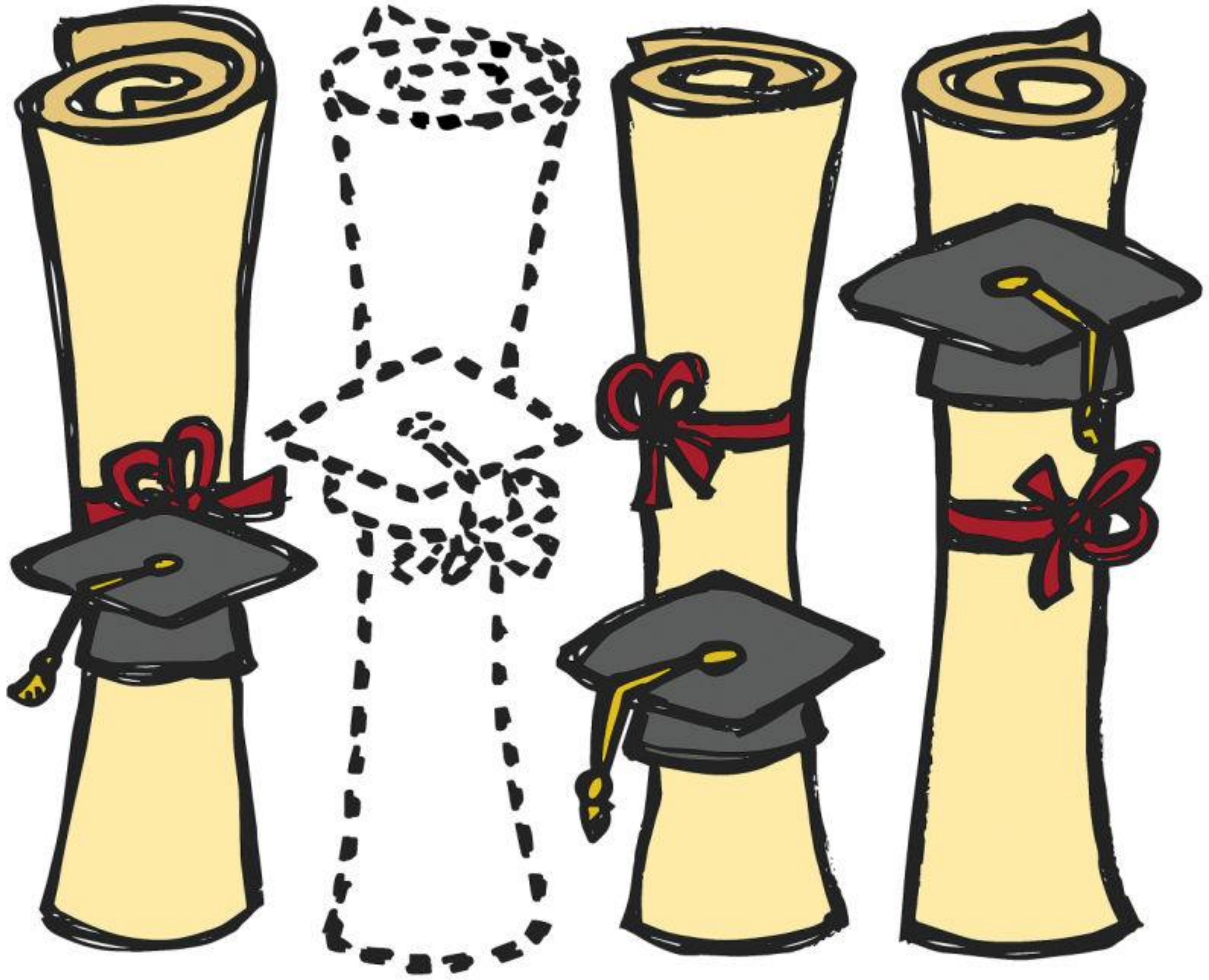
JIGSAW ACADEMY
THE ONLINE SCHOOL OF ANALYTICS

# Capstone Project

By : Joohi Rana (ID : Jig20917)

FULL STACK DATA SCIENCE PROGRAM (FSDS) -2020



# Students' Early Attrition Modelling for Clearwater State University

## Business Problem

Objective of this study is  to Identify key drivers of early student attrition and Build a predictive model to identify students with higher early attrition risk

## Methodology

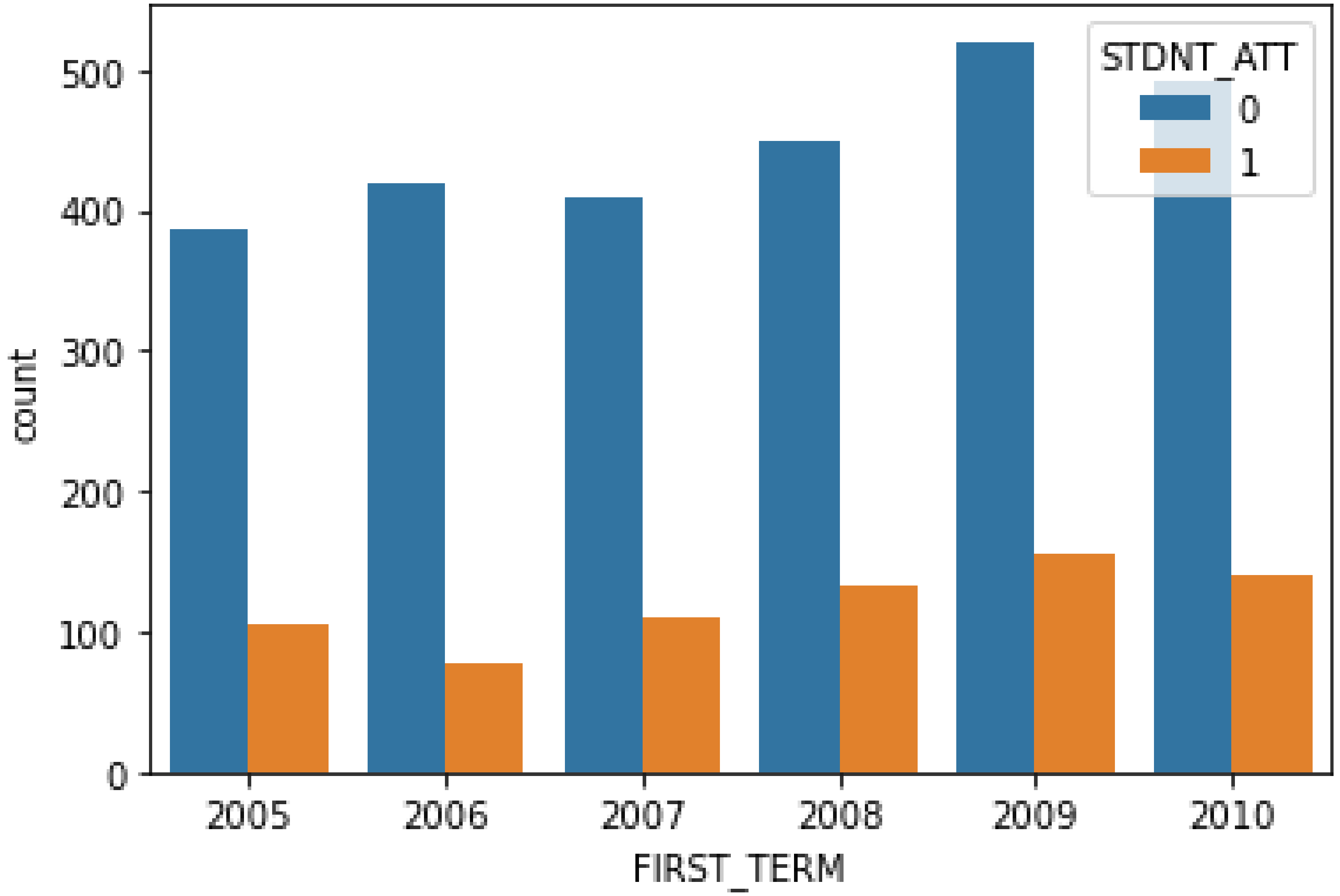
Data Pre –Processing
Handle the missing values Missing value Analysis
Exploratory Data analysis and Feature Engineering
Create an estimation sample and two validation samples by splitting the data into three groups.
Handling imbalanced datasets in machine learning
Set up the dependent variable, Student attrition (as a categorical 0-1 variable)
Create  the classification model using data, and interpret the results
Hyper parameter tuning for machine learning models.
Finally, assess the accuracy of classification

JIGSAW ACADEMY
THE ONLINE SCHOOL OF ANALYTICS

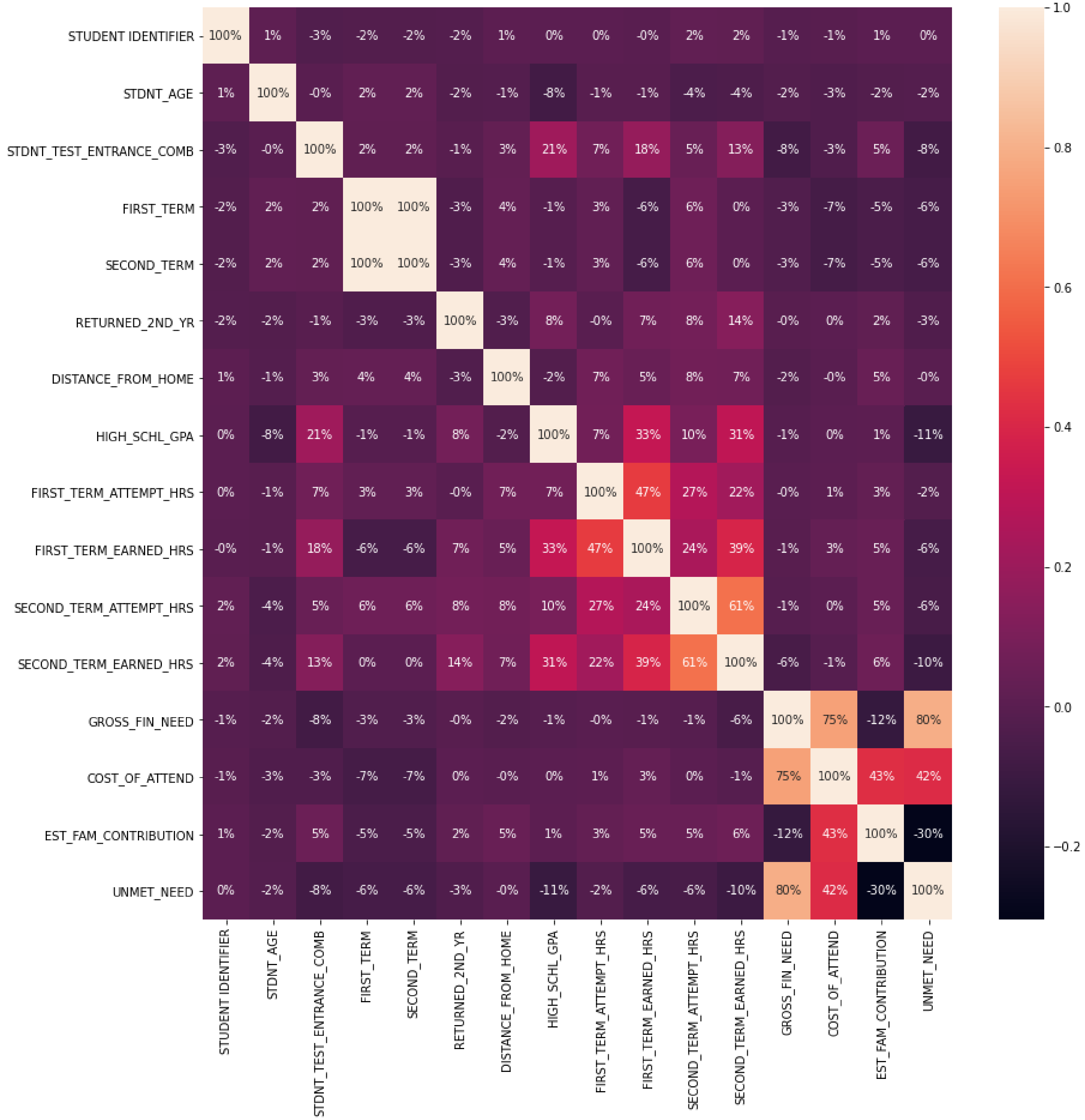
## Data Student Attrition 2005 - 2010

- The maximum attrition rate in 2009 is 4.5%
- In the three years from 2006 to 2009, the rate doubled from 2.2% to 4.5%
- 2010 It is declining but at a negligible rate
- In 6-year data (2005 to 2010), the attrition percentage is 21%
- The registration rate has been increasing from 12% from 2007 to 2009 to 15%
- As we can see there is not much increasing enrollment in the data.

- For more details, we have to check



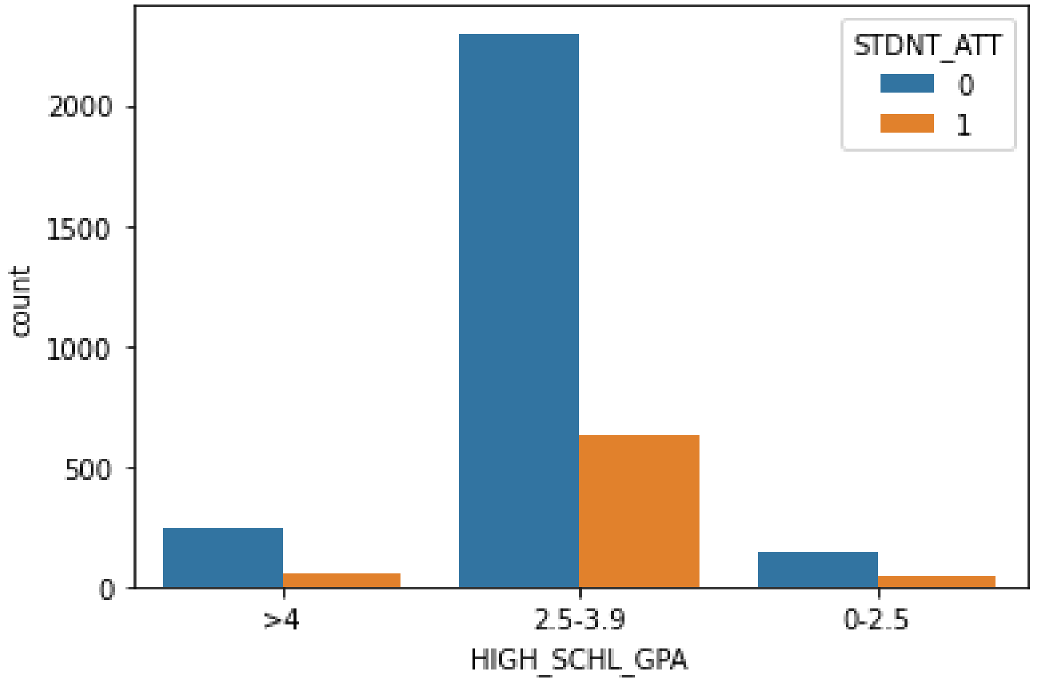
Note : given % is  in context of whole six years data

