Carnegie Mellon University

Pittsburgh Public Schools Retention/Mobility Research

Final Presentation of 36-726 Capstone Project

Client: Steven Greene from Pittsburgh Public Schools
Project Advisor: Professor Zach Branson
36-726 Instructor: Professor Brian Junker
Presenters: Huiyi Guo, Jenny Luo, Yuhang Ying

Agenda

- Introduction
- Data
- Research Question 1
  - Methods
  - Results
  - Discussion: Result Summary
- Research Question 2
  - Methods
  - Results
  - Discussion: Result Summary
- Discussion: Take-home Policy, Limitations, Next Steps
- Q&A
- Technical Appendix

Introduction

- Pittsburgh Public Schools: a public school district in Pennsylvania, US
- Funds a Promise scholarship for students' post-secondary education
  - Requirements
    - Cumulative GPA ≥ 2.5
    - Attendance Rate ≥ 90%
- Two research questions
  - **Scholarship Analysis:** investigate factors that influence whether students received Promise scholarships
  - **Retention Analysis:** evaluate factors that influence students' retention in college, and make comparisons between retention in different groups

Data

- 11 data sets ranging from 2014 to 2020
- Joined *Scholarship, School Enrollment, Attendance, Demographics, SAT, AP, GPA, Keystone,* and *CTE* together to conduct scholarship analysis
- Also joined NSC data to conduct retention analysis
- Data size
  - Scholarship Analysis: 1708 observations
  - Retention Analysis: 1378 observations

When to start college	2017	2018	2019	2020
Number of students	13	574	698	93

Research Question 1: Scholarship Analysis

Methods: Scholarship Analysis

- Logistic Regression
  - **Goal 1**: what factors would mostly influence whether Promise scholarship has been used by students?
    - **Dataset 1:** all students in Scholarship dataset (1708 observations)
  - **Goal 2**: among all qualified students, what factors influence their decision to use Promise scholarship?
    - **Dataset 2:** the subset of students who qualify for Promise in Scholarship dataset (1357 observations)

Methods: Scholarship Analysis

- Logistic Regression
  - Outcome variable: *EverReceivedPromiseAward*
  - Predictors: *AttendanceRate, Num_AP, Num_CTE, KeystoneMean, Race, Gender, ELLStatus, IEPGroup, EconDisab, SAT_Total, CumulativeGPA, MagnetInd*
  - Stepwise variable selection based on AIC
    - Null model: *EverReceivedPromiseAward ~ Race + Gender*
    - Full model: *EverReceivedPromiseAward ~ all predictors*

Results: Scholarship Analysis

- **Goal 1:** what factors would mostly influence whether Promise scholarship has been used by students?
- **Dataset 1:** all students in *Scholarship* dataset (1708)

Variable	Coefficient	P-Value
(Intercept)	-3.725	0.194
Race(Others)	0.056	0.767
Race(White)	-0.267	0.068 .
Gender(Male)	-0.091	0.435
Cumulative GPA	0.903	1.150e-08 ***
Attendance Rate	8.051	7.130e-05 ***
Mean Keystone Scores	-0.006	4.580e-04***
ELL Status(Not in ELL)	1.101	0.012 *
Magnet School Indicator	0.249	0.032 *

*Notice: significant variables are in **red**.*

Results: Scholarship Analysis

- **Analysis of all students**
- **Interpretations of significant variables:**
  - When holding everything else fixed:
    - the odds to receive Promise for **white** students are **30.6% lower** than **black** students
    - the odds to receive Promise for student who are in **magnet** school are **38.3% higher** than those in **non-magnet** school
    - the odds for student who are **not in ELL** group are **200.7% higher** than those in **ELL group**
    - **0.1 increase** in **GPA → 9.4% increase** in the odds of receiving Promise
    - **0.01 increase** in **attendance rate → 8.4% increase** in the odds of receiving Promise
    - **100 increase** in **Keystone mean → 55.5% decrease** in the odds of receiving Promise

Results: Scholarship Analysis

- **Goal 2:** among qualified students, what factors influence their decision to use Promise scholarship?
- **Dataset 2:** qualified students in *Scholarship* dataset (1357)

Variable	Coefficient	P-Value
(Intercept)	-4.917	0.140
Race(Others)	0.107	0.588
Race(White)	-0.205	0.185
Gender(Male)	-0.023	0.852
Cumulative GPA	0.475	0.015 *
Attendance Rate	10.402	5.380e-05 ***
Mean Keystone Scores	-0.005	0.002 **
ELLStatus(Not in ELL)	0.784	0.090 .
Magnet Schools Indicator	0.225	0.066 .

