

Helping Students Achieve Four-year Graduation Rates by Predicting Computer Science (CS) Program Pain Points in CS Course Sequence.



by Vanessa Gonzalez

Overview

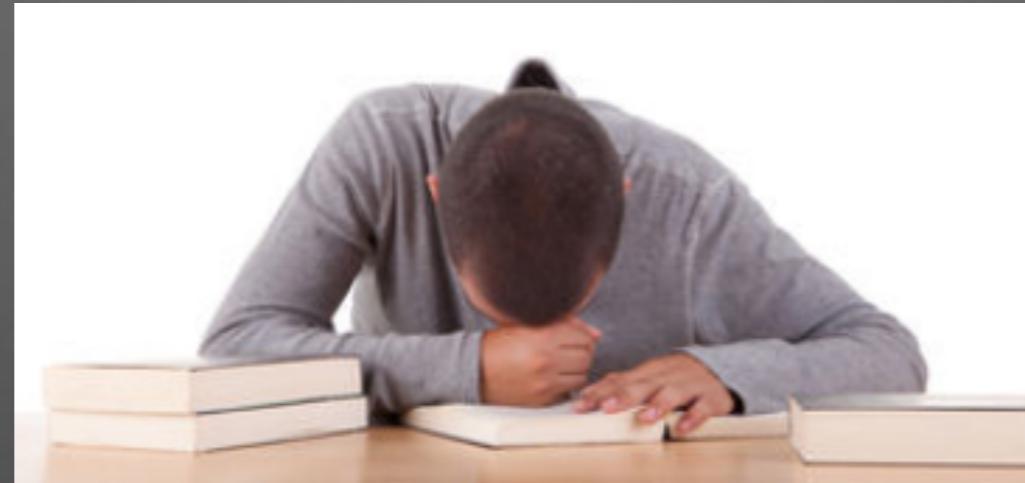
- Higher Education Institutions struggle with the problem of how to increase the graduation rates
- Federal graduation rate reporting:
 - 4 Year Graduation Rate
 - 6 Year Graduation Rate
- Graduation Rates are used to rank institutions and are important to maintain or increase enrollment and reputation.

4YGR
6YGR

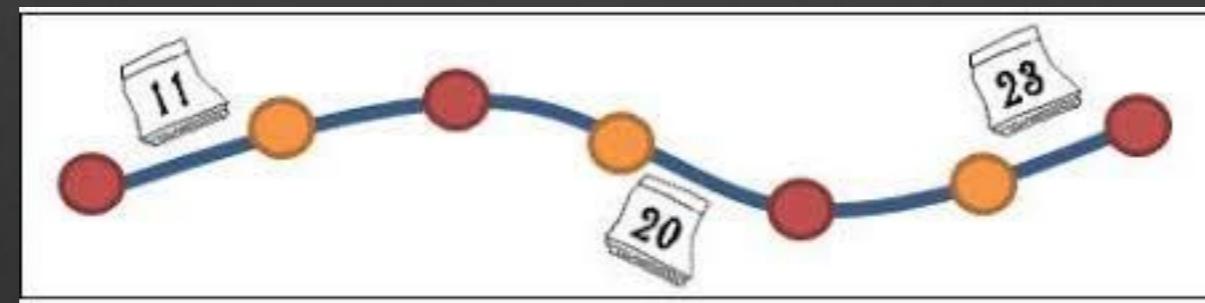


Questions to be Answered

- Which Computer Science (CS) students are at risk of leaving the program or the institution?



- Which are the main points of attrition in the CS course sequence conducive to the loss of students from the CS program or University?



Data

Query Data
from
University
Data Base



Combine and
Format Data



Data Cleaning
and
Preparation



Data
Modeling



Exploratory
Data Analysis



Data Sets

Data Set 1

- If original major = “CS” or undecided “(UN)”. If UN then first major =“CS”
- 536 observations
- 24 variables
- Year of original major date: 2008-2014
- Grades in GPA form as measurement of Courses

Data Set 1

```
'data.frame': 536 obs. of 24 variables:  
Year of OriginalMajorDate: int 2014 2008 2008 2011 2008 2...  
GraduationStatus : Factor w/ 3 levels "CurrentStudent",...: 2...  
YearsFromOMD : num 4 9.84 9.84 6.84 9.84 9.84 9.84 9.84...  
CsGrad : Factor w/ 3 levels "NG","OtherMajor",...: 3 2 1 1 2 3...  
4YG : Factor w/ 2 levels "No","Yes": 2 1 1 1 2 2 2 2 1 2 ...  
5YG : Factor w/ 2 levels "No","Yes": 2 2 1 1 2 2 2 2 1 2 ...  
6YG : Factor w/ 2 levels "No","Yes": 2 2 1 1 2 2 2 2 1 2 ...  
1_CSCI101 : num 4 NA NA NA NA 4 4 NA NA NA ...  
1_MATH111 : num 3 3 3 3 3 3 4 3 2 3 ...  
2_CSCI261 : num 4 4 4 3.3 4 3 4 3 3 3 ...  
2_MATH112 : num 2 2 2 4 3 3 4 4 2 2 ...  
2_MATH201 : num 3 3 2 NA 3 2 4 NA NA 1 ...  
3_CSCI262 : num 3 NA 1 3 NA 3 4 4 1 NA ...  
3_MATH213 : num 4 3 2 2 3 4 4 4 1 2 ...  
4_CSCI341 : num 2 NA 2 2 NA 3 4 NA 3 3 ...  
4_CSCI358 : num 4 NA 3 2 NA 2 4 NA NA NA ...  
4_MATH225 : num 4 3 1 4 4 3 4 4 NA 1 ...  
5_CSCI306 : num 3 NA NA 3.7 NA 4 4 4 NA NA ...  
5_CSCI403 : num 4 NA NA 3 NA 4 4 NA NA NA ...  
5_MATH332 : num 3 NA 2 3 NA 3 4 NA NA NA ...  
6_CSCI406 : num 2 NA NA 0.3 NA 2 4 NA NA NA ...  
7_CSCI370 : num 3.3 NA NA NA NA 4 4 NA NA NA ...  
8_CSCI400 : num 3.3 NA NA 3.3 NA 3 4 NA NA NA ...  
9_CSCI442 : num 2.3 NA NA NA NA 3 4 NA NA NA ...
```

Data Sets

Data Set 2

- If original major = “CS” or undecided “(UN)”. If UN then first major =“CS”
- 195 observations
- 25 variables
- Year of original major date: 2008-2018
- Date when course was taken as measurement of Courses
- CsGrad = “Other Major” or “NG”

Data Set 2

```
Classes 'tbl_df', 'tbl' and 'data.frame': 195 obs. of 25 variables:  
  UID : chr "12972" "12973" "41647" "98022" ...  
  Year of OriginalMajorDate: chr "2008" "2008" "2011" "2008" ...  
  YearsFromOMD : chr "9.88" "9.88" "6.88" "9.88" ...  
  4YG : Factor "No" "No" "No" "Yes" ...  
  5YG : Factor "Yes" "No" "No" "Yes" ...  
  6YG : Factor "Yes" "No" "No" "Yes" ...  
  GraduationStatus : chr "Graduated" "InactiveReg" "InactiveReg" ...  
  CsGrad : chr "OtherMajor" "NG" "NG" "OtherMajor" ...  
  Nine.CSCI442 : num NA NA NA NA NA ...  
  Eight.CSCI400 : num NA NA 2017 NA NA ...  
  Seven.CSCI370 : num NA NA NA NA NA ...  
  Six.CSCI406 : num NA NA 2018 NA NA ...  
  Five.CSCI403 : num NA NA 2017 NA NA ...  
  Five.MATH332 : num NA 2011 2015 NA NA ...  
  Five.CSCI306 : num NA NA 2017 NA 2011 ...  
  Four.CSCI358 : num NA 2012 2017 NA NA ...  
  Four.CSCI341 : num NA 2011 2015 NA NA ...  
  Four.MATH225 : num 2010 2011 2015 2010 2009 ...  
  Three.CSCI262 : num NA 2010 2015 NA 2010 ...  
  Three.MATH213 : num 2010 2010 2013 2009 2009 ...  
  Two.CSCI261 : num 2010 2010 2014 2010 2009 ...  
  Two.MATH201 : num 2011 2012 NA 2010 NA ...  
  Two.MATH112 : num 2009 2010 2013 2009 2009 ...  
  One.MATH111 : num 2009 2009 2013 2009 2009 ...  
  One.CSCI101 : num NA NA NA NA NA ...
```

Data Preparation



- Academic period format: Year-period to year-month.
- Course names: Number of Semester recommended plus course code.
- Double majors: Show just the CS record.
- Several CASE statements in Tableau were used to define depending registration in Spring 18 and Fall 18 if the students were current students or if the students had left the institution.
- CASE statements in Tableau were used to define the “Student group” as “CS students” if CS had been their original major or their first major.
- Calculations were added in Tableau to define the length between original major date and graduation date.
- Additional modifications and preparation of the data sets happened in R.