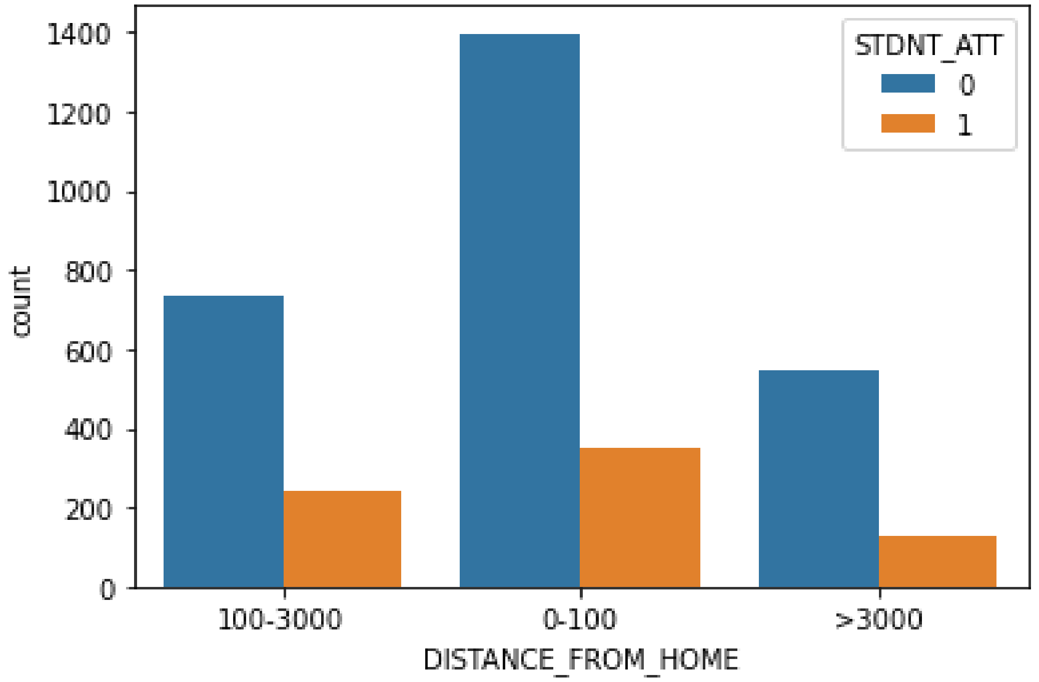


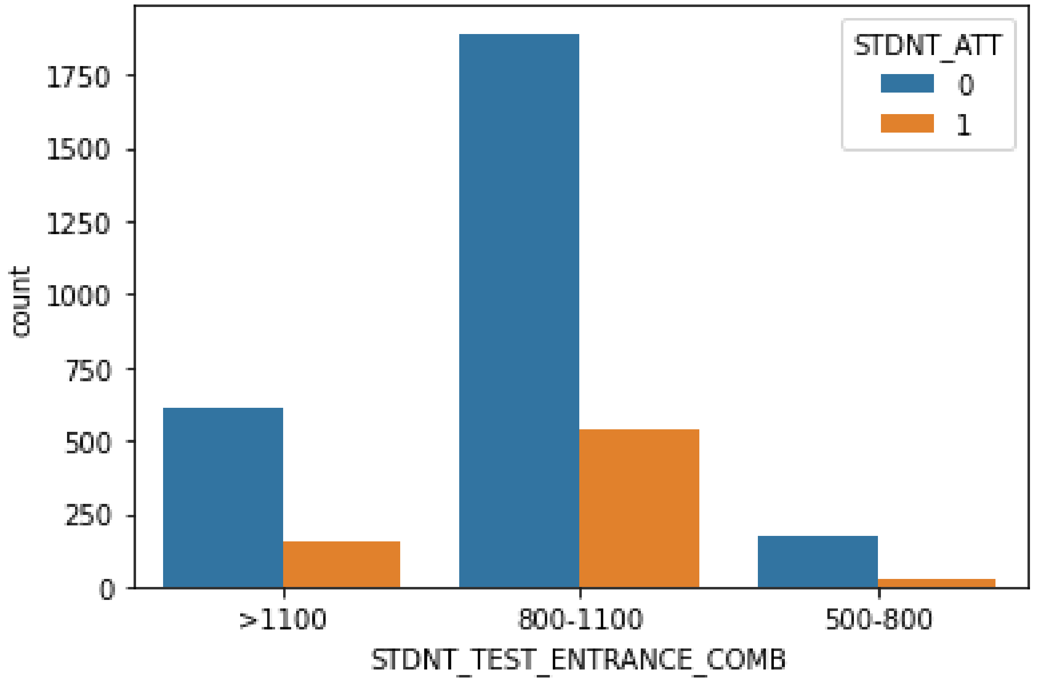
| | STUDENT IDENTIFIER | STDNT_AGE | STDNT_TEST_ENTRANCE_COMB | FIRST_TERM | SECOND_TERM | RETURNED_2ND_YR | DISTANCE_FROM_HOME | HIGH_SCHL_GPA | FIRST_TERM_ATTEMPT_HRS | FIRST_TERM_EARNED_HRS | SECOND_TERM_ATTEMPT_HRS | SECOND_TERM_EARNED_HRS | GROSS_FIN_NEED | COST_OF_ATTEND | EST_FAM_CONTRIBUTION | UNMET_NEED |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| STUDENT IDENTIFIER | 100% | 1% | -3% | -2% | -2% | -2% | 1% | 0% | 0% | -0% | 2% | 2% | -1% | -1% | 1% | 0% |
| STDNT_AGE | 1% | 100% | -0% | 2% | 2% | -2% | -1% | -8% | -1% | -1% | -4% | -4% | -2% | -3% | -2% | -2% |
| STDNT_TEST_ENTRANCE_COMB | -3% | -0% | 100% | 2% | 2% | -1% | 3% | 21% | 7% | 18% | 5% | 13% | -8% | -3% | 5% | -8% |
| FIRST_TERM | -2% | 2% | 2% | 100% | 100% | -3% | 4% | -1% | 3% | -6% | 6% | 0% | -3% | -7% | -5% | -6% |
| SECOND_TERM | -2% | 2% | 2% | 100% | 100% | -3% | 4% | -1% | 3% | -6% | 6% | 0% | -3% | -7% | -5% | -6% |
| RETURNED_2ND_YR | -2% | -2% | -1% | -3% | -3% | 100% | -3% | 8% | -0% | 7% | 8% | 14% | -0% | 0% | 2% | -3% |
| DISTANCE_FROM_HOME | 1% | -1% | 3% | 4% | 4% | -3% | 100% | -2% | 7% | 5% | 8% | 7% | -2% | -0% | 5% | -0% |
| HIGH_SCHL_GPA | 0% | -8% | 21% | -1% | -1% | 8% | -2% | 100% | 7% | 33% | 10% | 31% | -1% | 0% | 1% | -11% |
| FIRST_TERM_ATTEMPT_HRS | 0% | -1% | 7% | 3% | 3% | -0% | 7% | 7% | 100% | 47% | 27% | 22% | -0% | 1% | 3% | -2% |
| FIRST_TERM_EARNED_HRS | -0% | -1% | 18% | -6% | -6% | 7% | 5% | 33% | 47% | 100% | 24% | 39% | -1% | 3% | 5% | -6% |
| SECOND_TERM_ATTEMPT_HRS | 2% | -4% | 5% | 6% | 6% | 8% | 8% | 10% | 27% | 24% | 100% | 61% | -1% | 0% | 5% | -6% |
| SECOND_TERM_EARNED_HRS | 2% | -4% | 13% | 0% | 0% | 14% | 7% | 31% | 22% | 39% | 61% | 100% | -6% | -1% | 6% | -10% |
| GROSS_FIN_NEED | -1% | -2% | -8% | -3% | -3% | -0% | -2% | -1% | -0% | -1% | -1% | -6% | 100% | 75% | -12% | 80% |
| COST_OF_ATTEND | -1% | -3% | -3% | -7% | -7% | 0% | -0% | 0% | 1% | 3% | 0% | -1% | 75% | 100% | 43% | 42% |
| EST_FAM_CONTRIBUTION | 1% | -2% | 5% | -5% | -5% | 2% | 5% | 1% | 3% | 5% | 5% | 6% | -12% | 43% | 100% | -30% |
| UNMET_NEED | 0% | -2% | -8% | -6% | -6% | -3% | -0% | -11% | -2% | -6% | -6% | -10% | 80% | 42% | -30% | 100% |

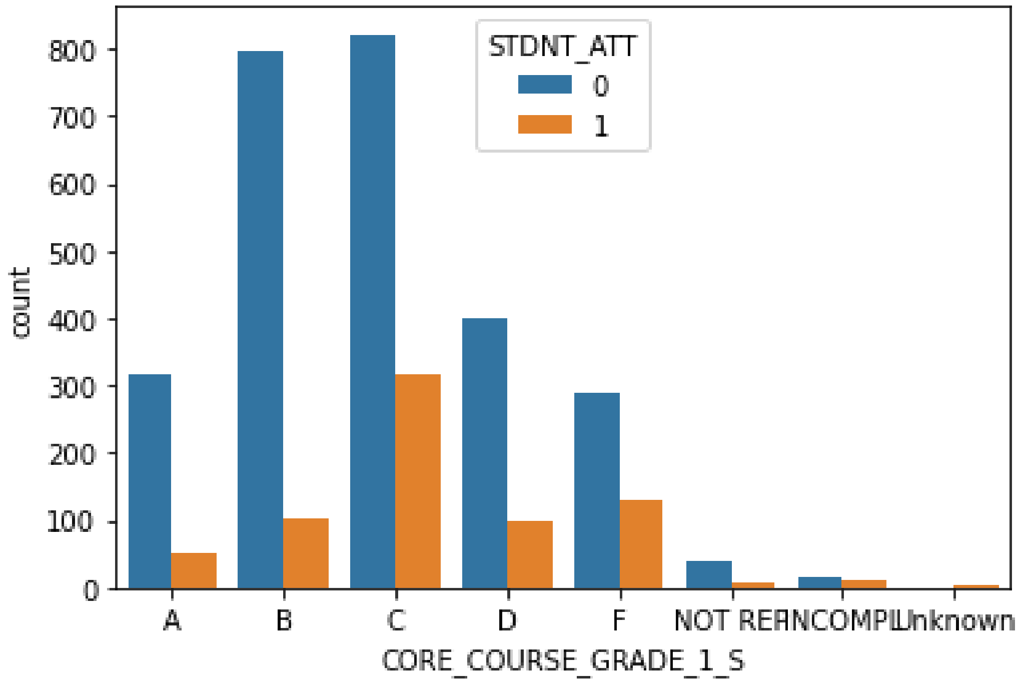
# Correlation

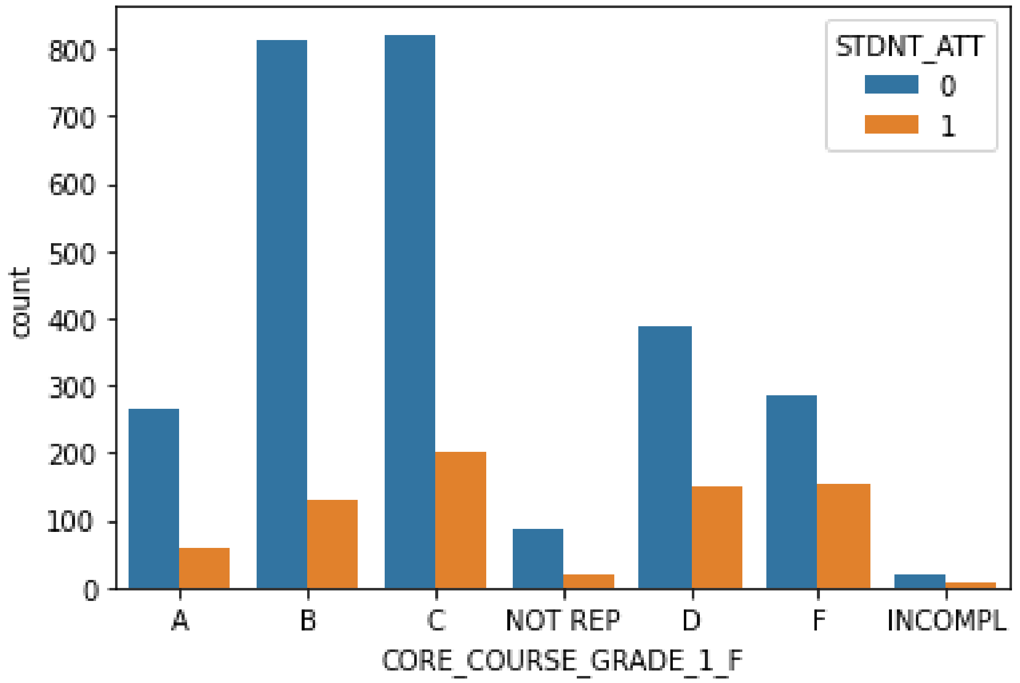
- By looking at this figure we can check the direction of the relationship
- From this figure we can see that the maximum variables have a positive relationship with each other
- As you can see Unmet_Need, Cost_Fay_Attend and Gross_Fin_Need are co-related so we can use any one of them to present the data - the rest we can leave to reduce the data.
- We can see that the variable in the first term is 100% equal to the second term.

