Quantifying Delaware's Secondary to Post-Secondary Student Pipeline: A Novel Open-source Approach Towards Estimating Workforce Eligible Populations

Abstract

For state education and labor policymakers and administrators, understanding the secondary education to eligible workforce pipeline would enable data-driven decision-making toward building a resilient and thriving future labor market. Moreover, it would assist variable stakeholders in understanding the effectiveness and return on investment (ROI) of the dollars attributed to different initiatives over time. A major barrier is the lack of multi-level longitudinal data that integrates high school and higher education journeys. To address this, we present a novel open-source approach that wrangles and integrates multi-source data to engineer the first quantitative representation of Delaware's (DE) secondary school-to-eligible workforce population pipeline. Data ranges from 2015 to 2022 and is labeled by graduation year-based cohorts so we can make inferences at the cohort and/or grade levels. Stratified by the Carnegie Classification of Institutions of Higher Education we quantify variable pathways towards workforce eligibility by making absolute and relative comparisons across the student journey. For example, of all 2015 DE 9th graders (cohort 2018; N=13,856): 1) 73 % graduated from 12th grade in DE, 2) 40% enrolled in a DE 4-year degree school, yet 3) only 22% (approximately one in five) earned a bachelor's degree from a DE institution. As high school cohorts graduate this data set will progressively expand eventually allowing us to map the average student trajectory to the workforce, accounting for variability over time. New space for collaboration can now exist for researchers and others to leverage and optimize this methodology for future comparative studies across states.
Keywords: *college readiness; enrolment; workforce; IPEDS; Delaware open-data portal.*

1 Introduction

In recent years, Delaware's education landscape has undergone significant transformations, with notable developments observed at both the high school and college levels. Much of this focus can be attributed to the variable challenges and opportunities that the state faces as it relates to its workforce Gardner (1984). It would be nearly impossible to engage in a discussion on the value of education, without seriously weighing the implications this has for the future of a given workforce. Siloed data on educational and labor outcomes have failed to provide a holistic view of the PreK through the workforce pipeline Venezia and Jaeger (2013). For the purposes of this study, when we refer to the workforce, we are referring to the population eligible to be accounted for in the active workforce. To predict and prepare for the demand of future growth jobs and sectors would be both a complicated and invaluable contribution to the workforce, regardless of the baseline assumptions made today Griffith and Wade (2001).

As a step towards, for example, multi-state (or even National) comparative studies on education to workforce pipelines, we propose a methodology that sources and integrates secondary and higher education data from longitudinal data systems. For Delaware, a

state with a population of just over 1 million persons (also comprised of only 3 counties), educational data is tracked at the student level and conveniently publicly reported in aggregate at the school district level yearly. As such, we propose an open-source approach valid for the state of Delaware and scalable to other states with personalized changes addressing differences in how school district-level count data is collected and/or reported.

The purpose of this paper is to propose a methodology that sources and integrates secondary and higher education data from longitudinal data systems. This approach, which leverages volumes of available data, is a first step towards filling the gaps in our knowledge about the education-to-workforce pipeline. We present our methodology in sequential order. First, we will describe the primary pathways into the workforce in our state and the relevance of the data we collect to said pathways. We then describe our theoretical framework for integrating secondary and higher education data from three different sources. The data is then wrangled, integrated, and analyzed. Visualizations at this stage are static but provide us with the foundations to scale our open-source dashboard prototype. We expect that major highlights will focus on the overall turnover of 9th graders into the eligible workforce after high school graduation and the differential rate of enrollment among Delaware high school students for the state of Delaware. The inferences that this study facilitates have a direct influence on educational administrators and policymakers. Understanding the student journey and time-varying trends to becoming eligible workforce, not only allows a state to reflect on the impact of its past and current educational initiatives, but it also aids them in developing an understanding of the changes that need to be made in order to ensure the needs of tomorrow's workforce are being introduced in educational spaces today. As a working draft, we host a public access interactive dashboard of our visualizations here [1]