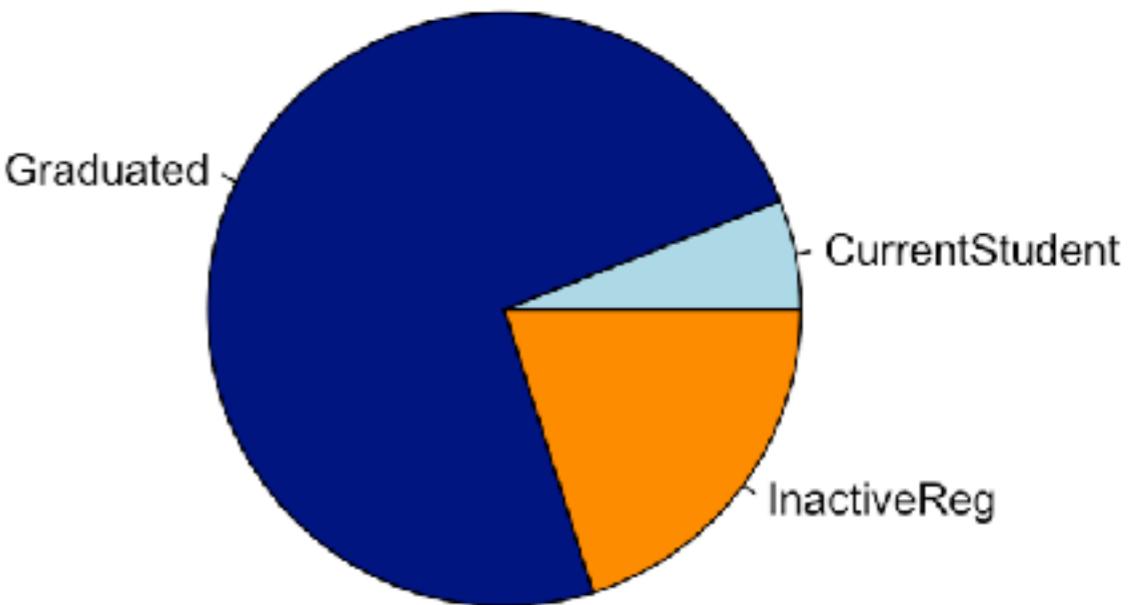
Exploratory Analysis (EDA)

536 Students in
the Data Set 1

Total Number of Students with a CS First Major		
	Number of Students	% of Students
Graduated	396	73.88%
Left Inst.	108	20.15%
Current	32	5.97%
Total CS Students	536	1