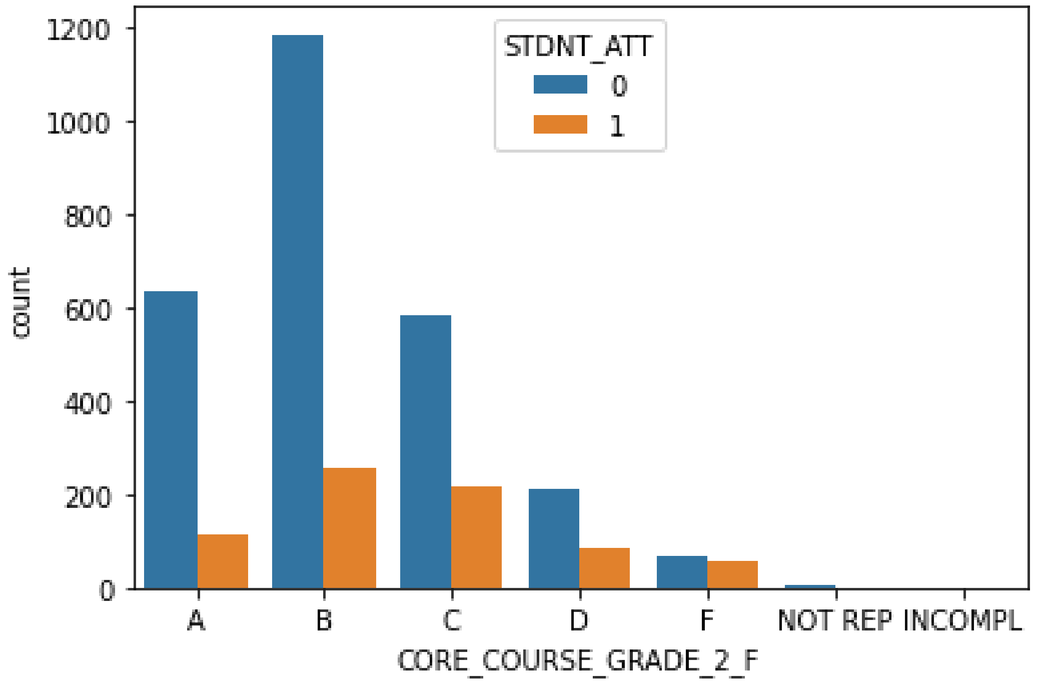

count
2000
1500
1000
500
0
>4
2.5-3.9
0-2.5
HIGH_SCHL_GPA
STDNT_ATT
0
1


count
1400
1200
1000
800
600
400
200
0
100-3000
0-100
>3000
DISTANCE_FROM_HOME
STDNT_ATT
0
1


count
1750
1500
1250
1000
750
500
250
0
>1100
800-1100
500-800
STDNT_TEST_ENTRANCE_COMB
STDNT_ATT
0
1


count
800
700
600
500
400
300
200
100
0
A
B
C
D
F
NOT REP
INCOMPL
Unknown
CORE_COURSE_GRADE_1_S
STDNT_ATT
0
1


count
800
700
600
500
400
300
200
100
0
A
B
C
NOT REP
D
F
INCOMPL
CORE_COURSE_GRADE_1_F
STDNT_ATT
0
1

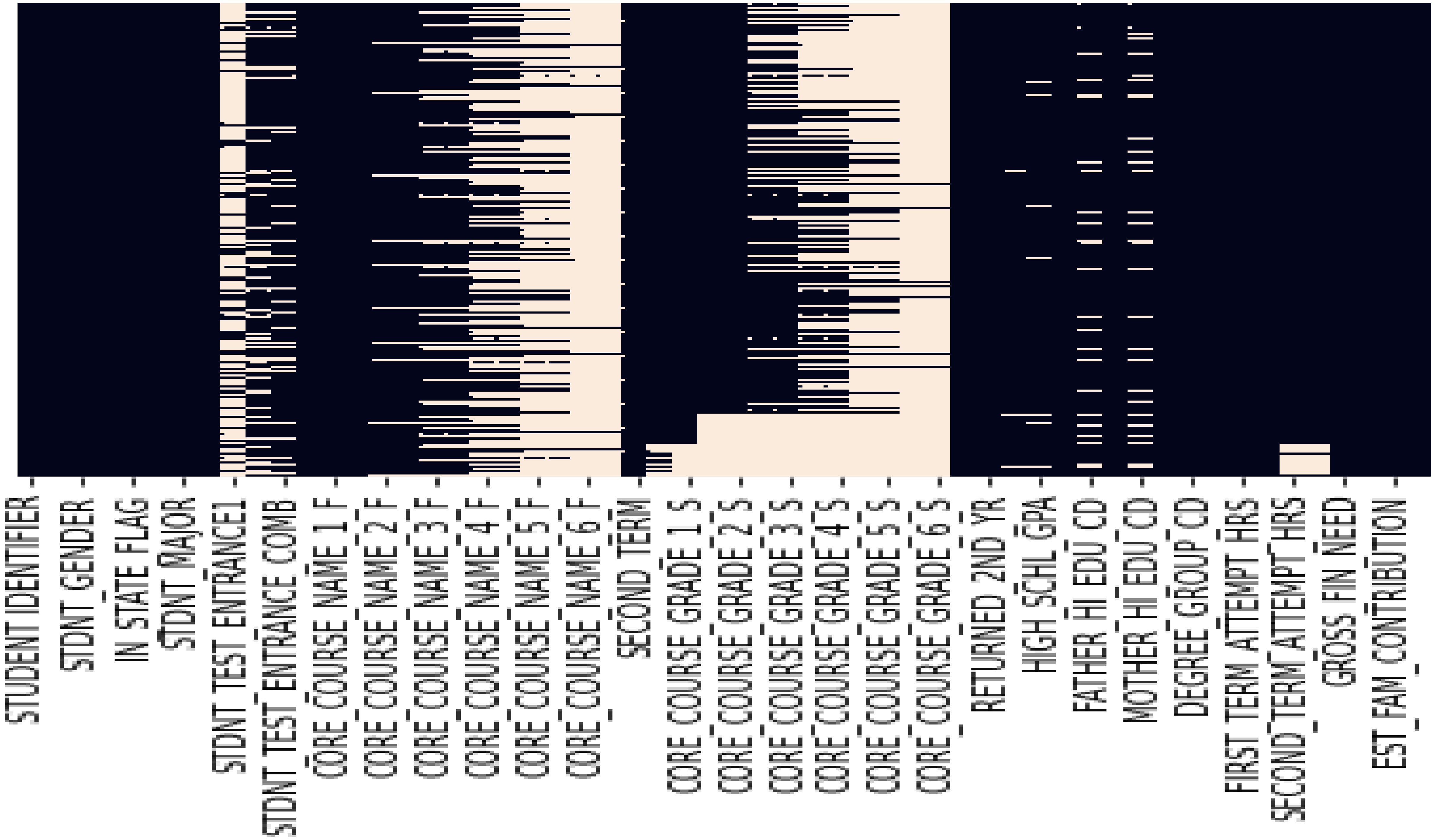

count
1200
1000
800
600
400
200
0
A
B
C
D
F
NOT REP
INCOMPL
CORE_COURSE_GRADE_2_F
STDNT_ATT
0
1

## Data Pre –Processing

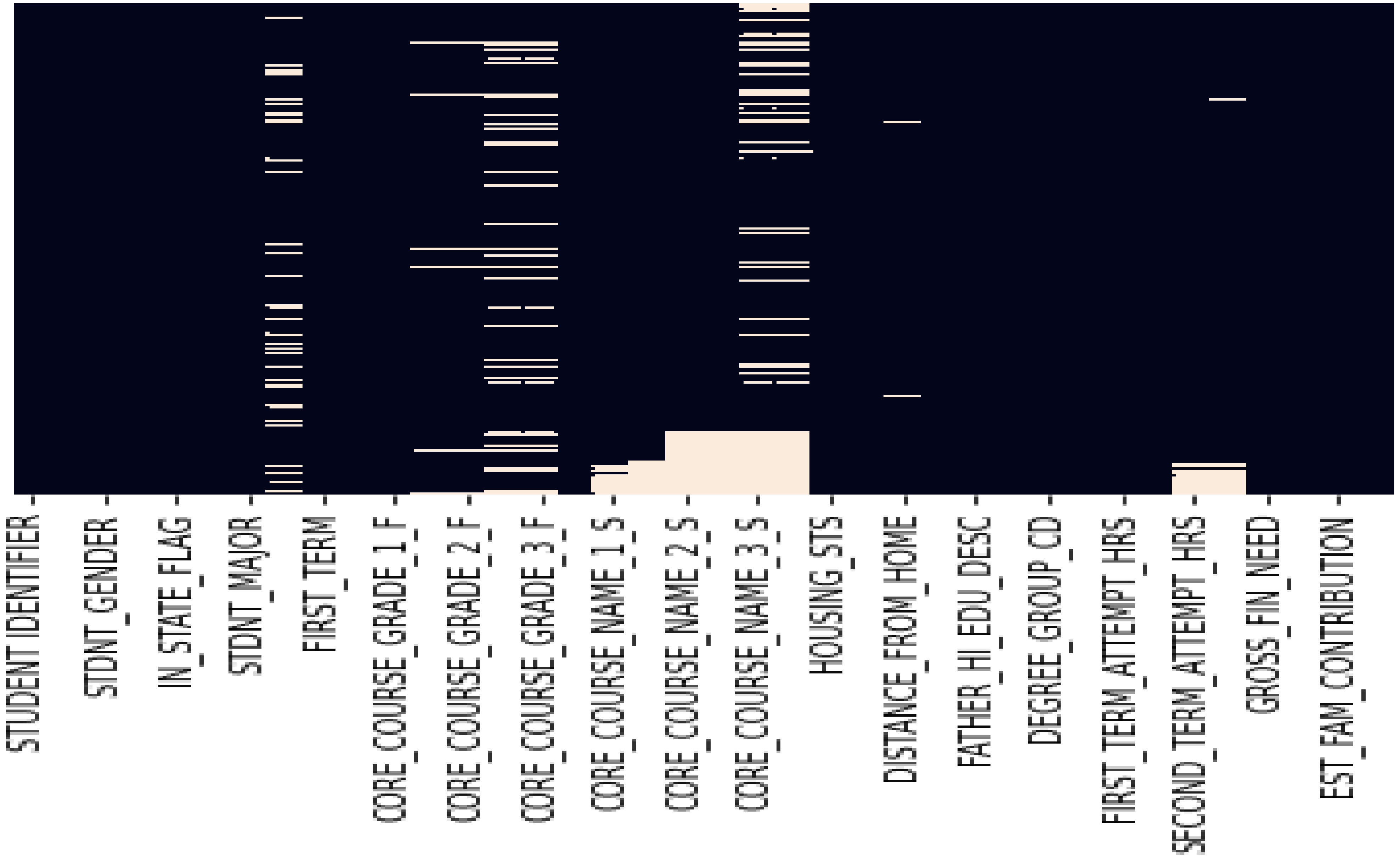
- Data shape : (3400, 56)

- Data Null Values :

Maximum numbers of data are Null, after evaluation of the whole data, left the columns with more than 17% missing values to avoid bias outcome



## Data Pre –Processing



**Missing values Replacement :**

Categorical values : Replaced by Mode

(as we can not calculate mean median of it)

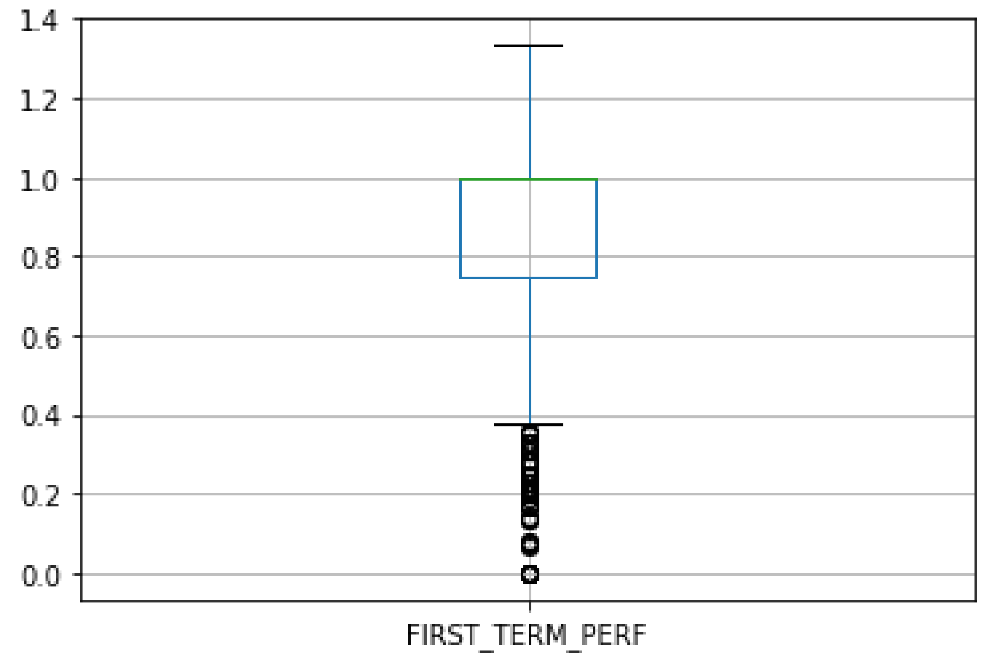
- CORE_COURSE_NAME_2_F 0.0291 %missing values
- CORE_COURSE_GRADE_2_F 0.0291 %missing values
- CORE_COURSE_NAME_3_F 0.1662 %missing values
- CORE_COURSE_GRADE_3_F 0.1662 %missing values

Numerical values : Replaced by Median

(As we don't want touch outliers)

- STDNT_TEST_ENTRANCE_COMB 0.1524 %missing values
- DISTANCE_FROM_HOME 0.0074 %missing values
- HIGH_SCHL_GPA 0.0156 %missing values
- SECOND_TERM_ATTEMPT_HRS 0.0606 %missing values
- SECOND_TERM_EARNED_HRS 0.0615 %missing values

