R Notebook

# DATA WRANGLING

### FINANCIAL AID DATA

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

Combine by summing the loan, scholarship, grant and work/study data from 2011 to 2017

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 : financial aid data

# 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

#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:/KAZI/UPWORK/KAGGLE COMPETITION/Student Retention Challenge Data/Student Static Data")  
  
# 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)

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

## [1] TRUE

JOIN the train labels and the static data

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

Progress data files loaded and merged

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:/KAZI/UPWORK/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

Summarise the data to get rid of duplicated StudentID

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(CompleteDevEnglish),  
 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 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

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','CompleteDevMath',  
 '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)

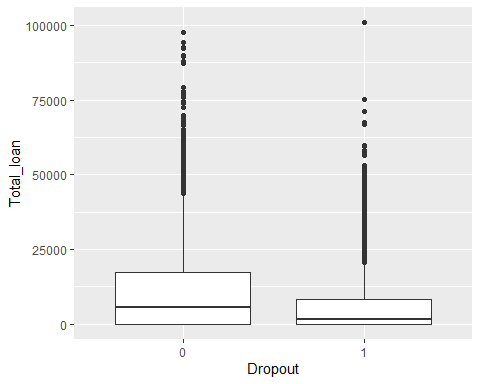
## 'data.frame': 12261 obs. of 59 variables:  
## $ StudentID : int 285848 302176 301803 302756 301067 297371 273211 302772 280023 300412 ...  
## $ Dropout : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 2 2 1 2 ...  
## $ cohort\_term : Factor w/ 2 levels "1","3": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Marital.Status : 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

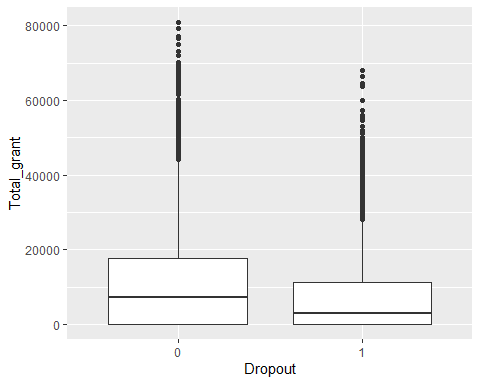
# Summarize the data  
summary(TRAIN\_DATA)

## 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   
##

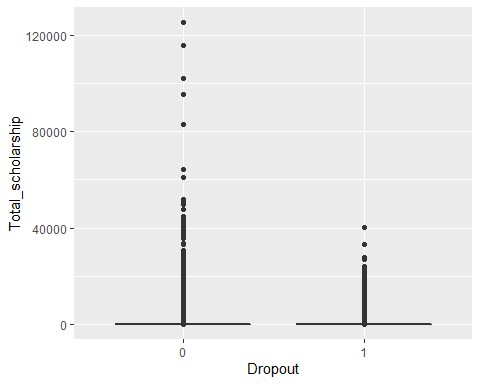
# visualize the financial information in relation to the dropout variable  
library(dplyr)  
ggplot(data = TRAIN\_DATA, mapping = aes(x =Dropout,y=Total\_loan))+  
 geom\_boxplot()



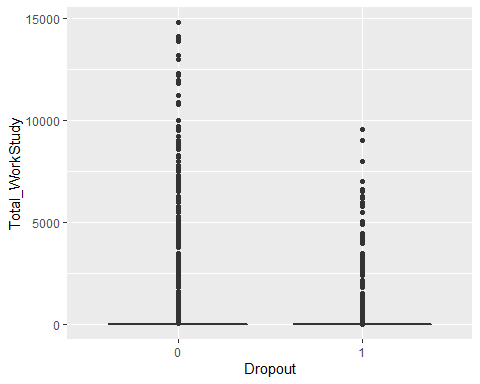
ggplot(data = TRAIN\_DATA, mapping = aes(x =Dropout,y=Total\_grant))+  
 geom\_boxplot()



ggplot(data = TRAIN\_DATA, mapping = aes(x =Dropout,y=Total\_scholarship))+  
 geom\_boxplot()



ggplot(data = TRAIN\_DATA, mapping = aes(x =Dropout,y=Total\_WorkStudy))+  
 geom\_boxplot()



