Pittsburgh Public Schools Retention\Mobility Research

slash goes the other way! ;)

Huiyi Guo, Jenny Luo, Yuhang Ying
Project Advisor: Zach Branson
Department of Statistics and Data Science
Carnegie Mellon University

Abstract:

Pittsburgh Public Schools funds a Promise scholarship for post-secondary education of qualified students in their district, and it hopes to evaluate factors that influence whether students received Promise awards, and factors related to students' retention in college study. Data sets ~~of~~ this study include students' demographic information, academic performance in high school, information related to Promise scholarships, and enrollment records of post-secondary education. We employed a logistic regression model to investigate factors related to whether students received Promise scholarships, and adopted exploratory data analysis and t tests to compare students' retention in different groups. Regarding the logistic regression analysis, we have coefficients of predictors in the model considering all students and the model including only eligible students, and model diagnostics will be offered in later drafts of the research paper. Regarding retention analysis, we find that among students who started their college in Pennsylvania in 2018, their retention differs significantly between students who received Promise scholarship and those who did not; also, retention in college study differed significantly between black and white students. With respect to the discussion part, the major limitations of this study are limited sample size and possibly invalid assumption about students who are not in the scholarship data set, and translation of results into a take-home message will be provided in later drafts of the paper.

for

please say which way the difference goes in each case.

Introduction:

Pittsburgh Public Schools is a public school district for pre-K 12 students in Pittsburgh, Pennsylvania, United States. This organization funds a Promise scholarship for post-secondary education of students who are qualified for the scholarship requirements. The major requirements of being qualified for Promise scholarship are graduating from a secondary school in Pittsburgh Public Schools, having a high-school cumulative GPA greater than or equal to 2.5, having a high-school attendance rate no less than 90%, and planning to enroll in a college or university in Pennsylvania. Although the goal of Promise scholarship is helping students from Pittsburgh Public Schools finish their college study, little is known about whether this award really helps or motivates students to pursue post-secondary education. Therefore, Pittsburgh Public Schools initiated a project to examine Promise scholarship use and

(give a reference

pursue and

Actually I think the client is PPS and the SPOC is Steven.

post-secondary retention of students from Pittsburgh Public Schools. The client of this project is Steven Greene from Pittsburgh Public Schools. The two major research questions in this project are as follows:

- Investigate factors that influence whether students received Promise scholarship.
- Evaluate factors that influence students' retention and make comparisons.

clarify this.

Data:

For this project, we have 11 data sets in total. The table below shows the basic information of these 11 data sets.

can you say what years these data sets cover?

Table 1: Basic Information of 11 Data Sets

Data	Meaning of Data	Number of Observations	Number of Variables
School Enrollment	All enrollment records to and from PPS schools	6833406	14
Course Enrollment	Courses students completed during their high-school careers	60778	12
Attendance	Attendance data of students in high schools	109428	10
Demographics	Demographic information of students in each semester in high school	19039	11
NSC	Semester college enrollment records of students	5629	11
SAT	Highest SAT scores for students	3143	6
AP	AP exams and scores taken by students	5352	5
GPA	All end-of-year cumulative GPAs during students' high school careers	19436	5
Keystone	Scores that students received on the Keystone Assessment based on different subject	37331	8
CTE	Career and Technical Education(CTE) certifications earned by students in high school	1179	6
Scholarship	Information about students eligibility for Promise scholarship and receipt of Promise scholarship	2265	8

please always say what the question is, rather than "first", "second". No one can remember....

Regarding the first research question, we joined School Enrollment, Attendance, Demographics, SAT, AP, GPA, Keystone, CTE, and scholarship data together to analyze factors related to students' receipt of Promise scholarship. The total number of observations in the joined data set for research question 1 is 1708. Regarding the second research question, we joined NSC, demographics, and scholarship data together to conduct retention analysis. The total number of observations in the joined data set for research question 2 is 1378. Among them, 13 observations initiated their college study in 2017, 574 observations initiated their college study in 2018, 698 students initiated their college study in 2019, and 93 students in 2020.

can you say something about who is included in the 1708 and the 1378? Help reader understand why these numbers make sense, and why they are not the same as any of the numbers in Table 1.

make a new paragraph here

The definitions of variables we used in the project are displayed as follows. Notice that black variables refer to the ones only used in the first research question; dark-red variables refer to the ones only used in the second research question; dark-green variables refer to the ones used in both research questions.

again, here and throughout, say what the question is, rather than "first" and "second"

Table 2: Variable Definition

Variable	Definition	Data Set
RandomID	Unique student ID	
QualifiedforCorePromise	Eligibility for Promise(binary)	Scholarship
EverReceivedPromiseAward	Whether students received Promise(binary)	Scholarship
Gender	Gender of students	Demographics
Race	Race of students	Demographics
ELLStatus	English language level of students	Demographics
IEPGroup	Whether students need special education	Demographics
EconDisab	Economic status of students	Demographics
Num_AP(created)	Number of AP tests taken	AP
CumulativeGPA(created)	Cumulative GPA	GPA
AttendanceRate(created)	1-("absent unexcused"/ "total days")	Attendance
KeystoneMean(created)	Average keystone scores	Keystone

how did you decide which variables go with which question(s)?

SAT_Total(created)	Highest SAT score	SAT
Num_CTE(created)	Number of Career and Technical Education(CTE) Certifications	CTE
MagnetInd	Whether students go to magnet schools(binary)	Enrollment
GradYear	Year in which students graduated from high school	Scholarship
Enrollment_Begin	When a student enrolled in a college semester	NSC
Enrollment_End	When the college semester ended	NSC
College_State	State where the college is located	NSC
Retention(created)	Enrollment_End-Enrollment _Begin	NSC
Start_College_Year(created)	Year in which a student first enrolled in college	NSC

The exploratory data analysis on students' eligibility for Promise scholarship and receipt of Promise scholarship is as follow:

Figure 1: Students' Eligibility for Promise Scholarship in Different Racial Groups

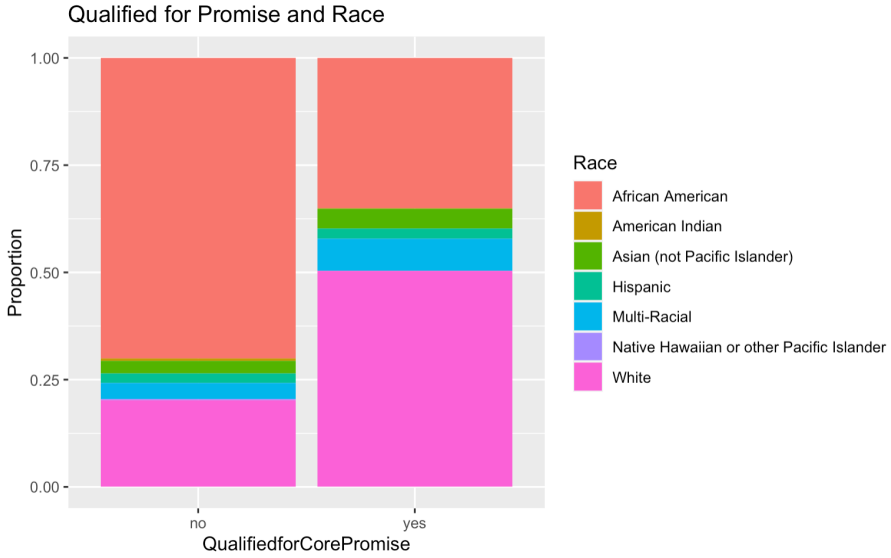
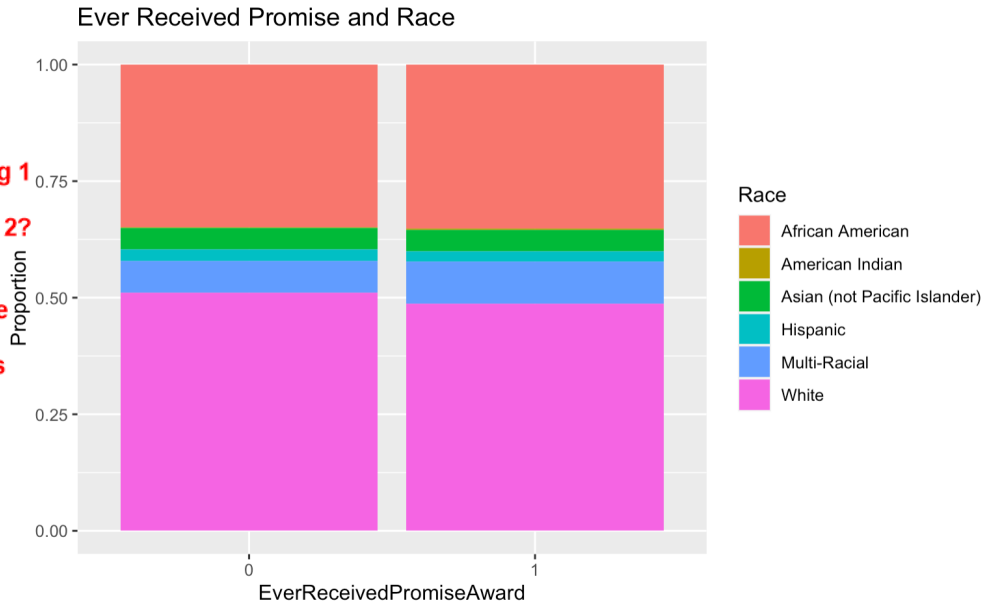


Figure 2: Students' Receipt of Promise Scholarship in Different Racial Groups



I don't understand the relationship between Fig 1 and Fig 2. What are the denominators in Figure 2?

Is it all students of the given race, or all eligible students? If it is all eligible students, why is it that not all eligible students receive the scholarship?

From figure 1, we observe that while the proportion of white students is highest among people who are eligible for Promise scholarship, the proportion of African American students is highest among people who are not eligible for Promise scholarship. From figure 2, we see that the distribution of race among students who received Promise scholarship is similar to that among students who did not receive.

captialize "Figure" when you use it with a figure number.

Figure 3: Students' Eligibility for Promise Scholarship in Different Gender Groups

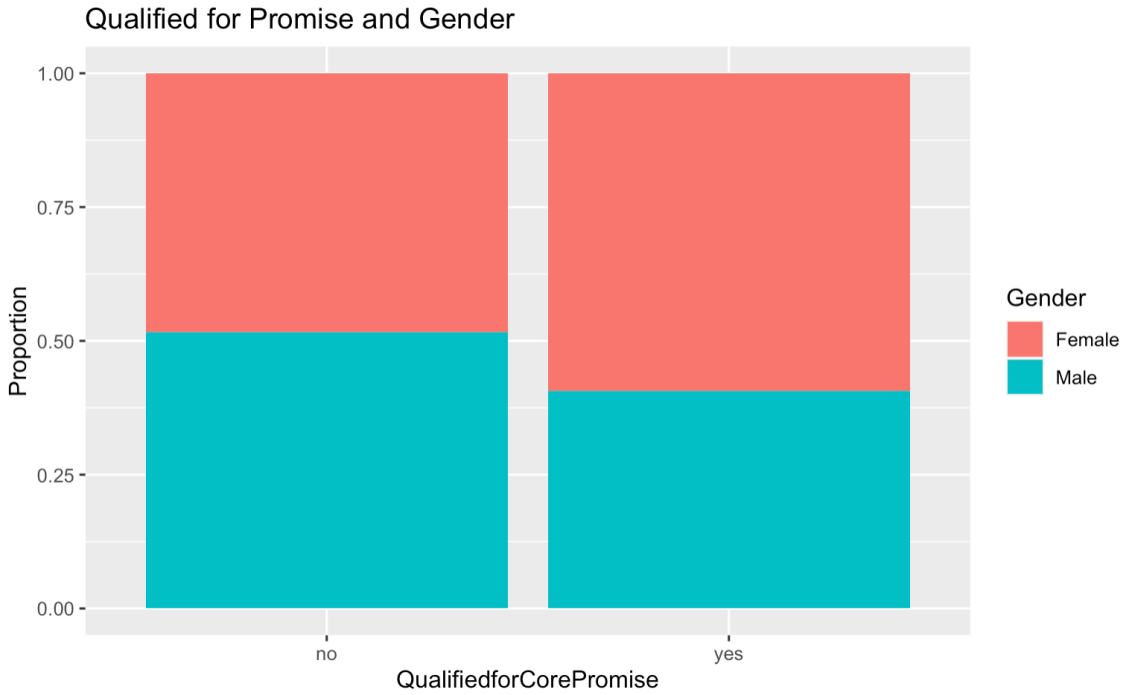
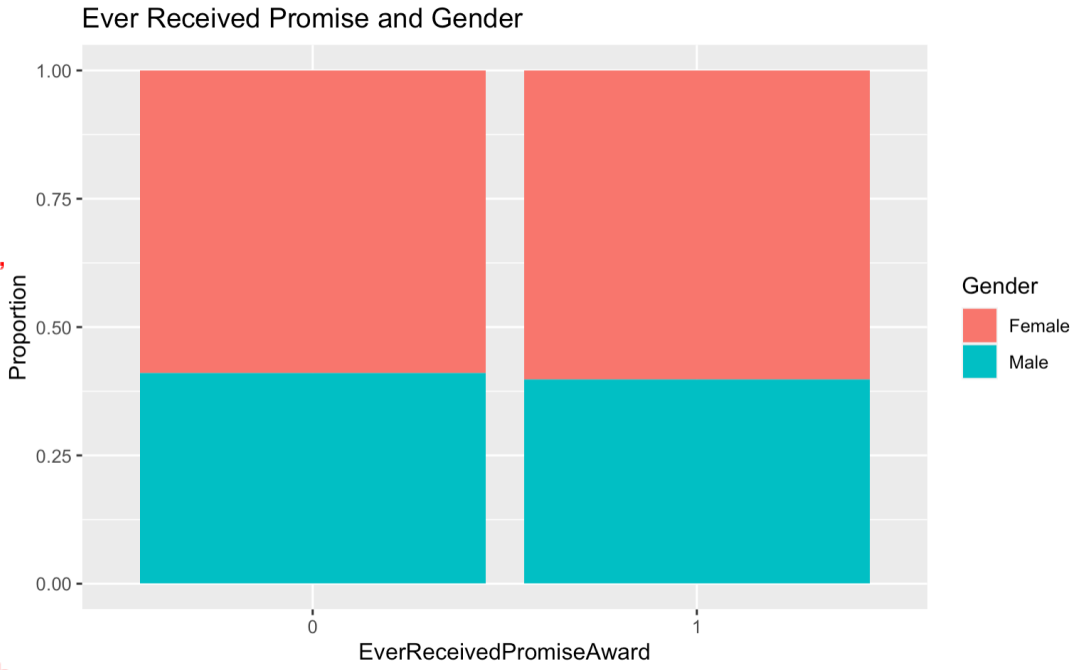


