Problem (Seed# 28)

The provided data is about university students, some of whom filled out the Student Experience in the Research University (SERU) survey. The university would like to be able to predict which students are most likely to complete the survey, and the features that drive this prediction.

Understanding Data

Based on the available data, the idea is to find the characteristics of students who are likely to complete the survey.

The 'target' value from the dataset was found to be RESPONDED_SURVEY. This is boolean attribute that tells whether the said student responded to the survey or not.

Evaluating the data, it was found that there were several data features that were duplicates or granular features of an existing feature that could be ignored. Some categorical attributes have a code as well as a related text attribute. Some of such attributes are given below

COLLEGE_CODE1, MAJOR_CODE1

COLLEGE_CODE2, MAJOR_CODE2

SATICR, SATIIM, SATIIW, SATIM, SATIW

ETHINICITY_LOC, DECLINETOSTATE

ACTR, ACTE

Identified several score related fields and it would be good to analyze the effect of scores by getting a cumulative score or grouping the similar set of scores

Since the data dictionary and the actual dataset had discrepancies, I have assumed the below attributes define the characteristics as mentioned here.

INVITED_PREVIOUS: The student was invited earlier to respond to this survey. Data Type: Boolean

SP16_ENRL_HRS_AT_CENSES: Students initial plan in hours to enroll before Spring2016 semester started

SP16_TERM_UI_GRADED_HRS: Actual enrolled hours in Spring 2016

SP16_TERM_UI_TOTAL_HRS: Not very sure. Why this would be different from graded hrs. or why is it sometimes lesser than graded hours.

Data Preparation

Since the different scores are with different scales, it would be better to transform all the scores to the same scale for doing a better analysis

Compute number of years of the student from the YEAR attribute and analyze the number of years in the university.

Some of the missing column values are inherently a boolean ‘N’. The attributes which needs to be treated in that fashion are given below. Update the NA’s with ‘N’

NONRESIDENT

PACIFICISLANDER

HISPANIC

ASIAN

AMERINDIAN

AFRICANAMERICAN

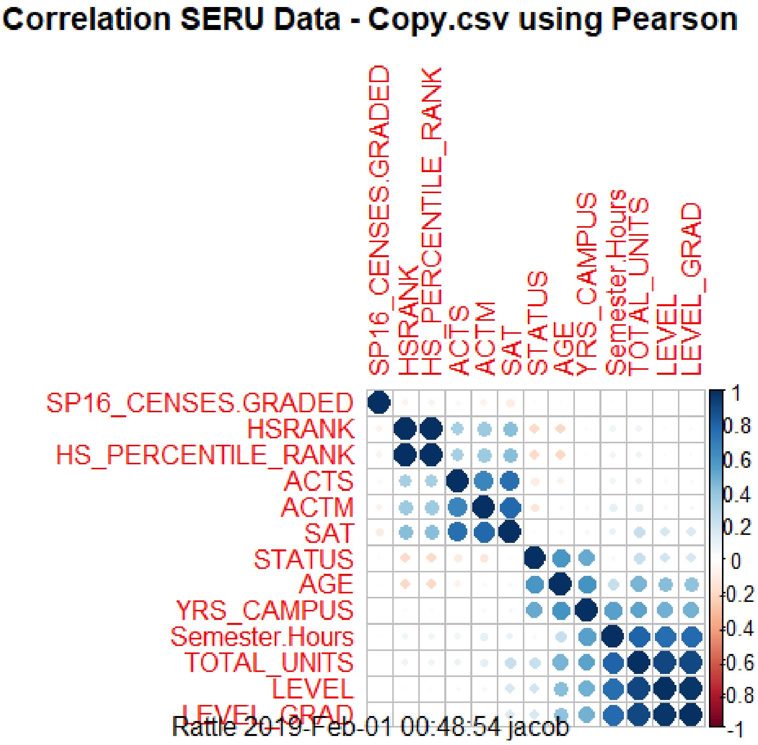
WHITE

A calculated field was created to see whether the students did take the number of courses they intended to take during that semester with the actual graded courses.