2 Research Questions

This study addresses two fundamental questions central to understanding the educational landscape in Delaware. Analyzing the current trends in education within the state, specifically exploring the methodologies, curriculum designs, and policies implemented at both high school and college levels, is essential, however, limited by the accessibility of data. This analysis aims to uncover the evolving patterns and innovations that have shaped students' educational experiences in Delaware. This research delves into the outcomes of these educational initiatives and provides evidence of its value by providing insights as to the percentage of 9th graders who progress to higher education within the state. Our overarching research questions are:

1. What are the compositions of the degree and non-degree pathways into Delaware's potential workforce?

2. How do post-secondary trends vary by cohort?

[1] `https://innovation-education.pages.dev/student-level`

Engineering a high school to post-secondary degree completion data set we can generate answers to standing questions that have implications not only for our education and labor sectors but also for our economy. The value of this work, if scaled, is not limited to educational data sets. One can source additional data sources, that if aligned with educational ones, will provide deeper insights into greater good social challenges and advancements (e.g. exploring the housing cost burden among millennials who attended post-graduate education).

3 Methods

We propose a step-wise approach towards building a joint educational data source and visualizing findings, while simultaneously documenting each one of our assumptions and choices for inclusion/exclusion in the data set.

3.1 Theoretical framework of workforce pathways from high school graduation

Figure 1 provides a theoretical overview of the different pathways to entering the workforce population post-high school graduation that we document and represent in this methodology.

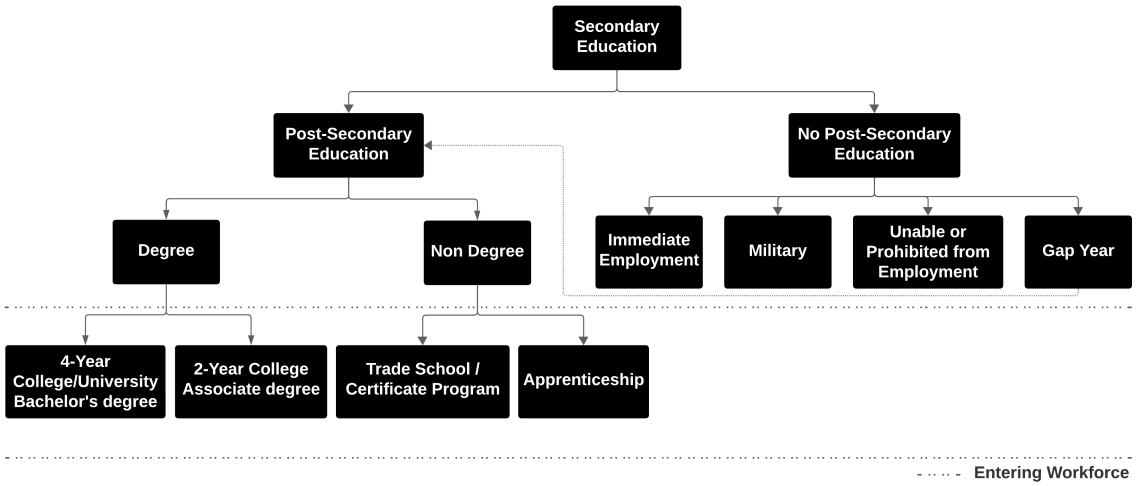

Figure 1: The theoretical framework for workforce pathways from high school graduation follows two main trajectories.

3.2 Data Collection

We employ three independent, publicly available, non-mutually exclusive data sources developing a method to sequentially wrangle, clean, match, merge, and store a joint secondary and post-secondary data set. The data sources are Delaware Open Data Portal (DODP), Delaware Department of Education (DDEO), and the Integrated Post-secondary Education Data System (IPEDS), NCES (2023).

3.2.1 Secondary Education Data

9th grade through 12th grade graduation data was gathered from two primary sources: DODP (2023) and DDEO (2023). The publicly available data on the state's open portal reports yearly grade-level data only for public schools. In this report when we refer to Nonpublic schools, we are referring to private institutions and at-home education (both single- and multi-family). Also, the public school includes charter schools and magnet schools. Nonpublic school data is not accessible through the portal or generally to the public. Private institutions follow reporting guidelines and not mandates based on their individual funding models, and whether they receive or administer any sort of public assistance. As such, to avoid systematically introducing selection bias into our sample, we separate and document the two different processes for sourcing and estimating public and private high school data, individually. They are documented as below:

- Public School: To access the public high school data for the state of Delaware we sourced the "Student Enrollment" dataset from DODP using the Socrata Open Data API (SODA). This dataset covers K-12 data and is available from 2015 onwards. It offers a comprehensive breakdown of enrollment by grade, public school district, gender, race, and special populations of interest. As noted by DDOE, "the total number of students is not meant to reflect the actual number of students enrolled at any point-in-time," it is rather the ceiling of students enrolled at least one day by the end of the academic year. This represents the maximum total population eligible to complete the school year and is the consistent approach across the year.