Figure 4: Students' Receipt of Promise Scholarship in Different Gender Groups



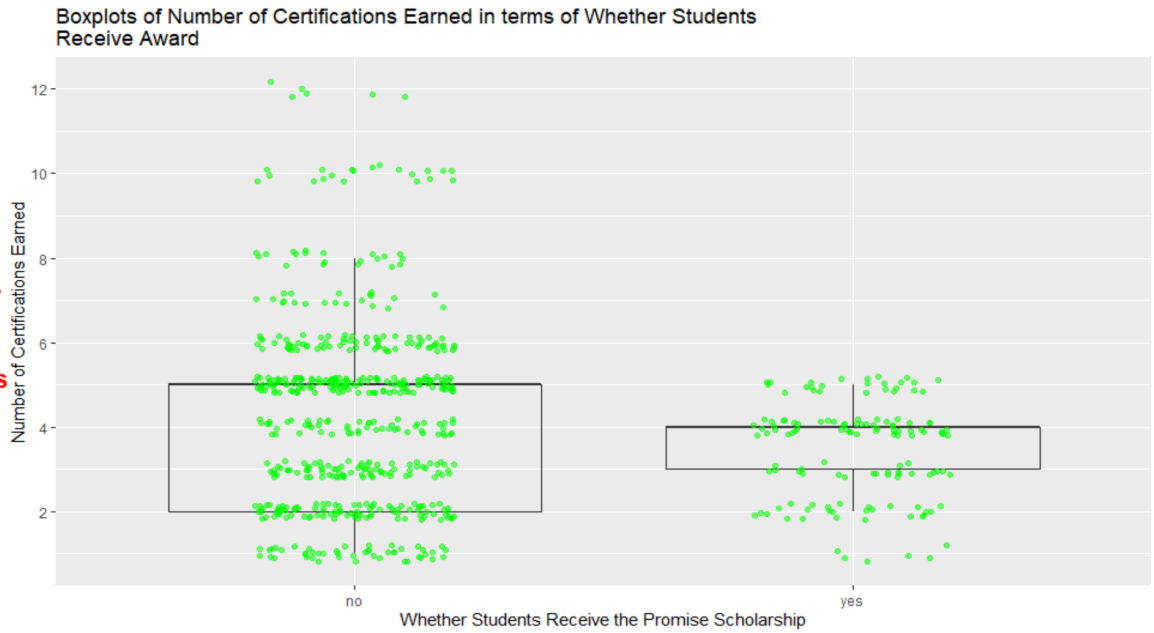
same question for the relationship between Figs 3 and 4, as for between Figures 1 and 2.

From figure 3, we find the proportion of females among students who are qualified for Promise scholarship is higher than that of males. Also, from figure 4, we learn that the proportion of females among students who received Promise scholarship is higher than that of males.

not much! Is this a big deal? How can we tell?

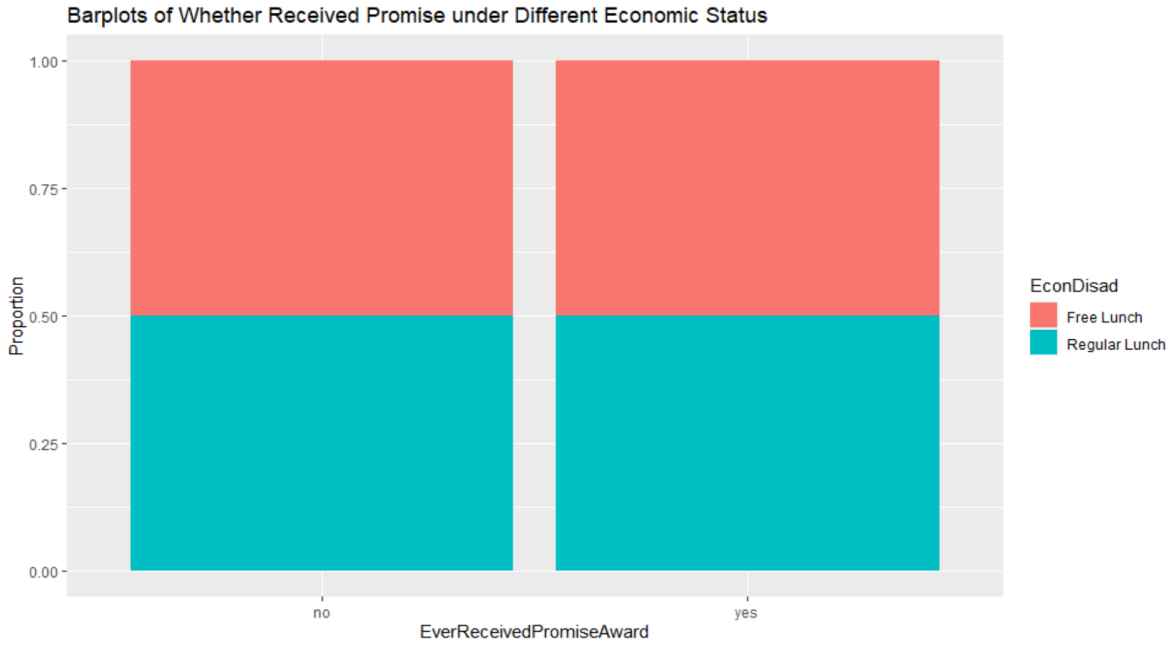
Career

Figure 5: Box Plot of Number of Certifications Earned vs. Receipt of Promise Scholarship



Explain to the reader why this plot looks different from the others

Figure 6: Students' Receipt of Promise Scholarship Under Different Economic Status



Explain why this makes sense.

From figure 5, we observe that students who received Promise scholarship on average earned less career certifications in high school than students who did not receive Promise scholarship. From figure 6, we learn that the distribution of students' economic status among students who received Promise scholarship is similar to that among students who did not receive. Thus, students' economic status might be unrelated to their likelihood of receiving a Promise scholarship.

This is really interesting, and somewhat surprising! Usually in the US we expect that higher economic status goes with higher academic achievement. What's going on here that makes it different from what we would expect?

Methods:

All analyses in this project were carried out with R programming language and environment. Also, the analyses of this project consists of two parts, one part for each research question. (insert reference)

Research Question 1: Logistic Regression Analysis

this is great; thanks for reminding reader!

To answer the first research question, we conducted two logistic regression analyses. The variables used for the two logistic regressions are the same. The response binary variable is *EverReceivedPromiseAward* in scholarship data. We used the scholarship data as base table, and left joined predictor variables *AttendanceRate, Num_AP, Num_CTE, KeystoneMean, Race, Gender, ELLStatus, IEPGroup, EconDisad, SAT_Total, CumulativeGPA, MagnetInd*. These chosen variables are the most related indicators that would affect students' enrollment in postsecondary attendance in PA based on EDA. After combining the scholarship data and predictor variables, 1708 records were left after omitting NA values.

Because the qualifications for scholarship are attendance and GPA, I'm surprised to see those variables in this model.

For the first logistic regression, we used all 1708 records which include all qualified/unqualified students for Promise. Then stepwise variable selection based on AIC is used in order to find

the relationship between the selected variables and student's enrollment in post-secondary institutions. For stepwise variable selection, the base model included *Race and Gender*. Because based on EDA, we found there were discrepancies between different races and gender on whether they received Promise scholarship. Plus, these two aspects are what our client is most concerned with. The full model included every variable we mentioned above. Then, we conducted forwards, backwards, and both-ways variable selection on AIC; all three methods gave the same results.

ok this will be interesting to see in the final paper!

However, the above analysis only gives a general understanding of how each variable is related to students' post-secondary enrollment. To understand what factors would affect students' enrollment in post-secondary institutions among those who already qualified for Promise, we will run another logistic regression on the subset of students who are marked as "yes" for binary variable *QualifiedforCorePromise*. After filtering the qualified students, 1357/1708 observations were preserved for further analysis. Then, the same procedure for the first logistic regression is performed. The base model had variables *Race* and *Gender* and the full model had all variables we mentioned before. We conducted forwards, backwards, and both-ways stepwise variable selection on AIC and all methods gave the same results.

not many unqualified students in merged data set - why is that? Can it be fixed some-how?

Research Question 2: Retention Analysis

again, thanks!

To answer the second research question, we conducted an exploratory data analysis and created box plots of students' retention in different groups. Because the Promise scholarship is designed for students who chose colleges in Pennsylvania, we restrict our analysis only students who went to colleges in Pennsylvania. A students' retention is calculated as the cumulative sum of the difference between *Enrollment_Begin* and *Enrollment_End* in the NSC data. In other words, retention refers to the total number of days a student has stayed in a college. For fair comparisons of students retention since the beginning of their colleges, we grouped students by the first year of their college enrollments.

If the data is available, a comparison of promise students with students who go to college out of state would also be very interesting.

another way to do this might be to look at "retention after 1 year", "retention after 2 years", etc., regardless of which calendar year was the students' first year of college.

In this analysis, we first compared retention between students who received Promise scholarships and who did not. To conduct the first comparison, we joined NSC and scholarship data by students' random ID, and used the variable *EverReceivedPromiseAward* to flag whether students received scholarships or not. Then, we created paired box plots to investigate whether college retention in days would differ between students who received promise scholarships and those who did not. To account for the effect that students with better academic performance might have better retention, we also compared students who were around the Promise Award qualification cutoff: GPA(2.0-3.0) and Attendance rate(0.85-0.95).

nice. almost(?) a regression discontinuity analysis.

We also compared retention among students in different racial groups. To conduct the second comparison, we joined NSC and demographics data by students' random ID. The racial information of each student is provided by the variable *Race* in the demographics data. Since the majority of students are either black or white, and we found noticeable differences in whether students are eligible for or received Promise scholarships from initial EDA, we only

this makes sense for a primary analysis.

a secondary analysis for all races might be nice, if time permits

focused on black and white students in this comparison. Similarly, we used paired box plots to observe racial differences in retention.

Finally, we investigate the interaction between receipt of scholarship and race. We redid our analysis for racial difference separately for students who received the scholarship and students who did not.

To validate our insights, we applied Welch t-tests to examine statistical significance for both retention comparisons. We also conducted Bartlett tests to first check whether variance of retention differs between groups, and chose either equal-variance t test or unequal-variance t test.

Results:

remind reader what the research question is.

Research Question 1: Logistic Regression Analysis

1. Analysis on Qualified and Unqualified Students

I would say add a column with the SE's here also.

Put the text before the table.

Table 3: Modeling Results of Analysis on Qualified and Unqualified Students

Variable	Coefficient	P-Value
(Intercept)	$\hat{\beta}_0$ = -3.627995	0.207566
RaceAmerican Indian	$\hat{\beta}_1$ = 0.448721	0.703162
RaceAsian (not Pacific Islander)	$\hat{\beta}_2$ = -0.192318	0.542354
RaceHispanic	$\hat{\beta}_3$ = -0.295090	0.454827
RaceMulti-Racial	$\hat{\beta}_4$ = 0.267703	0.243622
RaceNative Hawaiian or other Pacific Islander	$\hat{\beta}_5$ = -9.993213	0.975451
RaceWhite	$\hat{\beta}_6$ = -0.275851	0.060262
GenderMale	$\hat{\beta}_7$ = -0.092398	0.430571
CumulativeGPA	$\hat{\beta}_8$ = 0.919479	7.93e-09 ***

remind reader that the base model included race and gender since they were of interest to PPS

AttendanceRate	$\hat{\beta}_9$= 8.202987	5.79e-05 ***
KeystoneMean	$\hat{\beta}_{10}$= -0.005981	0.000389 ***
ELLStatusNot in ELL	$\hat{\beta}_{11}$= 0.952589	0.035397 *
MagnetInd1	$\hat{\beta}_{12}$= 0.232204	0.046337 *

I question whether this is meaningful, since high GPA and high attendance are required to get the scholarship.

Or is there another path to the scholarship that I am not thinking of?

using AIC as a model fit criterion

The table shows the result after we performed stepwise variable selection on AIC. By observing the coefficients, we find that students who have a high GPA and Attendance rate will be more likely to receive the Promise, which means they successfully enrolled in a PA college. Also, students who are not in ELL group and students who ever attended a magnet school have a higher rate in receiving Promise scholarship. But one odd thing is the keystone score showed a negative relationship with students' enrollment in PA college. This is very different from our intuition that students who have higher keystone scores means they have better academic performance at school; in this case, they will be more likely to successfully enroll in a college. So we are going to do more exploration on this, that is one of our next-steps. Also, race and gender here are not significant. Thus we consider conducting ANOVA tests to do further variable selection.

This is why it's also useful to look at students who went to college out of state. I suspect that higher-achieving students did go to college, just out of state and not in PA.

It's also worth noting that the effect is not very large.

2. Analysis on Qualified Students only

Again, text should precede table.

column of SEs

Table 4: Modeling Results of Analysis on Qualified Students Only

Variable	Coefficient	P-Value
(Intercept)	$\hat{\beta}_0$= -4.824993	0.14902
RaceAmerican Indian	$\hat{\beta}_1$= 1.282735	0.38592
RaceAsian (not Pacific Islander)	$\hat{\beta}_2$= -0.247677	0.41762
RaceHispanic	$\hat{\beta}_3$= -0.267055	0.50568
RaceMulti-Racial	$\hat{\beta}_4$= 0.263061	0.27133
RaceWhite	$\hat{\beta}_5$= -0.222510	0.15149

remind reader that the base model included race and gender since they were of interest to PPS

GenderMale	$\hat{\beta}_6$= -0.025427	0.83676
AttendanceRate	$\hat{\beta}_7$= 10.556032	4.54e-05 ***
KeystoneMean	$\hat{\beta}_8$= -0.005168	0.00364 **
CumulativeGPA	$\hat{\beta}_9$= 0.504048	0.01062 *
MagnetInd1	$\hat{\beta}_{10}$= 0.224038	0.06890 .

using AIC as the model fit criterion

The table shows the result after we performed stepwise variable selection on AIC. By observing the coefficients, we find that students who have a high GPA will be more likely to receive Promise scholarship. Also, students who have a high GPA also add likelihood to PA college enrollment. However, different from the results of the first logistic mode, GPA becomes less important affecting PA college enrollment for students who qualified for Promise. Same as GPA, the MagnetInd becomes less important affecting student's enrollment in PA college, but they still have a positive relationship. Like the first model, ANOVA tests are needed to do further variable selection since the variable *Race* and *Gender* are not significant.

model

Similar to

"they" or "it"?

will these come in the final paper?

Research Question 2: Retention Analysis

We compare students retention from two perspectives: whether a student received the scholarship and the student's race. We only conduct statistical tests for year 2018 and 2019, due to limited observations for year 2017, and insufficient time lag for year 2020. For racial comparison, we mainly focus on black and white, as they constitute the majority of students. Our results suggest interaction between racial difference and whether a student received the scholarship.

give the sample sizes for each year

explain this to the reader

1. Retention between students who received the scholarship and who did not.

text before figure

Figure 7: Students' Retention by the Receipt of Scholarship

There's a lot of right skewing in this data;

did you consider either
(a) a log or other transformation; or
(b) a poisson regression or gamma regression (both are kinds of glm's)?