Convert wrong numerical types to proper categorical type and fix wrong categories to numeric data type

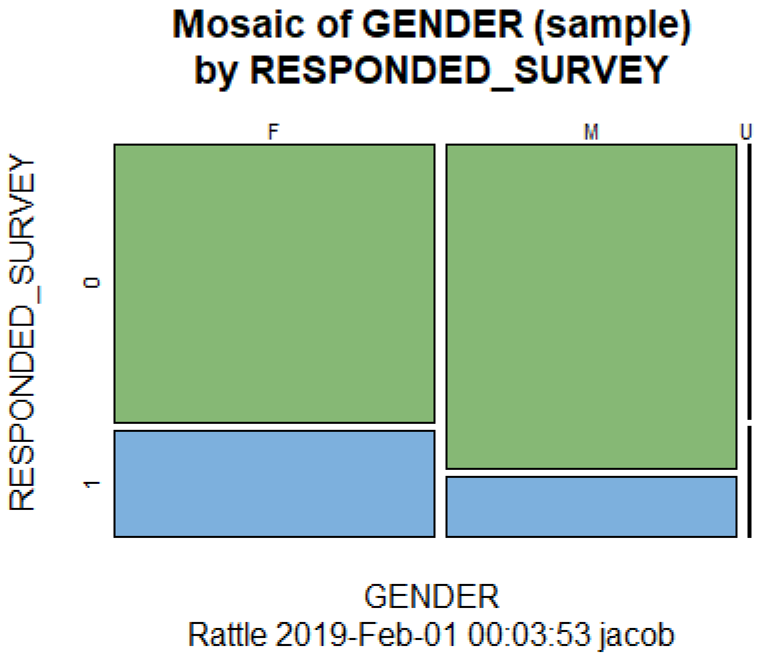
Data Evaluation

One of the first tasks was to evaluate the correlation graph using all available features. Several correlations between different features of the dataset were identified. Wanted to run this before removing any of the attributes; but ran after removing all the above-mentioned attributes in ‘Understanding Data’ because the plot wasn’t readable with all that data.

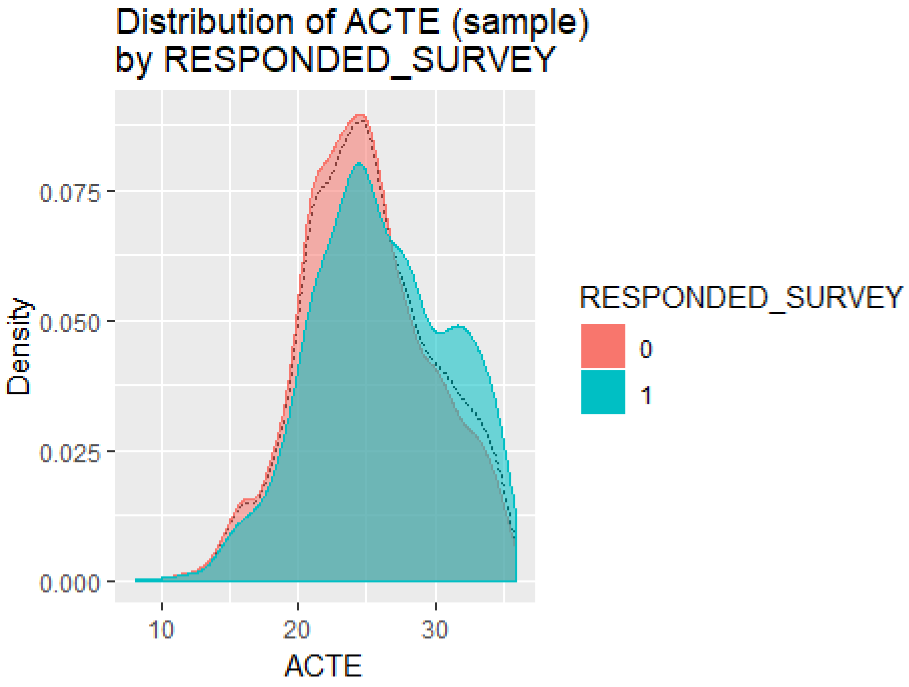


Here we do see a clear correlation between the different scores (ACTS, ACTM, SAT). HSRANK and HS_PERCENTILE_RANK have almost 100% correlation since one is a percentile of the other attribute. STATUS, AGE and YRS_CAMPUS is very much correlated because STATUS and YRS_CAMPUS relates to student AGE. Other highly related fields are SEMESTER_HOURS, TOTAL_UNITS, LEVEL and LEVEL_GRAD