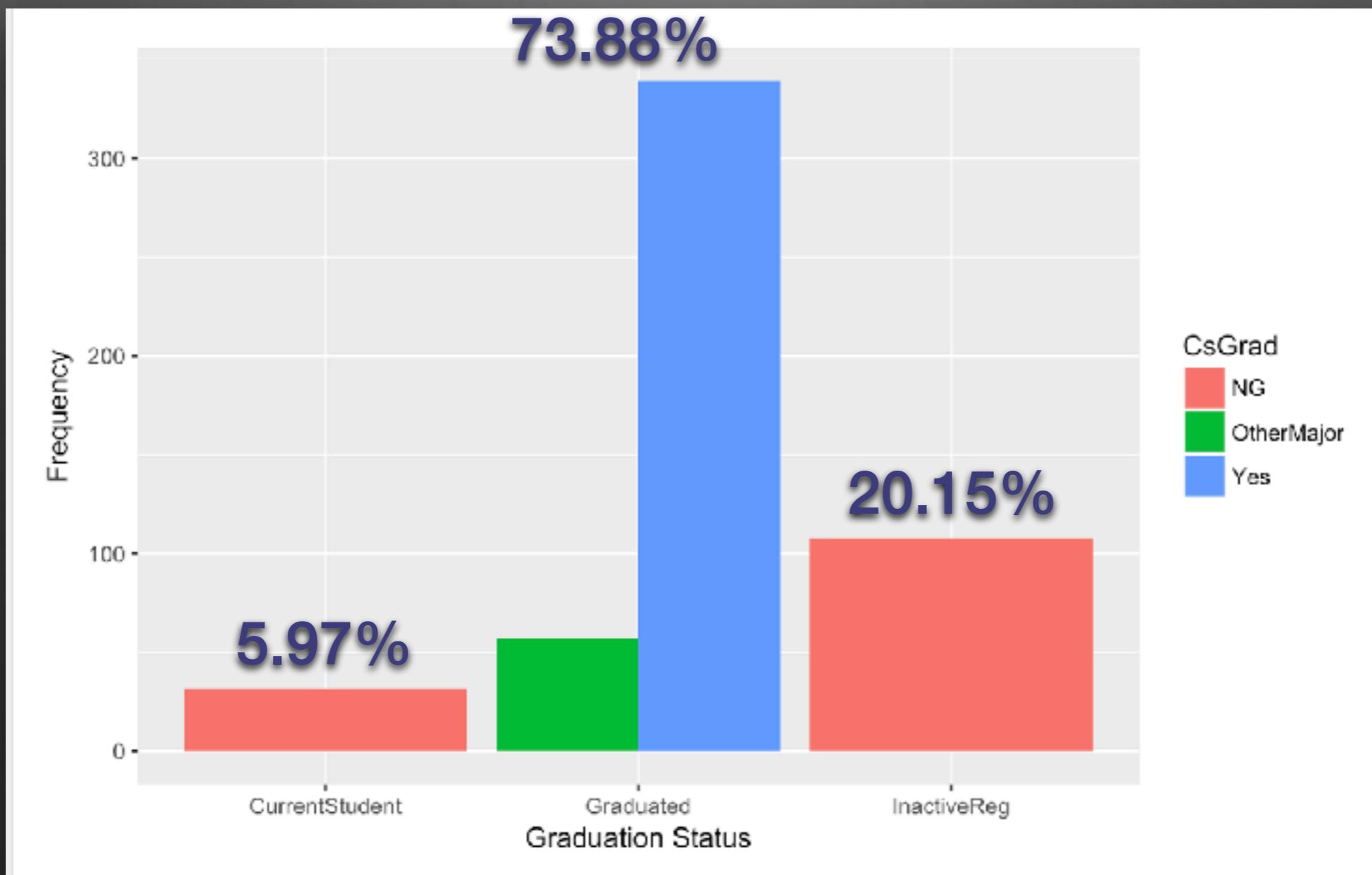
2008 - 2014

CS Students Status that Started 2008-2014

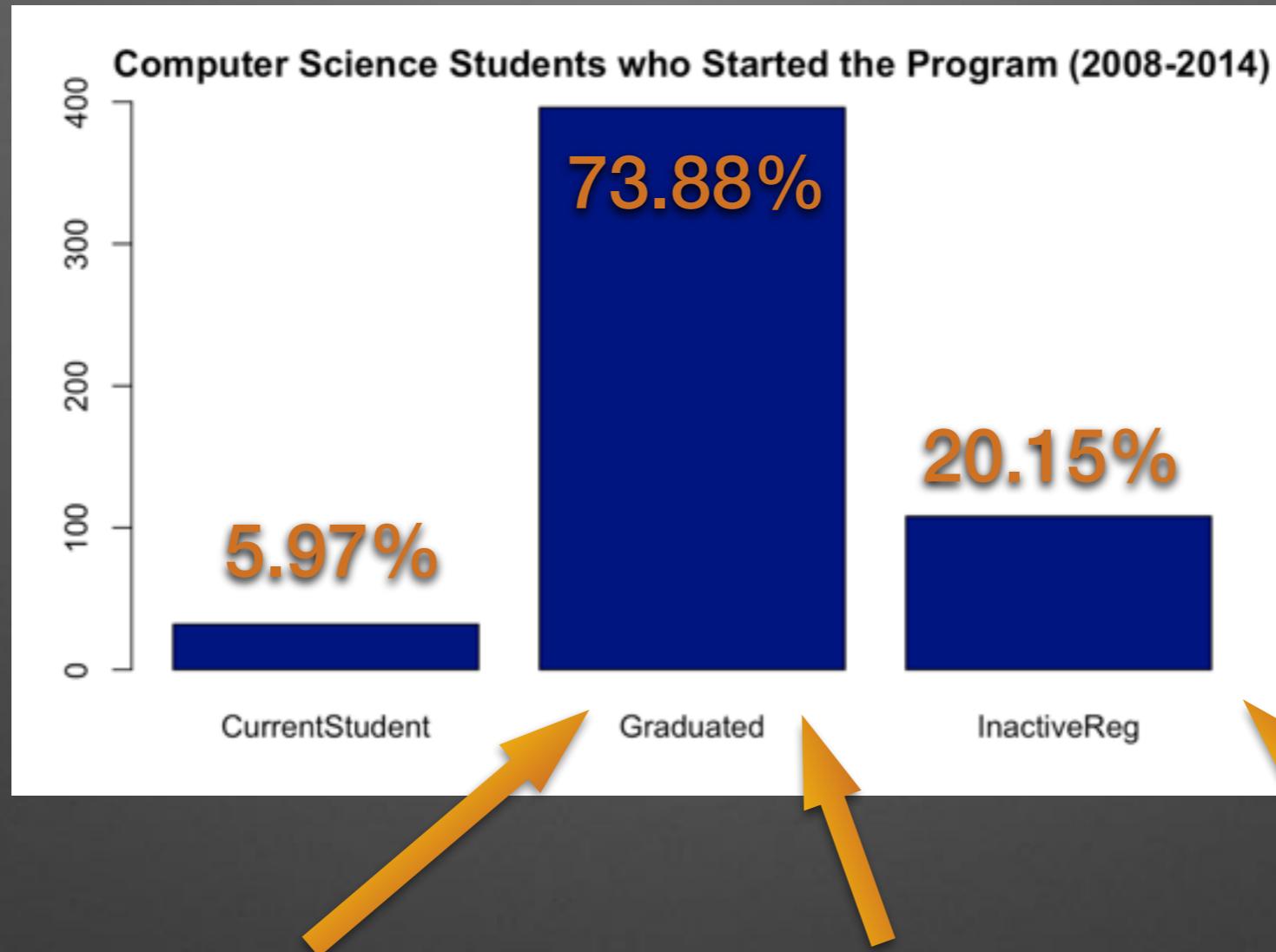


Explanatory Analysis (EDA)

What Happened to CS Students?



Exploratory Analysis (EDA)



Students who
Graduated from CS

Graduation Rates of Students that Graduated from CS

CS Major	Male	Female
4 year	72.00%	71.79%
5 year	93.67%	87.18%
6 year	97.67%	97.44%

Students who left to
other Majors

Graduation Rates of CS students that left to other Majors

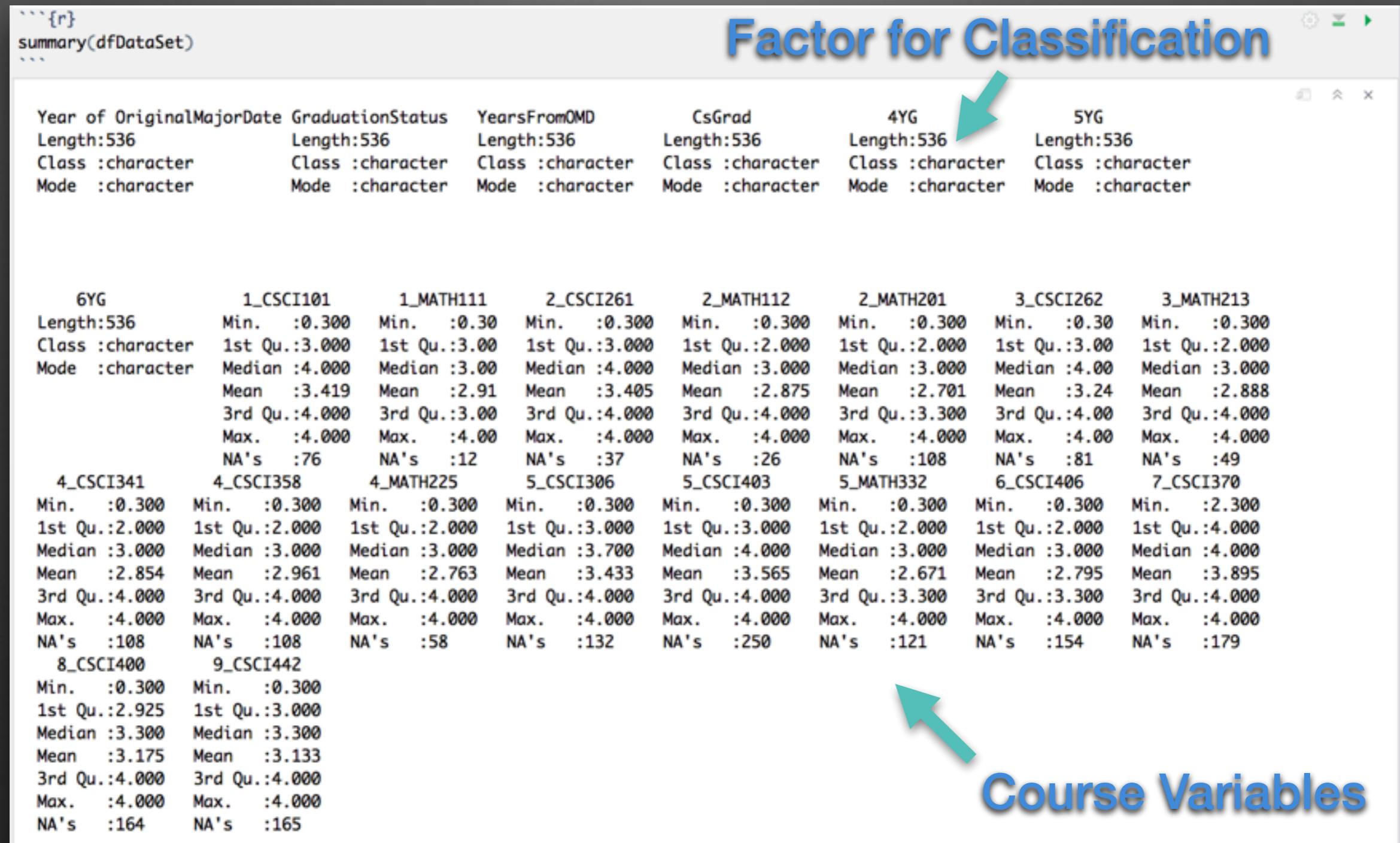
Other Major	Male	Female
4 year	70.83%	66.67%
5 year	91.67%	88.89%
6 year	97.92%	88.89%

Students who
left the
Institution

Exploratory Analysis (EDA)

