US Education Enrollment Status vs Ethnicity from years 1955 -2019

W200 Project 2
By David Trinidad, Nicholas Brown

Background:
The data used for this project was obtained from the **United States Census Bureau** (**USCB**). USCB is an Agency of the U.S. Federal Statistical System, responsible for producing data about the American people and economy. The USCB offers a variety of data sets in education.

Project Scope: For this project, we will focus on the United States Enrollment Status between 1955 -2019. This data set is available as a CSV file which can be downloaded through the following link https://www.census.gov/data/tables/time-series/demo/school-enrollment/cps-historical-time-series.html.

. Below is a brief outline of the dataset:

- Data Keys: Nursery, Kindergarten, Elementary, Highschool, College
  - Sub-keys: Public, Private
- Data Values: Total enrollment
- Data Variables: Year
  - Sub Variable: Ethnicity

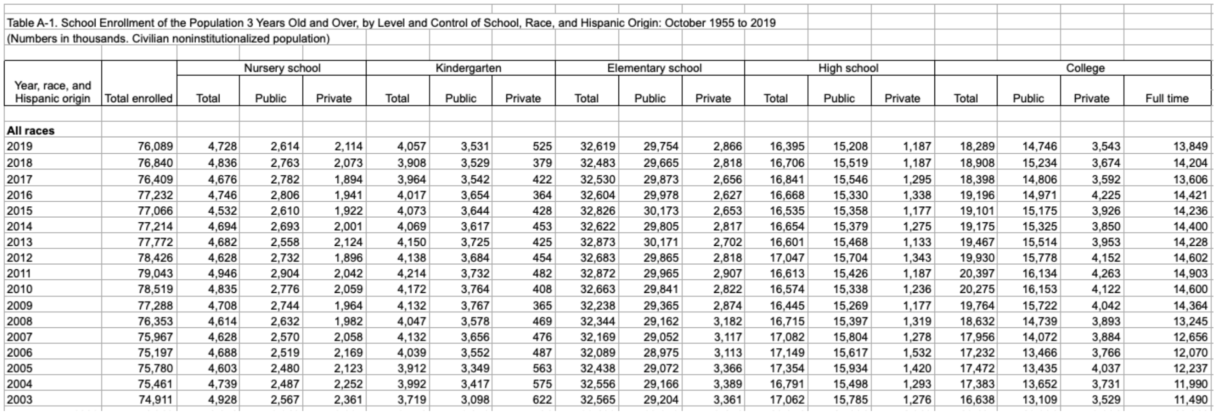


Table A-1. School Enrollment of the Population 3 Years Old and Over, by Level and Control of School, Race, and Hispanic Origin: October 1955 to 2019
(Numbers in thousands. Civilian noninstitutionalized population)

