- preProcess(TEST_Features[,1:52], method=c("center",
"scale"))
# summarize transform parameters
print(preprocessParams)
```

```
## Created from 1000 samples and 52 variables
##
## Pre-processing:
##   - centered (52)
##   - ignored (0)
##   - scaled (52)
```

```
# transform the dataset using the parameters
transformed_test <- predict(preprocessParams, TEST_Features[,1:52])
# summarize the transformed dataset
summary(transformed_test)
```

```
##  cohort_term.1     cohort_term.3     Marital.Status.Divorced
##  Min.   :-2.1048   Min.   :-0.4746   Min.   :-0.1353
##  1st Qu.: 0.4746   1st Qu.:-0.4746   1st Qu.:-0.1353
##  Median : 0.4746   Median :-0.4746   Median :-0.1353
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.4746   3rd Qu.:-0.4746   3rd Qu.:-0.1353
##  Max.   : 0.4746   Max.   : 2.1048   Max.   : 7.3825
##  Marital.Status.Married Marital.Status.Separated Marital.Status.Single
##  Min.   :-0.27          Min.   :-0.08976         Min.   :-3.1030
##  1st Qu.:-0.27          1st Qu.:-0.08976         1st Qu.: 0.3219
##  Median :-0.27          Median :-0.08976         Median : 0.3219
##  Mean   : 0.00          Mean   : 0.00000         Mean   : 0.0000
##  3rd Qu.:-0.27          3rd Qu.:-0.08976         3rd Qu.: 0.3219
##  Max.   : 3.70          Max.   :11.12996         Max.   : 0.3219
##  Adjusted.Gross.Income Parent.Adjusted.Gross.Income
##  Min.   :-0.5076       Min.   :-2.2776
##  1st Qu.:-0.4528       1st Qu.:-0.6290
##  Median :-0.4528       Median :-0.4771
##  Mean   : 0.0000       Mean   : 0.0000
##  3rd Qu.: 0.0329       3rd Qu.: 0.2477
##  Max.   : 8.7652       Max.   : 9.7366
```

```
##  Father.s.Highest.Grade.Level.College Father.s.Highest.Grade.Level.High
School
##  Min.   :-0.5848                      Min.   :-1.03
##  1st Qu.:-0.5848                      1st Qu.:-1.03
##  Median :-0.5848                      Median : 0.97
##  Mean   : 0.0000                      Mean   : 0.00
##  3rd Qu.: 1.7084                      3rd Qu.: 0.97
##  Max.   : 1.7084                      Max.   : 0.97
##  Father.s.Highest.Grade.Level.Middle School
##  Min.   :-0.3387
##  1st Qu.:-0.3387
##  Median :-0.3387
##  Mean   : 0.0000
##  3rd Qu.:-0.3387
##  Max.   : 2.9496
##  Father.s.Highest.Grade.Level.Unknown Mother.s.Highest.Grade.Level.College
##  Min.   :-0.3812                      Min.   :-0.554
##  1st Qu.:-0.3812                      1st Qu.:-0.554
##  Median :-0.3812                      Median :-0.554
##  Mean   : 0.0000                      Mean   : 0.000
##  3rd Qu.:-0.3812                      3rd Qu.:-0.554
##  Max.   : 2.6205                      Max.   : 1.803
##  Mother.s.Highest.Grade.Level.High School
##  Min.   :-1.1005
##  1st Qu.:-1.1005
##  Median : 0.9077
##  Mean   : 0.0000
##  3rd Qu.: 0.9077
##  Max.   : 0.9077
##  Mother.s.Highest.Grade.Level.Middle School
##  Min.   :-0.3442
##  1st Qu.:-0.3442
##  Median :-0.3442
##  Mean   : 0.0000
##  3rd Qu.:-0.3442
##  Max.   : 2.9027
##  Mother.s.Highest.Grade.Level.Unknown Housing.Off Campus
##  Min.   :-0.3532                      Min.   :-0.719
##  1st Qu.:-0.3532                      1st Qu.:-0.719
##  Median :-0.3532                      Median :-0.719
##  Mean   : 0.0000                      Mean   : 0.000
##  3rd Qu.:-0.3532                      3rd Qu.: 1.389
##  Max.   : 2.8286                      Max.   : 1.389
##  Housing.On Campus Housing Housing.With Parent   Total_loan
##  Min.   :-0.4066           Min.   :-1.0341     Min.   :-0.7157
##  1st Qu.:-0.4066           1st Qu.:-1.0341     1st Qu.:-0.7157
##  Median :-0.4066           Median : 0.9661     Median :-0.4342
##  Mean   : 0.0000           Mean   : 0.0000     Mean   : 0.0000
##  3rd Qu.:-0.4066           3rd Qu.: 0.9661     3rd Qu.: 0.3597
##  Max.   : 2.4569           Max.   : 0.9661     Max.   : 5.5309
```

