Carnegie Mellon University

PPS Retention/Mobility Research

Client: Steven Greene from Pittsburgh Public Schools
Project Members: Huiyi Guo, Jenny Luo, Yuhang Ying
Project Advisor: Zach Branson

Contents

Introduction

- **Research Question 1:** Investigate factors that influence whether students received Promise scholarship
- **Research Question 2:** Compare students' retention (stays in college) and evaluate factors that influence students' retention

Data: Data Sets Overview

11 Data Sets:

- School Enrollment
- Course Enrollment
- Attendance
- Demographics
- NSC
- SAT
- AP
- GPA
- Keystone
- CTE
- Scholarship

Data: Sample Size

- Sample Size
  - Data for research question 1: 1708 students
  - Data for research question 2:

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

Data: Variable Definitions

Variable Name	Definition	Dataset
RandomID	Unique student ID	
QualifiedforCorePromise	Eligibility to 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

Notice: black variables used only in research question 1; red variables used only in research question 2; green variables used in both questions.

Carnegie Mellon University

Data: 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

Notice: black variables used only in research question 1; red variables used only in research question 2; green variables used in both questions.

Carnegie Mellon University

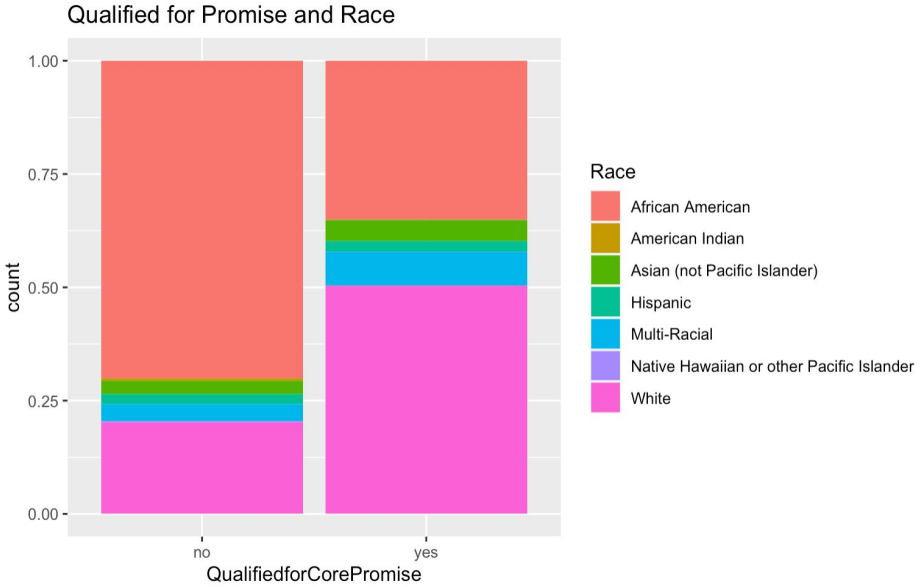
Data: Variable Definitions

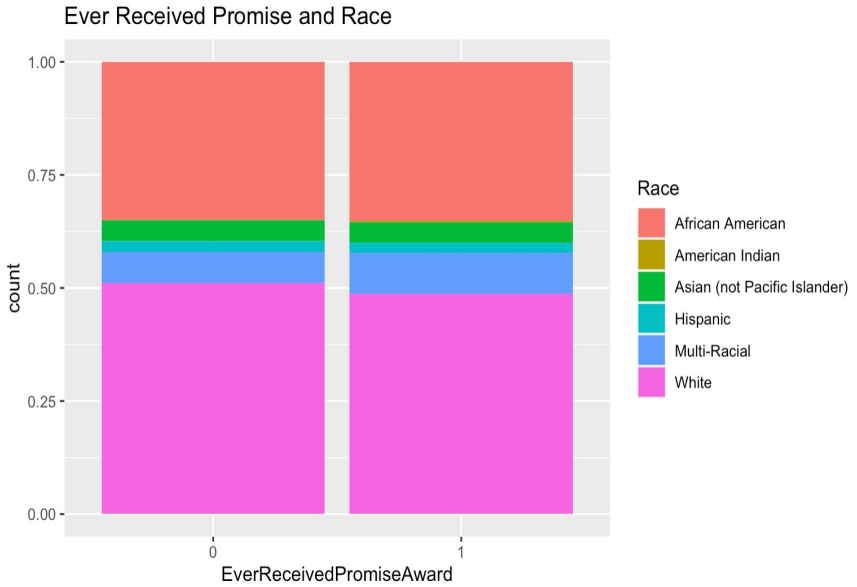
Variable Name	Definition	Dataset
MagnetInd	Whether students go to magnet schools(binary)	Enrollment
GradYear	Year in which students graduated from high school	Scholarship
Enrollment_Begin	When a student enrolled in a college semester	NSC
Enrollment_End	When the college semester ended	NSC
College_State	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

Notice: black variables used only in research question 1; red variables used only in research question 2; green variables used in both questions.