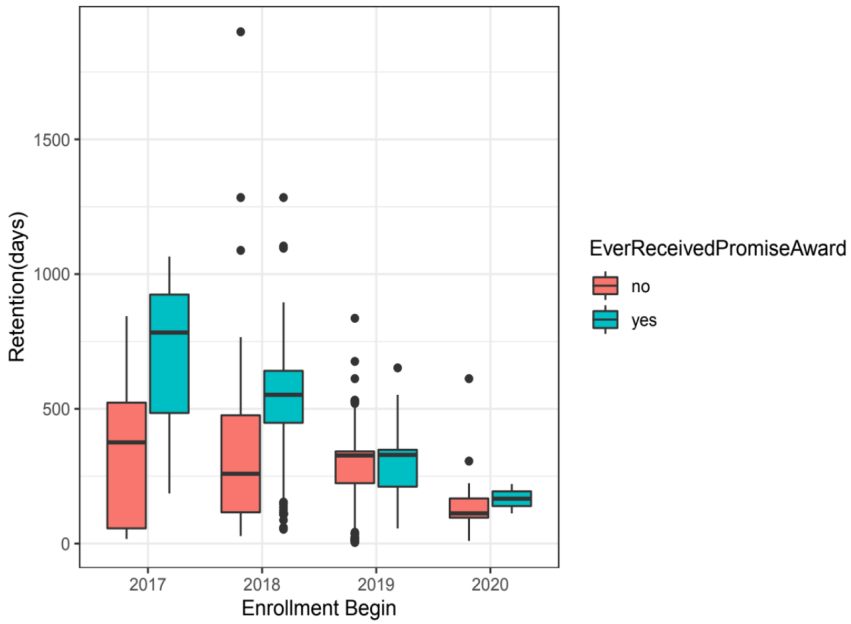
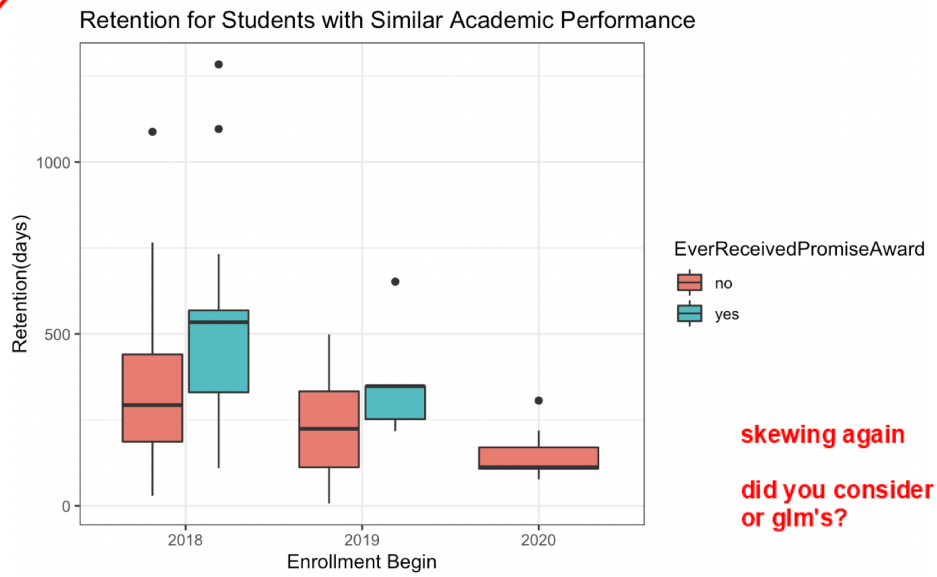


Figure 7 shows that in general, students who received the scholarship have higher retention on average except for year 2019. After checking the equal variance assumption with the Bartlett test, the Welch t test shows that the difference is significant for year 2018(p-value = 5.98e-10) but not for 2019(p-value = 0.60). We conclude for the year 2018, students who received the scholarship have better retention.

is this another case of insufficient time lag?

Figure 8: Retention by the Receipt of Scholarship

text before figure



skewing again

did you consider transformations or glm's?

show table of coefficients?

Our logistic regression indicates that students with higher GPA and attendance rate are more likely to get Promise Award. To account for the effect that the retention difference between students who received the scholarship and who didn't may come from different academic

capabilities, we restrict our analysis on the range of students with GPA of 2.0-3.0 and attendance rate of 0.85-0.95. Figure 8 shows that among students with similar academic performance, and our results show that the difference is significant for both 2018(p-value = 0.01) and 2019(p-value = 0.02).

2. Retention between different students' races.

text before figure

Figure 9: Students' Retention by Races

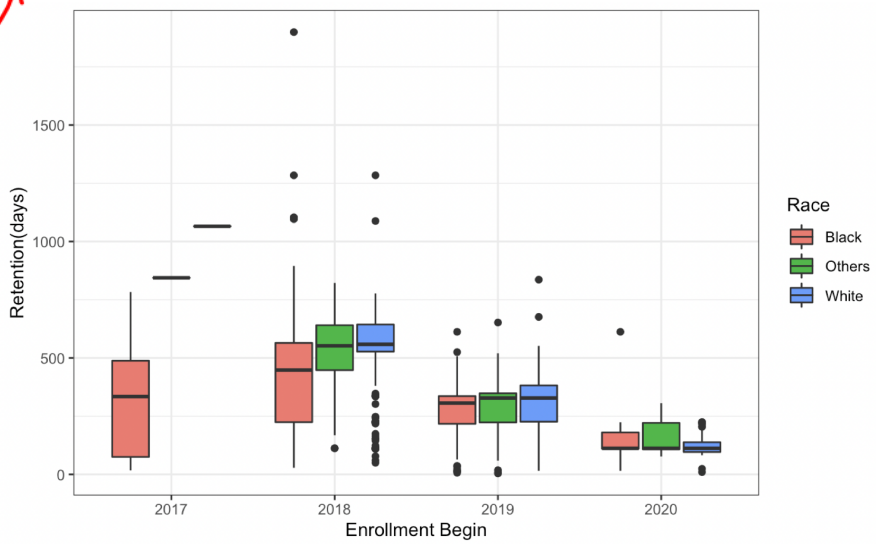


Figure 9 shows a comparison of retention between different races. The "others" includes Multi-Racial, Asian, Hispanic, American Indian, and Native Hawaiian or other Pacific Islander. We group these races together because they only constitute 12.9% of the observations. From the plot we observe a big retention difference between races for year 2018, and slight difference for 2019. Our one way anova test shows that the difference for both 2018 and 2019 are significant. We conclude that for both 2018 and 2019, the mean retention between races are not equal.

Next, we further investigate the differences between black and white students. After checking the equal variance assumption with the Bartlett test, the Welch t test shows that the difference is significant for both 2018(p-value = 8.48e-06) and 2019(p-value = 8.78e-08).

3. Interaction between receipt of scholarship and race.

text before figure

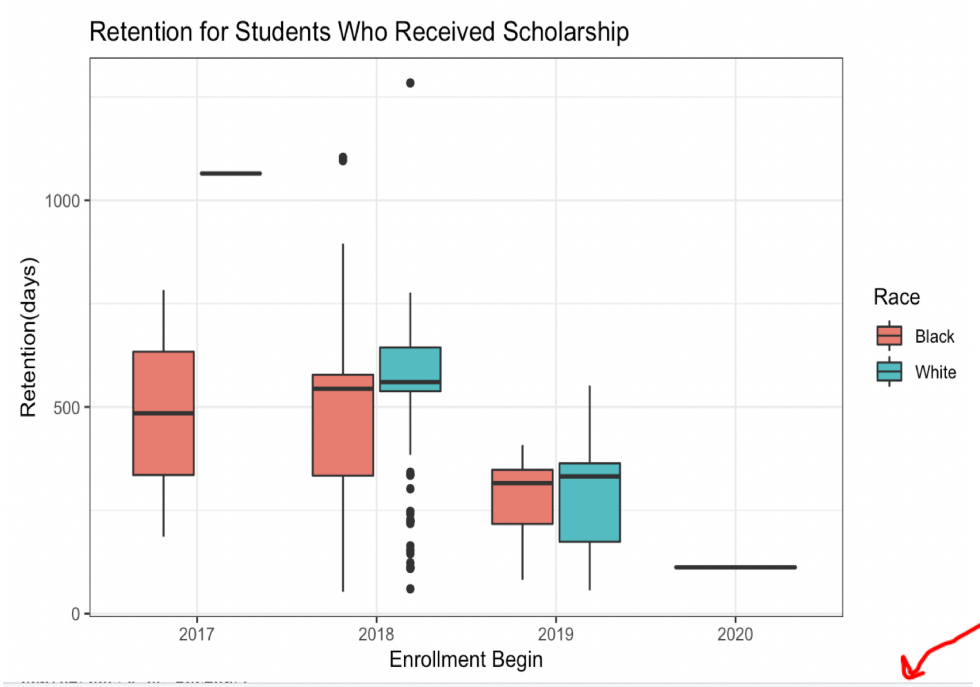
Table 5: P-values for Racial Difference in Retention

Show full table of coef estimates, SE's and pvalues

Group	2018	2019
Received Scholarship	3.4e-04	0.51
NOT Receive Scholarship	0.54	8.63e-08

Text before figure

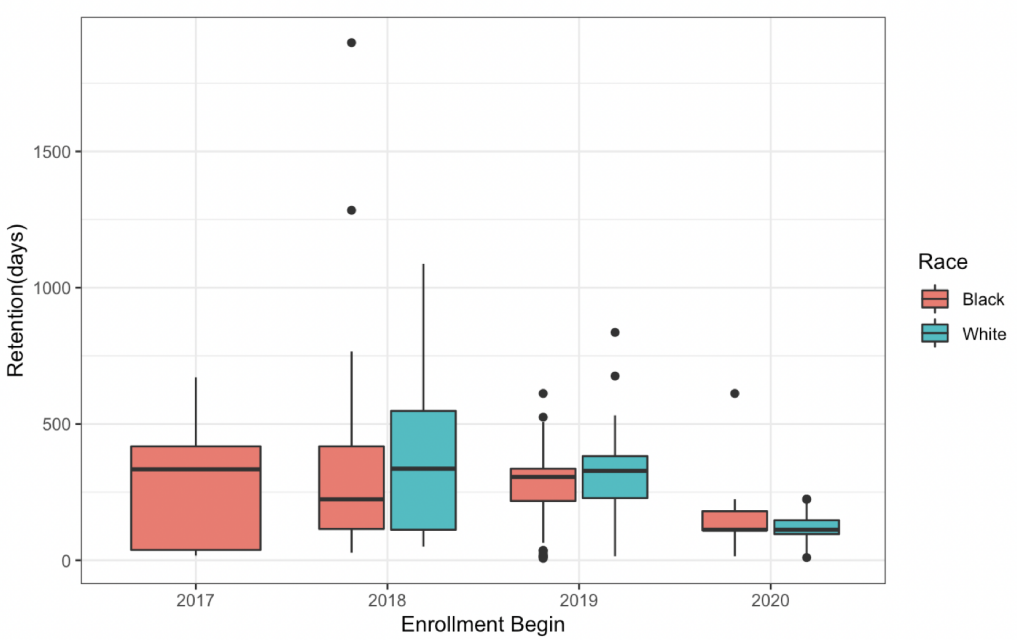
Figure 10: Racial difference in Retention and Receipt of Scholarship



remove line

text before figure

Figure 11: Racial difference in Retention and Receipt of Scholarship



Given we found some significant difference for receipt of scholarship and race, we will investigate their interactions by accessing the racial difference separately for students who received the scholarship and students who didn't. Figure 10 shows the comparison of racial

differences for students who received the scholarship. Our analysis shows that for 2018, the difference is significant, but not for 2019. For students who didn't receive the scholarship in Figure 11, our analysis shows that for 2018, the difference is not significant, but the difference is significant for 2019. The relevant p-values are shown in Table 5.

Discussion:

good start!

Looking forward to a complete discussion in final paper.

1. Although we don't see a significant difference in 2019, one interesting finding is that the gap between students' retention enlarges for more senior students. We suspect that the difference may become more apparent for junior and senior students. To understand better what other factors also affect retention, we will conduct a multivariate linear regression with the interaction between race and receipt of scholarship.
2. Limitations: First, for both research questions, we have a limited number of observations (e.g. 1708 for the first research question and 1378 for the second research question), and this limitation might affect the generalizability of our findings. Second, we assume that students who are not included in the scholarship data are students who did not receive Promise scholarships in our analyses. Nevertheless, this assumption might not be the reality, and the credibility of results we obtained from this project might be affected by this possibly invalid assumption.
3. Translation of results into a take-home policy/message for the client

No references??

Technical Appendix:

Logistic Regression Analysis

Good start

Don't forget to put english before and after each thing you've done, so reader undestands what, why, and what the results are

Load library

```
library(tidyverse)
library(dplyr)
library(visdat)
library(stats)
library(ggplot2)
```

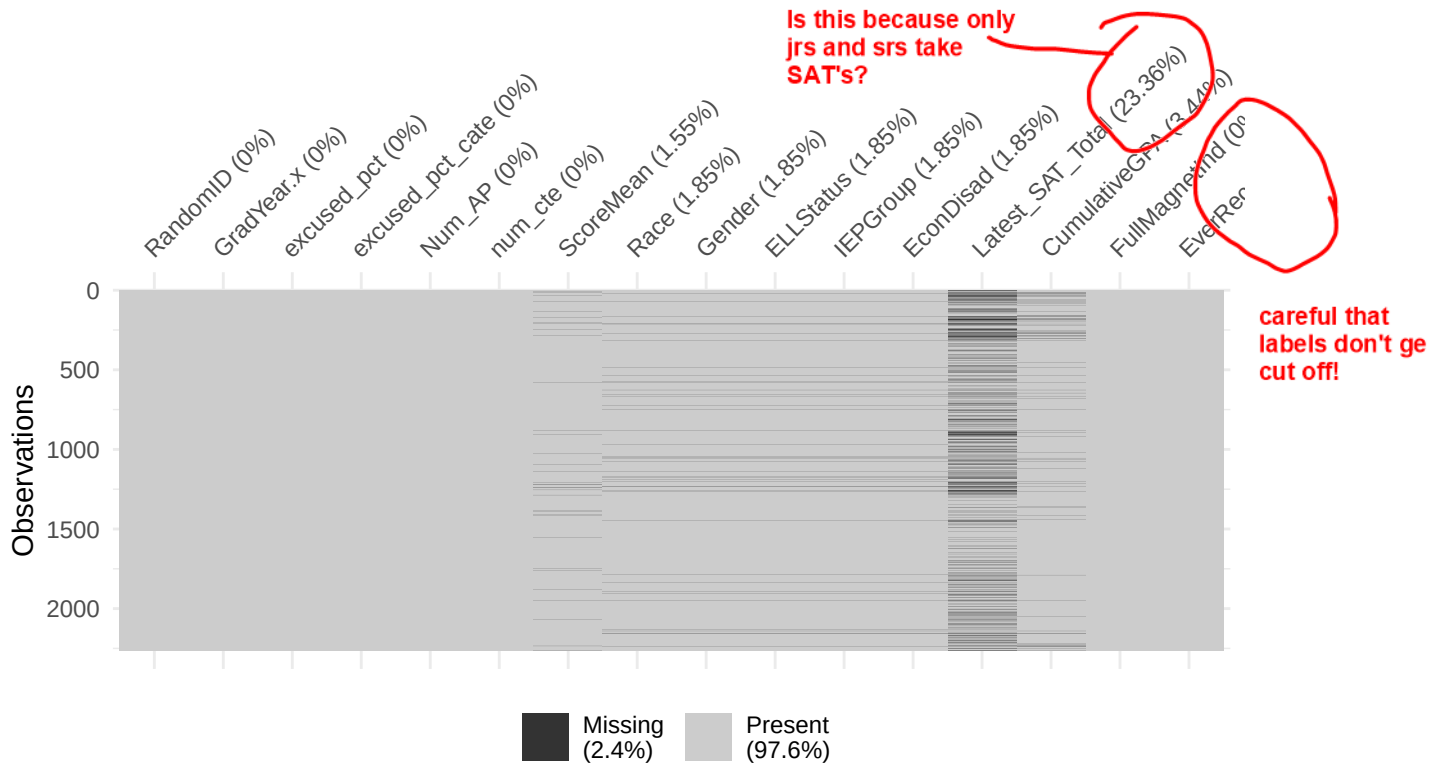
Load data

```
ap_model <- read.csv("ap_scholarship.csv")
attendance_model <- read.csv("attendance_rate.csv")
cte_model <- read.csv("CTE_scholarship.csv")
keystone_model <- read.csv("keystone_wide.csv")
nsc_model <- read.csv("nsc_scholarship_final.csv")
demographics_model <- read.csv("demo_scholarship_final.csv")
sat_model <- read.csv("sat_clean.csv")
gpa_model <- read.csv("senior.csv")
magnet_model <- read.csv("MagnetSchool_student.csv")
senior_gpa <- read.csv("senior_gpa.csv")
```