Summary of Data Set 1

```
```{r}
summary(dfDataSet)
```



Factor for Classification



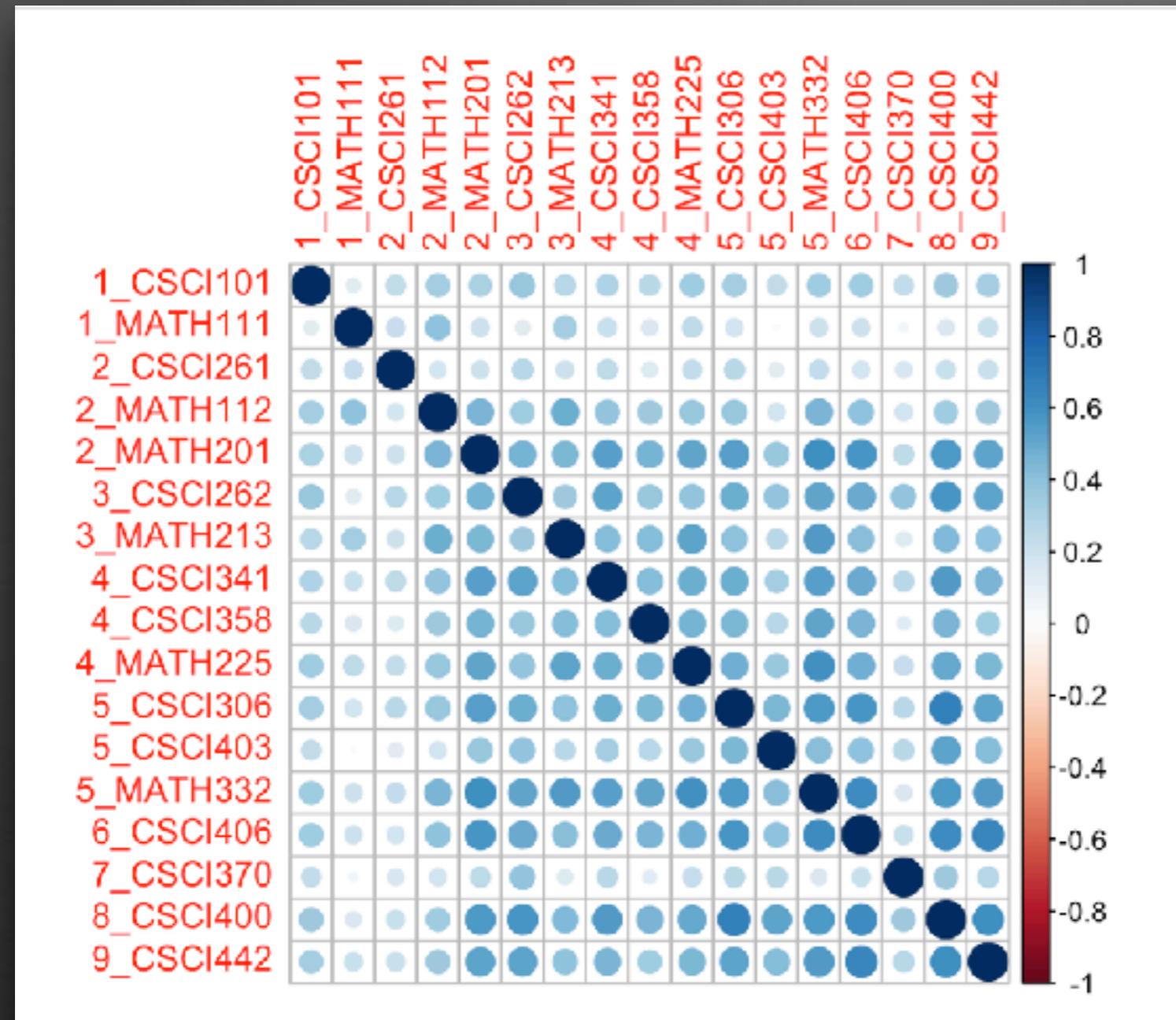
Year of OriginalMajorDate	GraduationStatus	YearsFromOMD	CsGrad	4YG	5YG
Length:536	Length:536	Length:536	Length:536	Length:536	Length:536
Class :character	Class :character	Class :character	Class :character	Class :character	Class :character
Mode :character	Mode :character	Mode :character	Mode :character	Mode :character	Mode :character


6YG	1_CSCI101	1_MATH111	2_CSCI261	2_MATH112	2_MATH201	3_CSCI262	3_MATH213
Length:536	Min. :0.300	Min. :0.30	Min. :0.300	Min. :0.300	Min. :0.300	Min. :0.30	Min. :0.300
Class :character	1st Qu.:3.000	1st Qu.:3.00	1st Qu.:3.000	1st Qu.:2.000	1st Qu.:2.000	1st Qu.:3.00	1st Qu.:2.000
Mode :character	Median :4.000	Median :3.00	Median :4.000	Median :3.000	Median :3.000	Median :4.00	Median :3.000
	Mean :3.419	Mean :2.91	Mean :3.405	Mean :2.875	Mean :2.701	Mean :3.24	Mean :2.888
	3rd Qu.:4.000	3rd Qu.:3.00	3rd Qu.:4.000	3rd Qu.:4.000	3rd Qu.:3.300	3rd Qu.:4.00	3rd Qu.:4.000
	Max. :4.000	Max. :4.00	Max. :4.000	Max. :4.000	Max. :4.000	Max. :4.00	Max. :4.000
	NA's :76	NA's :12	NA's :37	NA's :26	NA's :108	NA's :81	NA's :49
4_CSCI341	4_CSCI358	4_MATH225	5_CSCI306	5_CSCI403	5_MATH332	6_CSCI406	7_CSCI370
Min. :0.300	Min. :0.300	Min. :0.300	Min. :0.300	Min. :0.300	Min. :0.300	Min. :0.300	Min. :2.300
1st Qu.:2.000	1st Qu.:2.000	1st Qu.:2.000	1st Qu.:3.000	1st Qu.:3.000	1st Qu.:2.000	1st Qu.:2.000	1st Qu.:4.000
Median :3.000	Median :3.000	Median :3.000	Median :3.700	Median :4.000	Median :3.000	Median :3.000	Median :4.000
Mean :2.854	Mean :2.961	Mean :2.763	Mean :3.433	Mean :3.565	Mean :2.671	Mean :2.795	Mean :3.895
3rd Qu.:4.000	3rd Qu.:4.000	3rd Qu.:4.000	3rd Qu.:4.000	3rd Qu.:4.000	3rd Qu.:3.300	3rd Qu.:3.300	3rd Qu.:4.000
Max. :4.000	Max. :4.000	Max. :4.000	Max. :4.000	Max. :4.000	Max. :4.000	Max. :4.000	Max. :4.000
NA's :108	NA's :108	NA's :58	NA's :132	NA's :250	NA's :121	NA's :154	NA's :179
8_CSCI400	9_CSCI442						
Min. :0.300	Min. :0.300						
1st Qu.:2.925	1st Qu.:3.000						
Median :3.300	Median :3.300						
Mean :3.175	Mean :3.133						
3rd Qu.:4.000	3rd Qu.:4.000						
Max. :4.000	Max. :4.000						
NA's :164	NA's :165						


```

Exploratory Analysis (EDA)



Correlation
Between Course
Variables

Exploratory Analysis (EDA)

Summary of Data Set 2

| UID | Year of OriginalMajorDate | YearsFromOMD | 4YG | 5YG | 6YG | | |
|------------------|---------------------------|------------------|------------------|------------------|------------------|--------------|--------------|
| Length:195 | Length:195 | Length:195 | Length:195 | Length:195 | Length:195 | | |
| Class :character | Class :character | Class :character | Class :character | Class :character | Class :character | | |
| Mode :character | Mode :character | Mode :character | Mode :character | Mode :character | Mode :character | | |
| | | | | | | | |
| GraduationStatus | CsGrad | Nine.CSCI442 | Eight.CSCI400 | Seven.CSCI370 | Six.CSCI406 | Five.CSCI403 | Five.MATH332 |
| Length:195 | Length:195 | Min. :2010 | Min. :2010 | Min. :2011 | Min. :2010 | Min. :2010 | Min. :2010 |
| Class :character | Class :character | 1st Qu.:2012 | 1st Qu.:2014 | 1st Qu.:2012 | 1st Qu.:2013 | 1st Qu.:2014 | 1st Qu.:2012 |
| Mode :character | Mode :character | Median :2014 | Median :2015 | Median :2015 | Median :2015 | Median :2016 | Median :2014 |
| | | Mean :2014 | Mean :2015 | Mean :2014 | Mean :2015 | Mean :2016 | Mean :2014 |
| | | 3rd Qu.:2017 | 3rd Qu.:2016 | 3rd Qu.:2016 | 3rd Qu.:2017 | 3rd Qu.:2017 | 3rd Qu.:2015 |
| | | Max. :2018 | Max. :2017 | Max. :2018 | Max. :2019 | Max. :2019 | Max. :2018 |
| | | NA's :180 | NA's :177 | NA's :182 | NA's :172 | NA's :177 | NA's :125 |
| Five.CSCI306 | Four.CSCI358 | Four.CSCI341 | Four.MATH225 | Three.CSCI262 | Three.MATH213 | Two.CSCI261 | Two.MATH201 |
| Min. :2010 | Min. :2009 | Min. :2010 | Min. :2009 | Min. :2009 | Min. :2009 | Min. :2009 | Min. :2010 |
| 1st Qu.:2012 | 1st Qu.:2011 | 1st Qu.:2012 | 1st Qu.:2011 | 1st Qu.:2012 | 1st Qu.:2011 | 1st Qu.:2011 | 1st Qu.:2011 |
| Median :2014 | Median :2014 | Median :2014 | Median :2013 | Median :2014 | Median :2013 | Median :2013 | Median :2013 |
| Mean :2014 | Mean :2014 | Mean :2014 | Mean :2013 | Mean :2014 | Mean :2013 | Mean :2013 | Mean :2013 |
| 3rd Qu.:2016 | 3rd Qu.:2016 | 3rd Qu.:2016 | 3rd Qu.:2015 | 3rd Qu.:2016 | 3rd Qu.:2015 | 3rd Qu.:2015 | 3rd Qu.:2015 |
| Max. :2019 | Max. :2018 | Max. :2019 | Max. :2019 | Max. :2019 | Max. :2018 | Max. :2018 | Max. :2019 |
| NA's :153 | NA's :132 | NA's :123 | NA's :73 | NA's :94 | NA's :59 | NA's :37 | NA's :118 |
| Two.MATH112 | One.MATH111 | One.CSCI101 | | | | | |
| Min. :2009 | Min. :2009 | Min. :2011 | | | | | |
| 1st Qu.:2010 | 1st Qu.:2010 | 1st Qu.:2012 | | | | | |
| Median :2012 | Median :2012 | Median :2014 | | | | | |
| Mean :2013 | Mean :2012 | Mean :2014 | | | | | |
| 3rd Qu.:2015 | 3rd Qu.:2015 | 3rd Qu.:2015 | | | | | |
| Max. :2018 | Max. :2018 | Max. :2018 | | | | | |
| NA's :28 | NA's :10 | NA's :66 | | | | | |