*Notice: significant variables are in **red**.*

Results: Scholarship Analysis

- **Analysis of qualified students**
- **Interpretations of significant variables:**
  - When holding everything else fixed:
    - the odds to receive Promise for student who are in **magnet** school are **25.3% higher** than those in **non-magnet** school
    - the odds for student who are **not in ELL** group are 119% higher than those in **ELL group**
    - **0.1 increase** in **GPA → 4.9% increase** in the odds of receiving Promise
    - **0.01 increase** in **attendance rate → 11% increase** in the odds of receiving Promise
    - **100 increase** in **Keystone mean → 57.9% decrease** in the odds of receiving Promise

Discussion: Result Summary

- **Scholarship Analysis:**
  - Higher attendance rate → more likely to receive Promise
  - Higher GPA → more likely to receive Promise
  - Higher Keystone mean → less likely to receive Promise
  - Non-ELL group > ELL group
  - Magnet school > Non-magnet school

Research Question 2: Retention Analysis

Methods: Retention Analysis

- Conduct both EDA and multivariate regression analysis
- Restrict analysis to students who went to college in PA
- Retention in college = total number of days a student has stayed in college
  - Cumulative sum of (*Enrollment_End - Enrollment Begin*)
- Group students by their first years of college enrollments

Methods: Retention Analysis

- **Exploratory Data Analysis:**
  - Boxplots to compare retention in different groups
    - Received Promise vs. Not received Promise
    - Black vs. White
    - Interaction between race and scholarship receipt
      - Racial difference analysis only for students who received
      - Racial difference analysis only for students who did not
  - Welch t-test to examine statistical significance for retention comparisons

Methods: Retention Analysis

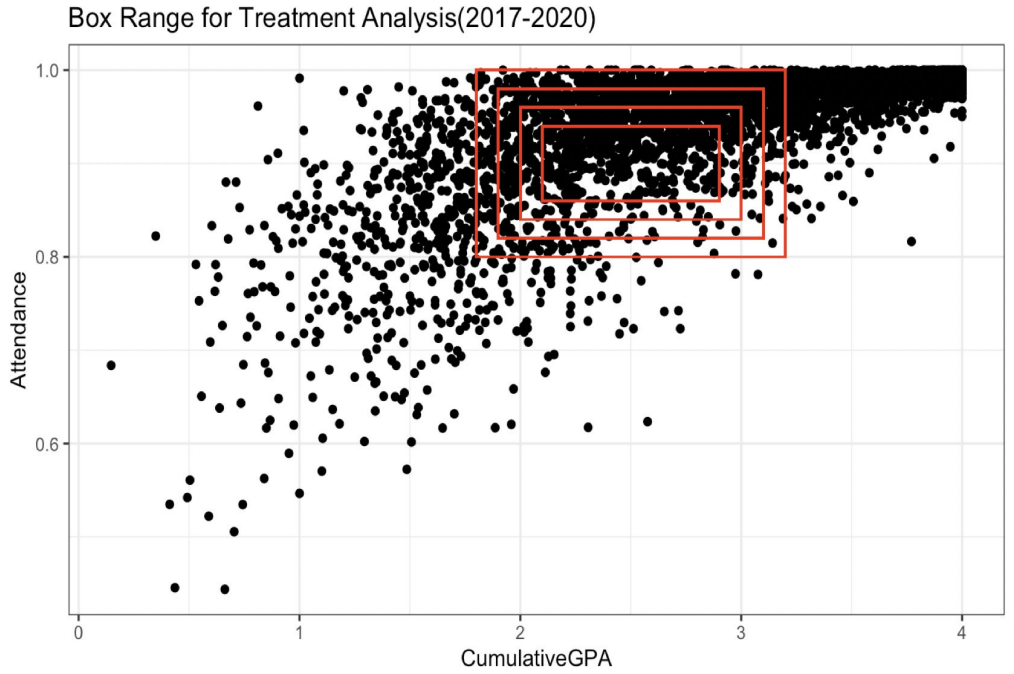
- **Multivariate Regression for Various Ranges:**
  - Construct one model for each college starting year (2018 & 2019)
  - Outcome variable: retention in college
  - All predictors: *AttendanceRate, Num_AP, Num_CTE, KeystoneMean, Race, Gender, ELLStatus, IEPGroup, EconDisad, SAT_Total, Cumulative GPA, MagnetInd, EverReceivedPromiseAward, Semester, EverReceivedPromiseAward*Race*
  - Stepwise selection on AIC to choose predictors
  - Null model: *Retention ~ EverReceivedPromiseAward +Race + EverReceivedPromiseAward*Race*

Methods: Retention Analysis

- **Multivariate Regression for Various Ranges**
  - **Whole range of GPA & attendance:**
    - Poisson regression models
  - **Box range of GPA & attendance:**
    - Multivariate linear regression to validate the effect of scholarship
    - Account for the effects from academic performance on retention
    - Treatment control test(t-test) for scholarship