- Nonpublic School: DDOE has released two annual reports for nonpublic schools in Delaware for the years 2021 and 2022 providing detailed enrollment breakdowns by grade, school district, and race. Additionally, the reports include data on the overall enrollment trends in nonpublic schools from 2010 to 2022. Given the total count and individual grade count of nonpublic secondary school students for two years, we calculated average proportional weights by grade. We then retroactively multiplied each weight for each year's (2015 to 2020) total count of nonpublic secondary school students to estimate their enrollment.

3.2.2 Post-Secondary Education Data

IPEDS is an exhaustive dataset encompassing information from over 7,000 US educational institutions, captured in 12 interconnected surveys annually collected. For the state of Delaware there are 15 degree granting institutions in operation from 2015 to 2022 for which data has been reported to the National Center of Education Statistics. Of these fifteen (15), by filtering with CCIHE labels, we are left with a final sample of five (5) four-year degree-granting institutions for the state of Delaware. Of the 12 surveys administered by NCES, four capture the enrollment and graduation data we leverage for our analysis:

1. Institutional Characteristics

2. Fall Enrollment: EFxxxx*A and EFxxxxC

3. Completions: Cxxxx_C (xxxx*denotes the individual cohort year)

Taken all together, these different sources make up the foundational individual data that is matched and merged to engineer the secondary education to eligible workforce pipeline. Detailed steps are presented next.

3.3 Data Integration

For simplicity purposes, we present our data wrangling and cleaning processes first and then provide our logic for merging the individual data sets. All data wrangling and analyses were conducted using Python v3.11.3, Pandas v1.5.3 and b eatifulsoup4 v4.9.3.