Course Variables

Analysis for Data Set 1



Which Computer Science (CS) students are at risk of leaving the program or the institution?

Machine Learning Algorithms



Machine Learning
Algorithms Used

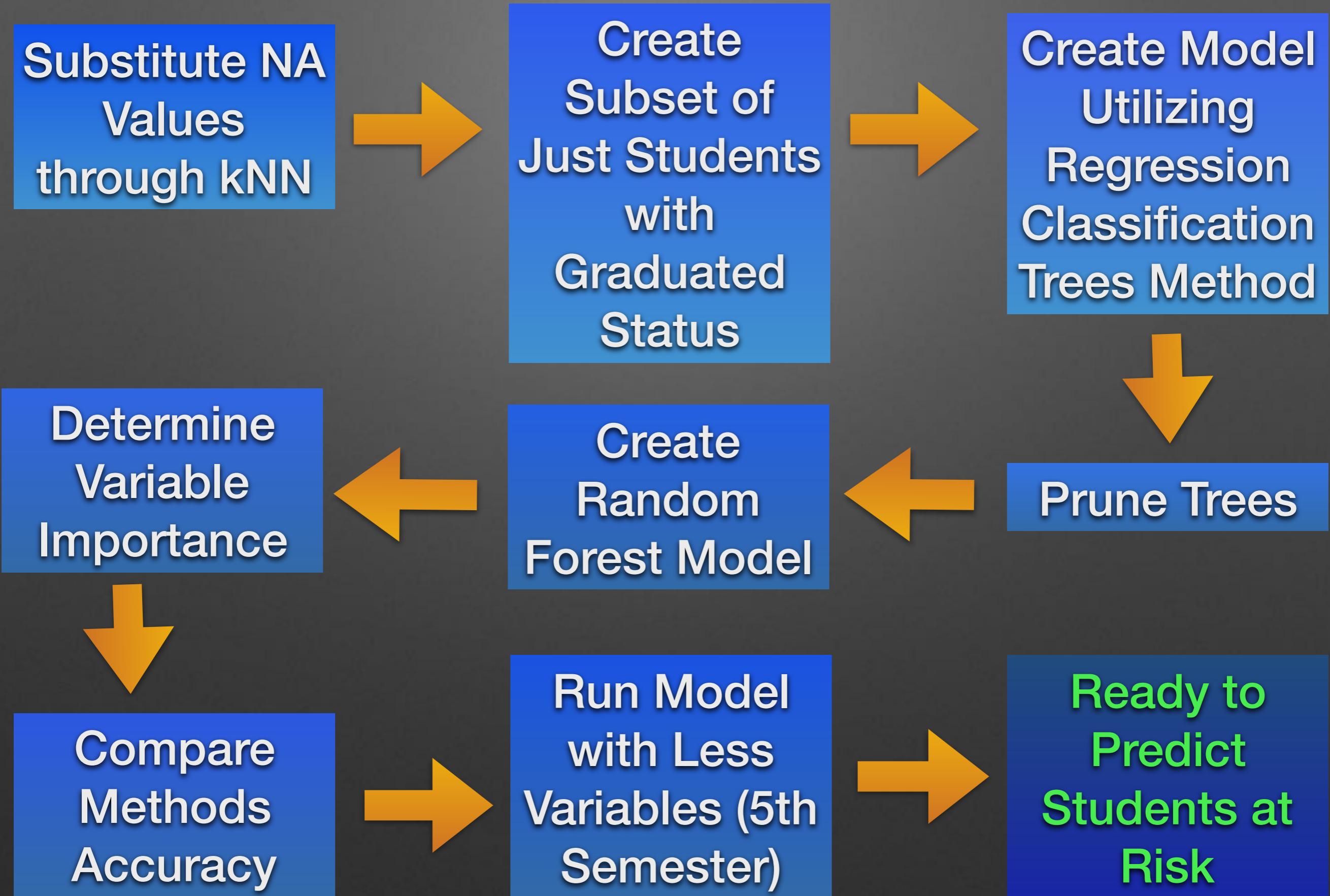
kNN

Regression
Classification Trees

Random Forest

Logistic Regression for
Variable Importance

Process of Analysis for Data Set 1

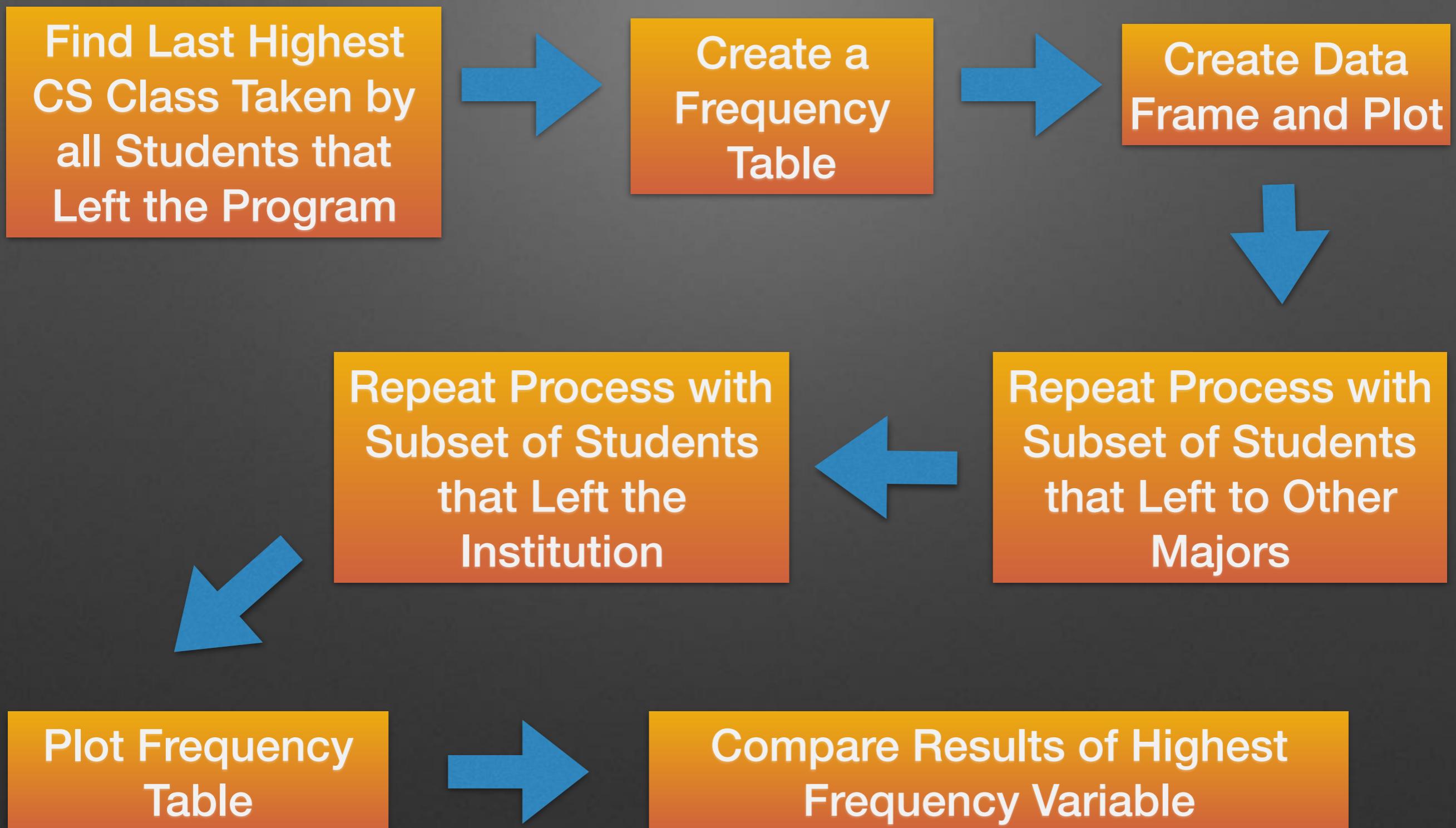


Analysis for Data Set 2



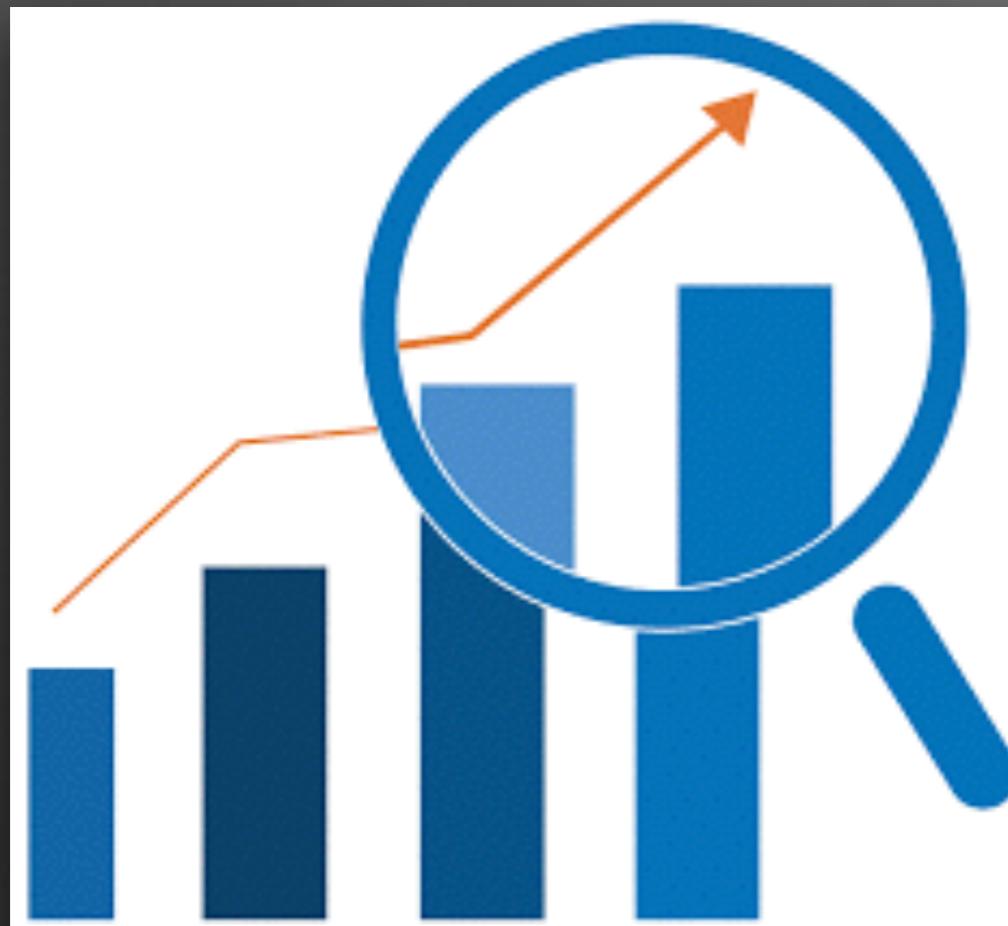
Which are the main points of attrition in the CS course sequence conducive to the loss of students from the CS program or University?

Process of Analysis



Results Data Set 1

Random Forest All Courses Analysis



80.77% Accuracy

Confusion Matrix and Statistics

FourYG.rf.prediction No Yes

| | | |
|-----|----|----|
| No | 8 | 1 |
| Yes | 14 | 55 |

Accuracy : 0.8077

95% CI : (0.7027, 0.8882)

No Information Rate : 0.7179

P-Value [Acc > NIR] : 0.047142

Kappa : 0.4214

McNemar's Test P-Value : 0.001946

Sensitivity : 0.3636

Specificity : 0.9821

Pos Pred Value : 0.8889

Neg Pred Value : 0.7971

Prevalence : 0.2821

Detection Rate : 0.1026

Detection Prevalence : 0.1154

Balanced Accuracy : 0.6729

'Positive' Class : No

Results Data Set 1

Random Forest Less Course Variables Analysis



79.49% Accuracy

| Confusion Matrix and Statistics | | |
|--------------------------------------|----|----|
| <pre>FourYG.rf.prediction.Less</pre> | | |
| No | 10 | 4 |
| Yes | 12 | 52 |
|
 | | |
| Accuracy : 0.7949 | | |
| 95% CI : (0.6884, 0.870) | | |
| No Information Rate : 0.7179 | | |
| P Value [Acc > NIR] : 0.08014 | | |
|
 | | |
| Kappa : 0.4307 | | |
| McNemar's Test P-Value : 0.08012 | | |
|
 | | |
| Sensitivity : 0.4545 | | |
| Specificity : 0.9286 | | |
| Pos Pred Value : 0.7143 | | |
| Neg Pred Value : 0.8125 | | |
| Prevalence : 0.2821 | | |
| Detection Rate : 0.1202 | | |
| Detection Prevalence : 0.1795 | | |
| Balanced Accuracy : 0.6916 | | |
|
 | | |
| 'Positive' Class : No | | |

Results Data Set 1

Random Forest 5th Semester Predictive Model



75.74% Accuracy

Confusion Matrix and Statistics

```
FourYG.rf.prediction.Less.Sem5 No Yes
                                No   11    8
                                Yes  11   48
```

```
Accuracy : 0.7564
95% CI  : (0.646, 0.8465)
No Information Rate : 0.7179
P-Value [Acc > NIR] : 0.2685
```

Kappa : 0.3726
Mcnemar's Test P-Value : 0.6464

```
Sensitivity : 0.5000
Specificity : 0.8571
Pos Pred Value : 0.5789
Neg Pred Value : 0.8136
Prevalence : 0.2821
Detection Rate : 0.1410
Detection Prevalence : 0.2436
Balanced Accuracy : 0.6786
```