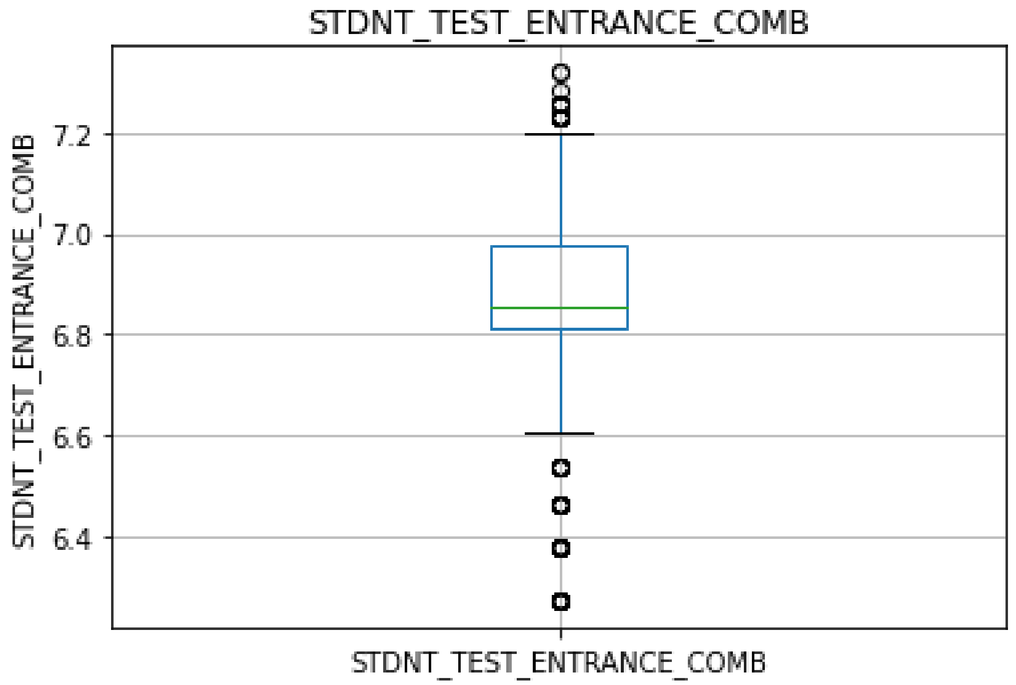
## Data Pre –Processing

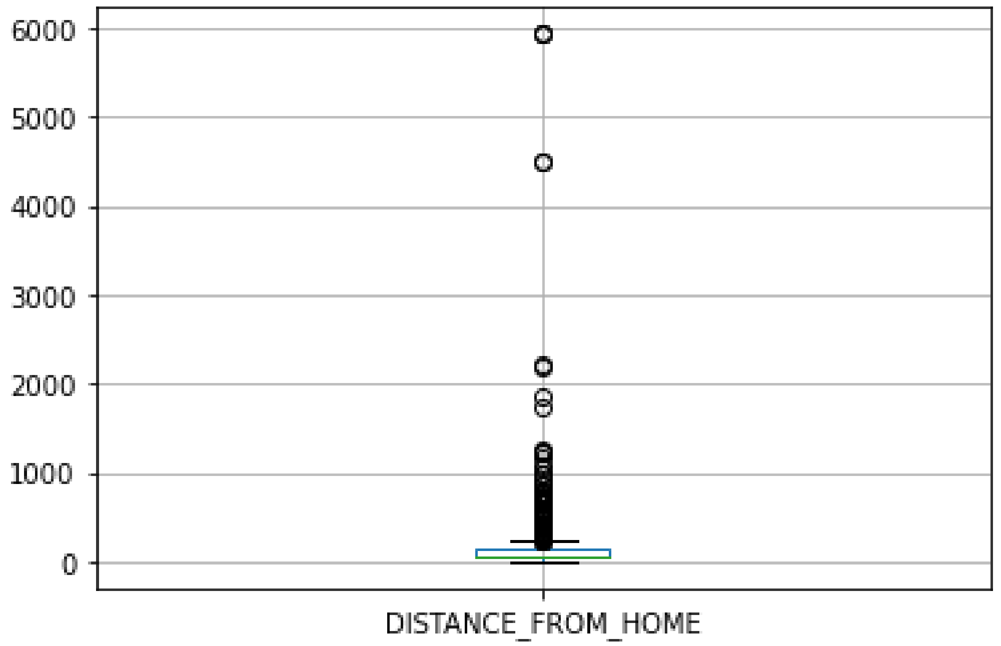


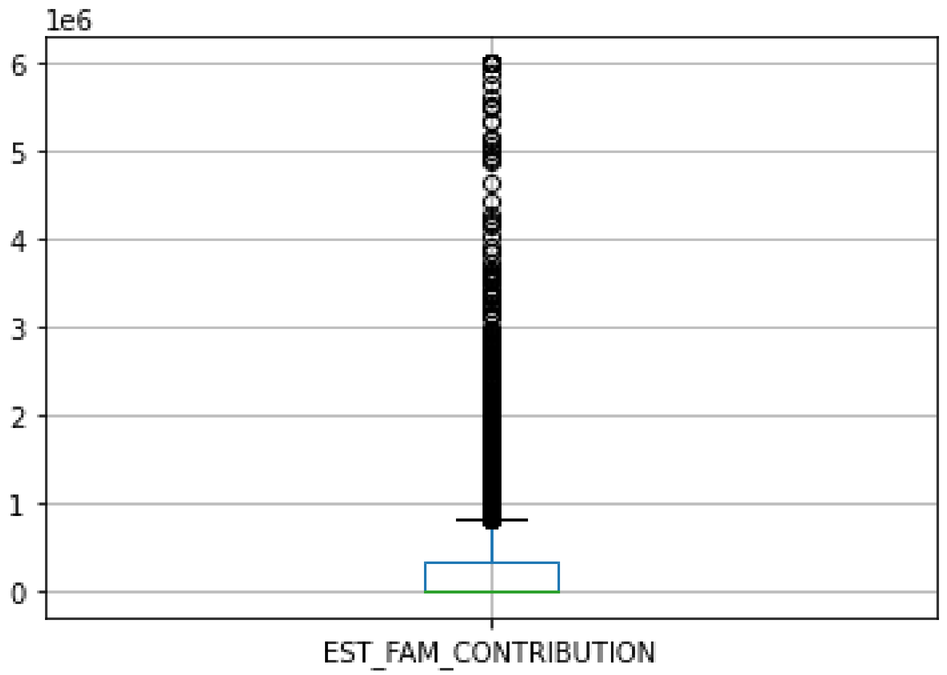
**Outliers :**

Outliers  can influence a data set, so it's important to keep them to better understand the big picture,

As we can use non sensitive outliers machine learning algorithms to create models







# Handle Categorical Data

## Reduced the number of levels from a categorical variable having large number of categories :

1. **STDNT_MAJOR**

- data['STDNT_MAJOR']=data['STDNT_MAJOR'].replace(['Political Science','English Language/Literature','History','Sociology','Liberal Arts','Communication'],'Arts')
- data['STDNT_MAJOR']=data['STDNT_MAJOR'].replace(['Art','Theatre Arts', 'Psychology','Music Performance','Theatre Education','Music Education','Music','Art Education'],'FineArts')
- data['STDNT_MAJOR']=data['STDNT_MAJOR'].replace(['Nursing', 'Biology','Pre-Nursing'],'Medical')
- data['STDNT_MAJOR']=data['STDNT_MAJOR'].replace(['Chemistry', 'Geology', 'General Studies/AS', 'Earth and Space Science'],'Other_Science')
- data['STDNT_MAJOR']=data['STDNT_MAJOR'].replace(['Pre-Engineering/RETP','Engineering Studies', 'Mathematics'],'Engineering ')
- data['STDNT_MAJOR']=data['STDNT_MAJOR'].replace([ 'Computer Science - Systems','Management Information Systems', 'Applied Computer Science', 'Information Technology', 'Computer Science - Games'],'Computer')
- data['STDNT_MAJOR']=data['STDNT_MAJOR'].replace(['Pre-Business', 'Management','Marketing', 'General Business', 'Finance', 'Accounting'],'Management')
- data['STDNT_MAJOR']=data['STDNT_MAJOR'].replace([ 'Spanish with Teacher Cert', 'Chemistry and Secondary Ed', 'Early Childhood Education', 'Middle Grades Education', 'French with Teacher Cert', 'History and Secondary Ed', 'English and Secondary Ed', 'Spec Ed: Gen. Curr. - Reading', 'Biology and Secondary Ed', 'Mathematics and Secondary Ed'],'Education')
- data['STDNT_MAJOR']=data['STDNT_MAJOR'].replace(['French', 'Spanish'],'Language')
- data['STDNT_MAJOR']=data['STDNT_MAJOR'].replace([ 'Exercise Science', 'Joint Enrollment - Accel', 'Health Science','Health and Physical Education', 'Early Admission - Accel'],'Health')

2. **CORE_COURSE_NAME_1_F**

- data['CORE_COURSE_NAME_1_F']=data['CORE_COURSE_NAME_1_F'].str.slice(0,4)

3. **CORE_COURSE_NAME_2_F**

- data['CORE_COURSE_NAME_2_F']=data['CORE_COURSE_NAME_2_F'].str.slice(0,4)

4. **CORE_COURSE_NAME_3_F**

- data['CORE_COURSE_NAME_3_F']=data['CORE_COURSE_NAME_3_F'].str.slice(0,4)

## Variables

- **Corrected the Data :**

```
data['FIRST_TERM']=data['FIRST_TERM'].replace({200508:2005,200608:2006,200708:2007,200808:2008,200908:2009,201008:2010})
data['SECOND_TERM']=data['SECOND_TERM'].replace({200602:2006,200702:2007,200802:2008,200902:2009,201002:2010,201102:2011})
```

- **Evaluated Student performance :**

```
    data['FIRST_TERM_PERF']=data['FIRST_TERM_EARNED_HRS']/data['FIRST_TERM_ATTEMPT_HRS']
    data['SECOND_TERM_PERF']=data['SECOND_TERM_EARNED_HRS']/data['SECOND_TERM_ATTEMPT_HRS']
```

Along with that rectified the performance by below :

```
data['SECOND_TERM_PERF']=data['SECOND_TERM_PERF'].values[data['SECOND_TERM_PERF'].values >1.0]=1.0
```

- **Changed Data type :**

```
data['GROSS_FIN_NEED']=pd.to_numeric(data['GROSS_FIN_NEED'],errors='coerce')
```

## Variables

**Dependent Variable :**

- Original provided in Data : Returned_2nd Year
- We need to convert the same in student attrition by lambda function :

```
data["STDNT_ATT"]=data["RETURNED_2ND_YR"].map(lambda x:1  if x==0 else 0)
```

**Independent Variables :** Data other than STDNT_ATT & RETURNED_2ND_YR

**Handled Independent Variables :**

- Created dummy variables to handle Categorical data :

```
data=pd.get_dummies(data,columns=['CORE_COURSE_NAME_1_F','CORE_COURSE_NAME_2_F','CORE_COURSE_NAME_3_F','CORE_COURSE_NAME_1_S','CORE_COURSE_NAME_2_S','CORE_COURSE_NAME_3_S‘, 'HOUSING_STS'])
```

## Balance Data :

- Let’s take a simple example if in our data set we have positive values which are approximately same as negative values. Then we can say our dataset in balance

- **Lets Check with our data :**

```
Attrition = data[data['STDNT_ATT']==1]
Retention = data[data['STDNT_ATT']==0]
```

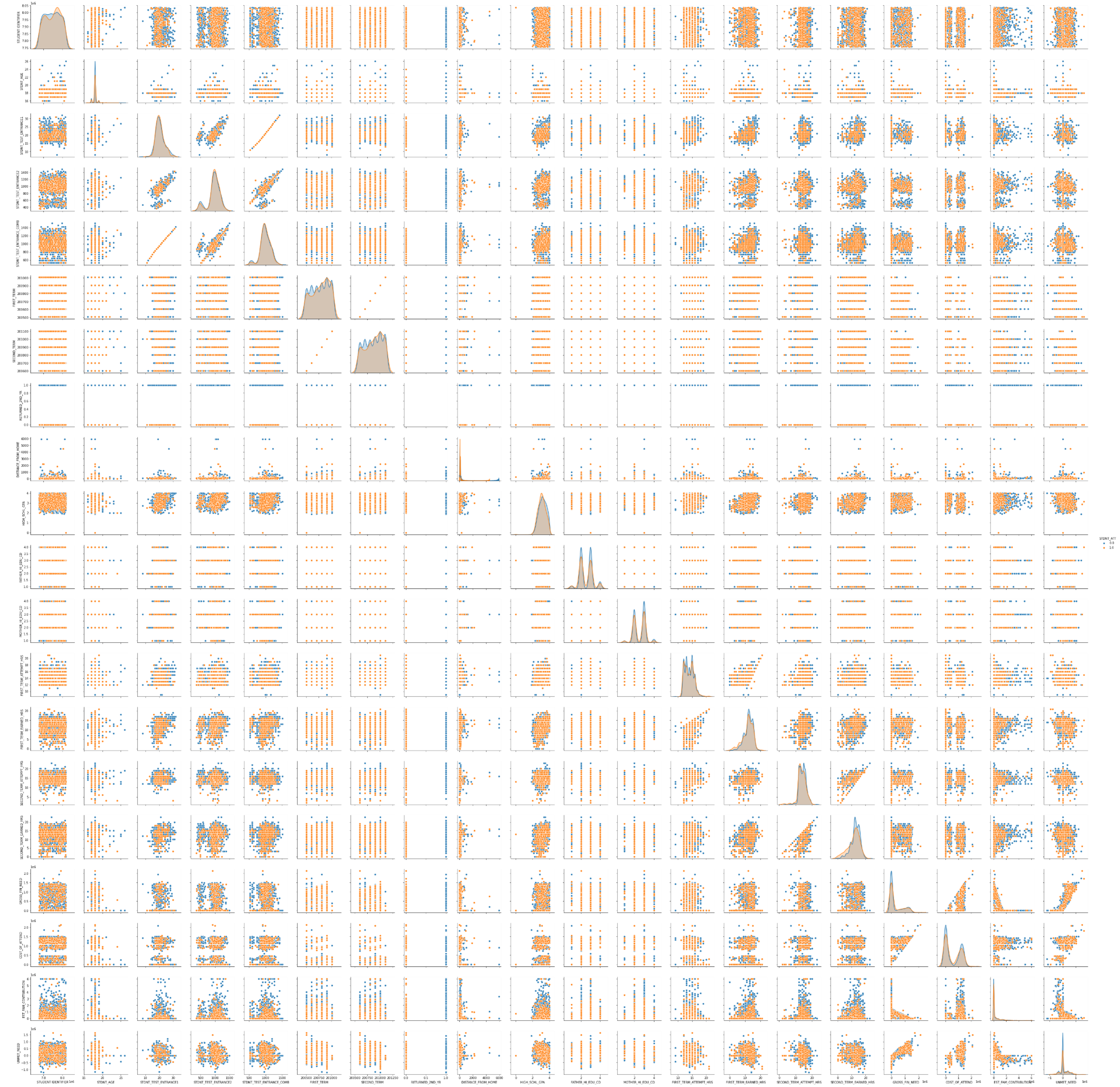
print(Attrition.shape,Retention.shape) : (723,117), (2677,117) – its turned out to be imbalanced dataset

**There are two technique for creating a balanced dataset – Under sampling and Over sampling**

**If we do under sampling it will become a very small dataset so we will implement oversampling to handle unbalanced data.**

```
from imblearn.combine import SMOTETomek
smk= SMOTETomek(random_state=200)
X_res,Y_res=smk.fit_sample(X,Y)
```

- Resampled dataset Shape : (2446,2446)



## Predictive Model

We can see in the figure that a lot of data overlaps so that we cannot draw a straight line or we can draw a bifurcated line of positive negative points.

With that, we can see that it contains a lot of categorical data

We can only conclude that we cannot use linear, KNN and logistic machine algorithms for this data.

So we have the option of choosing random forest and decision tree algorithms to check the main drivers of frustration of the early students.

As we are keeping outliers it is best to use non sensitive decision trees methods to predict model

## Hyper parameter tuning

hyperparameter is a parameter whose value is used to control the learning process.

Two Types : 1. Grid Search – Search each and every component of data to choose the best
2. Random Search – Estimate the best elements at random

We have used both in this project, 1st we do a random search after that grid search, By doing that we can save time and get the best elements for the model

```
import numpy as np
from sklearn.model_selection import RandomizedSearchCV
# Number of trees in random forest
n_estimators = [int(x) for x in np.linspace(start = 200, stop = 2000, num = 10)]
# Number of features to consider at every split
max_features = ['auto', 'sqrt','log2']
# Maximum number of levels in tree
max_depth = [int(x) for x in np.linspace(10, 1000,10)]
# Minimum number of samples required to split a node
min_samples_split = [2, 5, 10,14]
# Minimum number of samples required at each leaf node
min_samples_leaf = [1, 2, 4,6,8]
```

```
# Create the random grid
random_grid = {'n_estimators': n_estimators,'max_features':
max_features,
'max_depth': max_depth, 'min_samples_split': min_samples_split,
 'min_samples_leaf': min_samples_leaf, 'criterion':['entropy','gini']}
print(random_grid)
rf=RandomForestClassifier()
rf_randomcv=RandomizedSearchCV(estimator=rf,param_distributions=ra
ndom_grid,n_iter=100,cv=3,verbose=2,
                    random_state=100,n_jobs=-1)
### fit the randomized model
rf_randomcv.fit(X_train,Y_train)
rf_randomcv.best_params_
```

## Models comparison

| Model Name | Accuracy score | Precision score | Recall score |
|---|---|---|---|
| Random Forest Classifier | 88% | 89% | 87% |
| Decision Tree Classifier | 79% | 79% | 78% |
| XGBoost | 86% | 95% | 76% |

**Accuracy** - Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.
**Precision -** High precision relates to the low false positive rate.
**Recall -** Recall is the ratio of correctly predicted positive observations to the all observations in actual class

Although the results are good, from the table above, we can see that random forest classifiers have performed better.