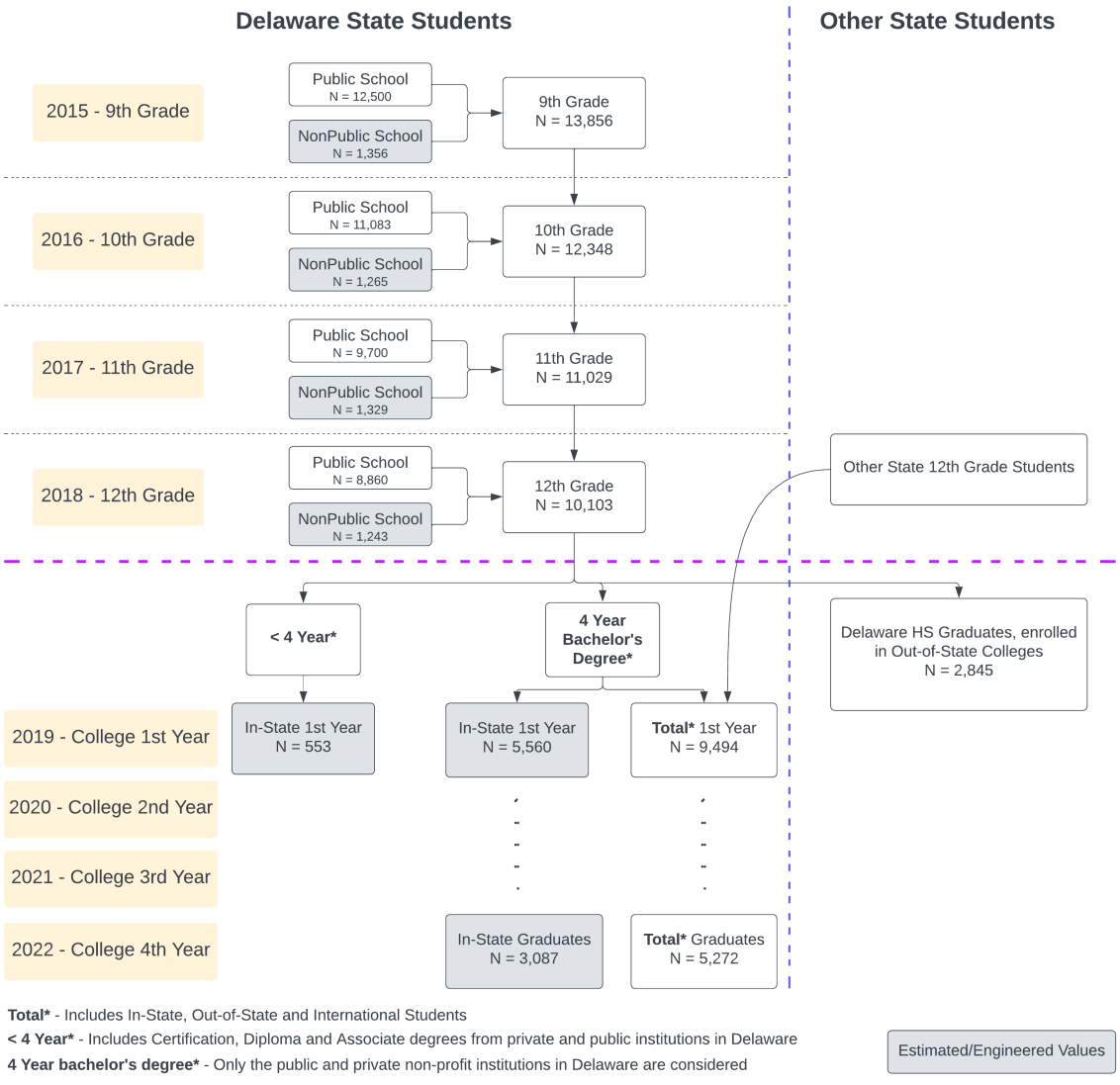

Figure 2: Data Integration Framework.

The progressive enrollment journey for the selected 2018 cohort of instate and out-of-state students as they advance through each grade level. Gray boxes represented estimated

or engineered counts. The horizontal, purple dashed line represents the high school graduation period. There is no count data for 2nd and 3rd year of college.

3.3.1 Individual data source pre-processing

- DODP : This data set included complete data from 2015 – 2022 with 0% missingness at the highest level of features (e.g. low count groups have been masked). No further data processing was needed.

- DDOE:

  - Nonpublic secondary school enrollment data broken down by grade was available only for the years 2021 and 2022. Total Nonpublic secondary school enrollment, however, is available from 2015 to 2022

  - For each year, we calculate the grade specific enrollment rate

  - These rates are then averaged for both years

  - The averaged weights are then multiplied to each year's total enrollment to obtain student count estimates

- IPEDS:

  - Extracting Delaware Institutions

  - Obtained the Fall enrollment data for the Delaware institutions.

  - Extracted the count of students enrolled in Delaware institutions and students who migrated from Delaware to out of states, to pursue their degrees after high school.

  - Obtained the count of students who Graduated from their 4 years bachelor's degree in period of 4 years. [2]

3.3.2 Data Merger

- We first defined nine cohorts spanning from 2015 to 2023, each corresponding to their high school graduation year.

- Next, we extracted high school and post-secondary student enrollment counts for each cohort from their respective datasets.

- To align each student class with their respective data across years the process involved extracting both public and non-public enrollment data. We assume that the 9th grade class is represented in each subsequent year. If a student class in 2015 was in 9th grade, they were matched to the 10th grade class data for 2016, and so on. This same approach current/lag approach was also applied to the higher education

[2] A fully detailed stepwise process for wrangling and processing the IPEDS data is included in Appendix 1 or can be accessed in Google Collab here: `https://tinyurl.com/ipedsprocessing`

cohort years, building the overarching education to workforce eligible population pipeline.

- Calculate the percentage change in student counts for each educational level within a cohort.
- Then compile the processed results into a master file, offering a comprehensive overview of enrollment and graduation trends across cohorts in a tidy data set format.
- Finally, we uploaded and hosted the master file into an Azure blob storage for secure and convenient internal access.