'Positive' Class : No

Results Data Set 1

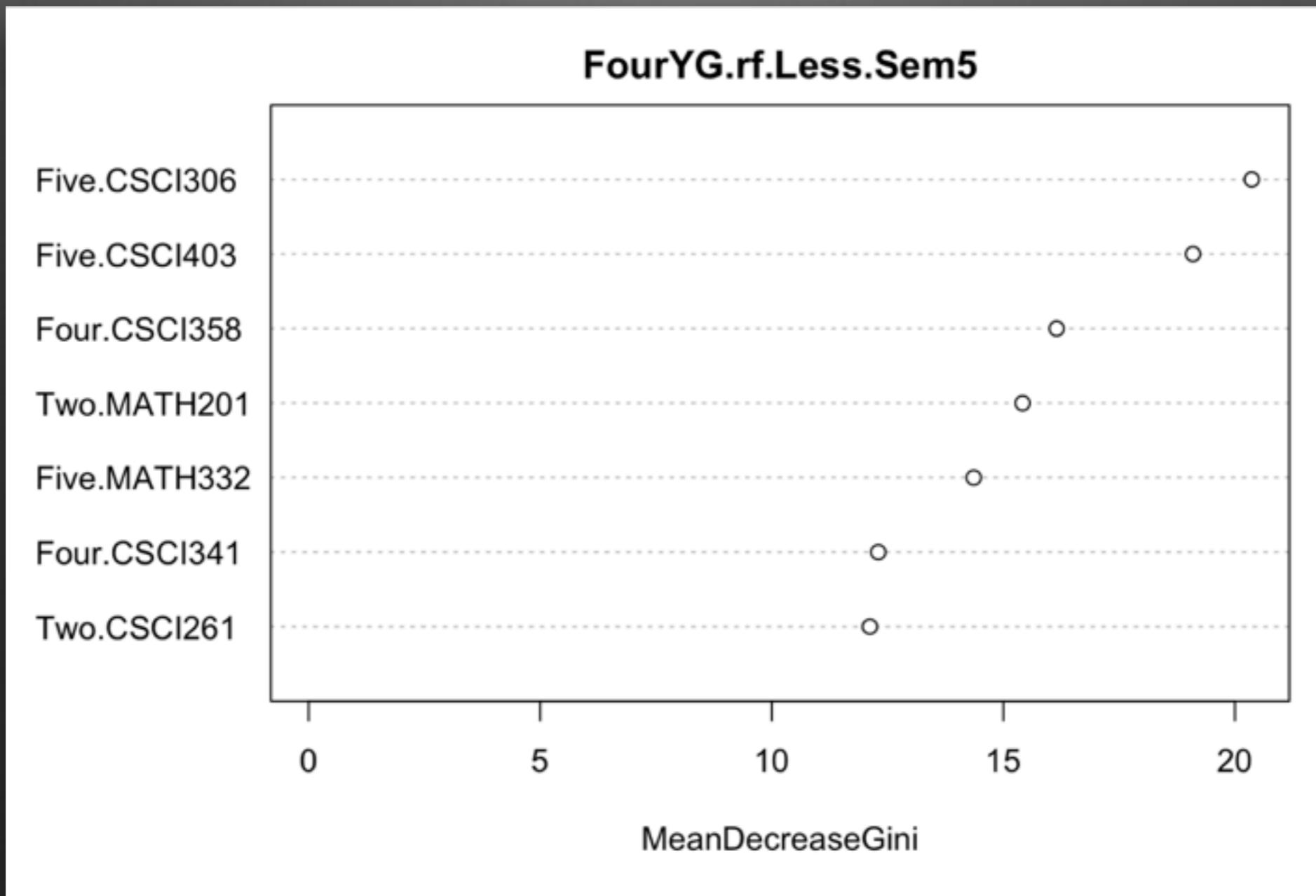
Variable Importance to
5th Semester
Prediction Model

CSCI306

| | MeanDecreaseGini |
|--------------|------------------|
| Two.CSCI261 | 12.11959 |
| Two.MATH201 | 15.41650 |
| Four.CSCI341 | 12.30205 |
| Four.CSCI358 | 16.15024 |
| Five.CSCI306 | 20.36310 |
| Five.CSCI403 | 19.09542 |
| Five.MATH332 | 14.36175 |