## Limitations

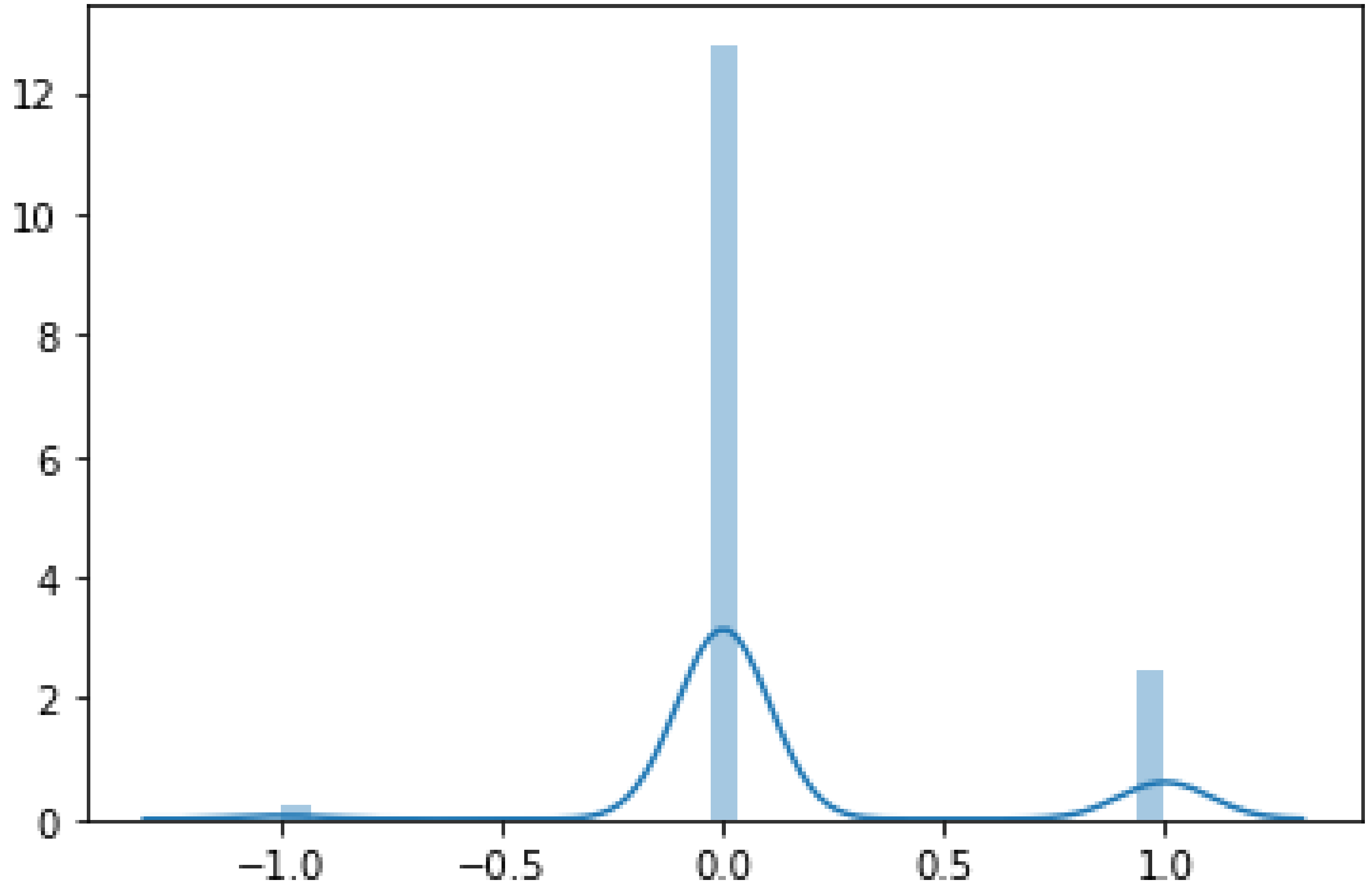
The data is not latest

## Findings

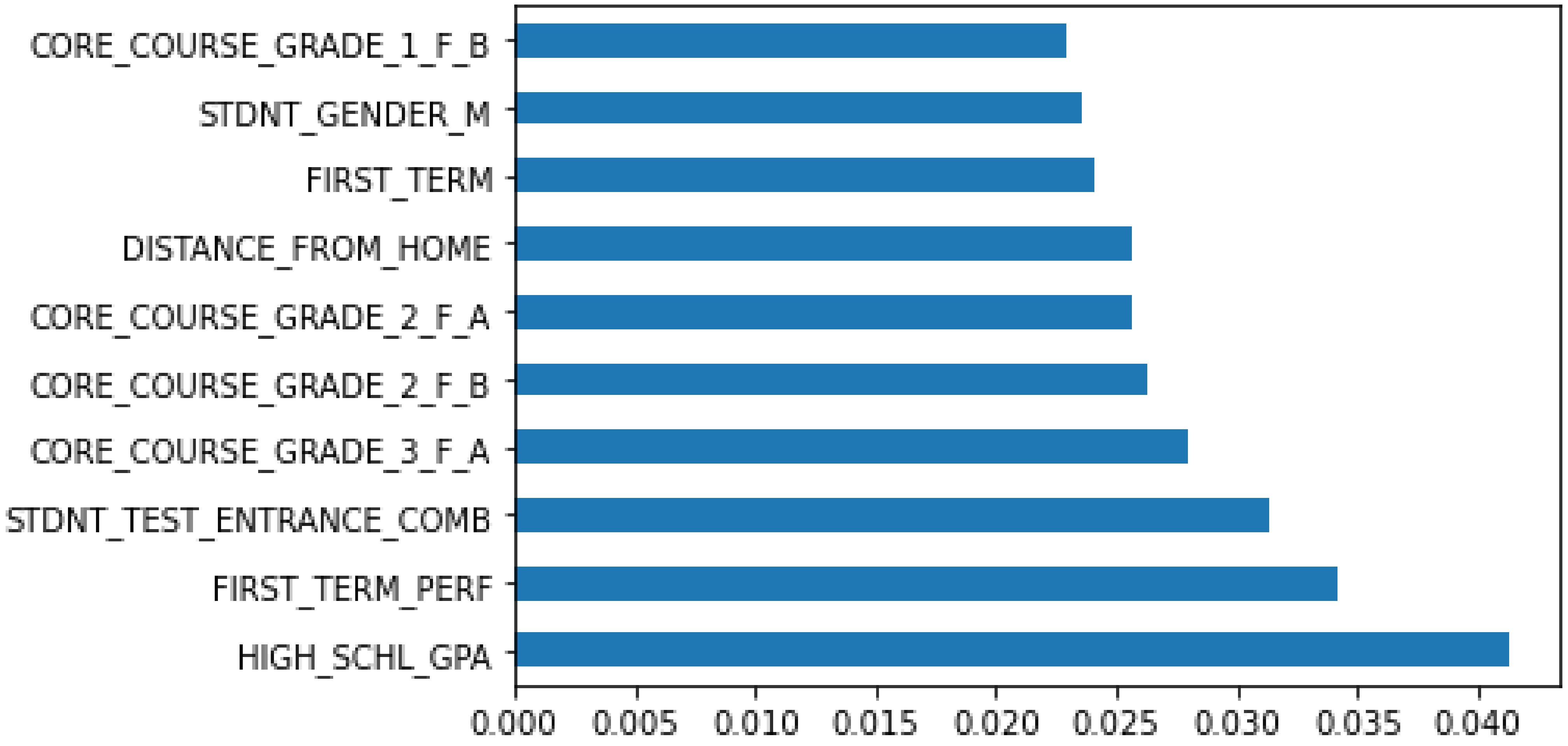
Performance is better in models developed from random forest along with accuracy rate than  decision tree and XGboost

The key drivers are  many factors as presented in the slide
Responsible for the attrition of students as well as many
Drivers who retain students

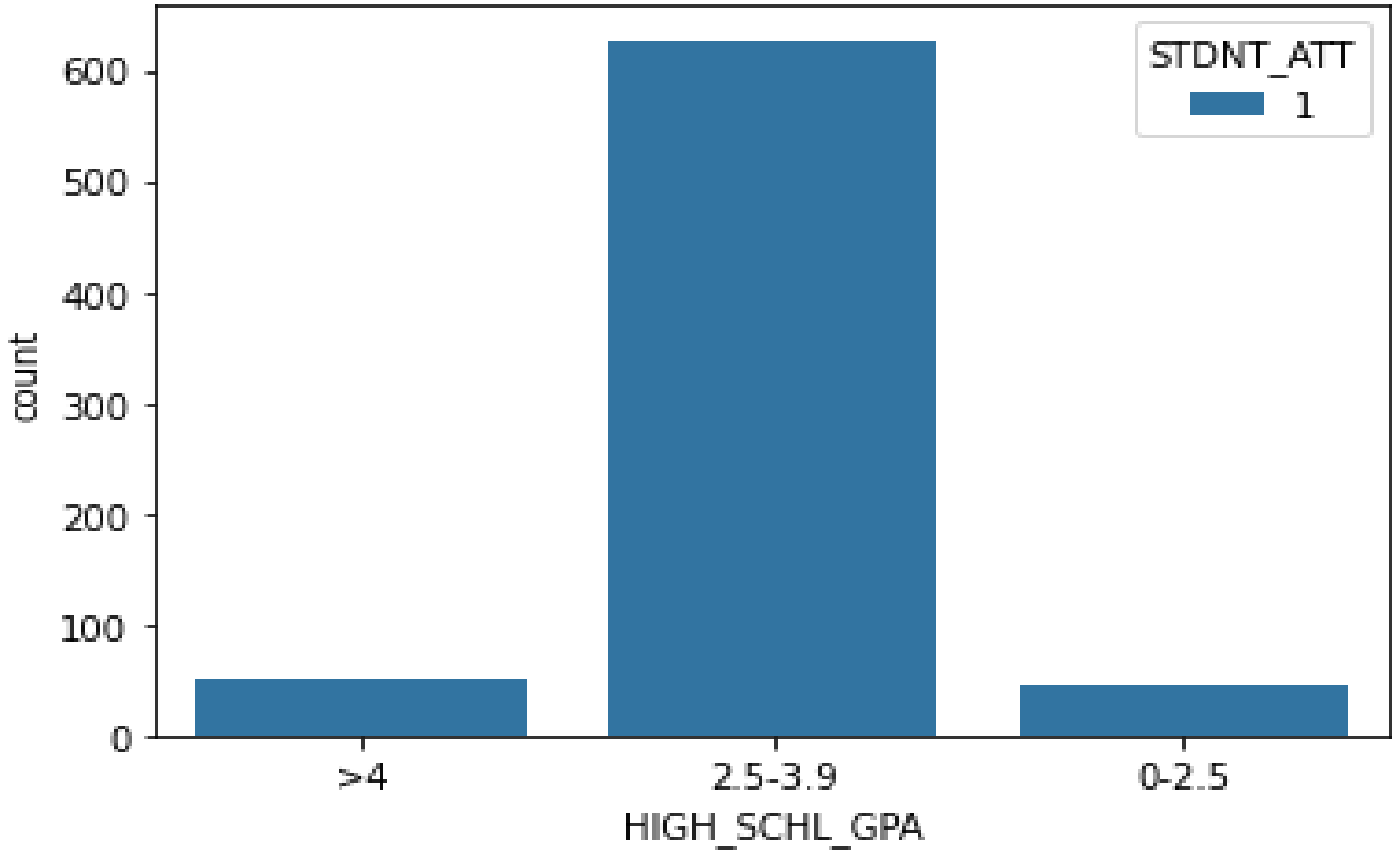
To solve the problem we need to evaluate the key drivers and take the necessary steps to retain the students