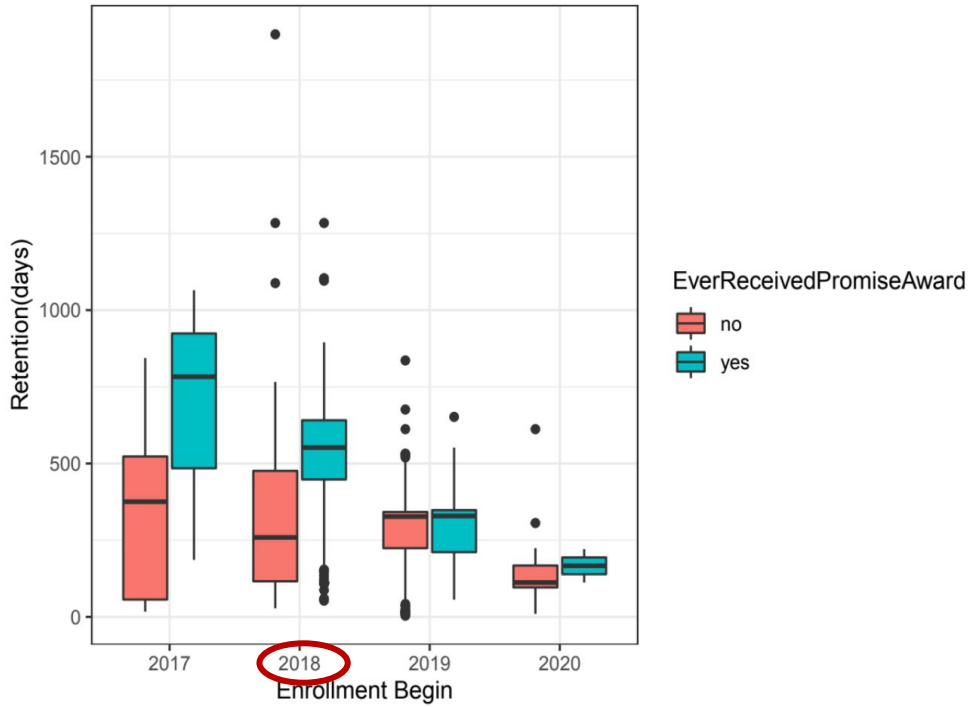
Methods: Retention Analysis

Attendance Lower	Attendance Upper	GPA Lower	GPA Upper
0.86	0.94	2.1	2.9
0.84	0.96	2.0	3.0
0.82	0.98	1.9	3.1
0.80	1.00	1.8	3.2



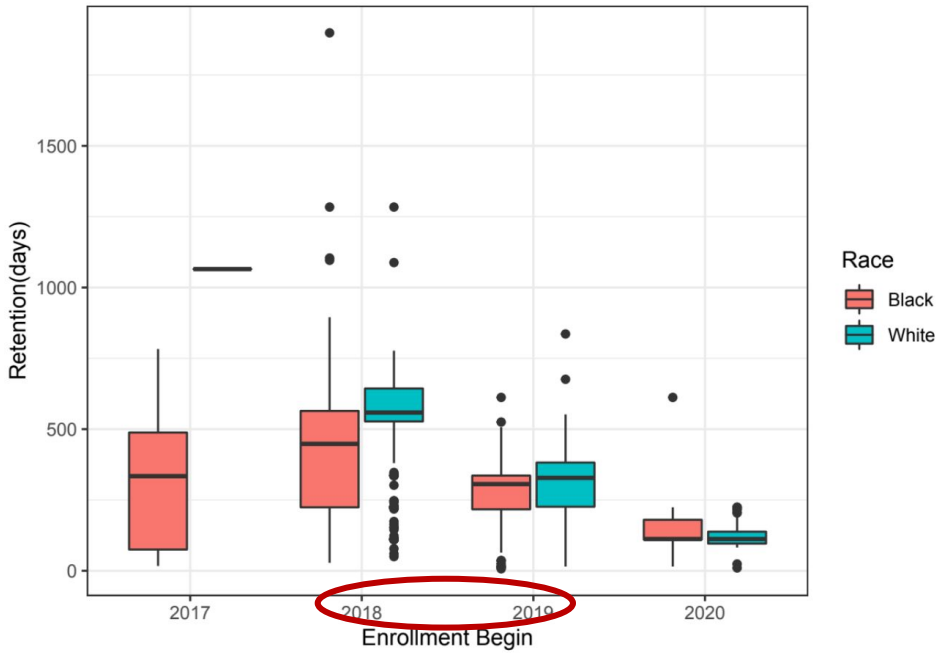
Results: Retention Analysis

- **EDA for retention and scholarship:**
  - Students who received scholarships tend to have better retention except for 2019
  - We focus on 2018 & 2019, because of limited observations in 2017 (13 obs) and 2020 is too recent.
  - Difference is significant for 2018 ($p = 5.98e\text{-}10$), but not for 2019 ($p = 0.60$).
  - Similar behavior for data separated by semester.