Year, race, and Hispanic origin	Total enrolled	Nursery school			Kindergarten			Elementary school			High school			College			
		Total	Public	Private	Total	Public	Private	Total	Public	Private	Total	Public	Private	Total	Public	Private	Full time
All races																	
2019	76,089	4,728	2,614	2,114	4,057	3,531	525	32,619	29,754	2,866	16,395	15,208	1,187	18,289	14,746	3,543	13,849
2018	76,840	4,836	2,763	2,073	3,908	3,529	379	32,483	29,665	2,818	16,706	15,519	1,187	18,908	15,234	3,674	14,204
2017	76,409	4,676	2,782	1,894	3,964	3,542	422	32,530	29,873	2,656	16,841	15,546	1,295	18,398	14,806	3,592	13,606
2016	77,232	4,746	2,806	1,941	4,017	3,654	364	32,604	29,978	2,627	16,668	15,330	1,338	19,196	14,971	4,225	14,421
2015	77,066	4,532	2,610	1,922	4,073	3,644	428	32,826	30,173	2,653	16,535	15,358	1,177	19,101	15,175	3,926	14,236
2014	77,214	4,694	2,693	2,001	4,069	3,617	453	32,622	29,805	2,817	16,654	15,379	1,275	19,175	15,325	3,850	14,400
2013	77,772	4,682	2,558	2,124	4,150	3,725	425	32,873	30,171	2,702	16,601	15,468	1,133	19,467	15,514	3,953	14,228
2012	78,426	4,628	2,732	1,896	4,138	3,684	454	32,683	29,865	2,818	17,047	15,704	1,343	19,930	15,778	4,152	14,602
2011	79,043	4,946	2,904	2,042	4,214	3,732	482	32,872	29,965	2,907	16,613	15,426	1,187	20,397	16,134	4,263	14,903
2010	78,519	4,835	2,776	2,059	4,172	3,764	408	32,663	29,841	2,822	16,574	15,338	1,236	20,275	16,153	4,122	14,600
2009	77,288	4,708	2,744	1,964	4,132	3,767	365	32,238	29,365	2,874	16,445	15,269	1,177	19,764	15,722	4,042	14,364
2008	76,353	4,614	2,632	1,982	4,047	3,578	469	32,344	29,162	3,182	16,715	15,397	1,319	18,632	14,739	3,893	13,245
2007	75,967	4,628	2,570	2,058	4,132	3,656	476	32,169	29,052	3,117	17,082	15,804	1,278	17,956	14,072	3,884	12,656
2006	75,197	4,688	2,519	2,169	4,039	3,552	487	32,089	28,975	3,113	17,149	15,617	1,532	17,232	13,466	3,766	12,070
2005	75,780	4,603	2,480	2,123	3,912	3,349	563	32,438	29,072	3,366	17,354	15,934	1,420	17,472	13,435	4,037	12,237
2004	75,461	4,739	2,487	2,252	3,992	3,417	575	32,556	29,166	3,389	16,791	15,498	1,293	17,383	13,652	3,731	11,990
2003	74,911	4,928	2,567	2,361	3,719	3,098	622	32,565	29,204	3,361	17,062	15,785	1,276	16,638	13,109	3,529	11,490

Figure 1. Sample of the School Enrollment 3-years and over by level 1955-2019

Primary Data Set

The main data set that we analyzed is historical data about school enrollment of the US population ages 3 years old and over by education level and race. The data includes school enrollment information from the years 1955 to 2019. The data contains information about school enrollment at the nursery school level, kindergarten level, elementary school level, high school level, and college level. The data also contains public school and private school enrollment information at each level. The data is organized in a way that you can obtain information about enrollment based on race. The different races that are included in the data are White alone, White alone non-Hispanic, Black Alone, Asian Alone, Hispanic, White alone or in combination, Black alone or in combination, Asian alone or in combination. You can also view the total enrollment of all races combined. We chose not to use the attributes White alone or in combination, Black alone or in combination, Asian alone or in combination because most of the information provided for these attributes have already been provided for other attributes (i.e. most of the information provided for Black alone or in combination is provided in Black alone). **It is important to note that data represents numbers in thousands.**

The race definitions as defined by the census bureau are listed below:

White Alone – White Alone refers to people who reported White and did not report any other race category.

White Alone or in combination - White alone or in combination consists of those respondents who reported White, whether or not they reported any other races. In other words, people who reported only White or who reported combinations such as "White and Black or African American," or "White and Asian and American Indian and Alaska Native" are included in the White alone or in combination category.

Black Alone – Black alone refers to people who reported Black or African American and did not report any other race.
Asian Alone – Asian alone refers to people who reported Asian and did not report any other race.
Asian Alone or in combination - Asian alone or in combination consists of those respondents who reported Asian, whether or not they reported any other races. In other words, people who reported only Asian or who reported combinations such as "Asian and White," or "Asian and Black and NHOPI" are included in the Asian alone or in combination category.
White alone, not Hispanic or Latino are individuals who responded "No, not Spanish/Hispanic/Latino" and who reported "White" as their only entry in the race question.
Hispanic - People of Hispanic origin may be of any race. Hispanics can choose one or more race categories, including White, Black or African American, American Indian and Alaska Native, Asian, and Native Hawaiian and Other Pacific Islander. If someone does not identify with any of the specified race groups, he or she may mark the “Some other race” category and write in their race.
People who are of Hispanic origin are asked to indicate the specific group they belong to: Cuban, Mexican, Puerto Rican, or other groups, such as Spanish, Honduran, or Venezuelan.

footnotes:

- Starting in 2003 respondents could identify more than one race. Except as noted, the race data in this table from 2003 onward represent those respondents who indicated only one race category.
- Prior to 1994, total enrollment does not include the 35 and over population.
- Data shown for 1955 to 1966 for the Black population are for Black and Other races.