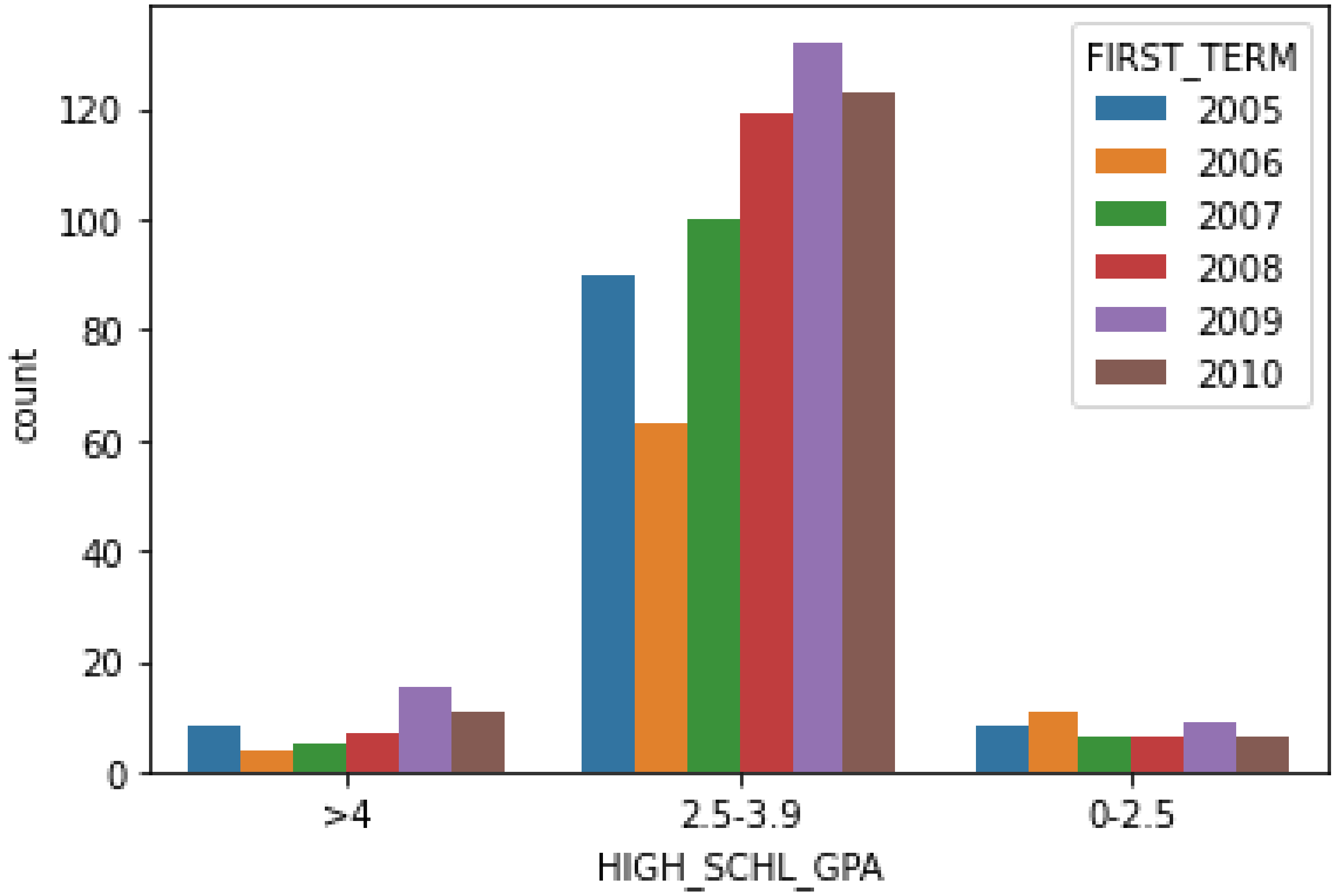


## Key drivers of early student attrition through Random Forest classification Model



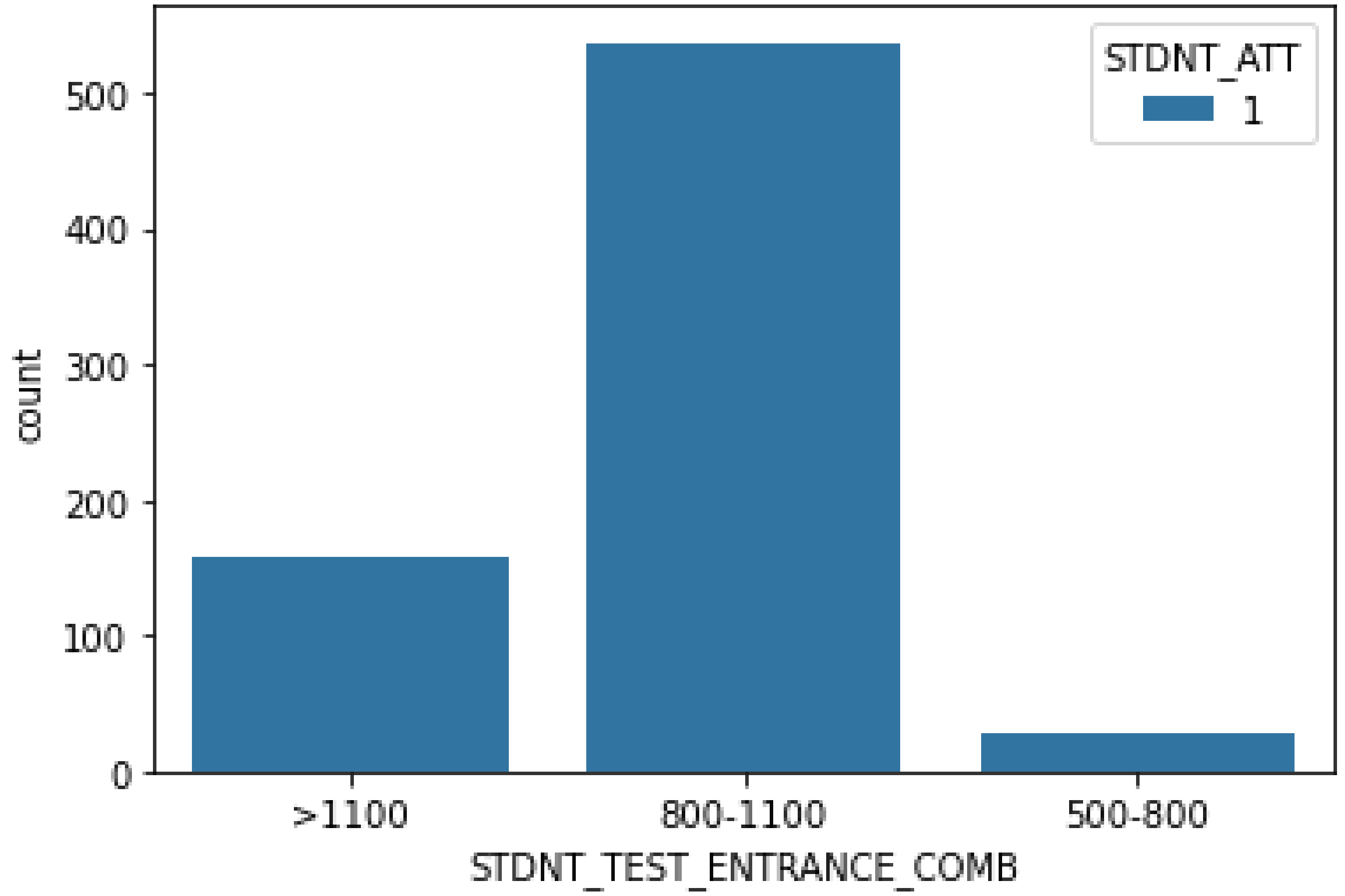
## Business Interpretations

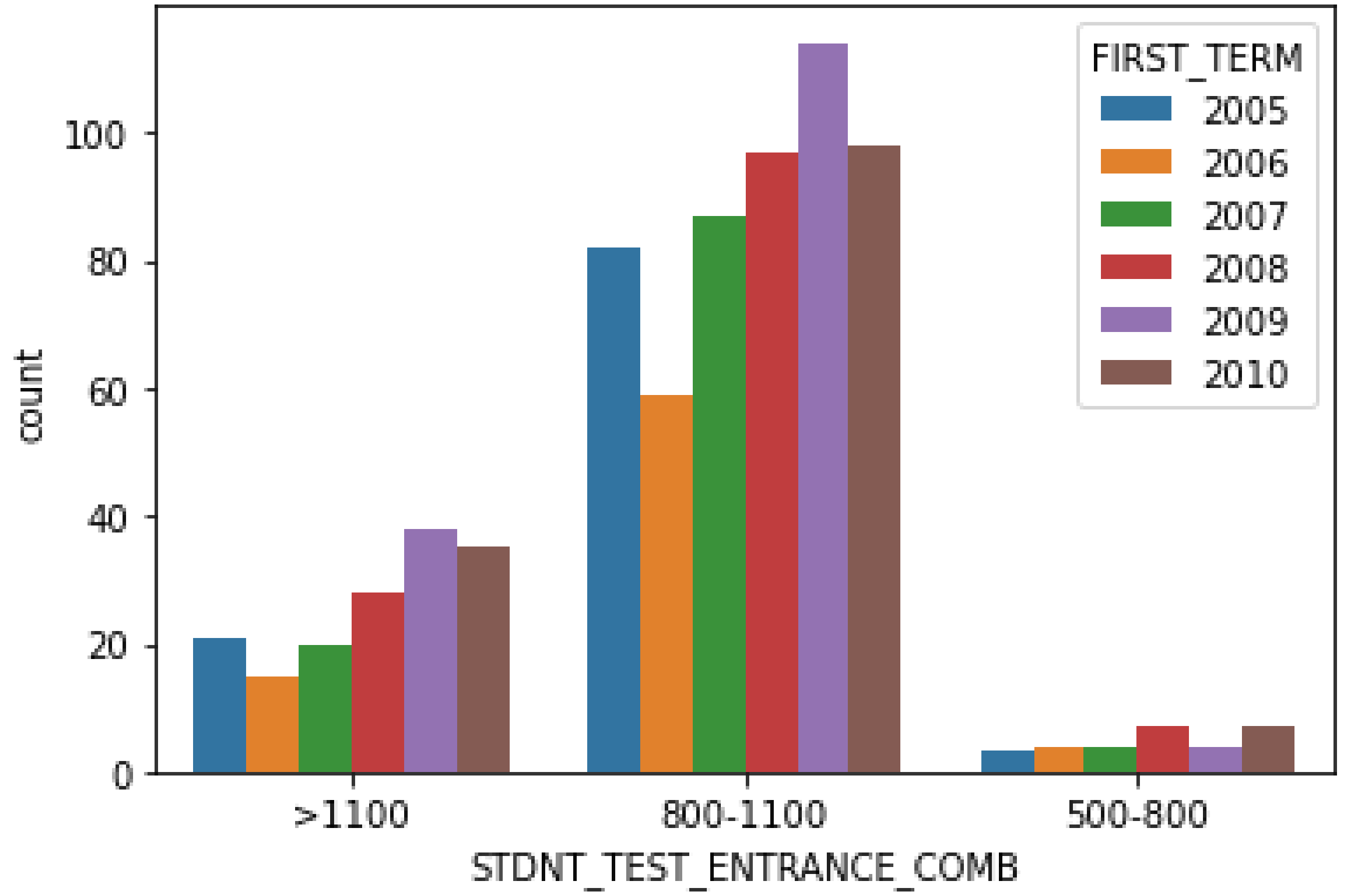




From the figure above, we can see over a six-year period that the maximum number of students dropping out of university is from average-performing in high school score.

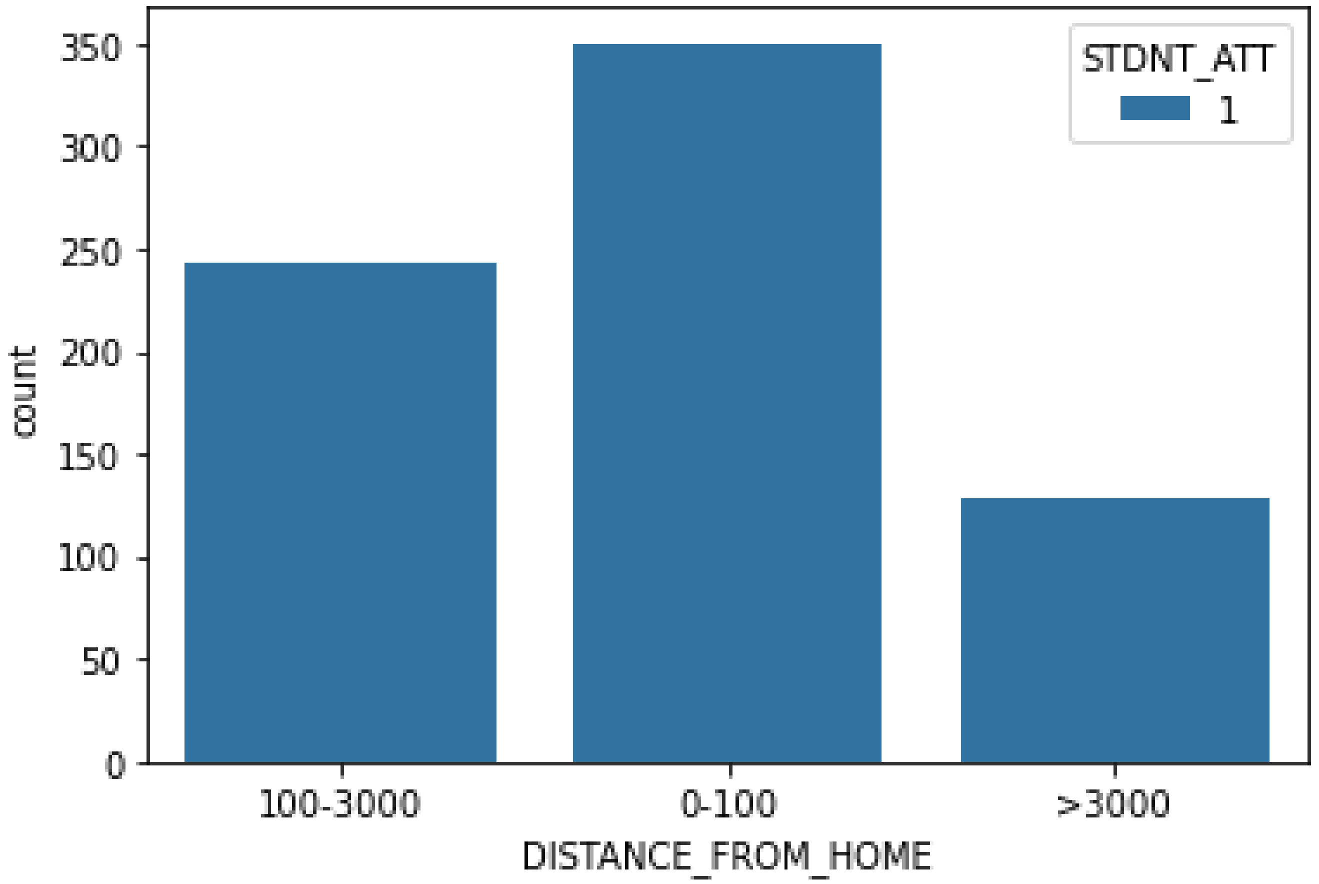
## Business Interpretations

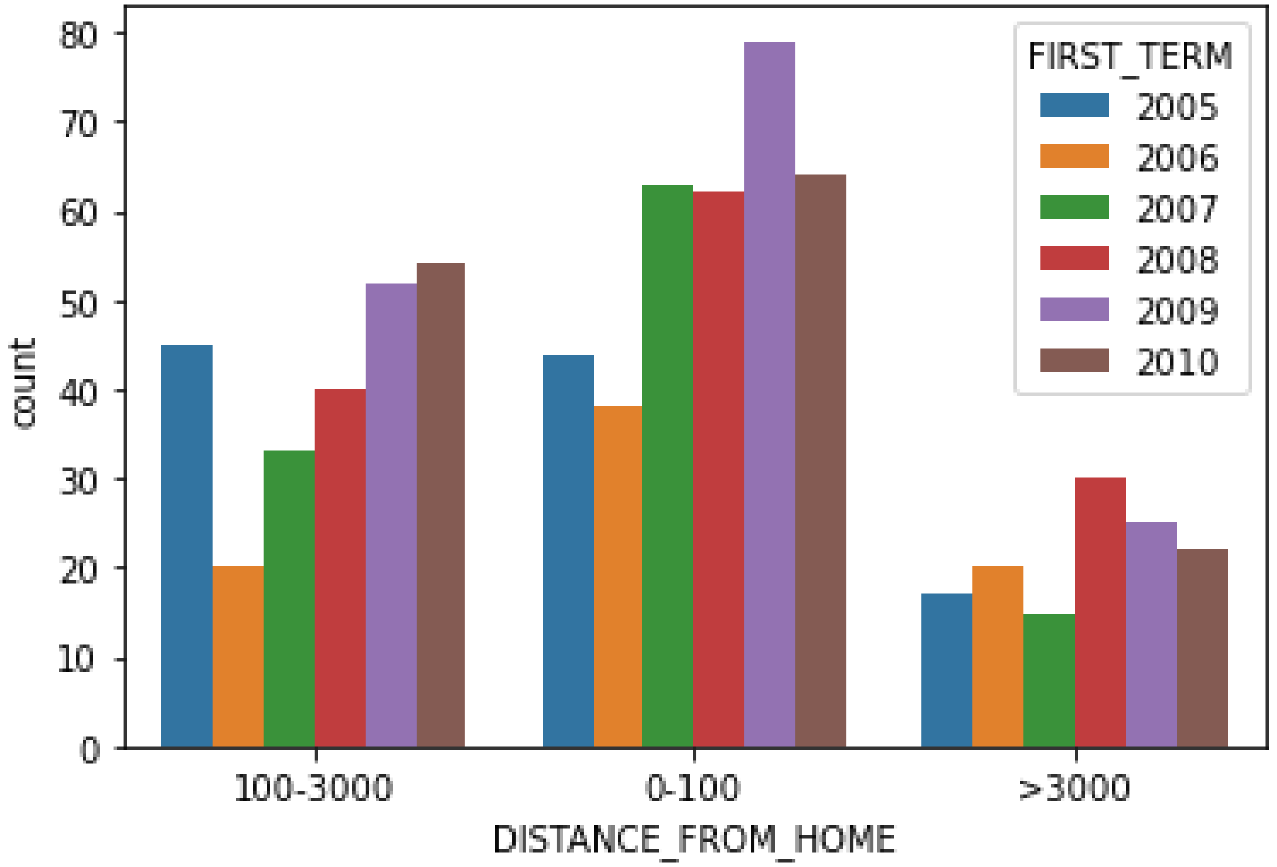




From the figure above, we can see over a six-year period that the maximum number of students dropping out of university is from average-performing in Entrance exam score.