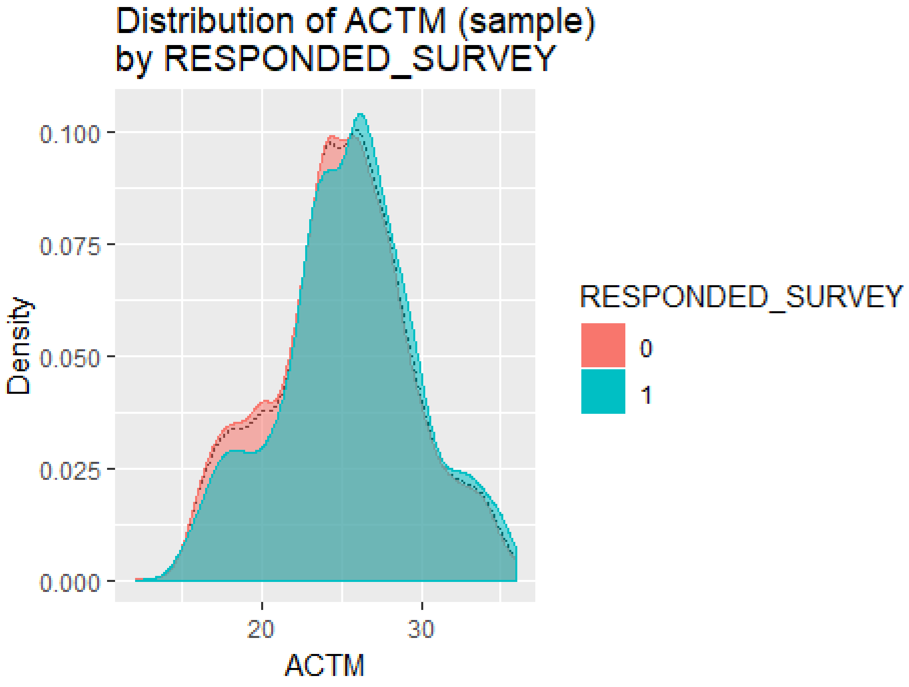
Now some distributions of certain features were analyzed against the target value. From the below plot we found that Females have a higher probability of responding to the survey than Men.



A study was done based on student’s ACT scores. I did a comparison between Science/Math related scores and English/Reading related scores. It was found that the students who got higher scores in English/Reading responded to the survey more than the ones who got higher scores in Science and Math. A similar pattern was also found when a comparison study was done with corresponding SAT scores. Since I don’t want to add all of those plots, I am only including couple of plots to show the pattern; one for English(ACTE) and other for Math(ACTM) can be seen below.