Carnegie Mellon University

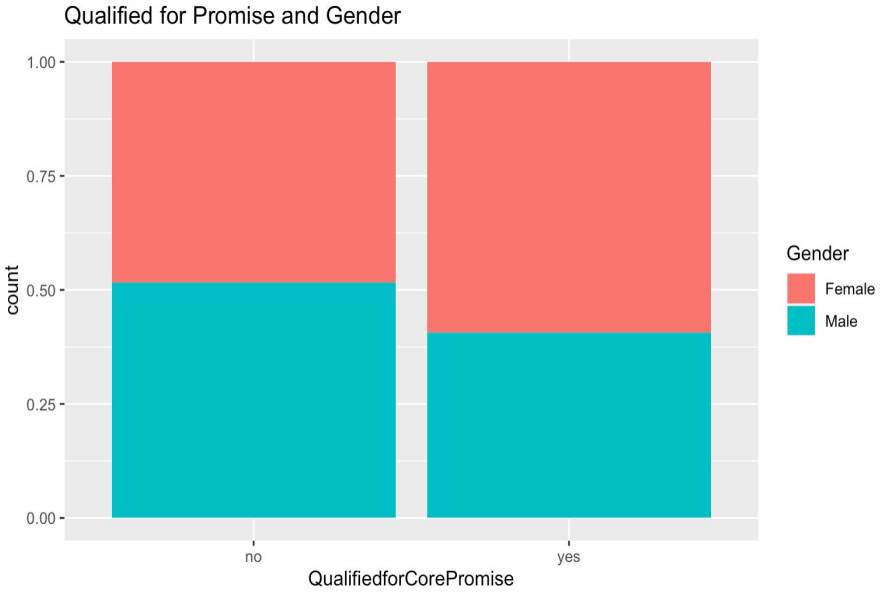
Data: Initial EDA

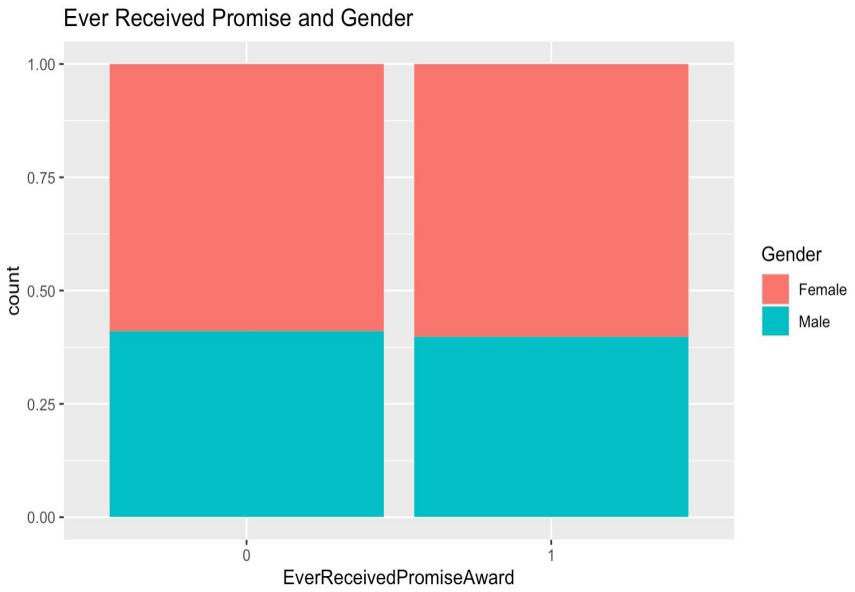




Carnegie Mellon University

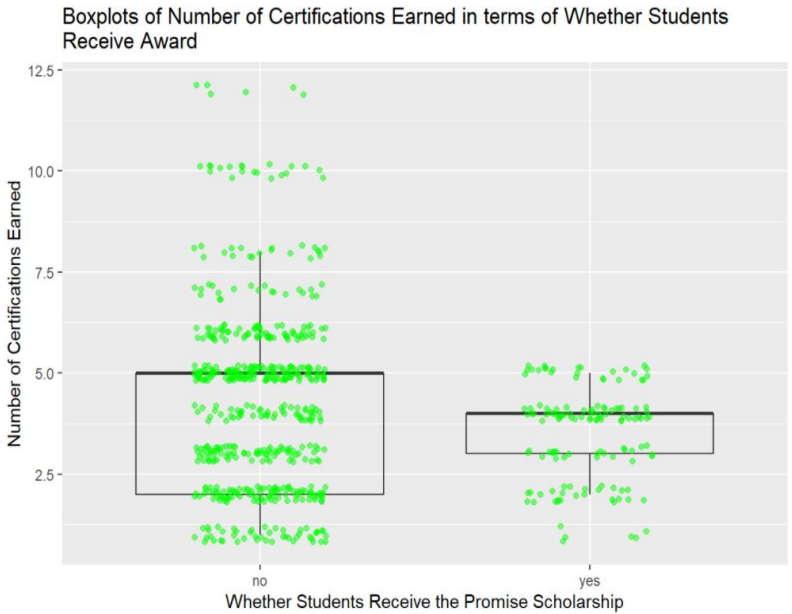
Data: Initial EDA

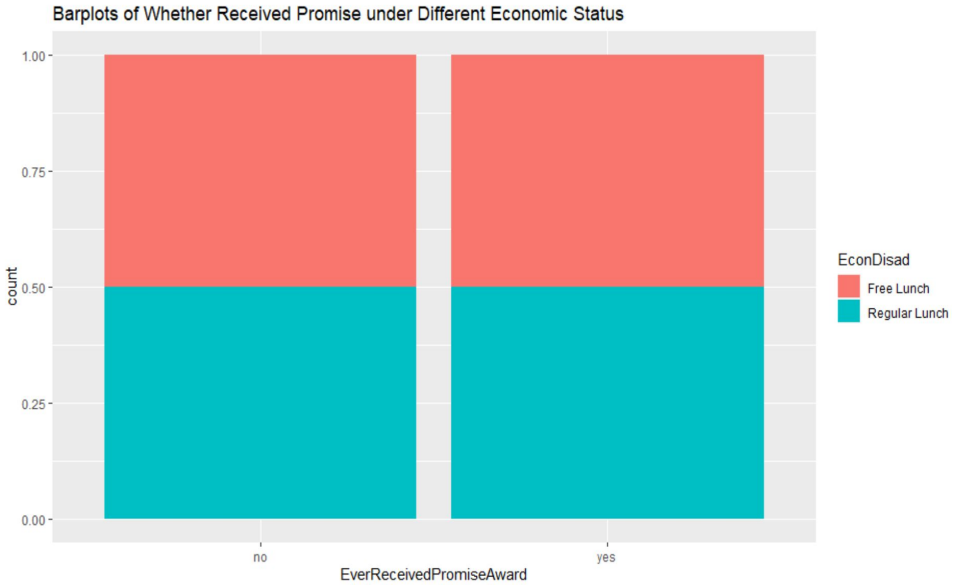




Carnegie Mellon University

Data: Initial EDA





Carnegie Mellon University

Methods

- Use R programming language and environment
- Preliminary EDA methods
  - Summarize all metrics for senior students, one row per student
  - GPA x Attendance x Scholarship analysis
    - Analyze students' eligibility cutoff
  - Demographic x Scholarship and CTE x Scholarship analyses
    - Evaluate how eligibility/received relate to different demographic groups
    - Evaluate how eligibility/received relate to #career certifications

Methods