Results: Retention Analysis

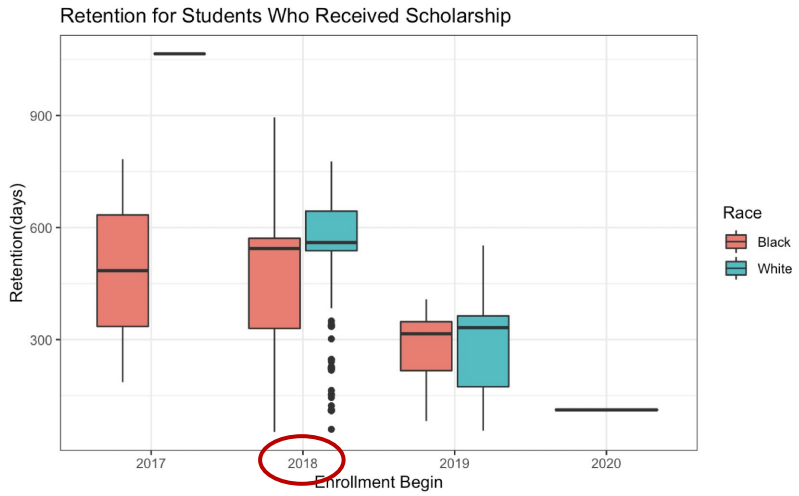
- **EDA for race and retention:**
  - White students tend to have better retention than black students.
  - The differences in retention are both significant for 2018 (p=8.48e-06) & 2019 (p=8.78e-08).

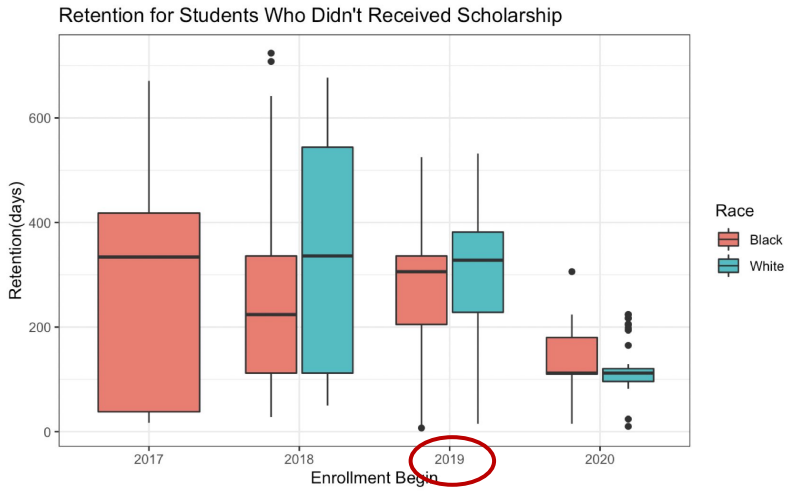


Results: Retention Analysis

- **Interaction(race and scholarship)**

Group	2018	2019
Received Scholarship	**8.96e-05**	0.51
NOT Receive Scholarship	0.1	**6.20e-09**





Results: Retention Analysis

- **Multivariate Regression for Various Ranges**
- **Whole range of GPA & attendance**
  - **Poisson regression for starting college in 2018**
    - Selected predictors: *AttendanceRate, Num_AP, Num_CTE, KeystoneMean, Race, Gender, ELLStatus, IEPGroup, EconDisad, SAT_Total, CumulativeGPA, MagnetInd, EverReceivedPromiseAward, QualifiedforPromise, Semester, Race*EverReceivedPromiseAward*
    - Insignificant variables: *Race (Other)*EverReceivedPromiseAward*
  - **Poisson regression for starting college in 2019**
    - Selected predictors: Same as above except for *IEPGroup*
    - Insignificant variables: *Race (White), EverReceivedPromiseAward*

Results: Retention Analysis

- **Poisson regression for starting college in 2018**
- **Interpretations of important variables:**
  - **10% increase in attendance rate** in high school → **25% increase** in mean retention
  - **10% increase in cumulative GPA** in high school → **1.7% increase** in mean retention
  - Compared to **black** students, **white** students have **6.5% higher** mean retention
  - Students who **received Promise** scholarships have **37% higher** mean retention than students who did not
  - **Male** students have **7.4% lower** mean retention than **female** students
  - **Positive effect of receiving Promise** scholarships on retention for **white** students is **6.3% smaller** than that for **black** students

Results: Retention Analysis

- **Poisson regression for starting college in 2019**
- **Interpretations of important variables:**
  - **10% increase in attendance rate** in high school → **6.7% increase** in mean retention
  - **10% increase in cumulative GPA** in high school → **1.7% increase** in mean retention
  - Compared to **females**, **males** have **5.4% lower** mean retention
  - **Positive effect of receiving Promise** scholarships on retention for **white** students is **5.8% smaller** than that for **black** students
  - **Positive effect of receiving Promise** scholarships on retention for students in **other racial groups** is **37% larger** than that for **black** students

Results: Retention Analysis

- **Box-range Analysis(Year 2018):**
- **T test of retention and scholarship:** aligns with EDA
  - Scholarship has significant effects on all boxes except the smallest box (due to only 34 observations)
  - Interaction term selected by the model
- **Multivariate linear regression:** aligns with t test results

Box	Number of Observations	Variables Selected by Linear Regression(Year 2018)	Significant Effect of Promise
1	34	**Race*EverReceivedPromiseAward**, Gender, ELLStatus	No
2	75	AttendanceRate, CumulativeGPA, SAT_Total, Num_AP, Num_CTE, KeystoneMean, **Race*EverReceivedPromiseAward**, Gender, ELLStatus, EconDisad, MagnetInd, QualifiedforCorePromise	No, but close (p-value = 0.052)
3	150	Same as Box 2	Yes (p-value = 0.007)
4	235	Same as Box 2	Yes (p-value = 0.000)