3.3.3 Master file post-processing

As a final factor to engineer, once the three data sources at matched and linked across years, we are able to calculate the graduation rates for each cohort and across years. College graduation rates (CGR) instate graduation count (IGC) were calculated as follows:

$$CGR(in\%) = \frac{\text{\# Students completed their 4th year (Instate + Outstate)}}{\text{\# Students who enrolled in 1st year (Instate + Outstate)}} \cdot 100 \quad (1)$$

$$IGC = \text{\# Instate Students enrolled in first year} \cdot CGR \quad (2)$$

4 Preliminary Data and Results

After post-processing, we are left with cohort level data across variable years. It is worth nothing that a complete period for the purposes of this study is 9th grade to college graduation, as we were interested in the journey of high school students into the eligible workforce through obtaining 4-year degrees. To visualize the progression of students across pipeline, we select the 2018 cohort given that we have complete observed data. Each year, one new cohort will have a completed 9th grade to graduation data+ set. In short, by 2024, there will be data accessible to quantify the complete journey of three (3) full cohorts and so forth. Figure 2 showcases Cohort 2018 as an illustrative example. For high school data, the total student count for each grade is determined by aggregating the counts of both public and nonpublic school students.

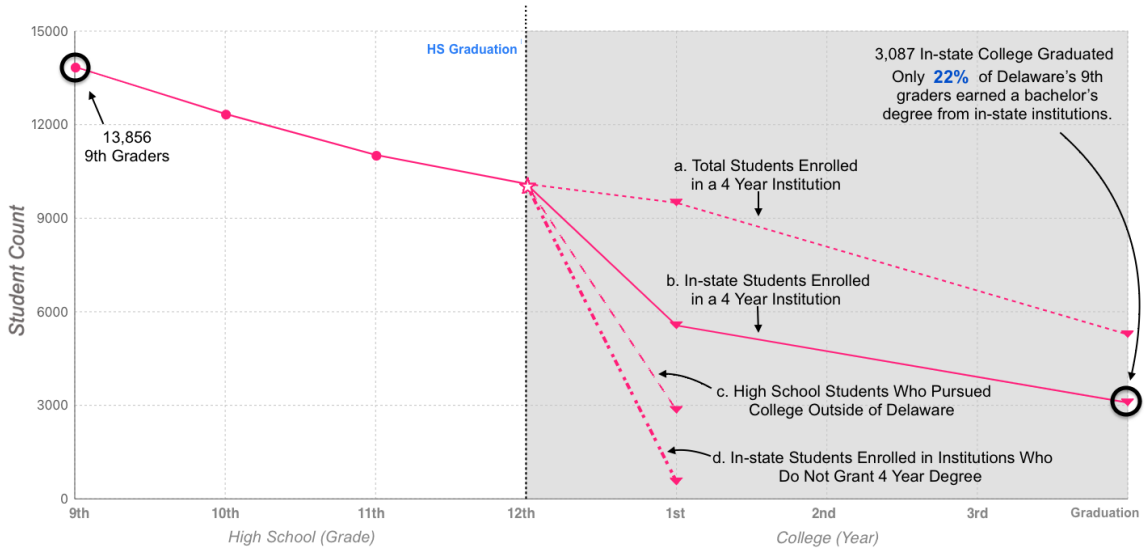

Figure 3: 2018 secondary to post-secondary enrollment and graduation counts for DE students: Depicts the enrollment journey for Delaware students towards entering the eligible workforce. There are variable pathways by which students enter the eligible workforce, each depicted as a line: a) Total of students from in-state, out-of-state, and international enrolled in a 4-year degree granting institutions, b) Total of in-state students enrolled in a 4-year degree granting institutions, c) Total of high school students who pursued college outside of Delaware, d) Total of in-state students enrolled in institutions who do not grant a 4-year degree. Notably, the solid line from High School to College is intended to represent the Delaware in-state enrollment continuum.

From figure 2, it is evident that there is attrition at each elapsed year. According to NCES, of all 12th grade graduates, 43% opt for a 1-year college degree program, while 19% prefer a 2-year college degree program. For Delaware, the 7-year average 12th grade student enrollment into a 4-year degree granting institution is 43.6%, 95% CI [39%, 48.2%], while the 3-year average for 12th grade student enrollment in a 2-year college is 4.7%, 95% CI [3.76%, 5.68%].