- Modeling methods for research question 1:
  - Logistic regression to analyze factors that influence whether students received scholarship
    - Outcome variable: EverReceivedPromiseAward
    - Predictors: AttendanceRate, Num_AP, Num_CTE, KeystoneMean, Race, Gender, ELLStatus, IEPGroup, EconDisab, SAT_Total, CumulativeGPA, MagnetInd
    - Stepwise variable selection on AIC
    - Only consider students going to PA colleges

Methods

- Modeling methods for research question 2:
  - Analyze students' retention between different groups
    - Only consider students who went to PA colleges
    - Retention = number of days enrolled in PA college (y-axis)
    - Group by the year a student started college (x-axis)
    - Two comparisons:
      - Between students who received scholarship and who did not
      - Between black and white students
  - Run statistical tests to evaluate the significance of differences
    - Bartlett tests to evaluate homogeneity of variances between groups
    - One-way t tests to check for significance of differences

Results: Research Question 1

- Logistic Regression
  - Base model:
    - Response variable: EverReceivedPromiseAward
    - Predictor variables: Race, Gender
  - Full model:
    - Response variable: EverReceivedPromiseAward
    - Predictor variables: AttendanceRate, Num_AP, Num_CTE, KeystoneMean, Race, Gender, ELLStatus, IEPGroup, EconDisad, SAT_Total, CumulativeGPA, MagnetInd
  - Stepwise variable selection on AIC

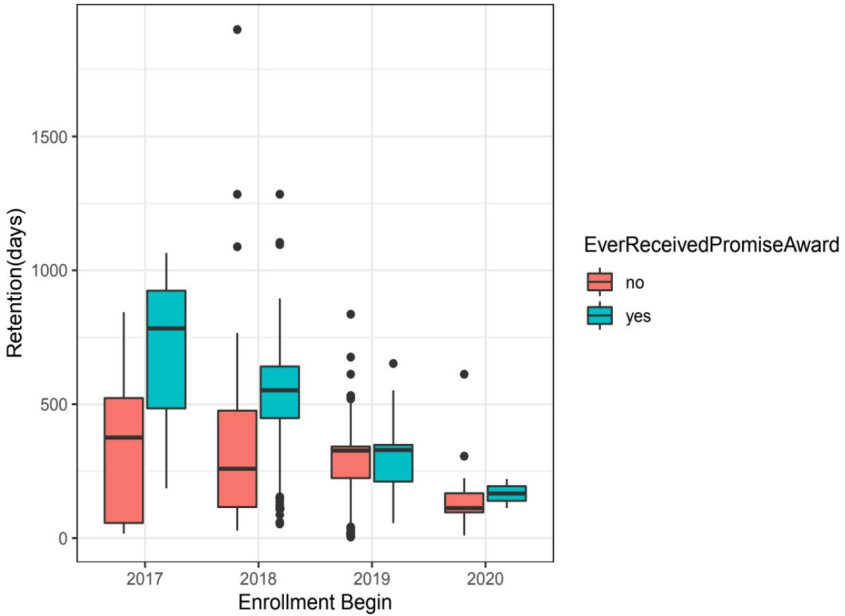
Results: Research Question 1

- Logistic Regression
  - After performing stepwise on AIC:

Variable	Coefficient	p-value
y-intercept	-3.6280	0.207566
RaceAmerican Indian	0.4487	0.703162
RaceAsian (not Pacific Islander)	-0.1923	0.542354
RaceHispanic	-0.2951	0.454827
RaceMulti-Racial	0.2677	0.243622
RaceNative Hawaiian or other Pacific Islander	-9.9932	0.975451
RaceWhite	-0.2759	0.060262
GenderMale	-0.0924	0.430571
CumulativeGPA	0.9195	7.93e-09 ***
AttendanceRate	8.2030	5.79e-05 ***
KeystoneMean	-0.0060	0.000389 ***
ELLStatusNot in ELL	0.9526	0.035397 *
MagnetInd1	0.2322	0.046337 *

Results: Research Question 2

(a) Compare retention between students who *received scholarships* and who *did not*



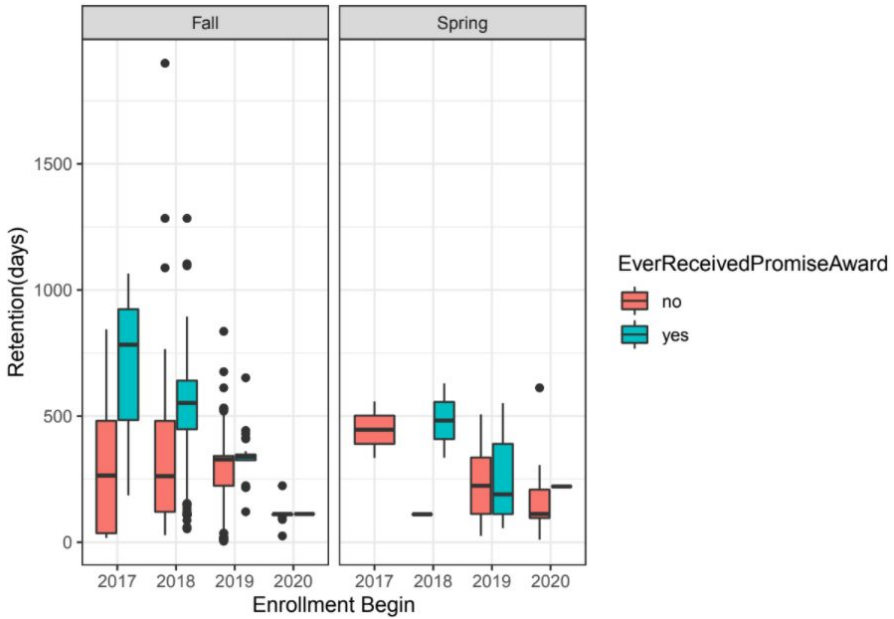
Findings:

1. Students who received scholarships tend to have better retention except for 2019.
2. Notice that we have only 13 obs for 2017 and 93 obs for 2020.
3. The difference in retention is significant for 2018($p = 5.98e\text{-}10$) but not for 2019($p = 0.60$).