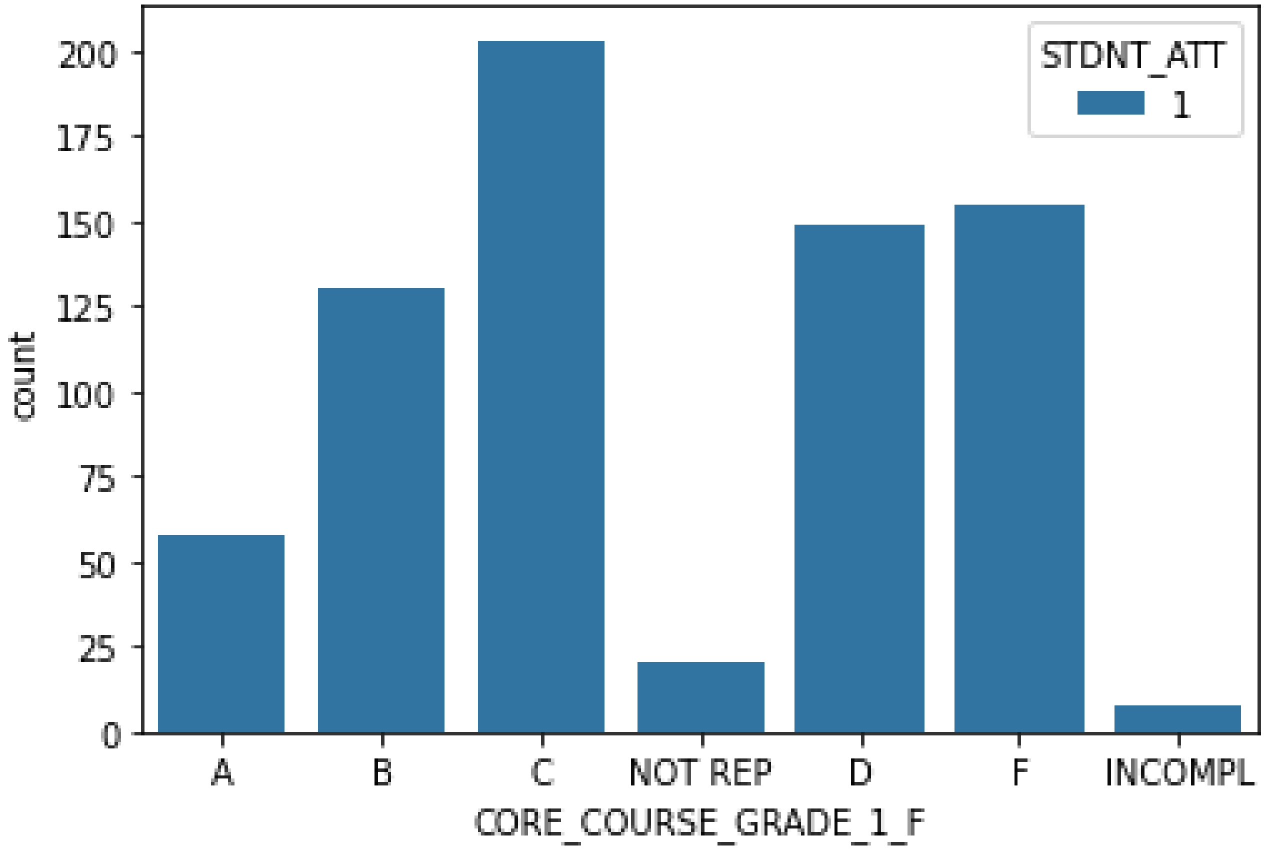
## Business Interpretations

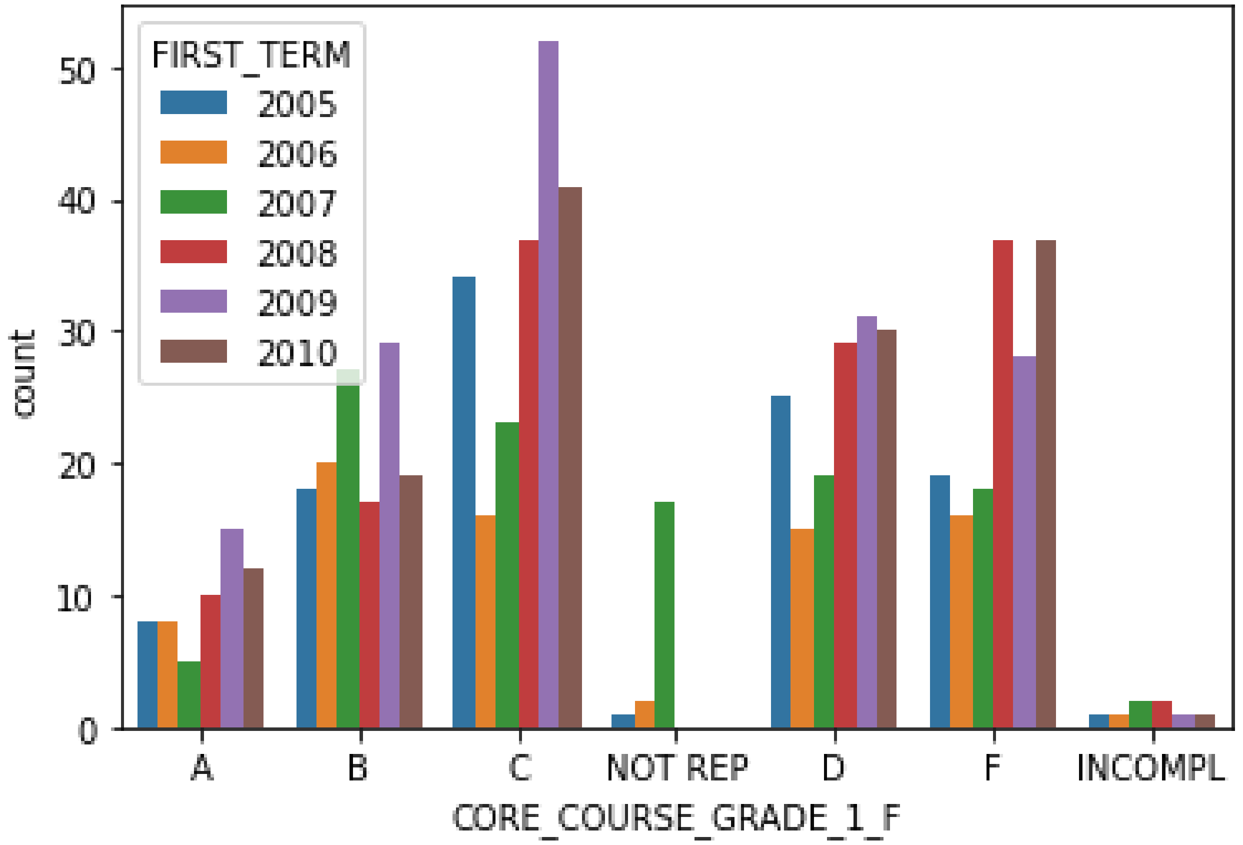




From the figure above, we can know that the maximum number of students living within 100 km does not join the university in the second year.

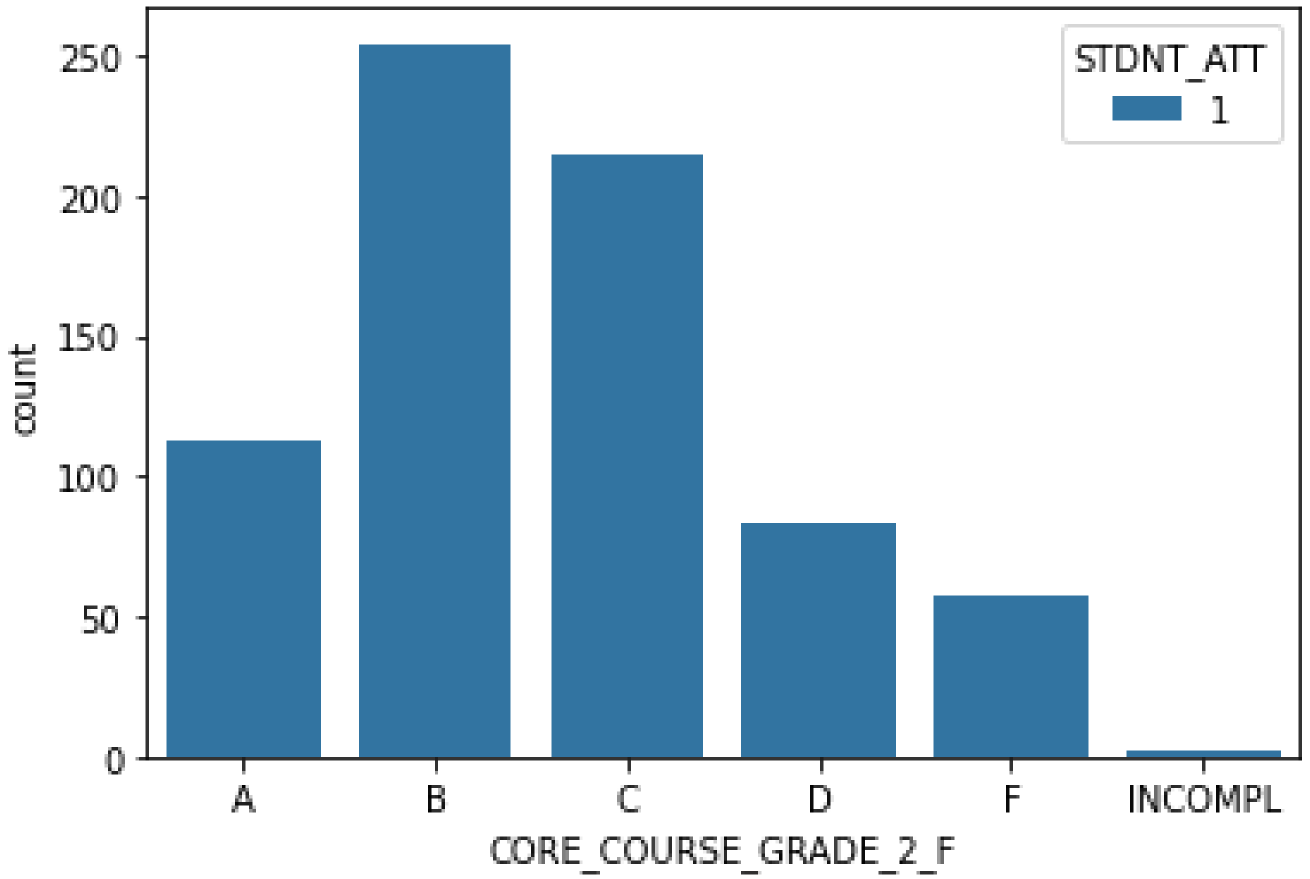
## Business Interpretations

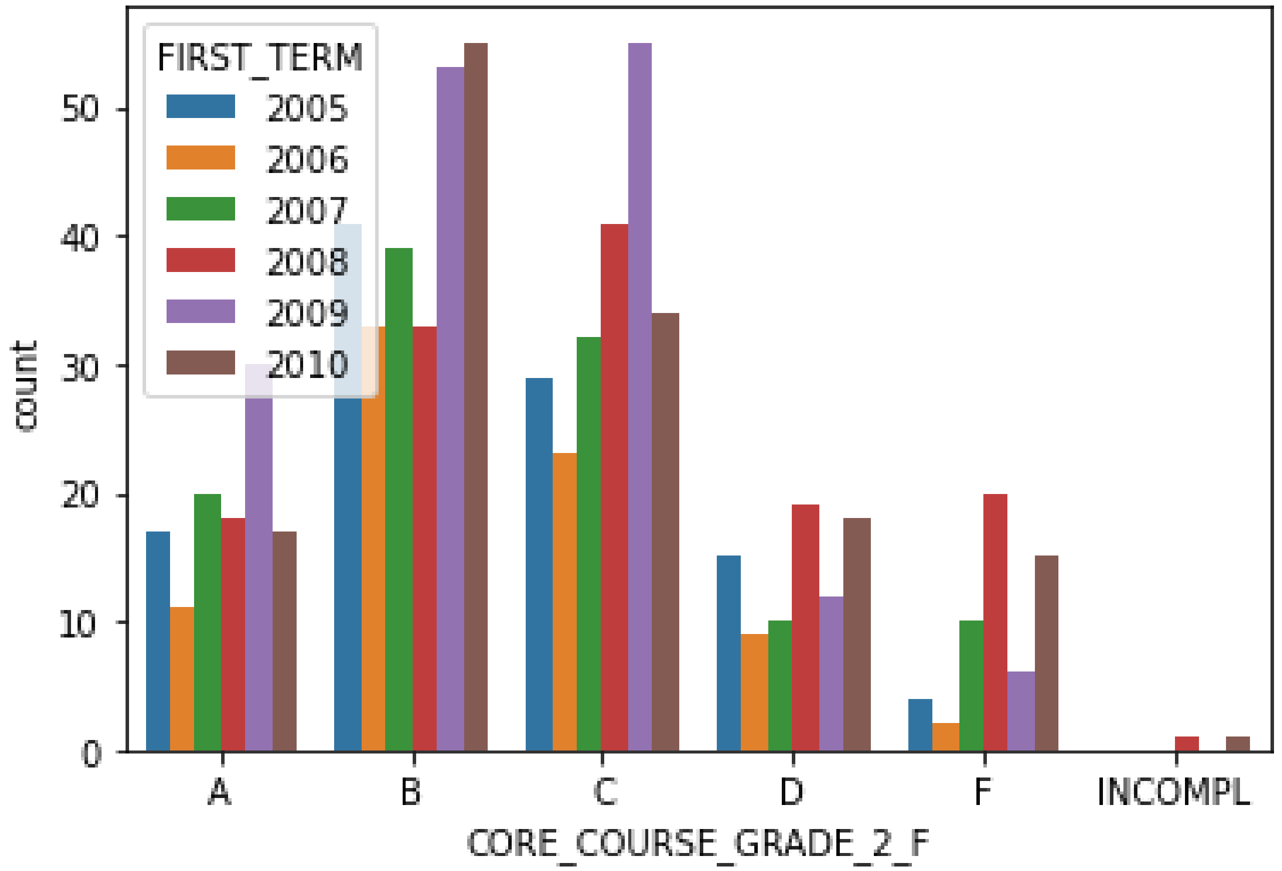




The maximum number of first-year students joining the Grade 1 course is leaving the institution, as can be seen from the figure above, the students with maximum low-performance are leaving the university.