Frequency analysis for categorical variables in relation to the Dropout variable

# Dropout  
table(TRAIN\_DATA$Dropout,exclude = NULL)

##   
## 0 1   
## 7527 4734

# 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

# 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

# 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

# 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

# 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

# CohortTerm  
table(TRAIN\_DATA$CohortTerm,TRAIN\_DATA$Dropout,exclude = NULL)

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

# MathPlacement  
table(TRAIN\_DATA$MathPlacement,TRAIN\_DATA$Dropout,exclude = NULL)

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

# 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

# Gender  
table(TRAIN\_DATA$Gender,TRAIN\_DATA$Dropout,exclude = NULL)

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

# Hispanic  
table(TRAIN\_DATA$Hispanic,TRAIN\_DATA$Dropout,exclude = NULL)

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

# AmericanIndian  
table(TRAIN\_DATA$AmericanIndian,TRAIN\_DATA$Dropout,exclude = NULL)

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

# Asian  
table(TRAIN\_DATA$Asian,TRAIN\_DATA$Dropout,exclude = NULL)

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

# Black  
table(TRAIN\_DATA$Black,TRAIN\_DATA$Dropout,exclude = NULL)

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

# NativeHawaiian  
table(TRAIN\_DATA$NativeHawaiian,TRAIN\_DATA$Dropout,exclude = NULL)

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

# White  
table(TRAIN\_DATA$White,TRAIN\_DATA$Dropout,exclude = NULL)

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

# TwoOrMoreRace  
table(TRAIN\_DATA$TwoOrMoreRace,TRAIN\_DATA$Dropout,exclude = NULL)

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

# HSDip  
table(TRAIN\_DATA$HSDip,TRAIN\_DATA$Dropout,exclude = NULL)

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

# TwoOrMoreRace  
table(TRAIN\_DATA$TwoOrMoreRace,TRAIN\_DATA$Dropout,exclude = NULL)

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

# HighDeg  
table(TRAIN\_DATA$HighDeg,TRAIN\_DATA$Dropout,exclude = NULL)

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

# EngPlacement  
table(TRAIN\_DATA$EngPlacement,TRAIN\_DATA$Dropout,exclude = NULL)

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

# GatewayMathStatus  
table(TRAIN\_DATA$GatewayMathStatus,TRAIN\_DATA$Dropout)

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

Correlation matrix of the financial aid data to check for multicollinearity among the continuous independent variables

# 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",  
 "Adjusted.Gross.Income","Parent.Adjusted.Gross.Income","CumLoanAtEntry")  
ggpairs(TRAIN\_DATA[vars])

## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :  
## Removed 1841 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :  
## Removed 1841 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :  
## Removed 7309 rows containing missing values

## Warning in cor(x, y, method = method, use = use): the standard deviation is zero

## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :  
## Removed 1841 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :  
## Removed 1841 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :  
## Removed 7309 rows containing missing values

## Warning in cor(x, y, method = method, use = use): the standard deviation is zero

## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :  
## Removed 1841 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :  
## Removed 1841 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :  
## Removed 7309 rows containing missing values

## Warning in cor(x, y, method = method, use = use): the standard deviation is zero

## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :  
## Removed 1841 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :  
## Removed 1841 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :  
## Removed 7309 rows containing missing values

## Warning in cor(x, y, method = method, use = use): the standard deviation is zero

## Warning: Removed 1841 rows containing missing values (geom\_point).  
  
## Warning: Removed 1841 rows containing missing values (geom\_point).  
  
## Warning: Removed 1841 rows containing missing values (geom\_point).  
  
## Warning: Removed 1841 rows containing missing values (geom\_point).

## Warning: Removed 1841 rows containing non-finite values (stat\_density).

## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :  
## Removed 1841 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :  
## Removed 7908 rows containing missing values

## Warning in cor(x, y, method = method, use = use): the standard deviation is zero

## Warning: Removed 1841 rows containing missing values (geom\_point).  
  
## Warning: Removed 1841 rows containing missing values (geom\_point).  
  
## Warning: Removed 1841 rows containing missing values (geom\_point).  
  
## Warning: Removed 1841 rows containing missing values (geom\_point).  
  
## Warning: Removed 1841 rows containing missing values (geom\_point).

## Warning: Removed 1841 rows containing non-finite values (stat\_density).

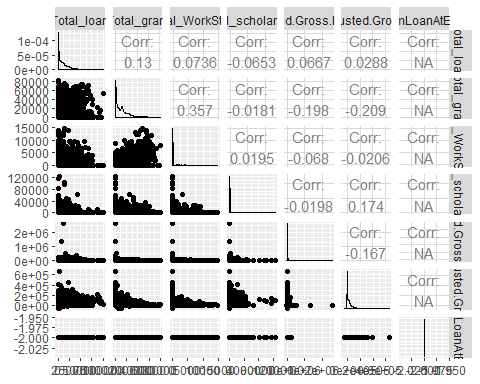
## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :  
## Removed 7908 rows containing missing values