Results: Retention Analysis

- **Box-range Analysis(Year 2019):**
- **T test of retention and scholarship:** aligns with EDA
  - No significant effect for all boxes
  - Interaction term selected by the model
- **Multivariate linear regression:** aligns with t test results

Box	Number of Observations	Variables Selected by Linear Regression(Year 2019)	Significant Effect of Promise
1	57	AttendanceRate, CumulativeGPA, SAT_Total, Num_AP, Num_CTE, KeystoneMean, **Race*EverReceivedPromiseAward**, Gender, EconDisad, MagnetInd, QualifiedforCorePromise	No
2	110	Box 1 Variables + ELLStatus	No
3	203	Box 1 Variables + ELLStatus	No
4	311	Box 1 Variables + ELLStatus	No

Discussion: Result Summary

- **Retention Analysis:**
  - **Overall**
    - The Promise scholarship tend to help students with their retention in college, the effectiveness varies by race (white < black)
    - Higher attendance rate in high school → higher retention
    - Higher GPA in high school → higher retention
    - Retention in college for female students > that for male students
  - **Students with similar academic performance**
    - The effectiveness seems to be more apparent for senior students

Discussion: Take-home Policy

- Current criteria for the Promise award (cumulative GPA ≥ 2.5 & attendance rate ≥ 90%) is valid → should continue
- The criteria for the Promise award shall be tailored for different racial groups

Discussion: Limitations

- Limited sample size for both research questions
  - Scholarship Analysis: 1708 & 1375
  - Retention Analysis: 1378
- Assume that students no in *Scholarship* data = students not received Promise awards but possibly invalid
- Fitness of poisson models in retention analysis is not ideal
- Limitations affect **generalizability/credibility** of results

Discussion: Next Steps

- **Scholarship Analysis:**
  - Include more variables in the logistic models
  - Include interaction terms in the logistic models
- **Retention Analysis:**
  - Perform two-stage least squares (TSLS) analysis
    - First stage: run a linear regression for *EverReceivedPromiseAward* with a binary indicator for students that passed the criteria & other variables
    - Second stage: replace the *EverReceivedPromiseAward* variable in the regression with the predicted *EverReceivedPromiseAward* in the last stage

Questions & Answers

Technical Appendix

Basic Information of 11 Datasets

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

Technical Appendix

Basic Information of 11 Datasets

Data	Meaning of Data	Number of Observations	Number of Variables
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

Technical Appendix

Variable Definitions

Variable Name	Definition	Dataset
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

Technical Appendix

Variable Definitions

Variable Name	Definition	Dataset
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
SAT_Total(created)	Highest SAT score	SAT
Num_CTE(created)	Number of Career and Technical Education(CTE) Certifications	CTE