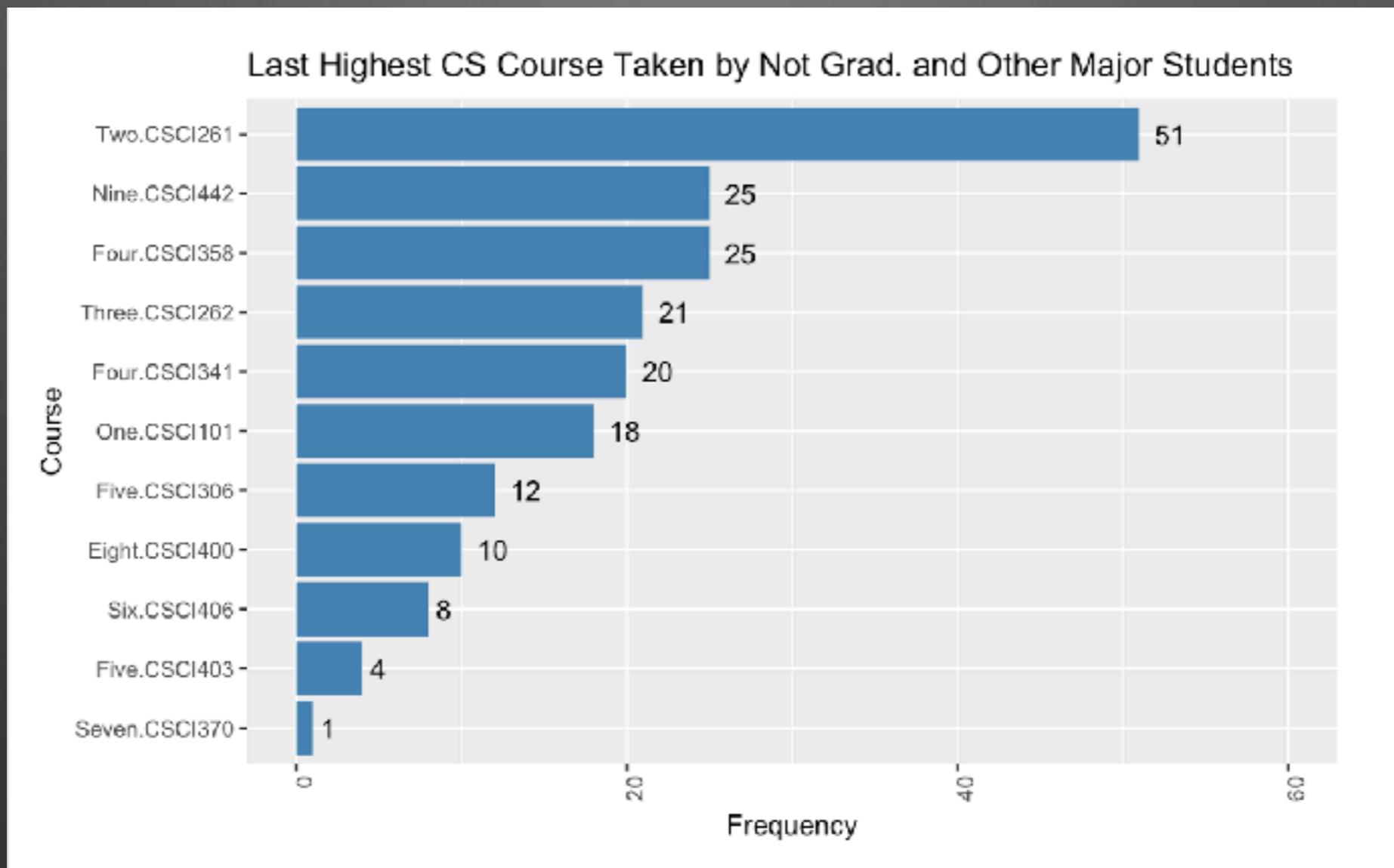
Results Data Set 1

Variables Importance Plot



Results Data Set 2

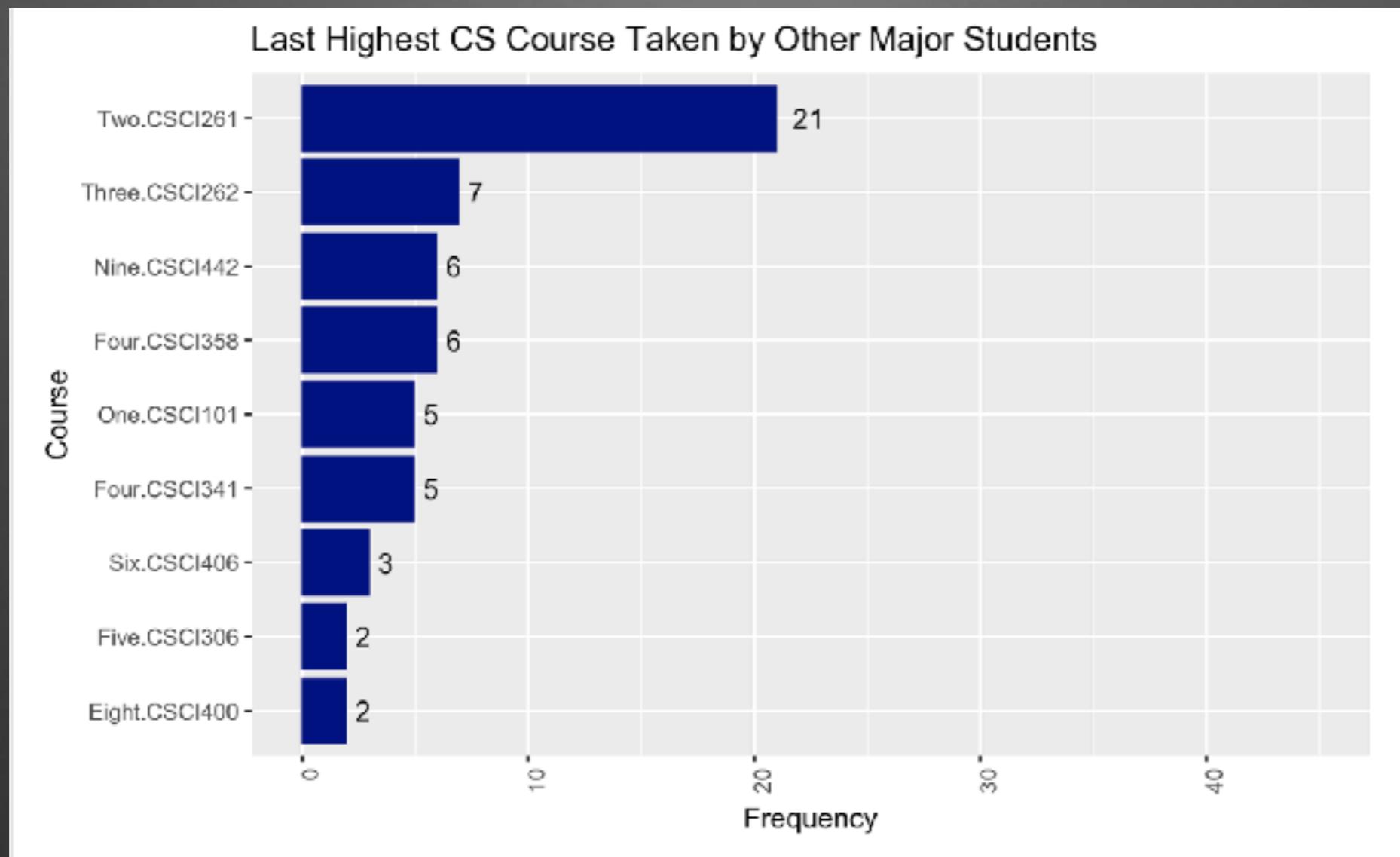
All Students that Left the Program



CSCI261- Highest Frequency Course

Results Data Set 2

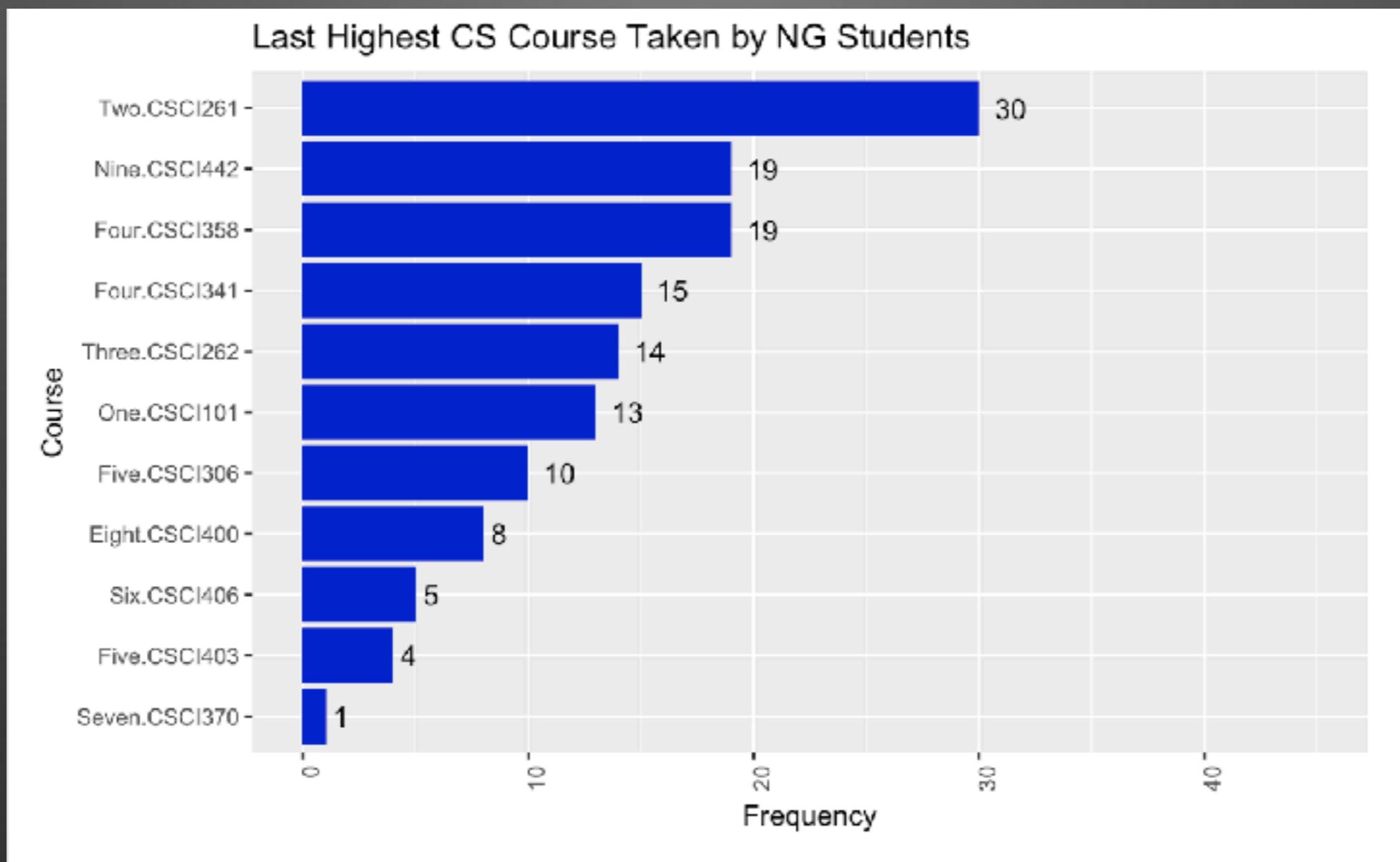
Students that Left to Other Majors



CSCI261- Highest Frequency Course

Results Data Set 2

Students that Left the Institution



CSCI261- Highest Frequency Course

Conclusions

- **536 Students** were registered with a CS first major in the 2008-2014 time frame. By Spring 2018, **73.88%** of those students graduated, **20.15%** left the institution and **5.97%** left the CS program and graduated from a different major.
- By using the 5th Semester Model at the start of students 6th Semester students at **risk of not graduating in four years** may be predicted with a **75.74% accuracy** and additional support may be provided to these students to increase the program four-year graduation rate.



