**NSC Subsequent Enrollment Data**

**Data Cleaning**

The below text outlines our methodology related to data cleaning for our analysis of National Student Clearinghouse Subsequent Enrollment data.

**Missing records.** Students with "No record found" by our National Student Clearinghouse Subsequent Enrollment file were removed (1.1% of total rows).

**Incorrect Institution type.** Any records of enrollment of 4-year ("4") higher education institutions (Column L) were retained; in contrast, enrollment records of students enrolled in 2-year (“2) or "less than 2 years" ("L" in Column L) were removed (6.2% of total rows). Future improvements to this analysis include retaining enrollment for 2-year institutions.

**Missing enrollment dates.** Any records that were missing enrollment start and/or end dates were removed (6.3% of total rows). Note that this percentage may be higher than expected; this is because it includes incomplete enrollment records from the most current semester, as of analysis (Spring 2023).

**Inappropriately early enrollment start date.** Student rows with enrollment start dates that chronologically occurred before the requested search start date (August 15 of student’s entry year) were excluded from analysis (0.3% of total rows). It appears as if search dates function not as “start dates” (i.e. only including records where enrollment started on or after the date), but rather as “enrollment dates” (i.e. returning enrollment data where the student was enrolled during the search date).

**First term earlier (or later) than anticipated first term.** Students who had no record of enrollment during their anticipated start term were removed from analysis (11.1% of total rows). This number was higher than expected; however, it may be accounted for by considering our analysis excludes student a) who chose to enroll in 2-year colleges instead of 4-year institutions and/or b) delayed enrollment, either via “gap year” or something similar (e.g. during COVID).

**Multiple enrollment records in one semester from same institution.** While it is common practice for institutions in NSC SE data to have one row per student per semester, some institutions will report more than one row of enrollment records during this time. In this situation, we utilized two different techniques, depending on whether the rows contained same or different enrollment statuses (e.g. full-time). See below:

***Different enrollment statuses.*** If a student had 2 or more rows indicating enrollment at the same institution in the same semester but with different enrollment status (Column P) (1.5% of total rows), a hierarchy was created to prioritize classifications. Enrollment hierarchy is listed in descending order of rank:

* + W (Withdraw)
  + F (Full-time)
  + Q (3/4 time)
  + H (1/2 time)
  + L (Less than ½ time)
  + A (Approved leave of absence)
  + D (Deceased)
  + [missing]

***Same enrollment status.*** When students had the same enrollment status at the same institution during the same semester (4.8% of total rows), the row with the earliest date was retained, while all other rows were deleted.

**Multiple enrollment records in one semester from different institutions.** In cases where there is a record of a student enrolling in more than one institution in the same semester, we retained the institution that had the earliest start date

**Determining semester dates.** Based on a descriptive analysis of numerous fall semester start dates across 4-year institutions, we selected August 15th as the default start date for fall enrollment in our analysis. We also selected December 31th as the default end date for fall enrollment. In other words, any enrollment record starting on or after August 15th andending on or before December 31st was classified as a “fall semester”. For spring, we decided on January 1st as a start date, and May 7th as an end date.

**Institution-Level Data**

**Methodology**

Institution-level data (e.g. net cost, median SAT score of enrollees) were obtained from College Scorecard[[1]](#footnote-1).

**Selected variables.** The following variables (with associated descriptions) from College Scorecard selected for inclusion are described below:

|  |  |
| --- | --- |
| **Variable** | **Description** |
| main\_campus | Flag for main campus |
| location.lat | Latitude |
| location.lon | Longitude |
| carnegie\_basic | Carnegie Classification – basic |
| carnegie\_undergrad | Carnegie Classification – undergraduate profile |
| carnegie\_size\_setting | Carnegie Classification – size and setting |
| minority\_serving.historically\_black | Flag for Historically Black College and University |
| minority\_serving.predominantly\_black | Flag for predominantly black institution |
| minority\_serving.hispanic | Flag for Hispanic-serving institution |
| men\_only | Flag for men-only college |
| women\_only | Flag for women-only college |
| religious\_affiliation | Religious affiliation of the institution |
| admission\_rate.overall | Admission rate |
| sat\_scores.midpoint.critical\_reading | Midpoint of SAT scores at the institution (critical reading) |
| sat\_scores.midpoint.math | Midpoint of SAT scores at the institution (math) |
| demographics.race\_ethnicity.white | Total share of enrollment of undergraduate degree-seeking students who are white |
| demographics.race\_ethnicity.black | Total share of enrollment of undergraduate degree-seeking students who are black |
| demographics.race\_ethnicity.hispanic | Total share of enrollment of undergraduate degree-seeking students who are Hispanic |
| demographics.race\_ethnicity.asian | Total share of enrollment of undergraduate degree-seeking students who are Asian |
| demographics.race\_ethnicity.non\_resident\_alien | Total share of enrollment of undergraduate degree-seeking students who are non-resident aliens |
| part\_time\_share | Share of undergraduate, degree-/certificate-seeking students who are part-time |
| tuition.in\_state | In-state tuition and fees |
| tuition.out\_of\_state | Out-of-state tuition and fees |
| instructional\_expenditure\_per\_fte | Instructional expenditures per full-time equivalent student |
| faculty\_salary | Average faculty salary |
| ft\_faculty\_rate | Proportion of faculty that is full-time |
| pell\_grant\_rate | Percentage of undergraduates who receive a Pell Grant |
| share\_first.time\_full.time | Share of entering undergraduate students who are first-time, full-time degree-/certificate-seeking undergraduate students |
| demographics.over\_23\_at\_entry | Percent of students over 23 at entry |
| demographics.female\_share | Share of female students |
| demographics.median\_family\_income | Median family income |
| demographics.share\_white.home\_ZIP | Percent of the population from students’ zip codes that is White, via Census data |
| demographics.share\_bachelors\_degree\_age25.home\_ZIP | Percent of the population from students’ zip codes with a bachelor’s degree over the age 25, via Census data |

**Geographic Data**

**Distance from institution**. We recorded applicants’ distances from their chosen institution (in miles) by computing distance between latitude/longitude of each applicant’s home zip code and latitude/longitude of the institution. Applicants from non-US countries were left blank.

**Socioeconomic variables.** We recorded several categories of data based on applicants’ home zip codes, obtained from Social Explorer and ESRI data. Applicants from non-US countries were left blank.

|  |  |
| --- | --- |
| **Variable** | **Description** |
| zip\_democrat | Area's level of Democratic political affiliation relative to the national level. Numbers higher than 100 represents higher affiliation than the national average, and a value of less than 100 represents lower Democratic affiliation than the national average. For example, an index of 120 implies that Democratic affiliation in the area is 20 percent higher than the US average. |
| zip\_pop\_density | Population density; persons per square mile in area. |
| zip\_wealth\_index | Area's level of wealth relative to the national level. Numbers higher than 100 represents higher wealth than the national average, and a value of less than 100 represents lower wealth than the national average. For example, an index of 120 implies that wealth in the area is 20 percent higher than the US average. |
| zip\_diversity\_index | 0 to 100 score, representing the likelihood that two persons, chosen at random from the same area, belong to different races or ethnic groups. |
| zip\_esri\_life\_mode | Esri generated classification of areas into 14 distinct "LifeMode" groups, based on a variety of behavioral and demographic characteristics. |
| zip\_esri\_segment | Esri generated classification of areas into 67 subgroups (compared to 14 main groups), based on similar criteria. |

**Primary citizenship.** We included information on applicants’ primary country of citizenship. For parsimony, we combined all countries with fewer than 20 applicants into a single category.

**Admissions Data**

Admissions data were used in this analysis. Unfortunately, due to confidentiality needs, details will not be provided.