Data cleaning/reformatting

```
# Left join all datasets
data_join <- attendance_model %>%
  left_join(ap_model, by = c("RandomID"="RandomID")) %>%
  left_join(cte_model, by = c("RandomID"="RandomID")) %>%
  left_join(keystone_model, by = c("RandomID"="RandomID")) %>%
  left_join(demographics_model, by = c("RandomID"="RandomID")) %>%
  left_join(sat_model, by = c("RandomID"="RandomID")) %>%
  left_join(senior_gpa, by = c("RandomID"="RandomID")) %>%
  left_join(magnet_model, by = c("RandomID"="RandomID"))

# colnames(data_join)

# Select variables by column names
# Delete duplicated rows
data_variables <- data_join %>%
  select(RandomID, GradYear.x, excused_pct, excused_pct_cate, Num_AP, num_cte,
         ScoreMean, Race, Gender, ELLStatus, IEPGroup, EconDisad, Latest_SAT_Total,
```

```
        CumulativeGPA, FullMagnetInd, EverReceivedPromiseAward.x) %>%
  distinct()

# Deal with NAs
data_na <- data_variables %>%
  replace_na(list(Num_AP=0, num_cte=0, FullMagnetInd=0 )) # Replace num_AP, num_cte with 0

# Visualize NAs
vis_miss(data_na)  # How to deal with missing values in SAT score?
```




```
# Omit all NAs
data_omit <- data_na %>%
  na.omit()

# Change response variable to 0,1
data_omit$EverReceivedPromiseAward.x <- as.numeric(as.factor(data_omit$EverReceivedPromiseAward.x))-1


# summary(data_omit)

# Change variables into appropriate format
data_changeFormat <- data_omit %>%
  mutate_at(c(2, 4,8:12,15:16), as.factor)

# Rename columns
```

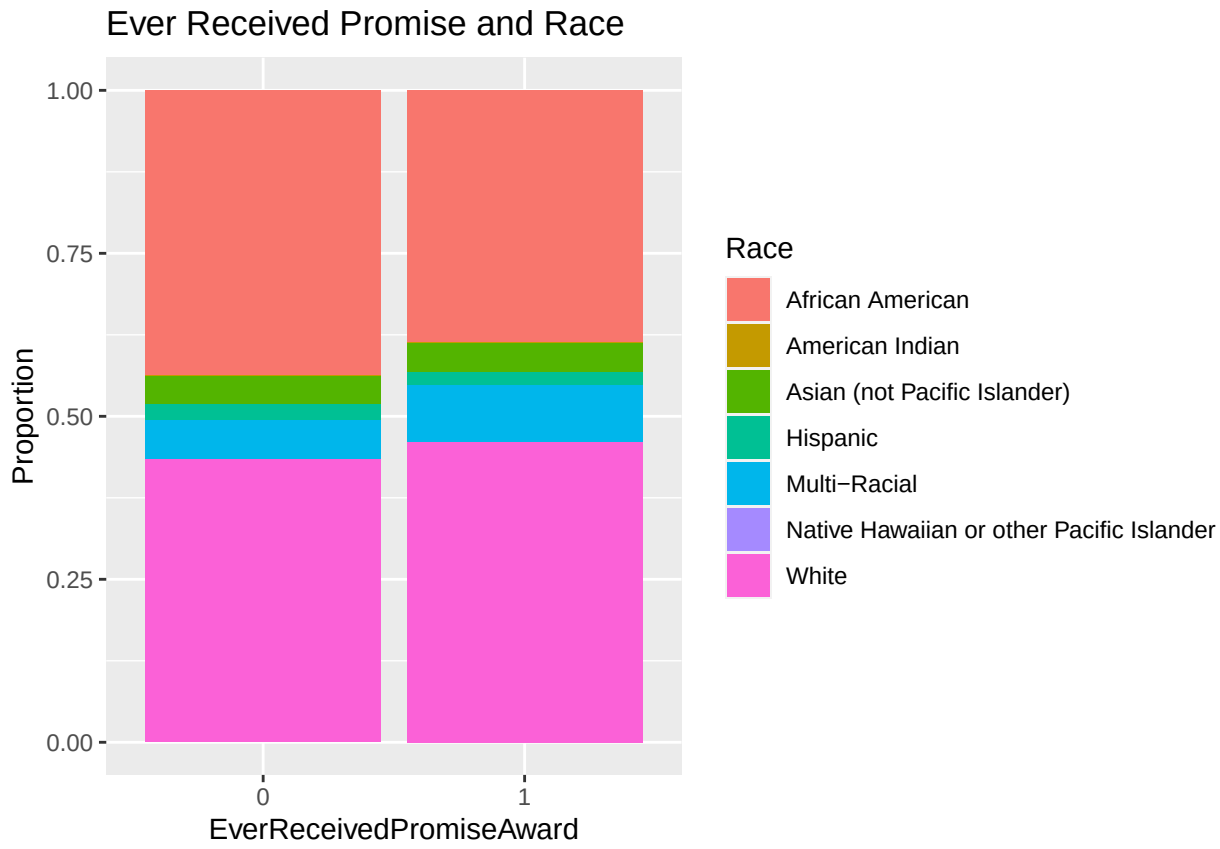
```r
colnames(data_changeFormat)<- c("RandomID","GradYear","AttendanceRate", "AttendaceRateCate",
                                "Num_AP", "Num_CTE", "KeystoneMean", "Race", "Gender",
                                "ELLStatus", "IEPGroup", "EconDisad", "SAT_Total",
                                "CumulativeGPA","MagnetInd", "EverReceivedPromiseAward")

summary(data_changeFormat)
```

```
##     RandomID       GradYear   AttendanceRate   AttendaceRateCate
##  Min.   :5841218   2018:735   Min.   :0.6574   <90% : 178
##  1st Qu.:6024748   2019:973   1st Qu.:0.9458   >=90%:1530
##  Median :6067092              Median :0.9750
##  Mean   :6122516              Mean   :0.9592
##  3rd Qu.:6130526              3rd Qu.:0.9902
##  Max.   :6832548              Max.   :1.0000
##
##      Num_AP          Num_CTE        KeystoneMean
##  Min.   : 0.000   Min.   :0.0000   Min.   :1375
##  1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.:1459
##  Median : 0.000   Median :0.0000   Median :1494
##  Mean   : 1.946   Mean   :0.2693   Mean   :1499
##  3rd Qu.: 3.000   3rd Qu.:0.0000   3rd Qu.:1537
##  Max.   :14.000   Max.   :7.0000   Max.   :1736
##
##                                        Race        Gender         ELLStatus
##  African American                        :721   Female:975   ELL       :  53
##  American Indian                         :  4   Male  :733   Not in ELL:1655
##  Asian (not Pacific Islander)            : 73
##  Hispanic                                : 40
##  Multi-Racial                            :115
##  Native Hawaiian or other Pacific Islander:  1
##  White                                   :754
##               IEPGroup            EconDisad     SAT_Total    CumulativeGPA
##  Gifted           : 351   Free Lunch   :880   Min.   : 490   Min.   :1.129
##  IEP              : 132   Regular Lunch:828   1st Qu.: 850   1st Qu.:2.593
##  Not IEP or Gifted:1225                       Median : 990   Median :3.110
##                                               Mean   :1021   Mean   :3.059
##                                               3rd Qu.:1170   3rd Qu.:3.597
##                                               Max.   :1600   Max.   :4.000
##
##  MagnetInd EverReceivedPromiseAward
##  0:825     0:1222
##  1:883     1: 486
##
##
##
##
##
```

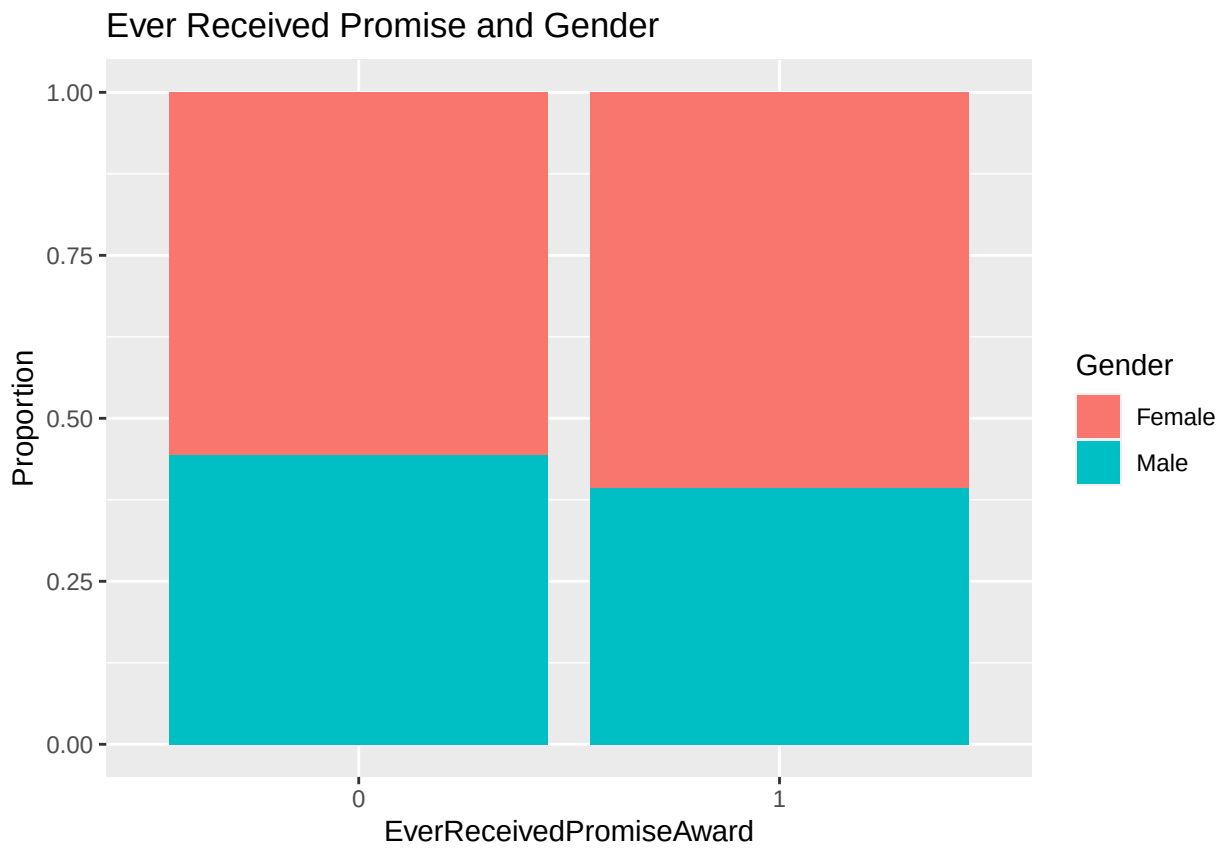
```r
data_everReceived <- data_changeFormat
#write_csv(as.data.frame(data_everReceived), "/Users/gloriaguo/Desktop/PPS/PPS_data_Processed/EverRecei

#head(data_changeFormat)
```