Carnegie Mellon University

Results: Research Question 2

(a) Get a closer look by the *semester* of enrollment



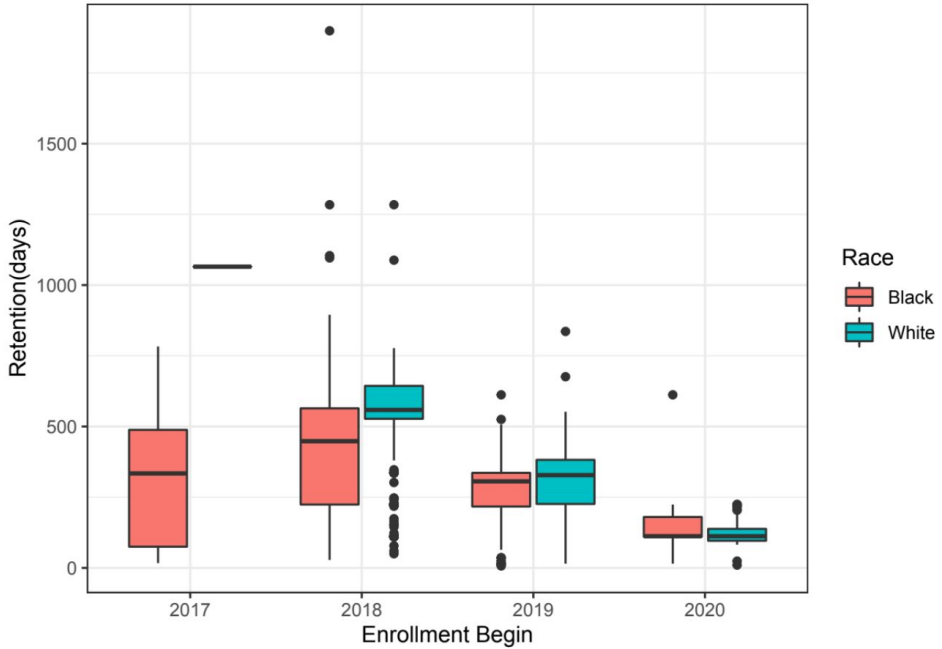
Findings:

1. Students who received scholarships tend to have better retention for Fall 2017 & Fall 2018.
2. We have limited obs for 2017 & 2020 and Spring semesters.
3. Similar behavior in retention difference for 2018 & 2019.

Carnegie Mellon University

Results: Research Question 2

(b) Compare retention between *black* students and *white* students



Findings:

1. White students tend to have better retention than black students except for 2020.
2. Similarly we don't have many obs for 2017 & 2020.
3. The differences in retention are both significant for 2018($p = 8.48e\text{-}06$) & 2019($p = 8.78e\text{-}08$).

Carnegie Mellon University

Discussion

- Next Steps:
  - Research question 1:
    - Include more predictor variables
    - Conduct ANOVA for further model selection
    - Explore the coefficient for KeystoneMean
    - Perform model diagnostics
  - Research question 2:
    - Compare retention difference for different races and receipt of the scholarship, conditioned on students around the qualification cutoff(GPA=2.5, Attendance = 0.9)
    - Linear regression to explore relationship between retention and other factors (e.g. GPA, attendance, gender...)

Discussion

- Limitations:
  - Small sample size for both research questions
  - Unsure about how to interpret students who are not in the scholarship data
    - Students not in the scholarship data = students did not receive Promise?

Questions?

Significance Tests for RQ2

- Retention vs. Receipt of Promise scholarship (by semester)
  - Bartlett test:
    - Same variance for students who started college in 2018 (p-value = 8.7e-10) → var.equal = TRUE in t test
    - Different variance for students who started college in 2019 (p-value = 0.1382) → var.equal = FALSE in t test
  - Welch t test:
    - Statistically significant difference in retention for students who started college in 2018 (p-value = 5.98e-10)
    - Not statistically significant difference in retention for students who started college in 2019 (p-value = 0.6023)

Significance Tests for RQ2

- Retention vs. Race
  - Only consider balck and white since they are the majority
  - Bartlett test:
    - Same variance for students who started college in 2018 (p-value = 0.0008769) → var.equal = TRUE in t test
    - Different variance for students who started college in 2019 (p-value = 0.7802) → var.equal = FALSE in t test
  - Welch t test:
    - Statistically significant difference in retention for students who started college in either 2018 or 2019 (p-value = 8.48e-6 or 8.782e-8 respectively)