**Analyzing National Student Clearinghouse’s Subsequent Enrollment File:**

**R Code for Subsequent Enrollment (SE) file**

library**(**tidyverse**)**

library**(**openxlsx**)**

library**(**janitor**)**

library**(**lubridate**)**

library**(**gtools**)**

library**(**rstudioapi**)**

# Getting the path of your current open file

current\_path **=** rstudioapi**::**getActiveDocumentContext**()$**path

setwd**(**dirname**(**current\_path **))**

print**(** getwd**()** **)**

#########################################################################

####Import National Student Clearinghouse data

#########################################################################

#Select location of NSC data file (e.g. Subsequent Enrollment)

file\_path **<-** file.choose**()**

#Read csv file

n **<-** read.csv**(**file\_path**)**

#n <- read.csv("C:/Users/sherrins/OneDrive - Wentworth Institute of Technology/Data Sources/NSC/Results from NSC/Subsequent Enrollment/Subsequent Enrollment - UG Admits (Fall 2015 to present).csv")

#Clean column names

n **<-** clean\_names**(**n**)**

#Select relevant columns only

n **<-** n %>%

filter**(**record\_found\_y\_n **==** "Y"**)** %>% #Only keep students whose records were found

select**(**requester\_return\_field, #Unique identifier of student

search\_date, #Subsequent Enrollment dates requested

enrollment\_begin, #Date student was enrolled at school

enrollment\_end,

enrollment\_status, #E.g. full-time

enrollment\_major\_1,

college\_code\_branch,

college\_name,

x2\_year\_4\_year**)**

#Keep only 4 year colleges

#(For ease in data analysis)

n **<-** n %>%

filter**(**x2\_year\_4\_year **==** "4"**)**

n **<-** n %>%

select**(-**x2\_year\_4\_year**)**

#Remove duplicate rows

n **<-** unique**(**n**)**

#Remove rows with missing enrollment dates

n **<-** n %>%

filter**(**is.na**(**n**$**enrollment\_begin**)** **==** **FALSE**,

is.na**(**n**$**enrollment\_end**)** **==** **FALSE)**

#Convert enrollment date to better format for analysis

convert\_date **<-** **function(**date\_string**){**

year **<-** substr**(**date\_string, 1, 4**)**

month **<-** substr**(**date\_string, 5, 6**)**

day **<-** substr**(**date\_string, 7, 8**)**

date\_string **<-** paste0**(**month,"-",day,"-",year**)**

date\_string **<-** mdy**(**date\_string**)**

return**(**date\_string**)**

**}**

n**$**search\_date **<-** convert\_date**(**n**$**search\_date**)**

n**$**enrollment\_begin **<-** convert\_date**(**n**$**enrollment\_begin**)**

n**$**enrollment\_end **<-** convert\_date**(**n**$**enrollment\_end**)**

#Find date in middle of beginning and end of enrollment

n**$**enrollment\_middle **<-** n**$**enrollment\_begin **+** **((**n**$**enrollment\_end **-** n**$**enrollment\_begin**)** **/** 2**)**

#Remove records where enrollment started before search date

n **<-** n %>%

filter**(**enrollment\_begin **>=** search\_date**)**

#Add date information re: academic terms

a **<-** read.xlsx**(**"data for code/Codebook.xlsx",

sheet **=** "Academic Terms - By Day"**)**

a **<-** clean\_names**(**a**)**

a**$**date\_by\_day **<-** mdy**(**a**$**date\_by\_day**)**

colnames**(**a**)** **<-** c**(**"date","term\_nsc"**)**

#Add date info to dataframe

n **<-** merge**(**n,a,

by.x **=** "enrollment\_begin",

by.y **=** "date",

all.x **=** **TRUE)**

a **<-** read.xlsx**(**"data for code/Codebook.xlsx",

sheet **=** "Academic Terms"**)**

a **<-** clean\_names**(**a**)**

n **<-** merge**(**n,a,

by.x **=** "term\_nsc",

by.y **=** "course\_term\_desc",

all.x **=** **TRUE)**

n **<-** n %>%

rename**(**nsc\_term\_sequence **=** term\_sequence**)**

rm**(**a**)**

#Remove rows with missing schools

#(This should be very rare.)