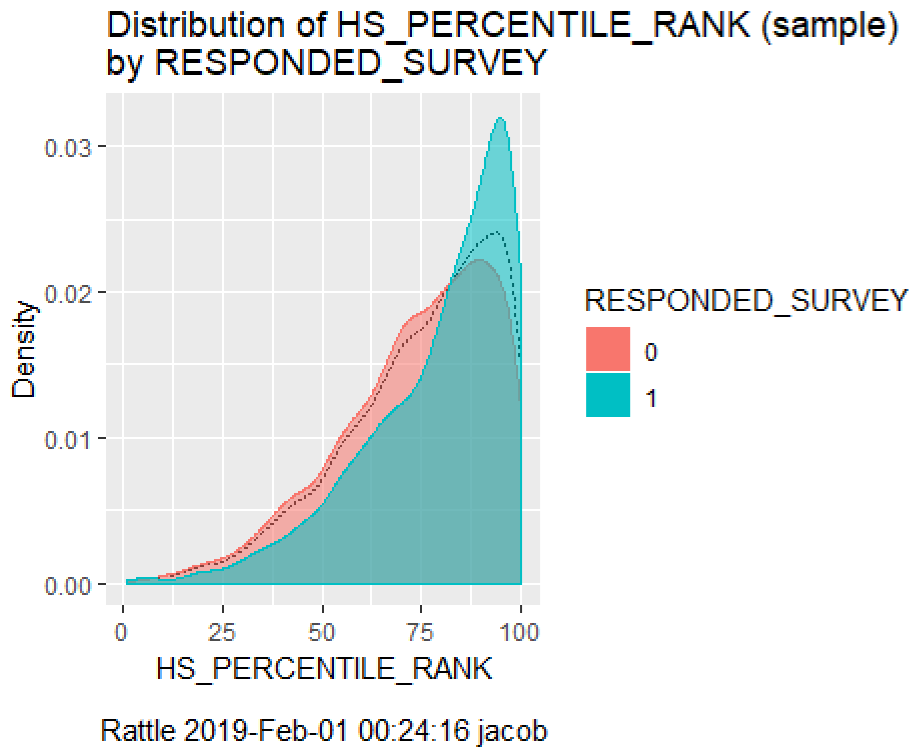

Distribution of ACTE (sample)
by RESPONDED_SURVEY

Rattle 2019-Feb-01 00:36:30 jacob


Distribution of ACTM (sample)
by RESPONDED_SURVEY

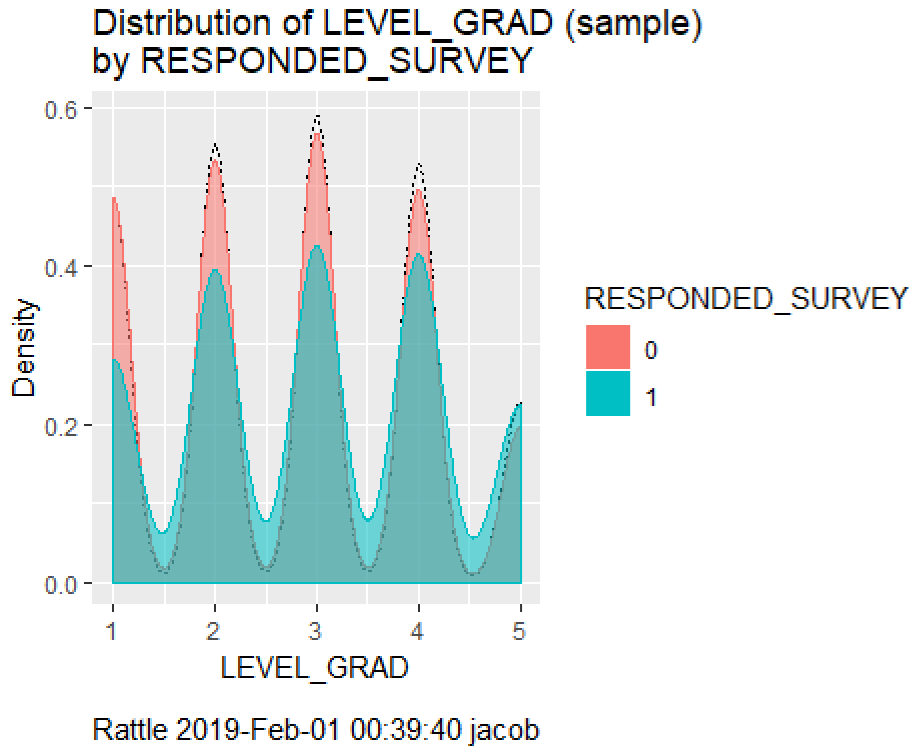
Rattle 2019-Feb-01 00:37:51 jacob

Based on the high school percentile, it looks like the students with higher percentile (70% - 100%) have a higher chance of responding to the survey.