Supplementary Data

Annual High School Dropout Rates of 15 to 24 Year Olds by Sex, Race, Grade, and Hispanic Origin: October 1967 to 2019.

Assumptions:

1. Due to the nature of survey data, we assume the data is not one hundred percent accurate.
2. Ethnicities are represented differently over time in the entire data set.
3. Since middle school is not mentioned, we can assume the data divided between elementary and high school.

Questions:

1. Are there any conspicuous trends in total enrollment?
2. Are the enrollment trends the same for all the races?
3. Looking at each individual race, are the enrollment trends the same at each education level?
4. Does a percentage increase or decrease in dropouts coincide with a percentage increase or decrease in high school enrollment?
5. Are there any historical events that may provide insight to the enrollment trends?

Pre-Data Cleaning

Data cleaning occurred both prior to uploading the data in python and after reading the data into python. One pre-cleaning method consisted of changing the structures of the headers so that data could be uploaded as a data frame in python with the columns set in place. The restructuring of the headers was necessary for both the primary and supplementary data sets. Another pre-cleaning method consisted of assigning numbers to identify the different races. The assigning of numbers to the different races could have been performed inside of python but it made more sense to make it part of the pre-cleaning process. Both the pre-cleaned data and the original data are included in the excel books. The pre-cleaned data is titled revised.

Sanity Check and Data Cleaning inside Jupyter Notebook

After viewing the column names, we noticed that a few of the columns had to be renamed. Some of the column titles contained spaces after the last letter, and some of the columns had spaces in between words. With the exception of the footnote column, we were also expecting all of the columns to be numerical. We noticed after checking the data types of the columns, not all of the expected columns were recognized as numerical, this was due to some of the NA values in excel having spaces before the NA.

We also noticed that data was missing for all 1980 high school and college, public and private enrollment. We also noticed that some races have two values for 1993; it is noted in the original dataset that one is revised and controlled to 1990 census-based population estimates. We decided to keep both 1993 values.

Data Analysis:

1. **Question 1:** Are there any conspicuous trends in total enrollment?