## Warning in cor(x, y, method = method, use = use): the standard deviation is zero

## Warning: Removed 7309 rows containing missing values (geom\_point).  
  
## Warning: Removed 7309 rows containing missing values (geom\_point).  
  
## Warning: Removed 7309 rows containing missing values (geom\_point).  
  
## Warning: Removed 7309 rows containing missing values (geom\_point).

## Warning: Removed 7908 rows containing missing values (geom\_point).  
  
## Warning: Removed 7908 rows containing missing values (geom\_point).

## Warning: Removed 7309 rows containing non-finite values (stat\_density).



Correlation matrix of the student performance to check for multicollinearity among the continuous independent variables

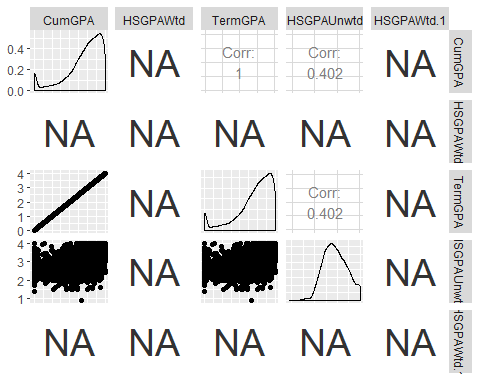
# Load the GGally package  
library(GGally)  
# Create a scatter plot matrix  
vars <- c("CumGPA", "HSGPAWtd","TermGPA","HSGPAUnwtd","HSGPAWtd")  
ggpairs(TRAIN\_DATA[vars])

## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :  
## Removed 8687 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :  
## Removed 8687 rows containing missing values

## Warning: Removed 8687 rows containing missing values (geom\_point).  
  
## Warning: Removed 8687 rows containing missing values (geom\_point).

## Warning: Removed 8687 rows containing non-finite values (stat\_density).



#### Hypotheses tests

# Test for association  
# 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

# 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

# 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

#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

# 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

# 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

### IMPUTATION OF MISSING VALUES

Summarize and lookj 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 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

drop.cols <- c('HSGPAWtd', 'FirstGen','TransferIntent','Campus',  
 'Address2','Major2')  
TRAIN\_DATA<-TRAIN\_DATA %>% select(-drop.cols)  
dim(TRAIN\_DATA)

## [1] 12261 53

str(TRAIN\_DATA)

## 'data.frame': 12261 obs. of 53 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 ...  
## $ Address1 : Factor w/ 12704 levels "","1 Brookside Ave",..: 916 279 168 954 673 370 769 1347 750 1275 ...  
## $ 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 ...  
## $ 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 ...  
## $ 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 ...  
## $ 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 ...

Impute categorical variables with mode

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.Highest.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.Grade.Level)]<-mode   
  
val<-unique(TRAIN\_DATA$Mother.s.Highest.Grade.Level[!is.na(TRAIN\_DATA$Mother.s.Highest.Grade.Level)])  
mode<-val[which.max(tabulate(match(TRAIN\_DATA$Father.s.Highest.Grade.Level, val)))]   
TRAIN\_DATA$Mother.s.Highest.Grade.Level[is.na(TRAIN\_DATA$Mother.s.Highest.Grade.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$Address1[!is.na(TRAIN\_DATA$Address1)])  
mode<-val[which.max(tabulate(match(TRAIN\_DATA$Address1, val)))]   
TRAIN\_DATA$Address1[is.na(TRAIN\_DATA$Address1)]<-mode   
  
val<-unique(TRAIN\_DATA$State[!is.na(TRAIN\_DATA$State)])  
mode<-val[which.max(tabulate(match(TRAIN\_DATA$State, val)))]   
TRAIN\_DATA$State[is.na(TRAIN\_DATA$State)]<-mode   
  
val<-unique(TRAIN\_DATA$City[!is.na(TRAIN\_DATA$City)])  
mode<-val[which.max(tabulate(match(TRAIN\_DATA$City, val)))]   
TRAIN\_DATA$City[is.na(TRAIN\_DATA$City)]<-mode   
  
val<-unique(TRAIN\_DATA$Zip[!is.na(TRAIN\_DATA$Zip)])  
mode<-val[which.max(tabulate(match(TRAIN\_DATA$Zip, val)))]   
TRAIN\_DATA$Zip[is.na(TRAIN\_DATA$Zip)]<-mode   
  
val<-unique(TRAIN\_DATA$BirthYear[!is.na(TRAIN\_DATA$BirthYear)])  
mode<-val[which.max(tabulate(match(TRAIN\_DATA$BirthYear, val)))]   
TRAIN\_DATA$BirthYear[is.na(TRAIN\_DATA$BirthYear)]<-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   
  