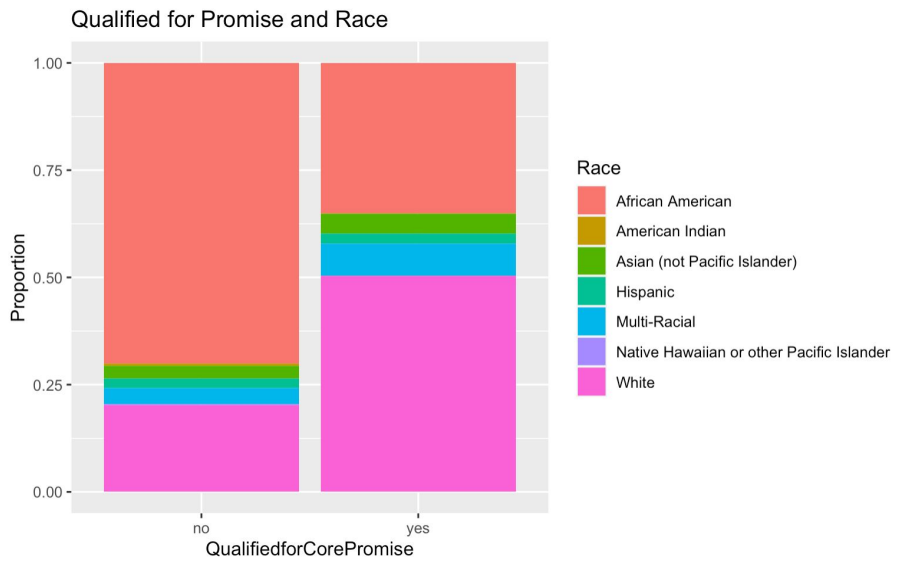
Technical Appendix

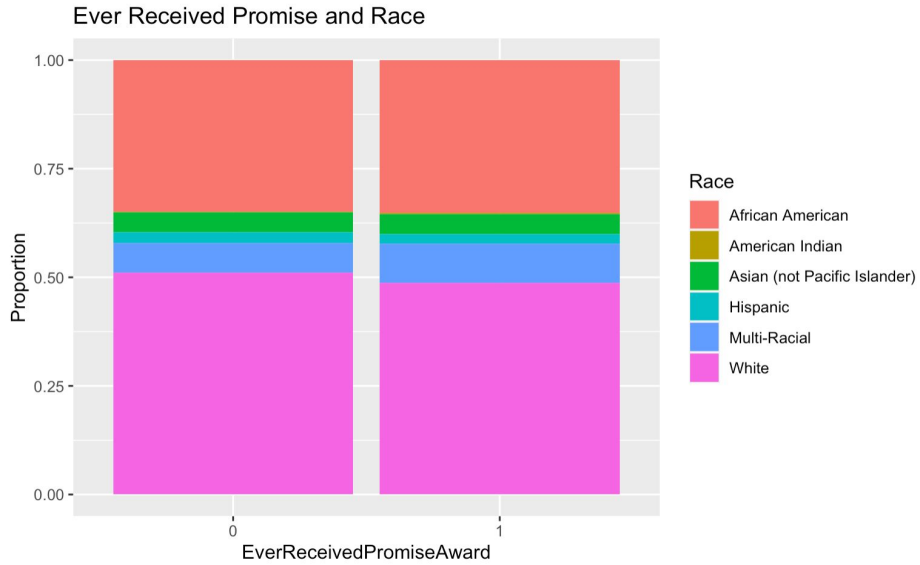
Variable Definitions

Variable Name	Definition	Dataset
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
Semester(created)	Started college in fall/spring semester(binary)	NSC

Technical Appendix

Initial EDA

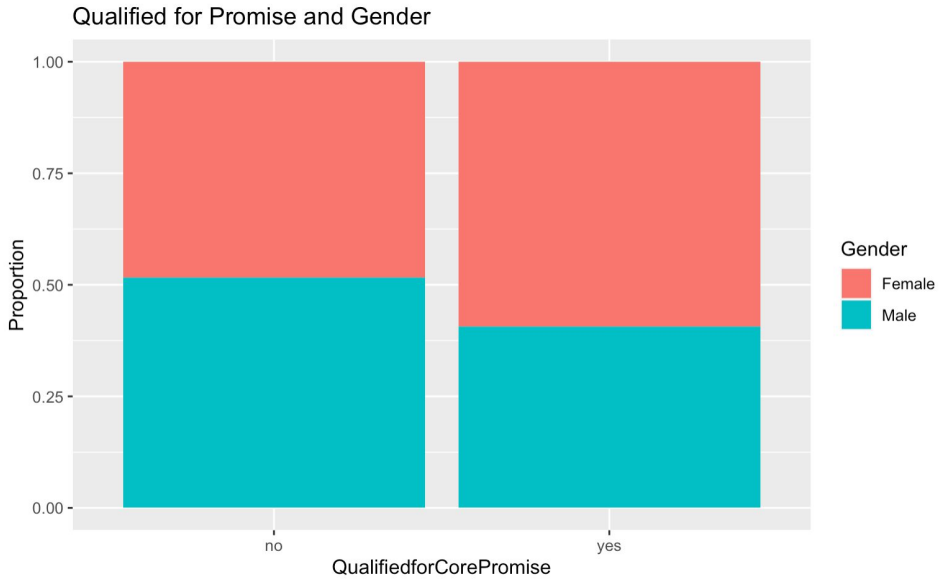


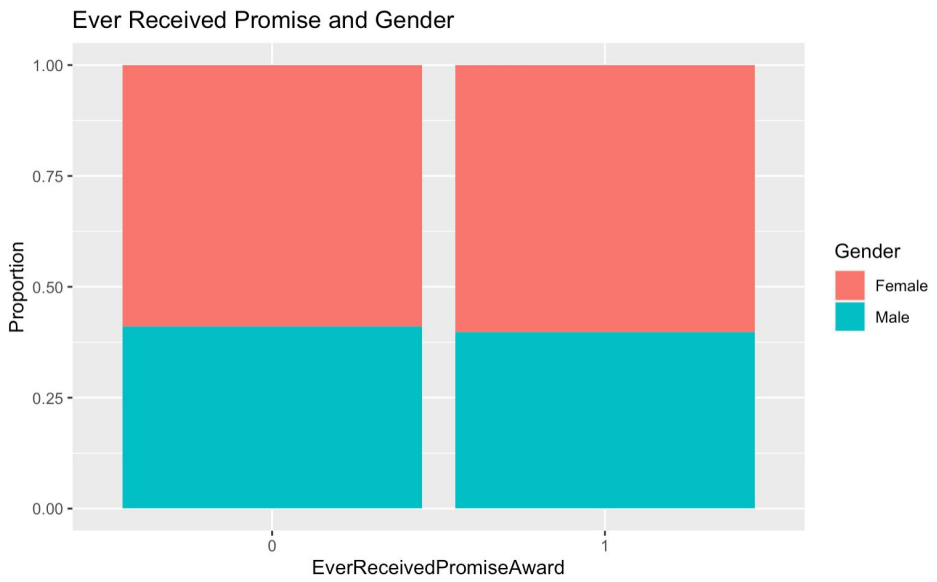


- **Left:** denominator = all students in the joined data set of *Demographics* and *Scholarship*
- **Right:** denominator = qualified students in the joined data set of *Demographics* and *Scholarship*