Furthermore, to test the validity of the DOPD high school student graduation counts, we were able to obtain observed yearly graduation counts aggregated from individual-level student data for DE. Comparing the observed data counts to those found in the DOPD, the 8-year average percent difference is 2.15%, 95% CI [0.983%, 3.31%]. These previous metrics, although valuable, provide cross-sectional and limit our ability to understand the student journey as a continuum. The formative outcomes are equally relevant as the relative ones. Figure 3 illustrates the education to workforce eligible pipeline and details major highlights.

Cohort 2018 reveals that only 22% of Delaware's 9th graders earned a bachelor's degree within the state's institutions. These individuals make up a potential workforce with roots in Delaware and are likely to remain in the state. Specifically, because we quantify those who earn bachelor's degrees, we can ideate around and project estimates for career pathways which require minimally a 4-year degree to attain.

5 Importance of the Research

The implications of the data made accessible through this work was the driving factor to develop it to begin with. We do not know, what we do not know is clearly depicted in our hazy understanding of the pipeline of students that enter our workforce eligible population through various pathways. One of the major challenges in the workforce today is that we are educating children for the jobs we need today, not the projected jobs we will need from then eventually. In part, this is due to an economic space that is driven by reactive and immediate demand as opposed to one that is preventive and predictive. Regardless of the limitations we may face today when merging education and workforce data and attempting to inference from it, any given state would benefit from being able to answer multiple questions around "how many" and "when are they entering the workforce" neighboring matters.

We designed this research study with the hopes to eventually scale this work. Leveraging IPEDS data provides us with an almost exhaustive list and data of all institutions in the United States. The ability to replicate this methodology for another state will depend on whether they host State DOE data in open-source format. States that do can adapt this methodology to generate similar outcomes. Those who do not, may create ad-hoc solutions to engineer the longitudinal data and/or may develop plans towards building an open data portal. Given that we only use open-source data and tools, the complete script for this project can be run without additional licensing or pre-built tooling. Second, policy implications are at the center of this work also. If a state chooses to develop a new data portal, that will have policy implications for the state. Namely, as it related to data sharing, governance and funding practices. Additionally, the nature of this population centric work makes it relevant for public health management and policy more broadly. The findings and subsequent inferences that can be made from the outcome of this work have implications for curriculum, operational, interventional and preventive resources, programs and policies to name a few. Lastly, social return on investment (sROI or ROI) is a relevant matter to public and private funders. Individuals are growingly interested in understanding the direct influence of their contributed dollars. For some, general positive trends may suffice, and for others, they are more interested in the direct dollars saved because of causes only possible through the initiative they support. The latter is much more complex and requires methods that expand on data outside of just educational context.

6 Limitations and Future Research

We built a baseline approach implying that variable statistical and methodological assumptions were made during the development of the pipeline. There were certain metrics that required additional engineering and/or estimation, future development would focus on obtaining completely observed rather than partially (although mostly observed).

Additionally, we are limited by two specific factors outside of our control: time and lag periods. If a cohort is younger, they might not have college enrollment and/or graduation

data yet. In terms of lag, many of the data we leverage are representative of some retrospective cross-section (although recent). 6 month and 1 year lag periods are not uncommon in open-source database management.

Thus far, we have been able to be creative about the directions in which we can advance this work. Some of the projects we have identified as potential subsequent studies are: 1) Predicting the supply-demand gap for advanced degree careers (e.g. medicine, law, dentistry), 2) Match and integrating the eligible workforce population to the current workforce labor market demands, 3) Evaluate whether the COVID-19 pandemic differentially affected higher education enrollment, 4) Align enrollment years with relevant educational milestones that resulted in mandates or policy changes, 5) Merging with Census data to evaluate educational attainment gender equity in Delaware and, 6) Merging with Census data to evaluate workforce racial and ethnic equity in Delaware.

There are various aspects of this study that can be modified to enhance the naïve inferences. Below are some we propose:

- Identify a process that allows for accurate stratification by disability status. Knowing that disabilities and individual needs are variable, we expect educational and labor attainment for disabled individuals to also be variable.
- We were unable to obtain the Sophomore through Senior year data for Delaware high school students who went out of state to complete higher education.
- We are unable to validate which out-of-state students stay in Delaware post-graduation because we have no 1-year follow-up metrics at present.