  
# which columns have missing data  
colnames(TRAIN\_DATA)[colSums(is.na(TRAIN\_DATA)) > 0]

## [1] "Adjusted.Gross.Income" "Parent.Adjusted.Gross.Income"  
## [3] "HSDipYr" "HSGPAUnwtd"   
## [5] "NumColCredAttemptTransfer" "NumColCredAcceptTransfer"   
## [7] "CumLoanAtEntry" "MathPlacement"   
## [9] "EngPlacement" "CompleteDevMath"   
## [11] "CompleteDevEnglish" "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)

#Density plot  
library(ggpubr)

## Loading required package: magrittr

##   
## Attaching package: 'magrittr'

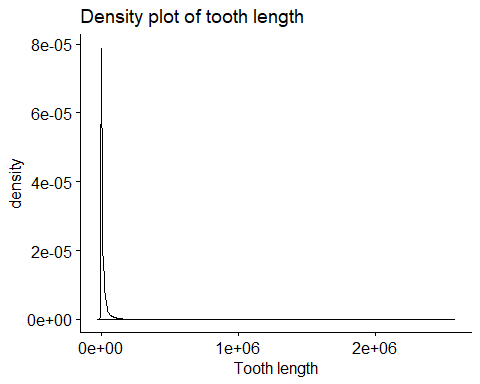
## The following object is masked from 'package:purrr':  
##   
## set\_names

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

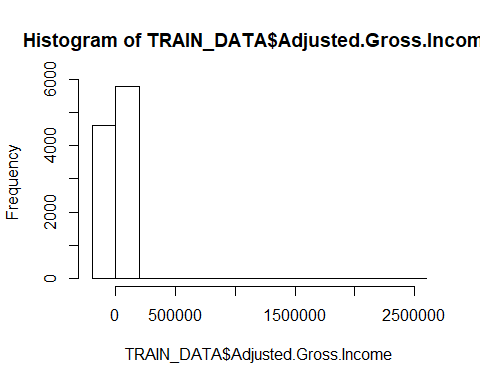
##   
## Attaching package: 'ggpubr'

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

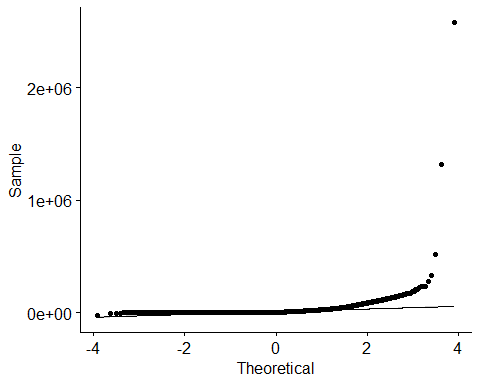
ggdensity(TRAIN\_DATA$Adjusted.Gross.Income,   
 main = "Density plot of tooth length",  
 xlab = "Tooth length")



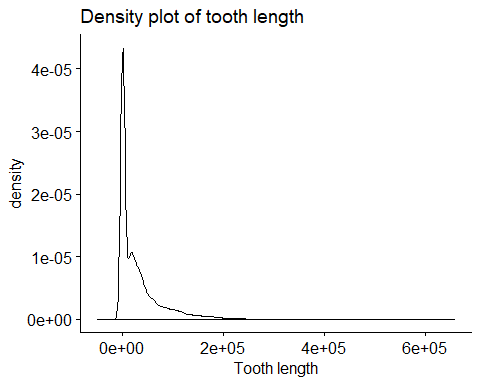
#Histogram  
  
hist(TRAIN\_DATA$Adjusted.Gross.Income) #histogram



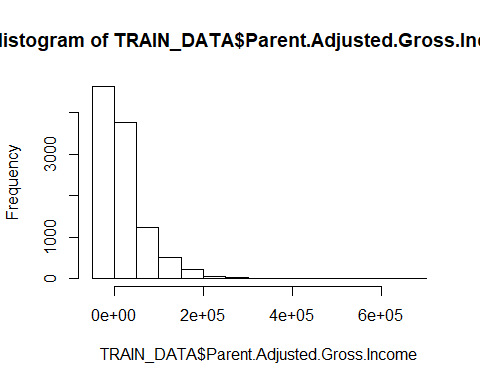
# QQplot  
ggqqplot(TRAIN\_DATA$Adjusted.Gross.Income) #qqplot