Distribution of HS_PERCENTILE_RANK (sample)
by RESPONDED_SURVEY

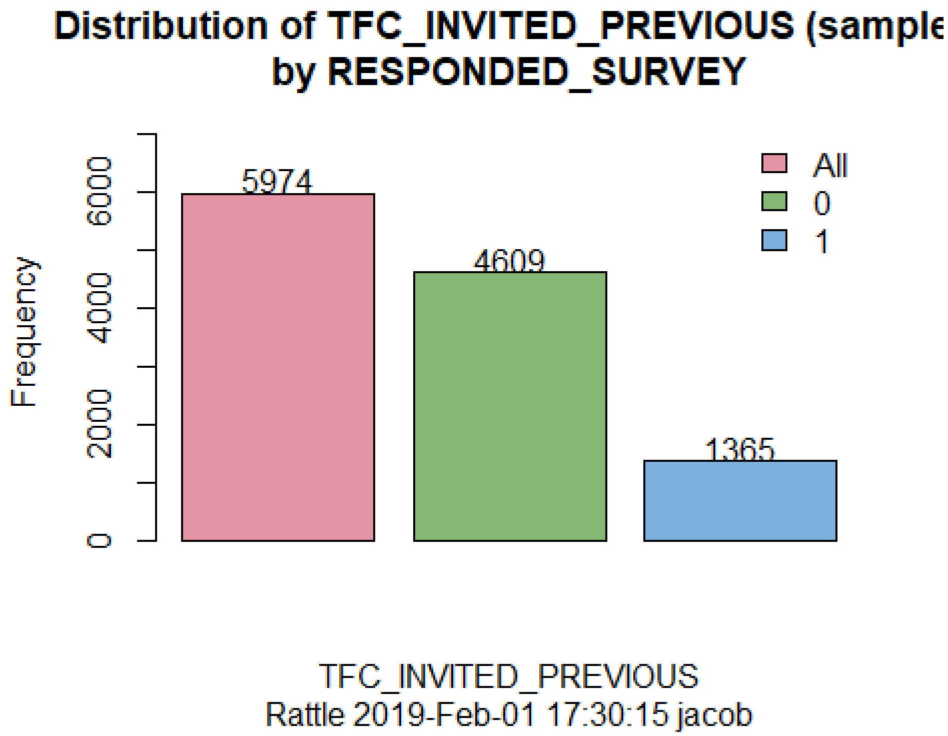
Rattle 2019-Feb-01 00:24:16 jacob

Freshman Students have a lesser chance of giving the survey compared to the students in other 3 gradulation levels. It was only 50% of the freshmen students roughly gave the survey.