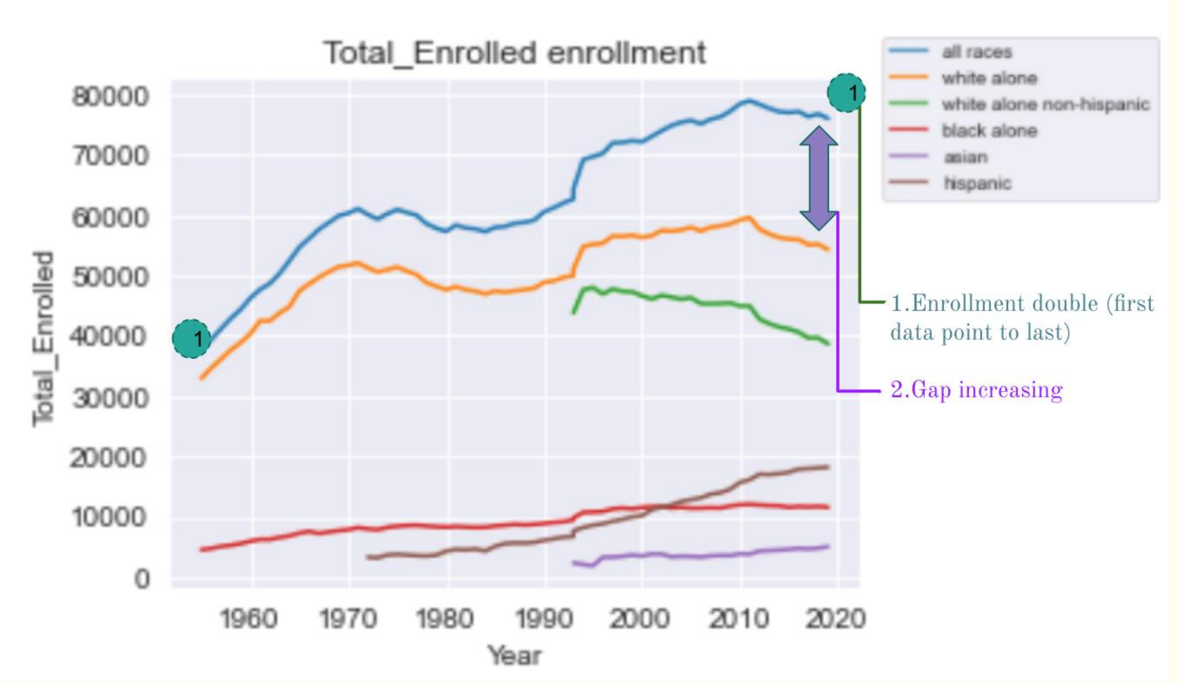


Figure 2: *Total enrollment for all races and all ethnicities from 1955 to 2019*

The purpose of this question is to provide a baseline; it helps to provide insight on questions that should be asked at the granular level. For this question we looked at the cumulative enrollment (nursery school through college) of each race over the time periods shown the figure 2 graph.

Key Points:

- The total enrollment approximately doubles from 1959 to 2019 (blue line).

- The majority of enrolled students are predominantly white, which can be inferred by observing how close the orange line (white alone) is to the blue line (all races) and by the similar patterns, but we see that as time passes, the gap between the lines widens. The widening of the gap indicates that the white alone category does not consist of as much of the total as it did in the past.

- All races have more people enrolled at the end of the timeframe than the beginning with the exception of White alone non-Hispanic (green line).

Question 2: Are the enrollment trends the same for all the races?

The objective of this question is to compare the trends of the different races to identify any similarities or differences.

Key Points:

- We see that from 1955 to approximately 1971 both the black alone and white alone enrollment is increasing (there is only data on these two races but no other races for this time period). We are also able to see that the growth rate for that time period is greater for the White alone race than the Black alone race. (Figure 2 listed above)

- For the same time period listed above, we see this spike occur at the kindergarten, elementary, high school, and the college level. (Figure 3 listed below)