References

DDEO. (2023). *Delaware department of education.* Retrieved from `https://education.delaware.gov/community/data/reports/nps/`

DODP. (2023). *Delaware open data portal.* Retrieved from `https://data.delaware.gov/`

Gardner, W. E. (1984). A nation at risk: Some critical comments. *Journal of Teacher Education*, *35*(1), 13-15. Retrieved from `https://doi.org/10.1177/002248718403500104` doi: 10.1177/002248718403500104

Griffith, J., & Wade, J. (2001). The relation of high school career- and work-oriented education to postsecondary employment and college performance: A six-year longitudinal study of public high school graduates. *The Journal of Vocational Education Research*, *26*, 328-365. Retrieved from `https://api.semanticscholar.org/CorpusID:155004814`

NCES. (2023). *National center for education statistics.* Retrieved from `https://nces.ed.gov/ipeds/use-the-data`

Venezia, A., & Jaeger, L. (2013). Transitions from high school to college. *The Future of children*, *23*(1), 117-36. doi: 10.1353/foc.2013.0004

Appendices

A IPEDS

IPEDS Process Followed:

A.1 Extracting the list of institutions in Delaware

- Query the institutional characteristics file for Delaware with Federal Information Processing Standards (FIPS) = 10 (10 is FIPS code for Delaware). UNITIDs are unique institution IDs defined in IPEDS.
- Classify all the extracted institutions into their type of highest-level offerings using Carnegie Classification and then cluster them into 2 groups.

  Group 1: Bachelors and Higher Degree Granting (4 years and above).

  University of Delaware

  Delaware State University

  Wilmington University

  Delaware Technical Community College-Terry

  Goldey-Beacom College

  Group 2: Certificates, Diploma and Associates granting (Less than 4 years).

A.2 Loading the Fall enrollment data for the Delaware institutions

- Fall Enrollment: EF2018A
- Get all Delaware institutions by UNITD's extracted from previous step.
- Query EFALEVEL = 4 to get the Total (Part time + Full time) count of students pursuing Undergraduate, Certificate/Diploma seeking for the FIRST – TIME. Same can also be achieved by querying EFALEVEL = 24 for Full time- first time and EFALEVEL = 44 for part time first time.
- EFTOTLT refers to the total number of students enrolled.

A.3 Obtaining the count of students enrolled in Delaware institutions to pursue their degrees after high school

- Use Fall Enrollment (EF2018C) aka Migration file has migration data for first – time freshman (part time + fulltime).
- Query the Delaware Institution by Unit IDs. Here's the python script for same. $students_DE = migration_2018_de.query("UNITID == @DE_institute_list")$

- EFCSTATE refers to Residence of First-time Freshmen. Notice EFCSTATE = 10 in the data frame as I.e. Delaware.
- Distribute the institutes as per our 2 Groups (4 years and above OR less than 4 years)
- Get the total count for each group from feature EFRES01.
- Furthermore, it can be distributed between group 1 and group 2.

A.4 Obtaining the count of students who migrated from Delaware to out of the state, to pursue their degrees after high school

- Using the same Fall Enrollment Migration file (EF2018C)
- Filter for EFCSTATE = 10 i.e. FIPS for Delaware.
- Query the file for all the institutions other than the Delaware institutes to get the list of institutes in US (to exclude Delaware). This is achieved by making a query - students_US = migration_2018_de.query("$UNITID != @DE\_institute\_list$")
- Get the total count for each group from feature EFRES01.

A.5 Obtaining the count of students who graduated from their bachelor's degree in a period of 4 years from the year of Enrollment

- To extract the graduation count use Completions 2022 (C2022_C) file and assume a student who is enrolled in 2018 will be graduated in 2022 i.e. 4 years
- Filter the Delaware institutes by UnitID (* Query for Group 1 i.e. 4 years degree granting institutes)
- Filter by AWLEVELC = 5
- AWLEVELC is Students receiving awards/degrees and 5 is to query for students who graduated with bachelor's degree
- Get the count from feature CS18_24 CS18_24 refers to the age category of students between 18 to 24. This way we can accurately track the enrolled students to their graduations.