EDA on full Scholarship data: EverReceived

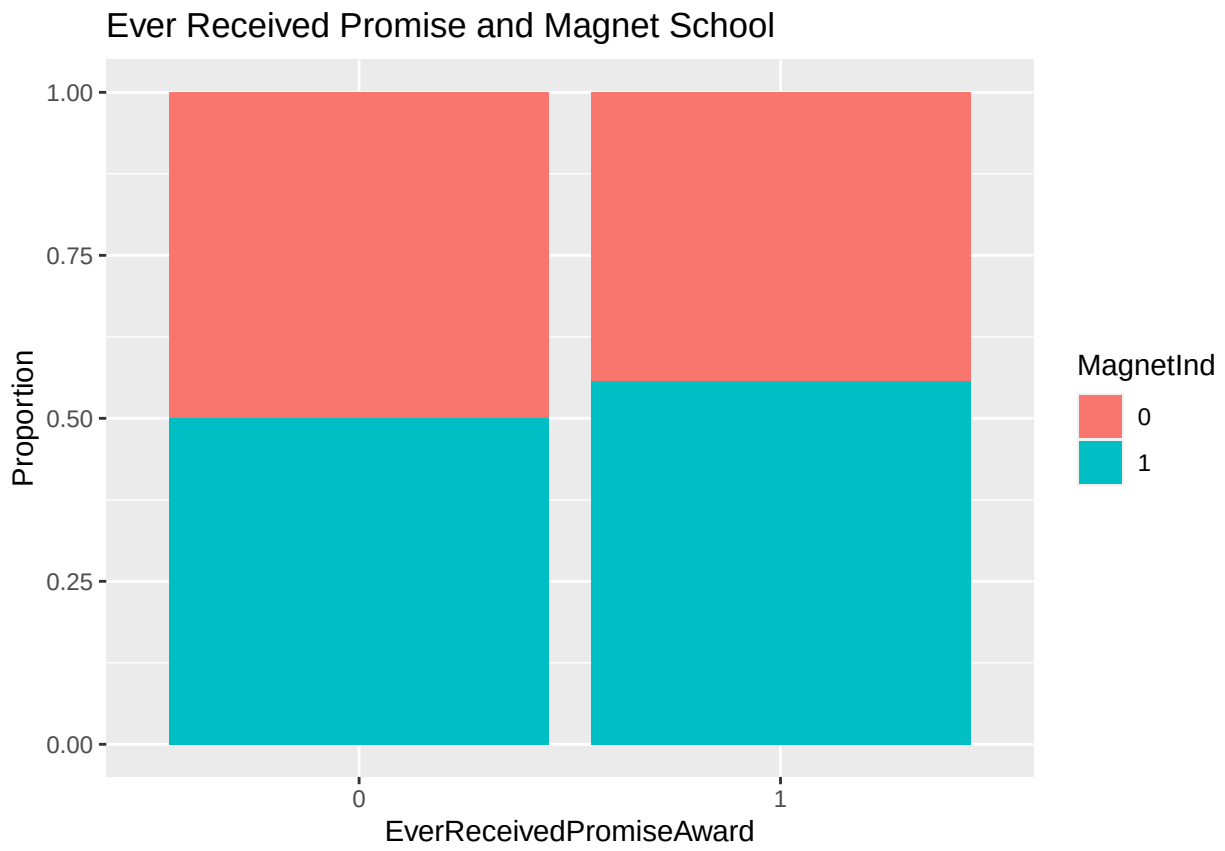
```
## Race
ggplot(data_everReceived, aes(fill=Race, x=EverReceivedPromiseAward)) +
  geom_bar(position="fill") +
  ggtitle("Ever Received Promise and Race") +
  labs(y="Proportion")
```




```
## Gender
ggplot(data_everReceived, aes(fill=Gender, x=EverReceivedPromiseAward)) +
  geom_bar(position="fill") +
  ggtitle("Ever Received Promise and Gender") +
  labs(y="Proportion")
```




```
## Magnet School
ggplot(data_everReceived, aes(fill=MagnetInd, x=EverReceivedPromiseAward)) +
  geom_bar(position="fill") +
  ggtitle("Ever Received Promise and Magnet School") +
  labs(y="Proportion")
```



Build logistic model on full Scholarship data

Select null model as base model

```
model_null <- glm(EverReceivedPromiseAward ~ 1, data = data_everReceived, family = "binomial")
model_full <- glm(EverReceivedPromiseAward ~ AttendanceRate+Num_AP+Num_CTE+KeystoneMean+Race
              +Gender+ELLStatus+IEPGroup+EconDisad+SAT_Total+CumulativeGPA+MagnetInd,
              data = data_everReceived, family = "binomial")
```

```
# Backwards selection on AIC
backwards <- step(model_full, trace = 0)
#summary(backwards)
```

```
forwards <- step(model_null, scope=list(lower=formula(model_null), upper=formula(model_full)),
                 direction="forward", trace = 0)
# summary(forwards)
```

```
bothways <- step(model_null, list(lower=formula(model_null), upper=formula(model_full)),
                 direction="both", trace=0)
summary(bothways)
```

```
##
```

```
## Call:
## glm(formula = EverReceivedPromiseAward ~ CumulativeGPA + AttendanceRate +
##     KeystoneMean + ELLStatus + MagnetInd, family = "binomial",
##     data = data_everReceived)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3566  -0.8912  -0.6618   1.2997   2.3777
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -1.872429   2.670453  -0.701 0.483200
## CumulativeGPA        0.896631   0.151639   5.913 3.36e-09 ***
## AttendanceRate       7.360586   1.963228   3.749 0.000177 ***
## KeystoneMean        -0.006718   0.001589  -4.227 2.36e-05 ***
## ELLStatusNot in ELL  1.004295   0.427269   2.350 0.018748 *
## MagnetInd1           0.285455   0.113404   2.517 0.011831 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2040.0  on 1707  degrees of freedom
## Residual deviance: 1925.5  on 1702  degrees of freedom
## AIC: 1937.5
##
## Number of Fisher Scoring iterations: 4
```

```
formula(backwards)
```

```
## EverReceivedPromiseAward ~ AttendanceRate + KeystoneMean + ELLStatus +
##     CumulativeGPA + MagnetInd
```

```
formula(forwards)
```

```
## EverReceivedPromiseAward ~ CumulativeGPA + AttendanceRate + KeystoneMean +
##     ELLStatus + MagnetInd
```

```
formula(bothways)
```

```
## EverReceivedPromiseAward ~ CumulativeGPA + AttendanceRate + KeystoneMean +
##     ELLStatus + MagnetInd
```

Select Race+Gender as base model

```
model_null <- glm(EverReceivedPromiseAward ~ Race+Gender, data = data_changeFormat,
                  family = "binomial")
model_full <- glm(EverReceivedPromiseAward ~ AttendanceRate+Num_AP+Num_CTE+KeystoneMean+Race
              +Gender+ELLStatus+IEPGroup+EconDisad+SAT_Total+CumulativeGPA+MagnetInd,
              data = data_changeFormat, family = "binomial")
```

```r
# Backwards selection on AIC
backwards <- step(model_full, scope=list(lower=formula(model_null),upper=formula(model_full)),
                  direction="backward", trace = 0)

# backwards <- step(model_full, trace = 0)
# summary(backwards)
```

```r
forwards <- step(model_null, scope=list(lower=formula(model_null), upper=formula(model_full)),
                 direction="forward", trace = 0)
```

```r
bothways <- step(model_null, list(lower=formula(model_null), upper=formula(model_full)),
                 direction="both", trace=0)
summary(bothways)
```

```
##
## Call:
## glm(formula = EverReceivedPromiseAward ~ Race + Gender + CumulativeGPA +
##     AttendanceRate + KeystoneMean + ELLStatus + MagnetInd, family = "binomial",
##     data = data_changeFormat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2949  -0.8875  -0.6620   1.2866   2.4004
##
## Coefficients:
##                                              Estimate Std. Error z value
## (Intercept)                                 -3.627995   2.878706  -1.260
## RaceAmerican Indian                          0.448721   1.177575   0.381
## RaceAsian (not Pacific Islander)            -0.192318   0.315659  -0.609
## RaceHispanic                                -0.295090   0.394827  -0.747
## RaceMulti-Racial                             0.267703   0.229595   1.166
## RaceNative Hawaiian or other Pacific Islander -9.993213 324.743830  -0.031
## RaceWhite                                   -0.275851   0.146817  -1.879
## GenderMale                                  -0.092398   0.117224  -0.788
## CumulativeGPA                                0.919479   0.159359   5.770
## AttendanceRate                               8.202987   2.040025   4.021
## KeystoneMean                                -0.005981   0.001686  -3.547
## ELLStatusNot in ELL                          0.952589   0.452798   2.104
## MagnetInd1                                   0.232204   0.116550   1.992
##                                             Pr(>|z|)
## (Intercept)                                 0.207566
## RaceAmerican Indian                         0.703162
## RaceAsian (not Pacific Islander)            0.542354
## RaceHispanic                                0.454827
## RaceMulti-Racial                            0.243622
## RaceNative Hawaiian or other Pacific Islander 0.975451
## RaceWhite                                   0.060262 .
## GenderMale                                  0.430571
## CumulativeGPA                               7.93e-09 ***
## AttendanceRate                              5.79e-05 ***
## KeystoneMean                                0.000389 ***
## ELLStatusNot in ELL                         0.035397 *
## MagnetInd1                                  0.046337 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2040.0  on 1707  degrees of freedom
## Residual deviance: 1916.5  on 1695  degrees of freedom
## AIC: 1942.5
##
## Number of Fisher Scoring iterations: 11
```

```
formula(backwards)
```

```
## EverReceivedPromiseAward ~ AttendanceRate + KeystoneMean + Race +
##     Gender + ELLStatus + CumulativeGPA + MagnetInd
```

```
formula(forwards)
```

```
## EverReceivedPromiseAward ~ Race + Gender + CumulativeGPA + AttendanceRate +
##     KeystoneMean + ELLStatus + MagnetInd
```

```
formula(bothways)
```

```
## EverReceivedPromiseAward ~ Race + Gender + CumulativeGPA + AttendanceRate +
##     KeystoneMean + ELLStatus + MagnetInd
```

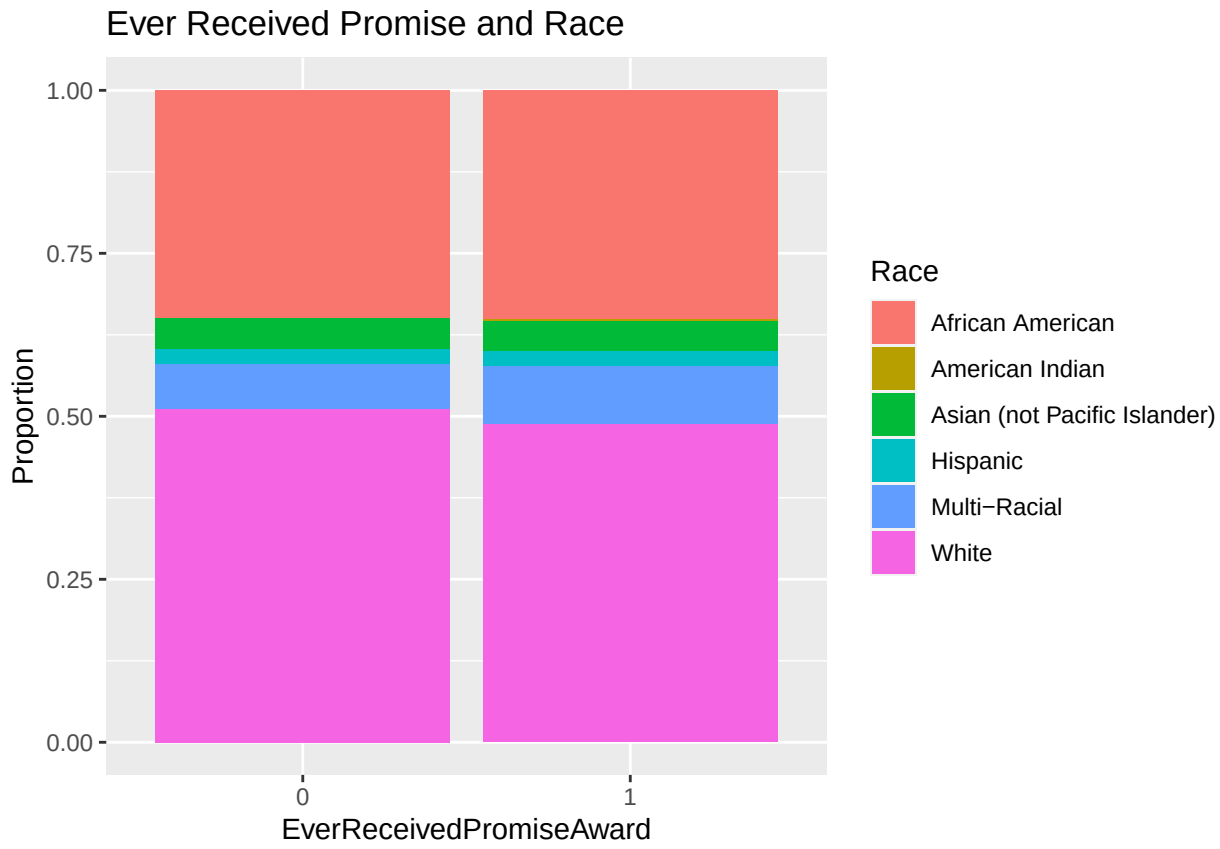
Build logistic model on qualified students

Construct dataframe

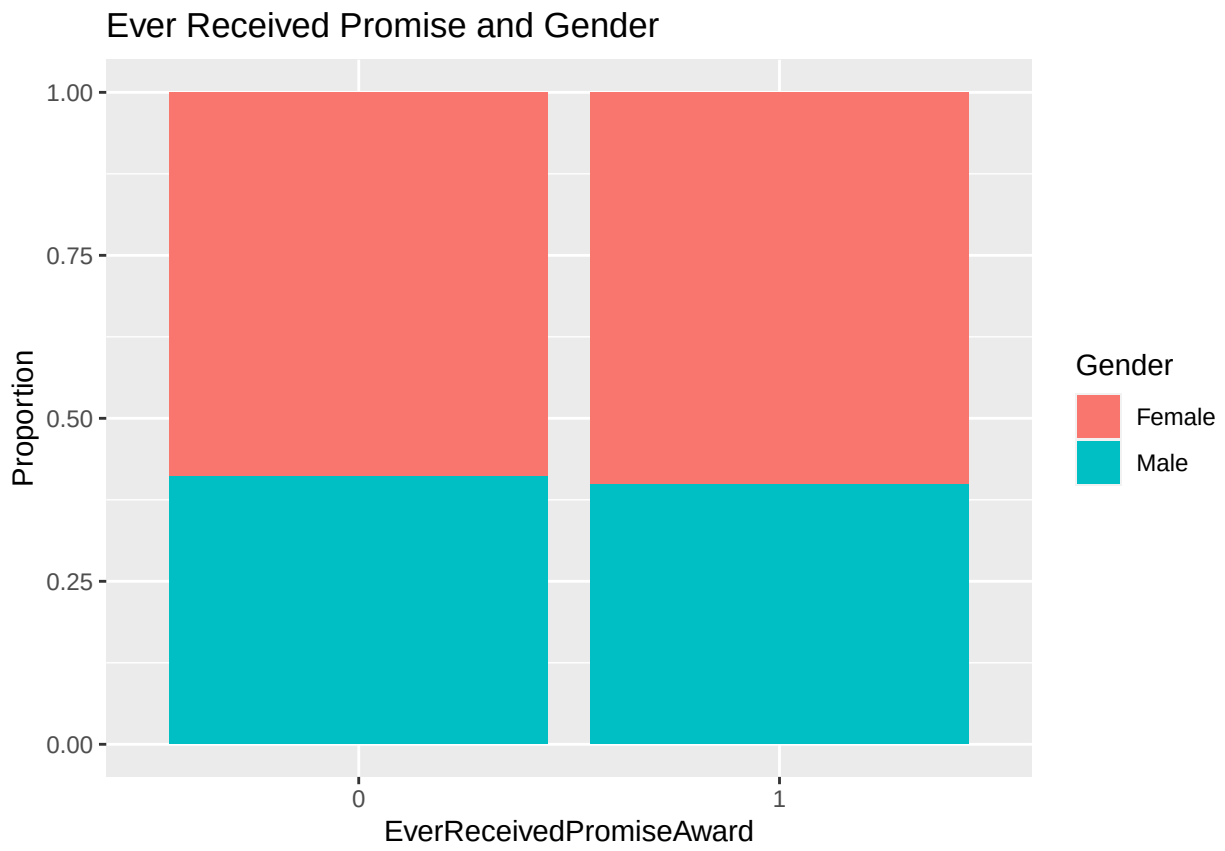
```
#colnames(data_join)

data_qualified <- data_join %>%
  select(RandomID, GradYear.x, excused_pct, excused_pct_cate, Num_AP, num_cte,
         ScoreMean, Race, Gender, ELLStatus, IEPGroup, EconDisad, Latest_SAT_Total,
         CumulativeGPA, FullMagnetInd, QualifiedforCorePromise.x, EverReceivedPromiseAward.x) %>%
  distinct()


data_qualified <- merge(data_changeFormat,
                        attendance_model[, c("RandomID", "QualifiedforCorePromise")], by="RandomID")

data_qualifiedYes <- data_qualified %>%
  filter(QualifiedforCorePromise == "yes")

#write_csv(as.data.frame(data_qualified), "/Users/gloriaguo/Desktop/PPS/PPS_data_Processed/data_qualifi
#write_csv(as.data.frame(data_qualifiedYes), "/Users/gloriaguo/Desktop/PPS/PPS_data_Processed/data_qual
```

