NSC Subsequent Enrollment Data

Data Cleaning

The below text outlines our methodology related to exclusion criteria 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

Data Methodology

The below text outlines our methodology related to data analysis of retention outcomes using the National Student Clearinghouse Subsequent Enrollment data.

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 and ending 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].

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

[1] Downloaded December 13, 2022 from https://collegescorecard.ed.gov/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

The below text outlines our methodology related to cleaning and analyzing our Admissions data.

Data manipulation. Applicant race/ethnicity data were reclassified into fewer categories, to increase statistical power for certain analyses. The old (IPEDS) and new classification scheme is shown below:

IPEDS CLASSIFICATION	NEW CLASSIFICATION
AMERICAN INDIAN OR ALASKA NATIVE	Student of Color
ASIAN	Asian
BLACK OR AFRICAN AMERICAN	Student of Color
HISPANIC OF ANY RACE	Student of Color
NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER	Student of Color
NONRESIDENT ALIEN	Nonresident Alien
RACE/ETHNICITY UNKNOWN	Other
TWO OR MORE RACES	Student of Color