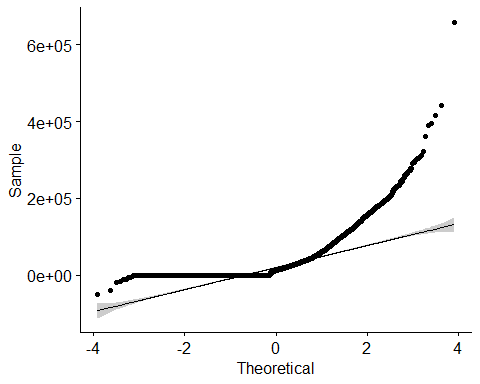
# Density plot  
ggdensity(TRAIN\_DATA$Parent.Adjusted.Gross.Income,   
 main = "Density plot of tooth length",  
 xlab = "Tooth length")



# Histogram  
hist(TRAIN\_DATA$Parent.Adjusted.Gross.Income) #histogram



# QQplot  
ggqqplot(TRAIN\_DATA$Parent.Adjusted.Gross.Income) #qqplot



* NOTE: the continous variables did not display normal distribution properties, thus imputation will be done using median

library(dplyr)  
  
# Imputation of multiple rows (i.e. the continuous variables) using median  
  
TRAIN\_DATA$Adjusted.Gross.Income[is.na(TRAIN\_DATA$Adjusted.Gross.Income)]=  
 median(TRAIN\_DATA$Adjusted.Gross.Income,na.rm = TRUE)  
  
sum(is.na(TRAIN\_DATA$Adjusted.Gross.Income))

## [1] 0

TRAIN\_DATA$Parent.Adjusted.Gross.Income[is.na(TRAIN\_DATA$ Parent.Adjusted.Gross.Income)]=median(TRAIN\_DATA$Parent.Adjusted.Gross.Income,na.rm = TRUE)  
sum(is.na(TRAIN\_DATA$Parent.Adjusted.Gross.Income))

## [1] 0

TRAIN\_DATA$HSDipYr[is.na(TRAIN\_DATA$HSDipYr)]=median(TRAIN\_DATA$HSDipYr,na.rm = TRUE)  
sum(is.na(TRAIN\_DATA$HSDipYr))

## [1] 0

TRAIN\_DATA$HSGPAUnwtd[is.na(TRAIN\_DATA$HSGPAUnwtd)]=median(TRAIN\_DATA$HSGPAUnwtd,na.rm = TRUE)  
sum(is.na(TRAIN\_DATA$HSGPAUnwtd))

## [1] 0

TRAIN\_DATA$NumColCredAcceptTransfer[is.na(TRAIN\_DATA$NumColCredAcceptTransfer)]=median(TRAIN\_DATA$NumColCredAcceptTransfer,na.rm = TRUE)  
sum(is.na(TRAIN\_DATA$NumColCredAcceptTransfer))

## [1] 0

TRAIN\_DATA$NumColCredAttemptTransfer[is.na(TRAIN\_DATA$NumColCredAttemptTransfer)]=median(TRAIN\_DATA$NumColCredAttemptTransfer,na.rm = TRUE)  
sum(is.na(TRAIN\_DATA$NumColCredAttemptTransfer))

## [1] 0

TRAIN\_DATA$CumLoanAtEntry[is.na(TRAIN\_DATA$CumLoanAtEntry)]=median(TRAIN\_DATA$CumLoanAtEntry,na.rm = TRUE)  
sum(is.na(TRAIN\_DATA$CumLoanAtEntry))

## [1] 0

TRAIN\_DATA$Major1[is.na(TRAIN\_DATA$Major1)]=median(TRAIN\_DATA$Major1,na.rm = TRUE)  
sum(is.na(TRAIN\_DATA$Major1))

## [1] 0

# check to see which columns have missing data  
colnames(TRAIN\_DATA)[colSums(is.na(TRAIN\_DATA)) > 0]

## [1] "MathPlacement" "EngPlacement" "CompleteDevMath"   
## [4] "CompleteDevEnglish"