EDA on full Scholarship data: Qualified

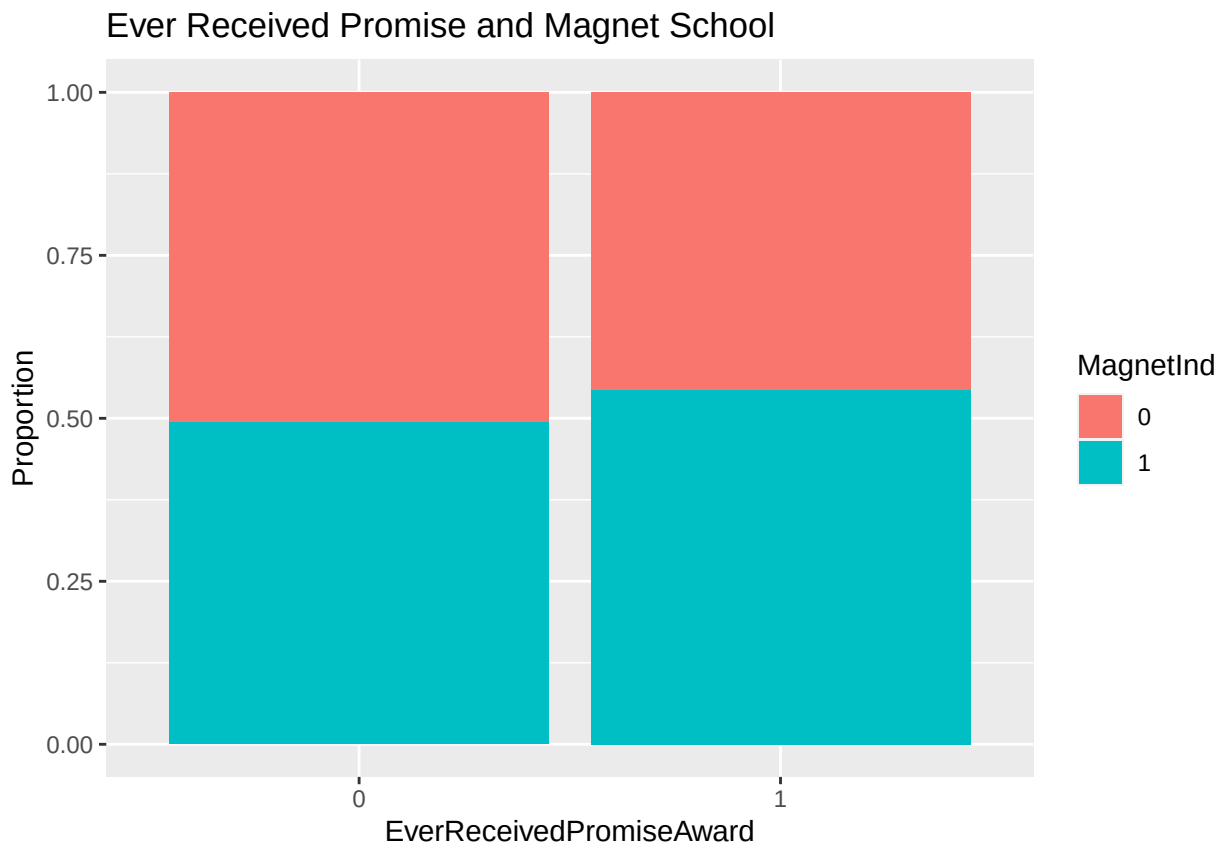
```
## Race
ggplot(data_qualifiedYes, aes(fill=Race, x=EverReceivedPromiseAward)) +
  geom_bar(position="fill") +
  ggtitle("Ever Received Promise and Race") +
  labs(y="Proportion")
```




```
## Gender
ggplot(data_qualifiedYes, aes(fill=Gender, x=EverReceivedPromiseAward)) +
  geom_bar(position="fill") +
  ggtitle("Ever Received Promise and Gender") +
  labs(y="Proportion")
```




```
## Magnet School
ggplot(data_qualifiedYes, aes(fill=MagnetInd, x=EverReceivedPromiseAward)) +
  geom_bar(position="fill") +
  ggtitle("Ever Received Promise and Magnet School") +
  labs(y="Proportion")
```



Select null model as base model

```
model_null <- glm(EverReceivedPromiseAward ~ 1, data = data_qualifiedYes, family = "binomial")
model_full <- glm(EverReceivedPromiseAward ~ AttendanceRate+Num_AP+Num_CTE+KeystoneMean+Race
              +Gender+ELLStatus+IEPGroup+EconDisad+SAT_Total+CumulativeGPA+MagnetInd,
              data = data_qualifiedYes, family = "binomial")
```

```
# Backwards selection on AIC
backwards <- step(model_full, trace = 0)
# summary(backwards)
```

```
# Forward selection on AIC
forwards <- step(model_null, scope=list(lower=formula(model_null),upper=formula(model_full)),
                 direction="forward", trace = 0)
# summary(forwards)
```

```
# Bothways selection on AIC
bothways <- step(model_null, list(lower=formula(model_null),upper=formula(model_full)),
                 direction="both", trace=0)
summary(bothways)
```

```
##
## Call:
```

```
## glm(formula = EverReceivedPromiseAward ~ AttendanceRate + KeystoneMean +
##     CumulativeGPA + MagnetInd + ELLStatus + Num_CTE, family = "binomial",
##     data = data_qualifiedYes)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3059  -0.9345  -0.8027   1.3549   2.1130
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -3.359592   3.113918  -1.079 0.280634
## AttendanceRate      10.008923   2.509608   3.988 6.66e-05 ***
## KeystoneMean        -0.006191   0.001714  -3.613 0.000303 ***
## CumulativeGPA        0.453072   0.190438   2.379 0.017355 *
## MagnetInd1           0.227239   0.120533   1.885 0.059392 .
## ELLStatusNot in ELL  0.698576   0.452300   1.544 0.122468
## Num_CTE             -0.097567   0.064626  -1.510 0.131113
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1722.8  on 1356  degrees of freedom
## Residual deviance: 1677.2  on 1350  degrees of freedom
## AIC: 1691.2
##
## Number of Fisher Scoring iterations: 4
```

```
formula(backwards)
```

```
## EverReceivedPromiseAward ~ AttendanceRate + Num_CTE + KeystoneMean +
##     ELLStatus + CumulativeGPA + MagnetInd
```

```
formula(forwards)
```

```
## EverReceivedPromiseAward ~ AttendanceRate + KeystoneMean + CumulativeGPA +
##     MagnetInd + ELLStatus + Num_CTE
```

```
formula(bothways)
```

```
## EverReceivedPromiseAward ~ AttendanceRate + KeystoneMean + CumulativeGPA +
##     MagnetInd + ELLStatus + Num_CTE
```

Select Race+Gender as base model

```
model_null <- glm(EverReceivedPromiseAward ~ Race+Gender, data = data_qualifiedYes, family = "binomial")
model_full <- glm(EverReceivedPromiseAward ~ AttendanceRate+Num_AP+Num_CTE+KeystoneMean+Race
              +Gender+ELLStatus+IEPGroup+EconDisad+SAT_Total+CumulativeGPA+MagnetInd,
              data = data_qualifiedYes, family = "binomial")
```

```r
# Backwards selection on AIC
backwards <- step(model_full, trace = 0)
# summary(backwards)
```

```r
# Forward selection on AIC
forwards <- step(model_null, scope=list(lower=formula(model_null),upper=formula(model_full)),
                 direction="forward", trace = 0)
# summary(forwards)
```

```r
# Bothways selection on AIC
bothways <- step(model_null, list(lower=formula(model_null),upper=formula(model_full)),
                 direction="both", trace=0)
summary(bothways)
```

```
##
## Call:
## glm(formula = EverReceivedPromiseAward ~ Race + Gender + AttendanceRate +
##     KeystoneMean + CumulativeGPA + MagnetInd, family = "binomial",
##     data = data_qualifiedYes)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2124  -0.9342  -0.7943   1.3560   2.1027
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -4.824993   3.343735  -1.443  0.14902
## RaceAmerican Indian               1.282735   1.479437   0.867  0.38592
## RaceAsian (not Pacific Islander) -0.247677   0.305563  -0.811  0.41762
## RaceHispanic                     -0.267055   0.401240  -0.666  0.50568
## RaceMulti-Racial                  0.263061   0.239144   1.100  0.27133
## RaceWhite                        -0.222510   0.155137  -1.434  0.15149
## GenderMale                       -0.025427   0.123407  -0.206  0.83676
## AttendanceRate                   10.556032   2.588334   4.078 4.54e-05 ***
## KeystoneMean                     -0.005168   0.001777  -2.908  0.00364 **
## CumulativeGPA                     0.504048   0.197292   2.555  0.01062 *
## MagnetInd1                        0.224038   0.123162   1.819  0.06890 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1722.8  on 1356  degrees of freedom
## Residual deviance: 1675.4  on 1346  degrees of freedom
## AIC: 1697.4
##
## Number of Fisher Scoring iterations: 4
```

```r
formula(backwards)
```

```
## EverReceivedPromiseAward ~ AttendanceRate + Num_CTE + KeystoneMean +
##     ELLStatus + CumulativeGPA + MagnetInd
```

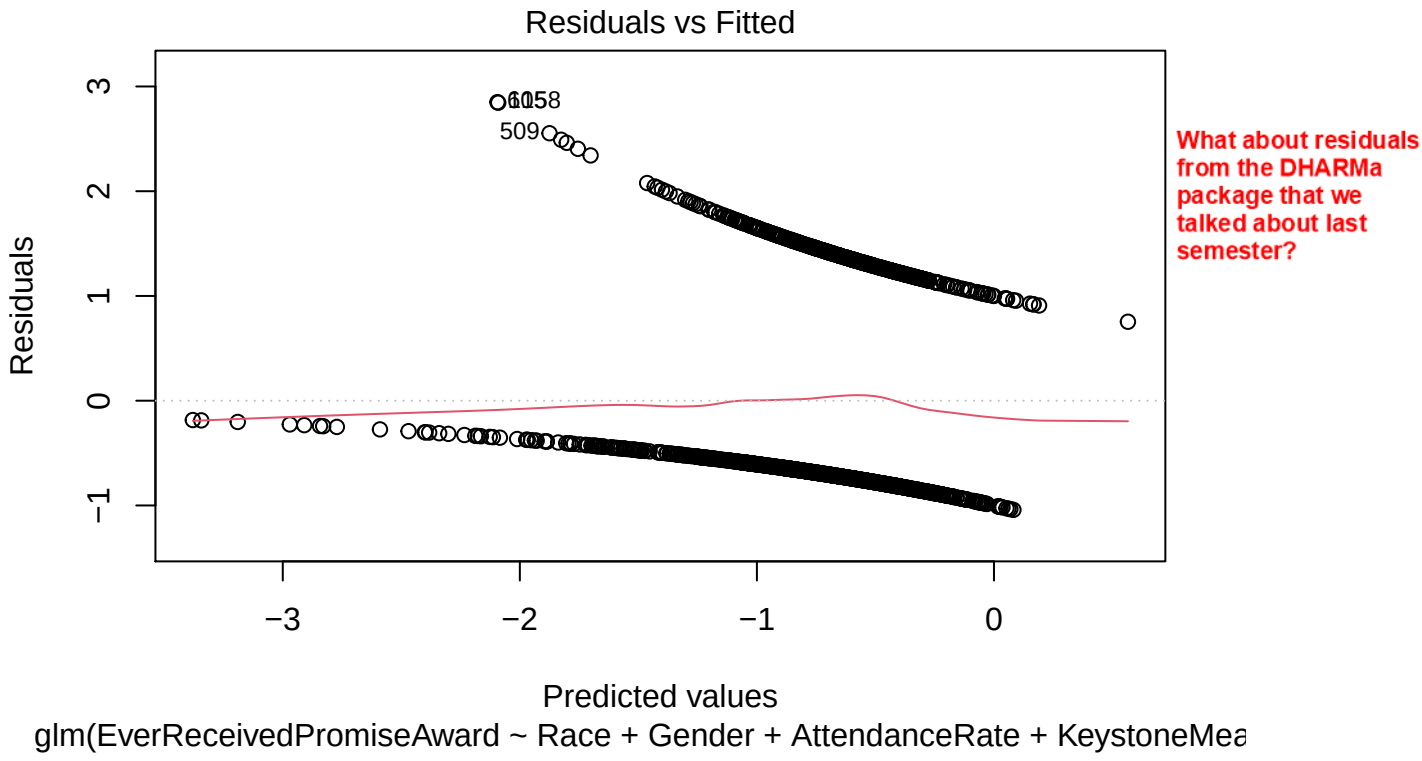
```
formula(forwards)
```

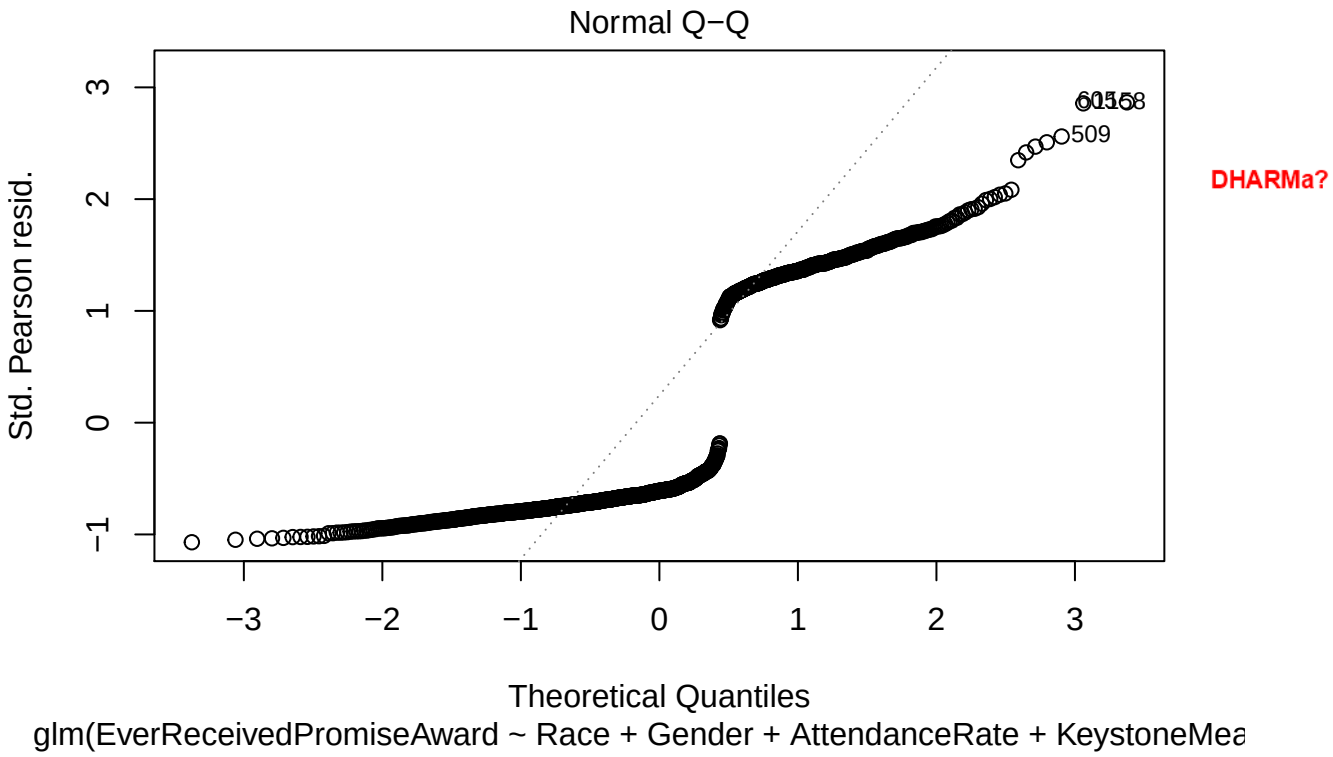
```
## EverReceivedPromiseAward ~ Race + Gender + AttendanceRate + KeystoneMean +
##     CumulativeGPA + MagnetInd
```