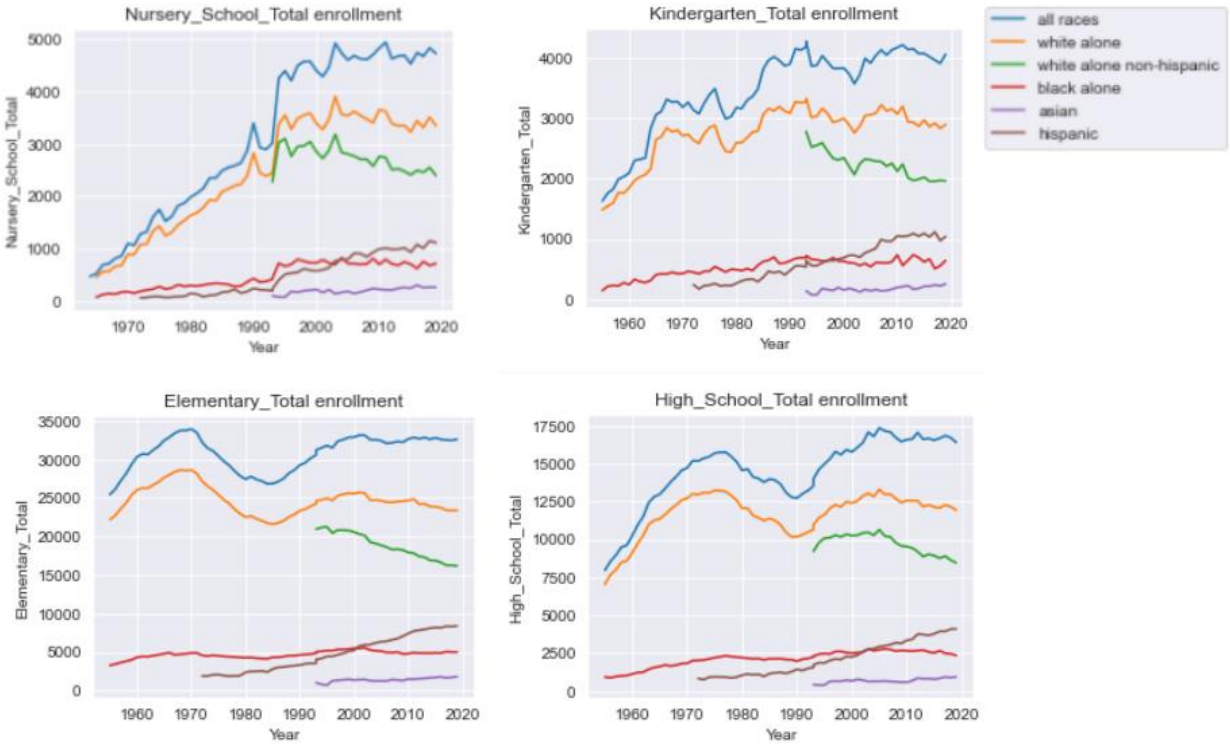


Figure 3. *Enrollment across each grade level over time.*

- Between 1971 and 1984, we see that total enrollment decreases for elementary and high school are decreasing for the all the races combined. But during this same time period, we see that enrollment is increasing in the Hispanic population for elementary and high school. We also see that after 1984, the total enrollment starts to increase again for the total races combined. (Figure 2 listed above)

- We are able to see that Non-Hispanic Whites are the only group whose total enrollment is steadily declining. (Figure 2 listed above)

Question 3: Looking at each individual race, are the enrollment trends the same at each education level?

The objective of this question is to look at each race individually, and identify any similarities that exist at the different education levels (e.g., Is the growth rate the same for Black alone at the elementary and high school level?).

The Hispanic population's enrollment appears to be increasing almost linearly at each level.

Key Points:

- The Hispanic population's enrollment appears to be increasing almost linearly at each level. (Figure 4 listed below)

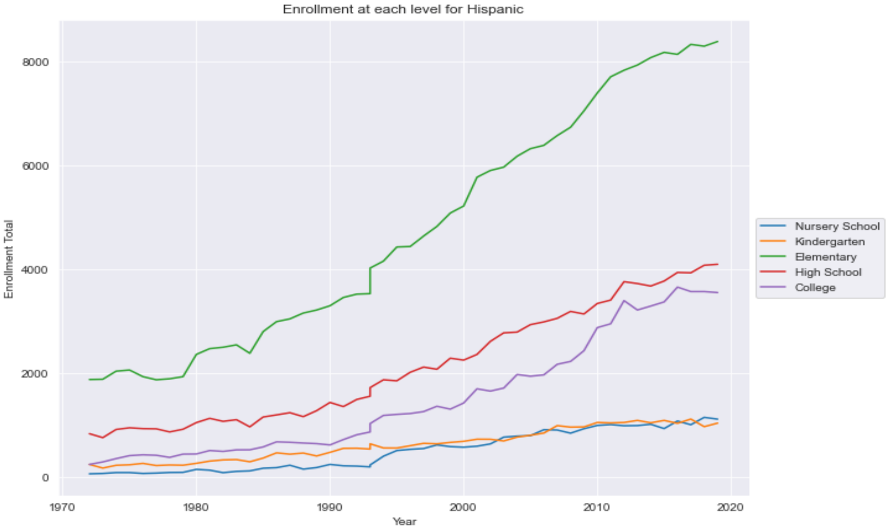


Figure 4. Hispanic enrollment at each education level.