n **<-** n %>%

filter**(**is.na**(**college\_code\_branch**)** **==** **FALSE)**

#######################################################################

###Calculate fall-to-fall retention

#######################################################################

#Create "enrollment status" hierarchy

#This will be important later

n**$**enroll\_hierarchy **<-** ifelse**(**n**$**enrollment\_status **==** "W",1, #Withdraw from institution

ifelse**(**n**$**enrollment\_status **==** "F", 2, #Full-Time

ifelse**(**n**$**enrollment\_status **==** "Q", 3, #3/4 time

ifelse**(**n**$**enrollment\_status **==** "H", 4, #1/2 time

ifelse**(**n**$**enrollment\_status **==** "L", 5, #Less than 1/2 time

ifelse**(**n**$**enrollment\_status **==** "A", 6, #Approved leave of absence

ifelse**(**n**$**enrollment\_status **==** "D", 7, #Deceased

ifelse**(**n**$**enrollment\_status **==** "",8, #No status

9**))))))))**

#If student has 2+ records in term with DIFFERENT ENROLLMENT STATUS from SAME school,

#only keep one row, according to "enrollment status" hierarchy.

n **<-** n %>%

group\_by**(**requester\_return\_field, #E.g. Jane Doe, Fall 2015, Worcester Polytechnic Institute

nsc\_term\_sequence,

college\_code\_branch**)** %>%

mutate**(**top\_status **=** min**(**enroll\_hierarchy**))** %>% #Find most important enrollment status for each student

filter**(**enroll\_hierarchy **==** top\_status**)** %>% #Keep only row(s) with most important enrollment status for each student

ungroup**()**

#Resolve instances where school reports same enrollment status

#multiple times for single student

n **<-** n %>%

group\_by**(**requester\_return\_field, #E.g. Jane Doe, Fall 2015, Worcester Polytechnic Institute

nsc\_term\_sequence,

college\_code\_branch**)** %>%

filter**(**row\_number**()** **==** 1**)** #keep only first row of each student

#Resolve instances where student attended more than one school in 1st term

#Let's keep only the earliest college attended in that term

#Get students' first college terms on record

n **<-** n %>%

group\_by**(**requester\_return\_field**)** %>%

mutate**(**nsc\_start\_term **=** min**(**nsc\_term\_sequence, na.rm **=** **TRUE))** %>%

ungroup**()**

n1 **<-** n %>%

filter**(**nsc\_term\_sequence **==** nsc\_start\_term**)** %>%

group\_by**(**requester\_return\_field**)** %>%

arrange**(**enrollment\_begin**)** %>%

filter**(**row\_number**()** **==** 1**)**

n1 **<-** n1 %>%

select**(**requester\_return\_field,

nsc\_first\_college\_attended **=** college\_code\_branch**)**

#Add students' first college to main data

n **<-** merge**(**n, n1,

by **=** "requester\_return\_field",

all.x **=** **TRUE)**

#Indicate whether student attended college during specific term indicating fall-to-fall retention

n**$**fall\_to\_fall **<-** ifelse**(**n**$**nsc\_term\_sequence **==** n**$**nsc\_start\_term **+** 3 **&** #If term is next fall...

n**$**college\_code\_branch **==** n**$**nsc\_first\_college\_attended **&** #...and they're still at same school...

n**$**enrollment\_status **!=** "W", #...and they're not recorded as Withdraw, then count them as retained.

1, 0**)**

#Indicate whether student persisted in college, but at DIFFERENT college than first one

n**$**fall\_to\_fall\_other\_school **<-** ifelse**(**n**$**nsc\_term\_sequence **==** n**$**nsc\_start\_term **+** 3 **&** #If term is next fall...

n**$**college\_code\_branch **!=** n**$**nsc\_first\_college\_attended **&** #...and they're NOT at same school...

n**$**enrollment\_status **!=** "W", #...and they're not recorded as Withdraw, then count them as retained.