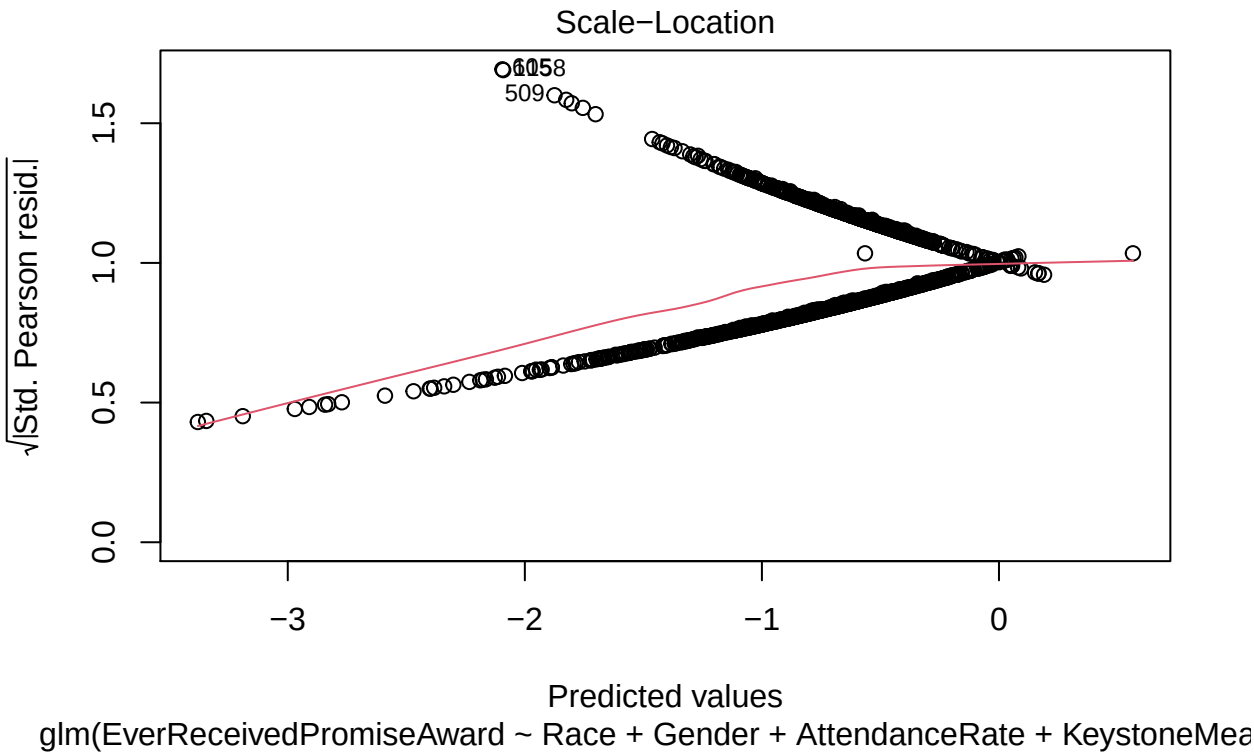
```
formula(bothways)
```

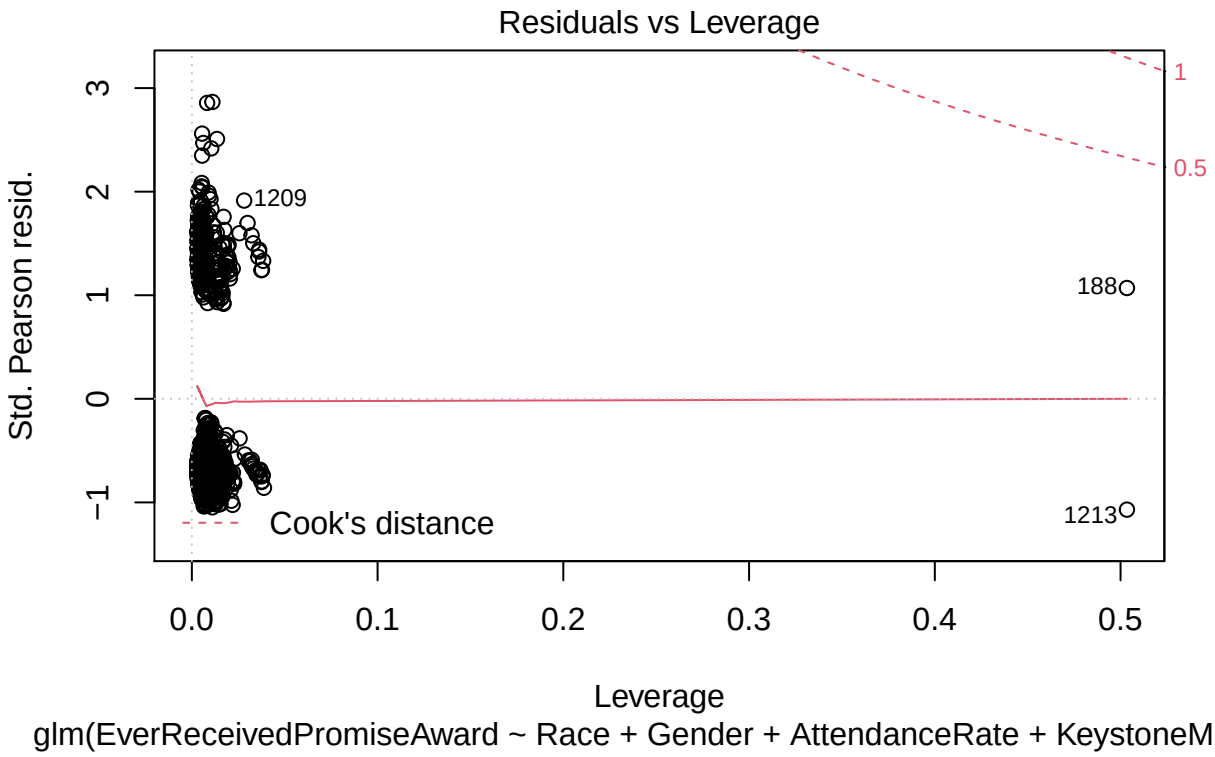
```
## EverReceivedPromiseAward ~ Race + Gender + AttendanceRate + KeystoneMean +
##     CumulativeGPA + MagnetInd
```

```
plot(bothways)
```




Normal Q−Q
Std. Pearson resid.
Theoretical Quantiles
glm(EverReceivedPromiseAward ~ Race + Gender + AttendanceRate + KeystoneMea
509
DHARMa?


Scale−Location
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
−3
−2
−1
0
Predicted values
glm(EverReceivedPromiseAward ~ Race + Gender + AttendanceRate + KeystoneMea



glm(EverReceivedPromiseAward ~ Race + Gender + AttendanceRate + KeystoneMea

Retention Analysis

Questions:

1. Only focus on students who went to PA college, compare student retention for received scholarship vs not received. Pay attention to sample size. Check for significance for year 2018 and 2019.
2. Only focus on students who went to PA college, compare student retention between whites and blacks. Check for significance for year 2018 and 2019.

- Definition: Retention is defined as the total number of days a students spent at colleges in PA(Sometimes students transferred to other state's colleges).
- Confirm with Steven: students who gets scholarship are definitely in the scholarship dataset, so that we assume students who are not in the scholarship dataset didn't receive the scholarship.

```
library(lubridate)
library(tidyverse)
promise_df = read.csv("Scholarship.csv")
promise_df <- promise_df %>%
  mutate(QualifiedforCorePromise = ifelse(QualifiedforCorePromise == "yes",1,0),
         QualifiedforExtensionPromise = ifelse(QualifiedforExtensionPromise == "yes",1,0),
         EverReceivedPromiseAward = ifelse(EverReceivedPromiseAward == "yes",1,0),
         StillReceivingAward = ifelse(StillReceivingAward == "yes",1,0),
         StillEligible = ifelse(StillEligible == "yes",1,0),
         HighSchool = as.factor(HighSchool)) %>%
  rename(RandomID = Random.ID)
head(promise_df)
```

```
##   RandomID GradYear QualifiedforCorePromise QualifiedforExtensionPromise
## 1  5829765     2018                       0                             0
## 2  5832055     2018                       0                             0
## 3  5833420     2018                       1                             0
## 4  5840516     2018                       0                             0
## 5  5841218     2018                       0                             0
## 6  5847024     2018                       1                             0
##   EverReceivedPromiseAward StillReceivingAward StillEligible
## 1                        0                   0             0
## 2                        0                   0             0
## 3                        0                   0             1
## 4                        0                   0             0
## 5                        1                   1             0
## 6                        0                   0             1
##                           HighSchool
## 1    Pittsburgh Allderdice High School
## 2   Pittsburgh UPrep 6-12 At Milliones
## 3 Pittsburgh Westinghouse Academy 6-12
## 4       Pittsburgh Carrick High School
## 5         Pittsburgh Perry High School
## 6    Pittsburgh Allderdice High School
```

```r
nsc = read.csv("NSC.csv")
head(nsc)
```

```
##   RandomID Cohort HIGH_SCHOOL_GRAD_DATE COLLEGE_STATE X2.YEAR_4.YEAR
## 1  5841218   1415              20180608            PA         4-year
## 2  5841218   1415              20180608            PA         4-year
## 3  5847024   1415              20180608            NY         4-year
## 4  5847024   1415              20180608            NY         4-year
## 5  5847024   1415              20180608            NY         4-year
## 6  5847024   1415              20180608            NY         4-year
##   PUBLIC_PRIVATE ENROLLMENT_BEGIN ENROLLMENT_END ENROLLMENT_STATUS GRADUATED
## 1         Public         20180827       20181214                 F         N
## 2         Public         20190122       20190510                 F         N
## 3        Private         20180904       20181221                 F         N
## 4        Private         20190128       20190521                 F         N
## 5        Private         20190903       20191220                 F         N
## 6        Private         20200127       20200519                 F         N
##   GRADUATION_DATE
## 1               0
## 2               0
## 3               0
## 4               0
## 5               0
## 6               0
```

```r
demo1415 = read.csv("Demographics_1415.csv")
demo1516 = read.csv("Demographics_1516.csv")
demo1617 = read.csv("Demographics_1617.csv")

demographics = rbind(demo1415,demo1516,demo1617)
demographics = demographics %>% select(RandomID, Race) %>% distinct()
#delete students who have more than 1 races
demographics <- demographics %>% filter(! RandomID %in% c(6008133,6040930,6126484,6137723,
                                                          6207255,6213261,6228923,6510341))
```