Technical Appendix

Initial EDA

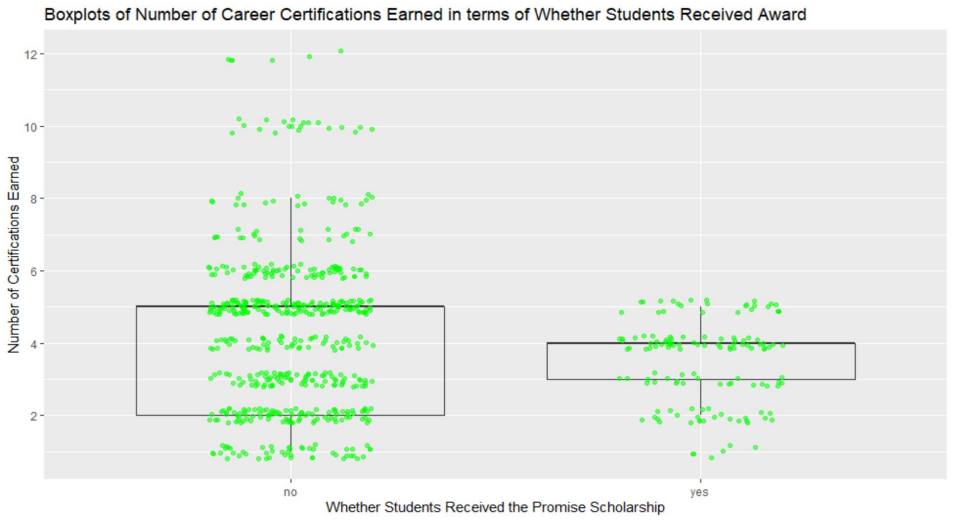


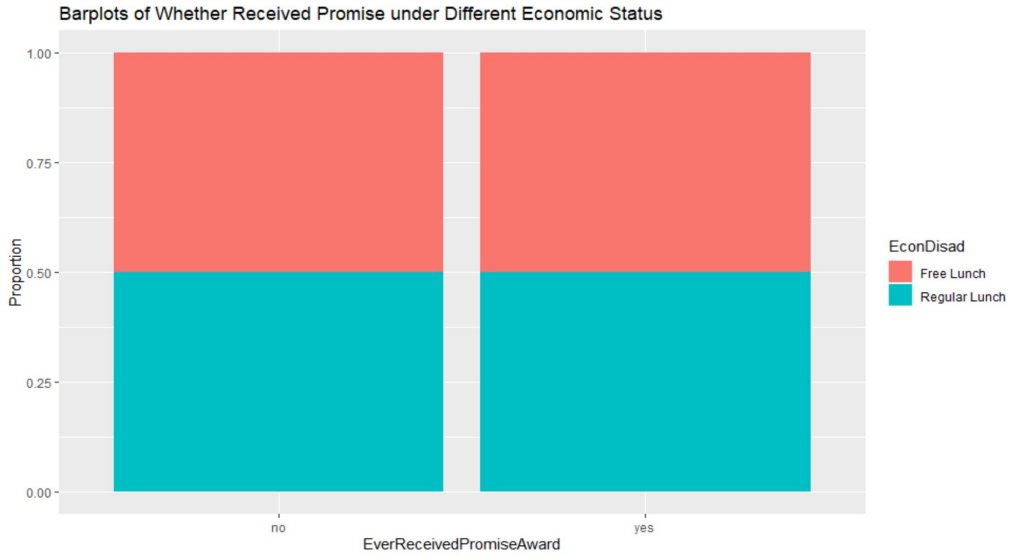


- **Left:** denominator = all students in the joined data set of *Demographics* and *Scholarship*
- **Right:** denominator = qualified students in the joined data set of *Demographics* and *Scholarship*