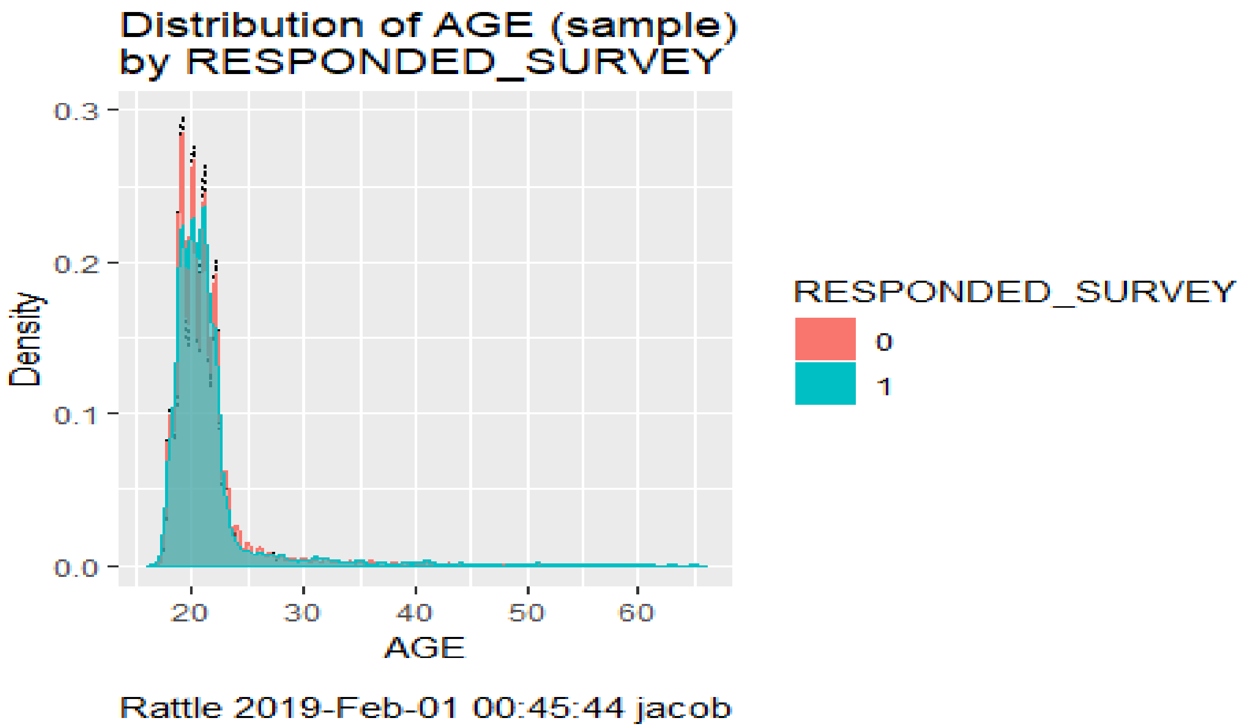

Distribution of LEVEL_GRAD (sample)
by RESPONDED_SURVEY

Rattle 2019-Feb-01 00:39:40 jacob

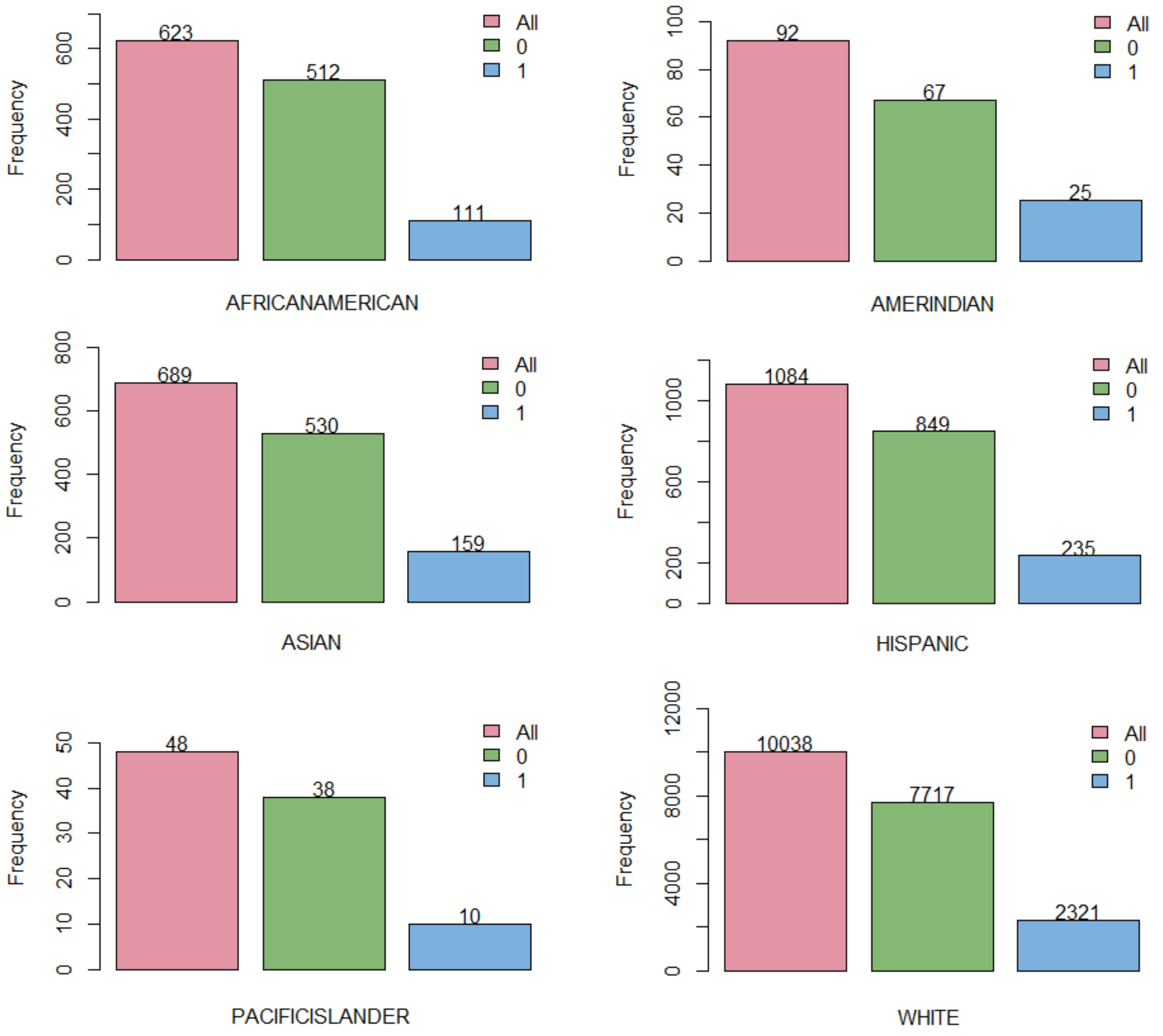
The below plot shows the distribution of students who were invited for the survey previously. Thought that I would see a better distribution since these students were invited for the survey earlier.



Older students have a higher probability of responding to the survey.



Study on Ethnicity of Students



Given below are the percentage of students in each category who responded to the survey

African American: 17.8% American Indian: 27%

Asian: 23% Hispanic: 22%

Pacific Islander: 21% White: 23%

Data Evaluation Summary

There are several attributes with low probability using which we could predict whether the student would respond to the survey or not. In my analysis the highest probability for a student to respond to the survey is with senior students and with students who do good in High School.

Thank you for your time.