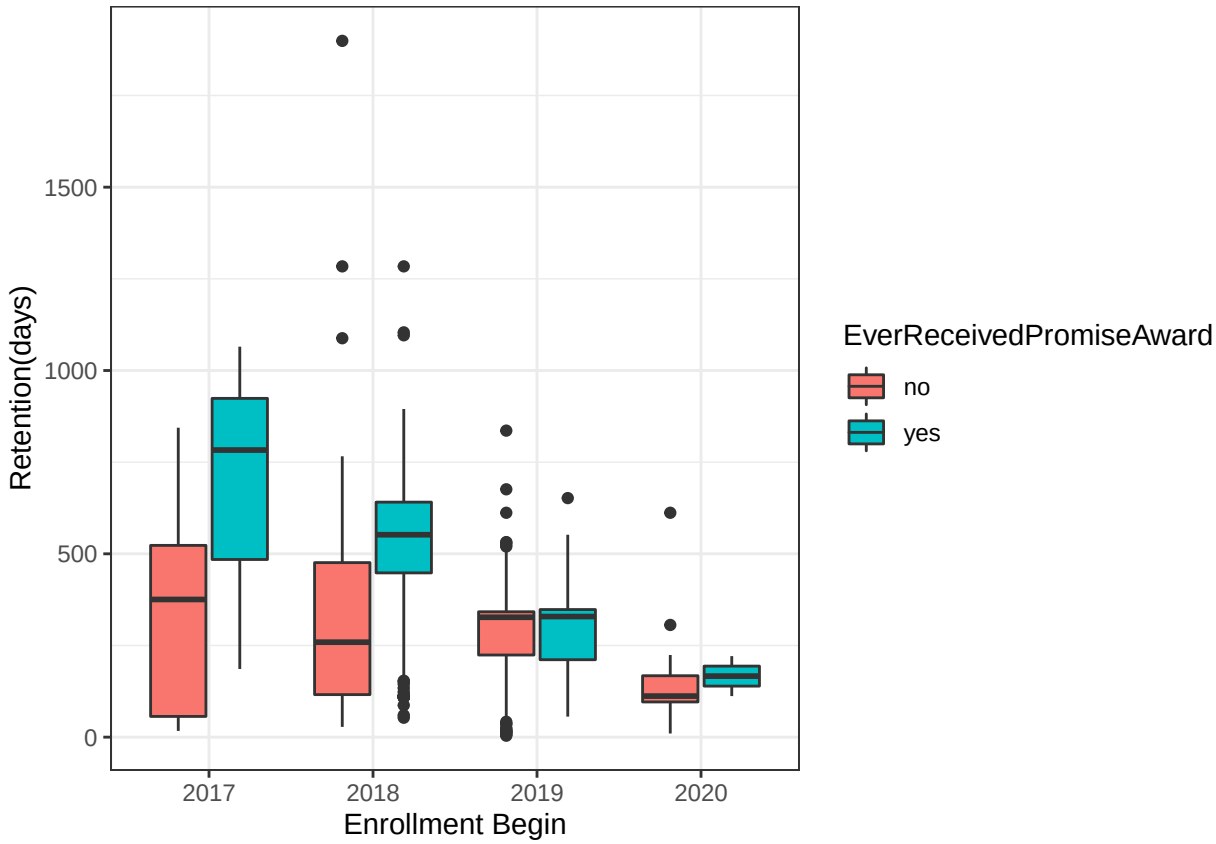
Question 1 Retention and Scholarship

```r
nsc_processed = nsc %>%
  filter(ENROLLMENT_BEGIN >0) %>%
  filter(ENROLLMENT_BEGIN  %/% 10000 != 2016) %>%
  filter(COLLEGE_STATE == "PA") %>%
  mutate(ENROLLMENT_BEGIN = as.Date(paste(substr(ENROLLMENT_BEGIN, 1, 4),
                                          substr(ENROLLMENT_BEGIN, 5, 6),
                                          substr(ENROLLMENT_BEGIN, 7, 8),
                                          sep = "-")),
         ENROLLMENT_END = as.Date(paste(substr(ENROLLMENT_END, 1, 4),
                                        substr(ENROLLMENT_END, 5, 6),
                                        substr(ENROLLMENT_END, 7, 8),
                                        sep = "-")),
         elapsed = ENROLLMENT_END - ENROLLMENT_BEGIN)

#issue: some students enroll in fall while others enroll in spring
nsc_retention = nsc_processed %>% group_by(RandomID) %>%
```

```
  summarise(retention = sum(elapsed),
            ENROLLMENT_BEGIN_year = year(min(ENROLLMENT_BEGIN)),
            semester = ifelse(month(min(ENROLLMENT_BEGIN))>6, "Fall","Spring"))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
nsc_promise = nsc_retention %>% left_join(promise_df, by = "RandomID") %>%
  mutate(EverReceivedPromiseAward = ifelse(EverReceivedPromiseAward==1,"yes","no"),
         EverReceivedPromiseAward = replace_na(EverReceivedPromiseAward,"no"),
         EverReceivedPromiseAward = as.factor(EverReceivedPromiseAward),
         retention = as.numeric(retention)) %>%
  select(-c(QualifiedforCorePromise,
            QualifiedforExtensionPromise,
            StillReceivingAward,
            StillEligible,HighSchool))

ggplot(nsc_promise, aes(x=as.factor(ENROLLMENT_BEGIN_year),
                        y=retention,
                        fill=EverReceivedPromiseAward)) +
  geom_boxplot() +
  labs(x = "Enrollment Begin", y = "Retention(days)") +
  theme_bw()
```



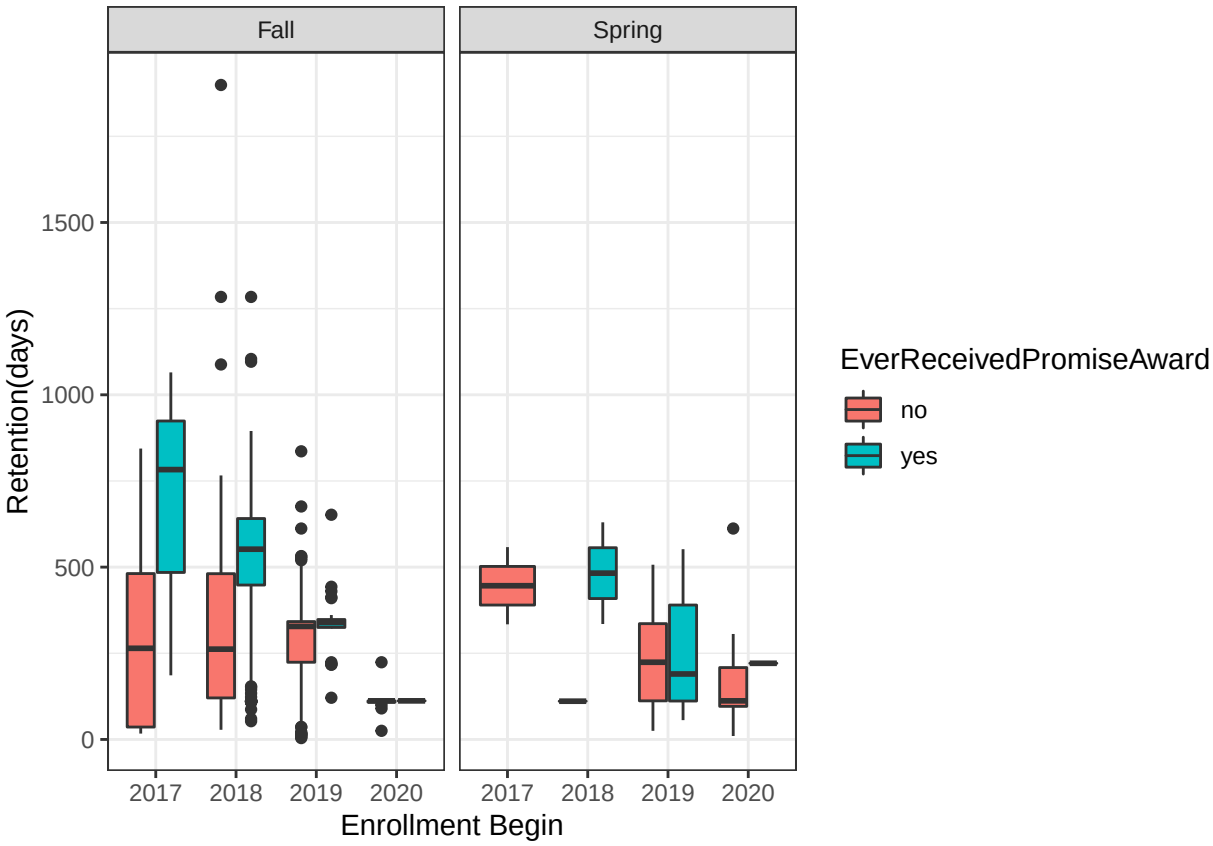
We see that students who recevied scholarship tend to have better retention except for year 2019. However, noticed that we only have 13 observations for year 2017, and 93 observations for year 2020.

```
nsc_promise %>% group_by(ENROLLMENT_BEGIN_year, semester) %>% count()
```

```
## # A tibble: 8 x 3
## # Groups:   ENROLLMENT_BEGIN_year, semester [8]
##   ENROLLMENT_BEGIN_year semester     n
##                   <dbl> <chr>    <int>
## 1                  2017 Fall        11
## 2                  2017 Spring       2
## 3                  2018 Fall       571
## 4                  2018 Spring       3
## 5                  2019 Fall       639
## 6                  2019 Spring      59
## 7                  2020 Fall        28
## 8                  2020 Spring      65
```

Take a closer look by the semester they enroll

```
ggplot(nsc_promise, aes(x=as.factor(ENROLLMENT_BEGIN_year),
                        y=retention,
                        fill=EverReceivedPromiseAward)) +
  geom_boxplot() +
  labs(x = "Enrollment Begin", y = "Retention(days)") +
  theme_bw() +facet_wrap(~semester)
```



This is just for reference, because we don't have many observations for year 2017 and all spring semesters.

Check for significance for year 2018 and 2019

```
nsc_promise_2018 = nsc_promise %>% filter(ENROLLMENT_BEGIN_year == 2018)
nsc_promise_2019 = nsc_promise %>% filter(ENROLLMENT_BEGIN_year == 2019)

bartlett.test(retention~EverReceivedPromiseAward, data = nsc_promise_2018)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  retention by EverReceivedPromiseAward
## Bartlett's K-squared = 37.596, df = 1, p-value = 8.7e-10
```

```
bartlett.test(retention~EverReceivedPromiseAward, data = nsc_promise_2019)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  retention by EverReceivedPromiseAward
## Bartlett's K-squared = 2.1978, df = 1, p-value = 0.1382
```

```
oneway.test(retention~EverReceivedPromiseAward, data = nsc_promise_2018, var.equal = FALSE)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  retention and EverReceivedPromiseAward
## F = 44.755, num df = 1.00, denom df = 130.29, p-value = 5.98e-10
```

```
oneway.test(retention~EverReceivedPromiseAward, data = nsc_promise_2019, var.equal = TRUE)
```

```
##
##  One-way analysis of means
##
## data:  retention and EverReceivedPromiseAward
## F = 0.27045, num df = 1, denom df = 696, p-value = 0.6032
```

We see that for year 2018, the difference is significant, but for year 2019 the difference is not significant. This align with what we see in the boxplot. We see that the difference is quite obvious in year 2018, whereas in year 2019 is not very obvious. Maybe the retention difference would become more obvious for more senior students.

Question 2 Retention and Race

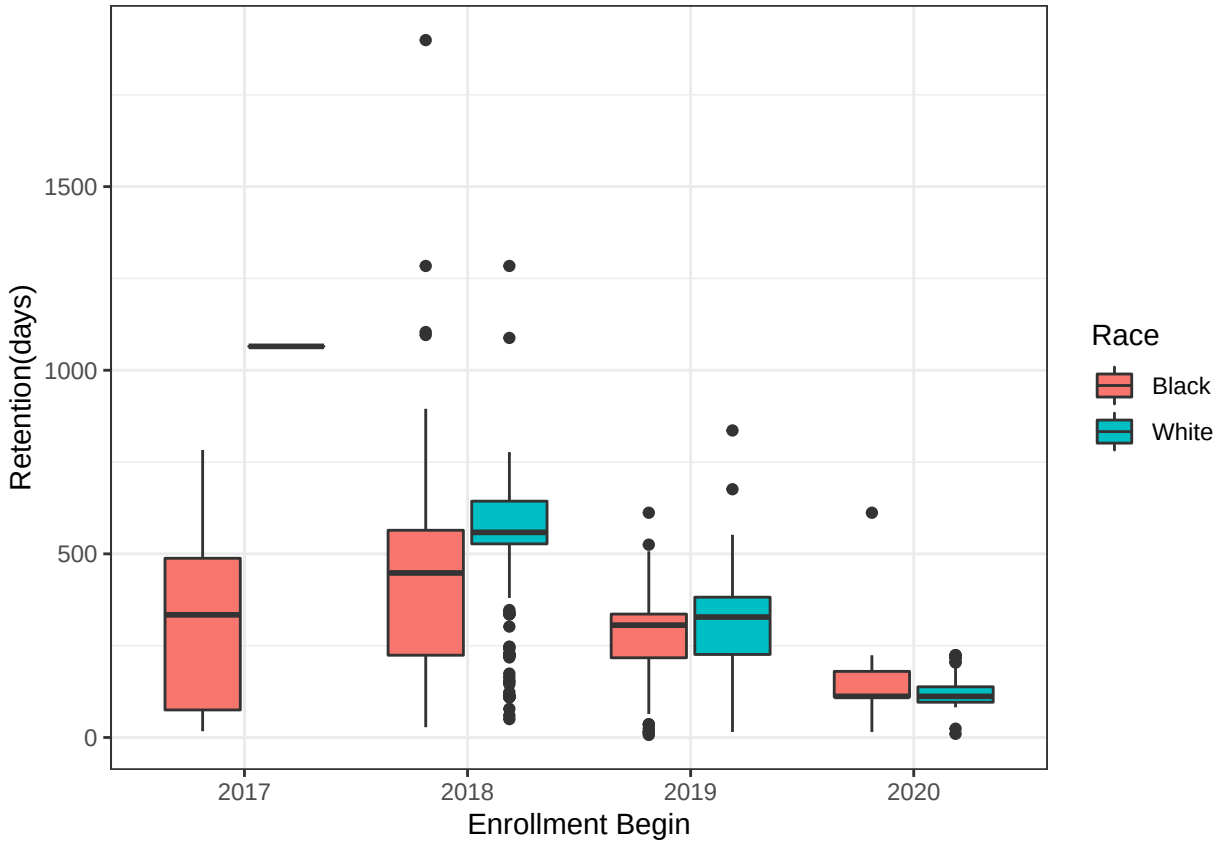
```
demographics %>% group_by(Race) %>% count() %>% arrange(-n)
```

```
## # A tibble: 7 x 2
## # Groups:   Race [7]
##   Race                                  n
##   <chr>                             <int>
## 1 African American                   2766
## 2 White                              1847
## 3 Multi-Racial                        327
## 4 Asian (not Pacific Islander)        180
## 5 Hispanic                            164
```

```
## 6 American Indian                                10
## 7 Native Hawaiian or other Pacific Islander      4
```

We see that the majority of students are blacks and whites. So we will explore the difference in student retention between these two races. Also, we restrict on students who went to college in PA.

```
nsc_demographics = nsc_promise %>%
  left_join(demographics, by = "RandomID") %>%
  filter(Race %in% c("White","African American")) %>%
  mutate(Race = ifelse(Race == "White", "White","Black"))

ggplot(nsc_demographics, aes(x=as.factor(ENROLLMENT_BEGIN_year), y=retention, fill=Race)) +
  geom_boxplot() +
labs(x = "Enrollment Begin", y = "Retention(days)") +
theme_bw()
```



We see that white students tend to have better retention than black students except for year 2020. Again, we don't have many observations for year 2017 and 2020.

```
nsc_demographics %>% group_by(ENROLLMENT_BEGIN_year) %>% count()
```

```
## # A tibble: 4 x 2
## # Groups:   ENROLLMENT_BEGIN_year [4]
##   ENROLLMENT_BEGIN_year     n
##                   <dbl> <int>
## 1                  2017    12
## 2                  2018   506
## 3                  2019   593
## 4                  2020    75
```

Check for significance for year 2018 and 2019

```
nsc_demographics_2018 = nsc_demographics %>% filter(ENROLLMENT_BEGIN_year == 2018)
nsc_demographics_2019 = nsc_demographics %>% filter(ENROLLMENT_BEGIN_year == 2019)

bartlett.test(retention~Race, data = nsc_demographics_2018)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  retention by Race
## Bartlett's K-squared = 11.071, df = 1, p-value = 0.0008769
```

```
bartlett.test(retention~Race, data = nsc_demographics_2019)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  retention by Race
## Bartlett's K-squared = 0.077868, df = 1, p-value = 0.7802
```

```
oneway.test(retention~Race, data = nsc_demographics_2018, var.equal = FALSE)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  retention and Race
## F = 20.256, num df = 1.00, denom df = 487.42, p-value = 8.48e-06
```

```
oneway.test(retention~Race, data = nsc_demographics_2019, var.equal = TRUE)
```

```
##
##  One-way analysis of means
##
## data:  retention and Race
## F = 29.355, num df = 1, denom df = 591, p-value = 8.782e-08
```

We see that for both 2018 and 2019 the differences are significant. We conclude that white students have better retention than black students for year 2018 and 2019.