```
##   Total_grant      Total_scholarship Total_WorkStudy   Cohort.2011-12
##  Min.   :-0.7952   Min.   :-0.2861   Min.   :-0.2608   Min.   :-0.4539
##  1st Qu.:-0.7952   1st Qu.:-0.2861   1st Qu.:-0.2608   1st Qu.:-0.4539
##  Median :-0.3840   Median :-0.2861   Median :-0.2608   Median :-0.4539
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.3815   3rd Qu.:-0.2861   3rd Qu.:-0.2608   3rd Qu.:-0.4539
##  Max.   : 4.6655   Max.   :11.8541   Max.   : 8.5800   Max.   : 2.2007
##  Cohort.2012-13    Cohort.2013-14    Cohort.2014-15    Cohort.2015-16
##  Min.   :-0.5122   Min.   :-0.4049   Min.   :-0.4427   Min.   :-0.4475
##  1st Qu.:-0.5122   1st Qu.:-0.4049   1st Qu.:-0.4427   1st Qu.:-0.4475
##  Median :-0.5122   Median :-0.4049   Median :-0.4427   Median :-0.4475
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.:-0.5122   3rd Qu.:-0.4049   3rd Qu.:-0.4427   3rd Qu.:-0.4475
##  Max.   : 1.9504   Max.   : 2.4670   Max.   : 2.2566   Max.   : 2.2323
##  Cohort.2016-17      CohortTerm      RegistrationDate     Hispanic
##  Min.   :-0.4182   Min.   :-0.4746   Min.   :-1.4823   Min.   :-0.7062
##  1st Qu.:-0.4182   1st Qu.:-0.4746   1st Qu.:-0.8603   1st Qu.:-0.7062
##  Median :-0.4182   Median :-0.4746   Median :-0.2573   Median :-0.7062
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.:-0.4182   3rd Qu.:-0.4746   3rd Qu.: 0.8975   3rd Qu.: 1.4146
##  Max.   : 2.3887   Max.   : 2.1048   Max.   : 1.5023   Max.   : 1.4146
##      Black          NativeHawaiian         White            HSDipYr
##  Min.   :-0.4935   Min.   :-0.03162   Min.   :-0.6032   Min.   :-16.653
##  1st Qu.:-0.4935   1st Qu.:-0.03162   1st Qu.:-0.6032   1st Qu.: -0.074
##  Median :-0.4935   Median :-0.03162   Median :-0.6032   Median : -0.074
##  Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000   Mean   :  0.000
##  3rd Qu.:-0.4935   3rd Qu.:-0.03162   3rd Qu.: 1.6561   3rd Qu.: -0.074
##  Max.   : 2.0243   Max.   :31.59115   Max.   : 1.6561   Max.   :  1.543
##    HSGPAUnwtd      EnrollmentStatus  NumColCredAttemptTransfer
##  Min.   :-1.3679   Min.   :-0.9995   Min.   :-0.7866
##  1st Qu.:-0.4940   1st Qu.:-0.9995   1st Qu.:-0.7866
##  Median :-0.4940   Median : 0.0000   Median :-0.7627
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.9764   3rd Qu.: 0.9995   3rd Qu.: 0.8810
##  Max.   : 2.1277   Max.   : 0.9995   Max.   : 2.8344
##  NumColCredAcceptTransfer CumLoanAtEntry       HighDeg
MathPlacement
##  Min.   :-0.8511          Min.   :-0.9745   Min.   :-0.5677   Min.   :-
0.8042
##  1st Qu.:-0.8511          1st Qu.:-0.9745   1st Qu.:-0.5677   1st Qu.:-
0.8042
##  Median :-0.8219          Median :-0.2937   Median :-0.5677   Median :-
0.8042
##  Mean   : 0.0000          Mean   : 0.0000   Mean   : 0.0000   Mean   :
0.0000
##  3rd Qu.: 1.1042          3rd Qu.: 1.0680   3rd Qu.:-0.5677   3rd Qu.:
1.2422
##  Max.   : 2.0089          Max.   : 1.7489   Max.   : 2.7849   Max.   :
1.2422
##   EngPlacement     GatewayMathStatus GatewayEnglishStatus CompleteDevMath
```

```
##  Min.   :-0.6156   Min.   :-0.3708   Min.   :-0.5432      Min.   :-0.7957
##  1st Qu.:-0.6156   1st Qu.:-0.3708   1st Qu.:-0.5432      1st Qu.:-0.7957
##  Median :-0.6156   Median :-0.3708   Median :-0.5432      Median :-0.7957
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000      Mean   : 0.0000
##  3rd Qu.: 1.6229   3rd Qu.:-0.3708   3rd Qu.:-0.5432      3rd Qu.: 1.0914
##  Max.   : 1.6229   Max.   : 2.6939   Max.   : 1.8392      Max.   : 1.9217
##  CompleteDevEnglish     Major1          Complete1        CompleteCIP1
##  Min.   :-0.6124    Min.   :-2.0334   Min.   :-0.5365   Min.   :-0.5136
##  1st Qu.:-0.6124    1st Qu.:-0.7152   1st Qu.:-0.5365   1st Qu.:-0.5136
##  Median :-0.6124    Median : 0.4311   Median :-0.5365   Median :-0.5136
##  Mean   : 0.0000    Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 1.3061    3rd Qu.: 0.9100   3rd Qu.: 0.4380   3rd Qu.: 0.2607
##  Max.   : 2.2653    Max.   : 1.0607   Max.   : 4.4753   Max.   : 4.6529
##     TermGPA
##  Min.   :-2.7566
##  1st Qu.:-0.3919
##  Median : 0.2765
##  Mean   : 0.0000
##  3rd Qu.: 0.7241
##  Max.   : 1.1522
```

```
dim(transformed_test)
```

```
## [1] 1000   52
```

```
# save the claen and transformed features into a local folder for future use
#write.csv(transformed_test,"D:/Hamed/KAGGLE
COMPETITION/FEATURES/clean_features/transformed_test.csv")
```