1, 0**)**

#Create summary dataframe with one row per student

nn **<-** n %>%

group\_by**(**requester\_return\_field**)** %>%

summarise**(**nsc\_fall\_to\_fall **=** max**(**fall\_to\_fall**)**,

nsc\_fall\_to\_fall\_other\_school **=** max**(**fall\_to\_fall\_other\_school**)**,

#nsc\_first\_college\_attended = nsc\_first\_college\_attended) %>%

nsc\_first\_college\_attended **=** nsc\_first\_college\_attended**)** %>%

unique**()**

###############################################################

#Add university/college data (from College Scorecard)

###############################################################

#Read College Scorecard

c **<-** read.xlsx**(**"data for code/college scorecard.xlsx"**)**

#Rename column

colnames**(**c**)[**4**]** **=** "name"

#Select columns

c **<-** c %>%

select**(**ope8\_id, name, main\_campus,

location.lat, location.lon,

carnegie\_basic, carnegie\_undergrad, carnegie\_size\_setting,

minority\_serving.historically\_black, minority\_serving.predominantly\_black, minority\_serving.hispanic,

men\_only, women\_only, religious\_affiliation,

sat\_scores.midpoint.math, sat\_scores.midpoint.critical\_reading,

admission\_rate.overall,

demographics.race\_ethnicity.white**:**demographics.race\_ethnicity.non\_resident\_alien,

part\_time\_share,

tuition.in\_state, tuition.out\_of\_state,

instructional\_expenditure\_per\_fte, faculty\_salary,

ft\_faculty\_rate,

pell\_grant\_rate,

share\_firstgeneration,

demographics.female\_share,

demographics.over\_23\_at\_entry,

demographics.median\_family\_income,

demographics.share\_white.home\_ZIP,

demographics.share\_bachelors\_degree\_age25.home\_ZIP**)**

#Change "NULL" to missing/na

c**[**c **==** "NULL"**]** **<-** **NA**

#plot\_missing(c)

# cc <- c %>%

# group\_by(ope8\_id) %>% tally()

# Add a new column "na" to `c` containing the count of missing values per row

c **<-** c %>%

mutate**(**na **=** rowSums**(**is.na**(**.**)))**

# Create dataframe with the minimum number of missing values ("fewest\_nas") and the number of rows ("n") per institution (some ope8\_ids have multiple institutions)

cc **<-** c %>%

group\_by**(**ope8\_id**)** %>%

summarise**(**fewest\_nas **=** min**(**na**)**,

n **=** n**())**

#Merge data

c **<-** merge**(**c, cc,

by **=** "ope8\_id",

all.x **=** **TRUE)**

#Keep rows with the minimum number of missing values

c **<-** c %>%

filter**(**na **==** fewest\_nas**)**

# Specify columns to be converted to factors and numeric

columns\_to\_factor **<-** c**(**"main\_campus", "carnegie\_basic", "carnegie\_size\_setting", "carnegie\_undergrad"**)**

columns\_to\_numeric **<-** c**(**"location.lat", "location.lon", "sat\_scores.midpoint.critical\_reading", "sat\_scores.midpoint.math", "admission\_rate.overall",

"demographics.race\_ethnicity.white", "demographics.race\_ethnicity.black", "demographics.race\_ethnicity.hispanic",

"demographics.race\_ethnicity.asian", "demographics.race\_ethnicity.non\_resident\_alien", "part\_time\_share", "tuition.in\_state",

"tuition.out\_of\_state", "instructional\_expenditure\_per\_fte", "faculty\_salary", "ft\_faculty\_rate", "pell\_grant\_rate",

"share\_firstgeneration", "demographics.female\_share", "demographics.median\_family\_income", "demographics.over\_23\_at\_entry",

"demographics.share\_bachelors\_degree\_age25.home\_ZIP", "demographics.share\_white.home\_ZIP"**)**

# Convert specified columns to factors and numeric

c **<-** c %>%

mutate**(**across**(**columns\_to\_factor, as.factor**)**,

across**(**columns\_to\_numeric, as.numeric**))**

# Merge with main data

nn **<-** merge**(**nn, c,

by.x **=** "nsc\_first\_college\_attended",

by.y **=** "ope8\_id",

all.x **=** **TRUE)**

# Export data

write.xlsx**(**nn, "results/nsc and college scorecard.xlsx"**)**

1. Downloaded December 13, 2022 from <https://collegescorecard.ed.gov/data> [↑](#footnote-ref-1)