Conclusions

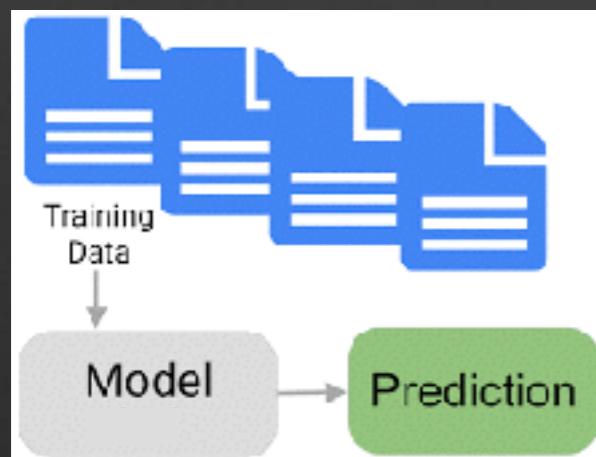
- Considering that the no **information rate was 71.8%** it is an acceptable result with a **Kappa of 0.37**.
- There is a **strong correlation between different CS courses** in the sequence but was interesting to find a strong correlation of the MATH201 (Statistics Course) with so many of the CS courses.
- With the second data set it was found that most students that leave the program do so after taking the **CSCI261** course followed by CSCI442, CSCI358, and CSCI262.



Steps Forward

- There is a lot more to be done. More questions to be answered and other angles to be explored. It would be interesting to add more variables to our data set including gender, nationality, instate or out of state tuition, and race.
- It would also be interesting to apply the same model and process to other programs course sequences and reach out to students at risk to provide them with additional support.

Thank you!



Student
Support



Thank You!