Technical Appendix

Initial EDA

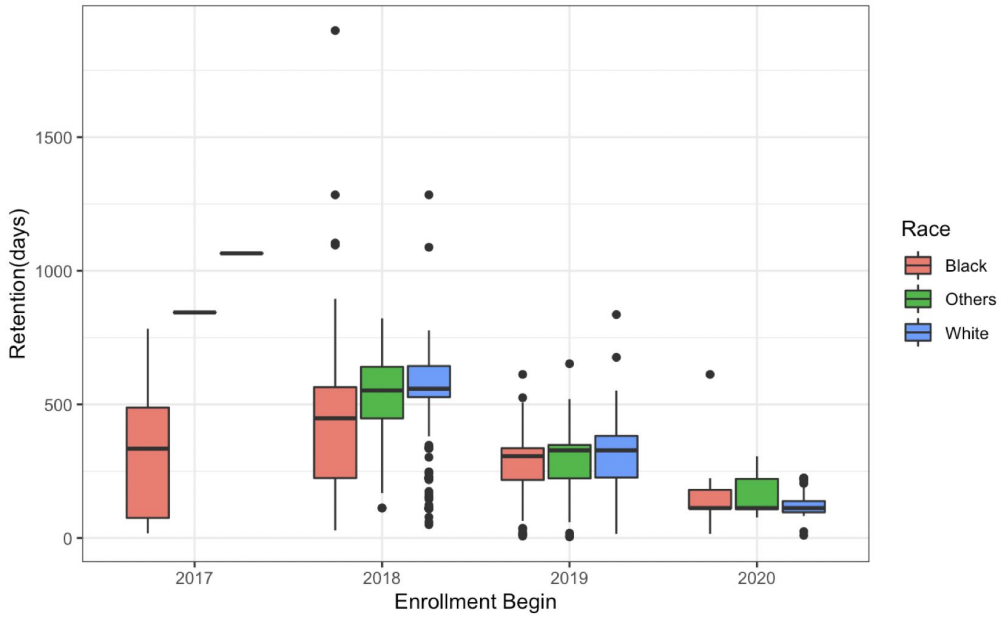




- **Left:** denominator = all students in the joined data set of *CTE* and *Scholarship*
- **Right:** denominator = all students in the joined data set of *Demographics* and *Scholarship*

Technical Appendix

Retention by Race(all 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.

Technical Appendix

T-test Results for Retention Analysis (2018)

	meanDiff	lower	upper	pval	significant
Box1	112.882	-233.952	8.187	0.067	FALSE
Box2	151.920	-228.868	-74.972	0.000	TRUE
Box3	143.685	-209.335	-78.035	0.000	TRUE
Box4	178.061	-229.340	-126.782	0.000	TRUE

T-test Results for Retention Analysis (2019)

	meanDiff	lower	upper	pval	significant
Box1	39.882	-142.535	62.771	0.344	FALSE
Box2	25.170	-94.225	43.885	0.441	FALSE
Box3	43.274	-96.661	10.112	0.105	FALSE
Box4	24.121	-67.447	19.205	0.265	FALSE

Technical Appendix

Whole-Range Retention Analysis: Poisson Regression Output (2018)

Variable	Coefficient	P-values
Intercept	6.110	<2e-16 ***
Attendance Rate	2.246	<2e-16 ***
Num_AP	4.585e-03	1.070e-05 ***
Num_CTE	-2.825e-02	<2e-16 ***
Mean Keystone Scores	-2.166e-03	<2e-16 ***
Race(Other)	4.463e-02	0.029 *
Race(White)	6.300e-02	2.680e-05 ***
Gender(Male)	-7.107e-02	<2e-16 ***
ELL Status(Not in ELL)	-5.165e-02	0.001 **
IEP Group(IEP)	2.181e-02	0.05217 .

 *Notice: significant variables are in **red**.*

Technical Appendix

Whole-Range Retention Analysis: Poisson Regression Output (2018)

Variable	Coefficients	P-Values
IEP Group(Not IEP or Gifted)	-5.117e-02	1.99e-15 ***
EconDisad(Regular Lunch)	1.093e-02	0.01822*
SAT_Total	4.265e-04	<2e-16 ***
Cumulative GPA	1.697e-01	<2e-16 ***
Magnet School Indicator	3.736e-02	1.47e-15 ***
Ever Received Promise Award	3.162e-01	<2e-16 ***
Qualified for Promise(yes)	-4.058e-02	1.76e-05 ***
semester(Spring)	1.279e-01	0.00147***
Race(Other)::Ever Received Promise Award	-4.254e-03	0.84068
Race(White)::Ever Received Promise Award	-4.717e-02	0.00219***

 *Notice: significant variables are in **red**.*

Technical Appendix

Whole-Range Retention Analysis: Poisson Regression Output (2019)

Variable	Coefficient	P-values
Intercept	4.432	<2e-16 ***
Attendance Rate	6.489e-01	2.110e-14 ***
Num_AP	3.395e-03	0.005 **
Num_CTE	-1.305e-02	2.070e-07 ***
Mean Keystone Scores	-2.601e-04	0.016 *
Race(Other)	-3.699e-02	8.710e-06 ***
Race(White)	9.875e-04	0.877
Gender(Male)	-5.275e-02	<2e-16 ***
ELL Status(Not in ELL)	1.950e-01	<2e-16 ***

 *Notice: significant variables are in **red**.*

Technical Appendix

Whole-Range Retention Analysis: Poisson Regression Output (2019)

Variable	Coefficients	P-Values
EconDisad(Regular Lunch)	2.151e-02	3.000e-05 ***
SAT Scores	2.291e-04	<2e-16 ***
Cumulative GPA	1.701e-01	<2e-16 ***
Magnet School Indicator	2.889e-02	1.740e-08***
Ever Received Promise Award	2.282e-02	0.150
Qualified for Promise(yes)	6.454e-02	2.690e-11 ***
Semester(Spring)	-2.563e-01	<2e-16 ***
Race(Other)::Ever Received Promise Award	3.113e-01	<2e-16 ***
Race(White)::Ever Received Promise Award	-5.807e-02	0.005 **

 *Notice: significant variables are in **red**.*