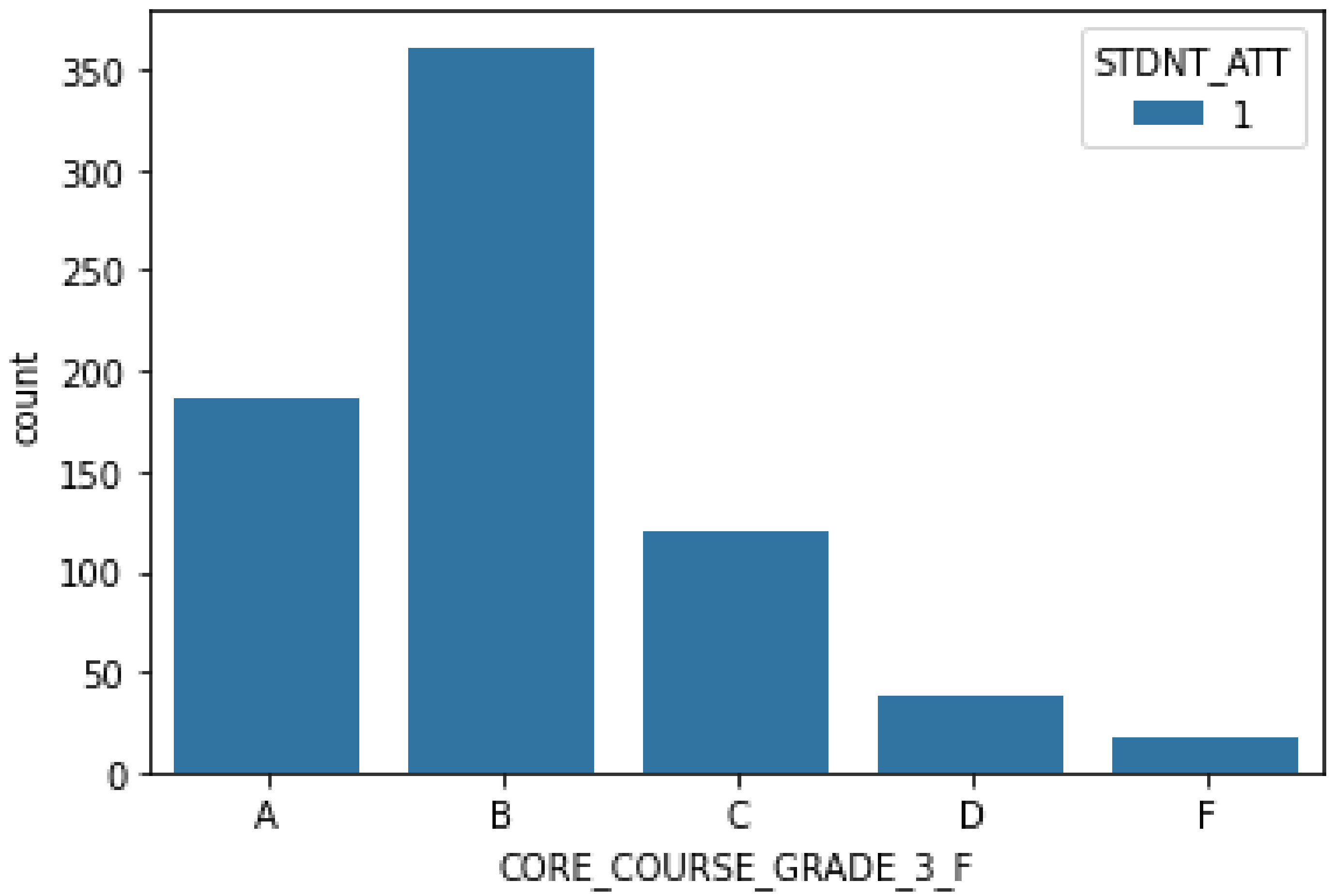
## Business Interpretations

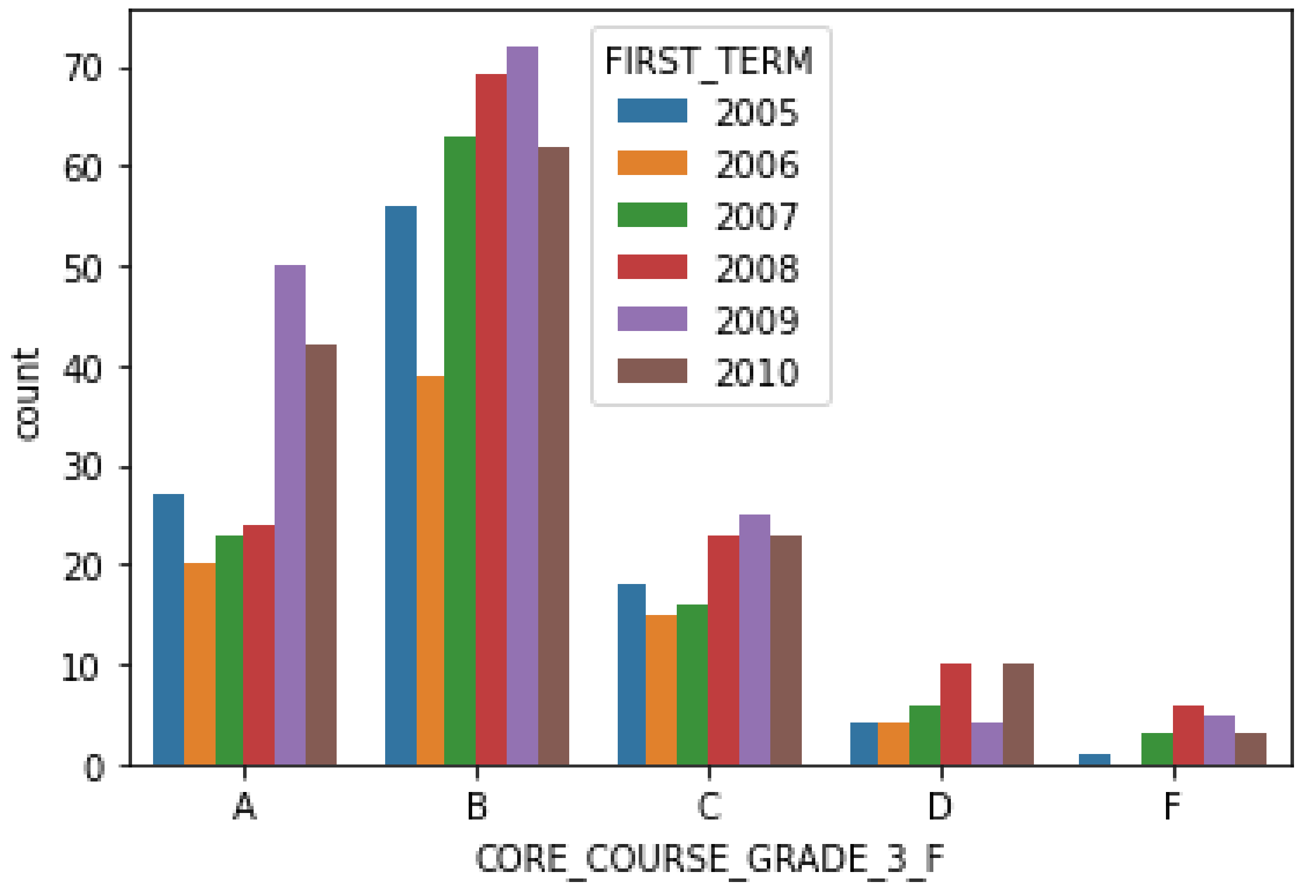




From the above visualization, we can see that the average performing student and high scoring students are leaving the institution who opted  the Core Course 2.

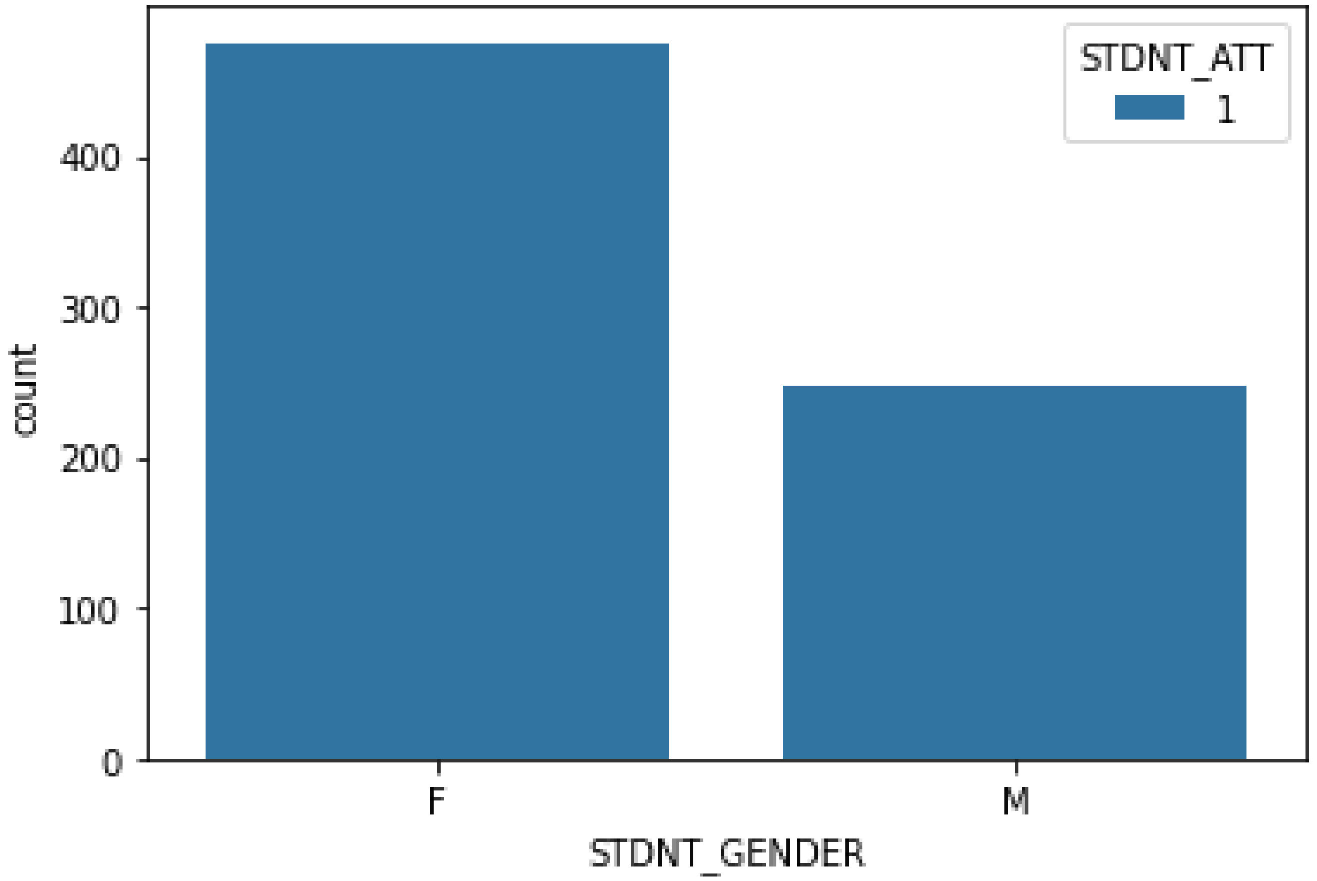
## Business Interpretations

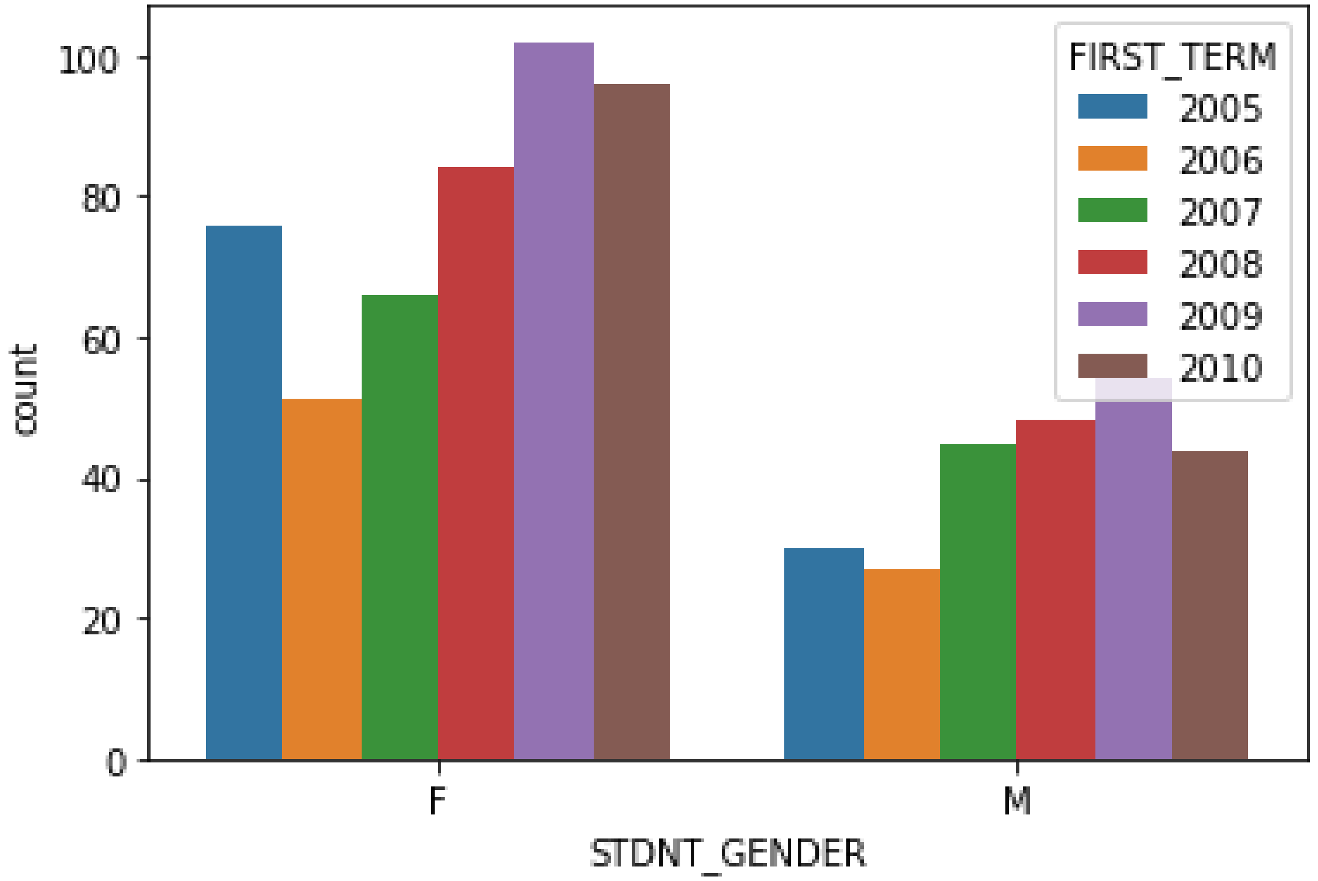




The maximum number of students joining Grade 3 does leave the institution, as shown in the figure above we can see that the above average performing students are leaving the university.

## Business Interpretations





From the above visualization, we can see that Female students are leaving the organization

## Business Interpretations

As we visualize key drive data, we can interpret it as follows: (21% of students left the institution in 2005-2010)

- 18% of students who get average in high school and 15% of students get average marks in entrance exam are leaving the University
- 21% of students who are leaving the university, 10% live within 100 kilometers
- 9% of students who are enrolled in a Grade 1 core course and leave the university with a score below average (C, D and F) grades
- 75% Students who are affiliated with a Grade 3 core course and leave the university with a good score (A and B Grade)
- 66% Female students who are are leaving the university

# Recommendations

As we visualize key drive data, we can interpret it as follows: (21% of students left the institution in 2005-2010)

- As we have seen that the average school performing student is leaving the university, for them we should plan the engagement of the faculty to understand the need of the students or improve the engagement of the teacher along with measuring the activity of the student.

- Take the opportunity to connect with alumni so that students get to know what happens at the end of the journey up and down graduation.

- For students who are performing below average, we may offer non-teaching services such as actively providing counseling and support services.

# Recommendations

- Provide special grants or financial aid programmers to sustain students who are scoring high

- Identified policies, strategies and interventions required to enhance participation, and retention of girls

- Achieving a good salary after graduation is the goal of every student, plays a key role in maintaining placement as well as enrollment, placement cell should be actively functioning in the university which can provide at least 99% placement. It should be stated in the whole market, mainly since the attrition rate is high in the local market, we have to recreate the image of the university in the local market.

- In today’s world, social media is playing a major role, each and every activity should be voiced through social platforms, which will create a different image across the market.

# THANK YOU