- The White Alone Non-Hispanic appear to be decreasing at each level with the exception of the college level, where it is increasing and then starts to decrease at approximately 2012. (Figure 5 listed below)

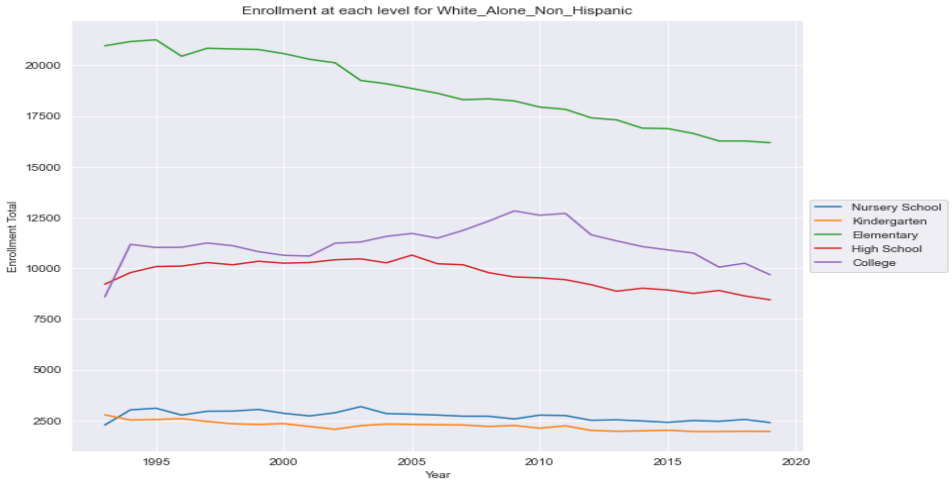


Figure 5. White alone Non-Hispanic enrollment at each education level.

- The Asian population is the only population whose College enrollment numbers are similar to the Elementary enrollment numbers. (Figure 6 listed below)

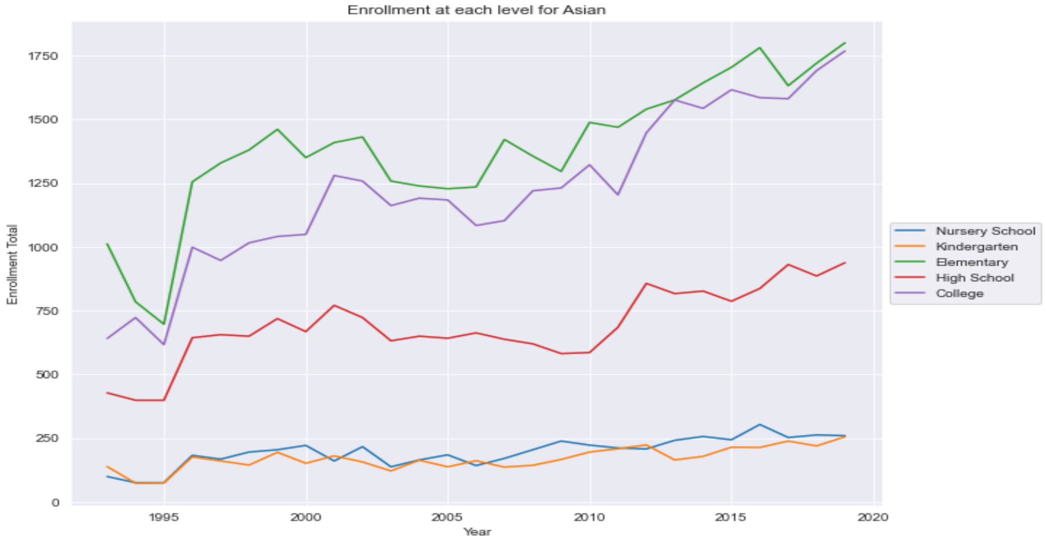


Figure 6. Asian enrollment at each education level.

Question 4: Does a percentage increase or decrease in dropouts coincide with a percentage increase or decrease in high school enrollment?

For this section we compared our main dataset to our supplemental dataset. We created an additional column for both the main and supplemental datasets. The columns created show the percentage increase or decrease for the number of students enrolled and dropping out each year. For the number of students dropping out, we looked at the number of students dropping out between 10-12 grade. We compared the number of students dropping out to the number of students enrolled in high school.

Key Points:

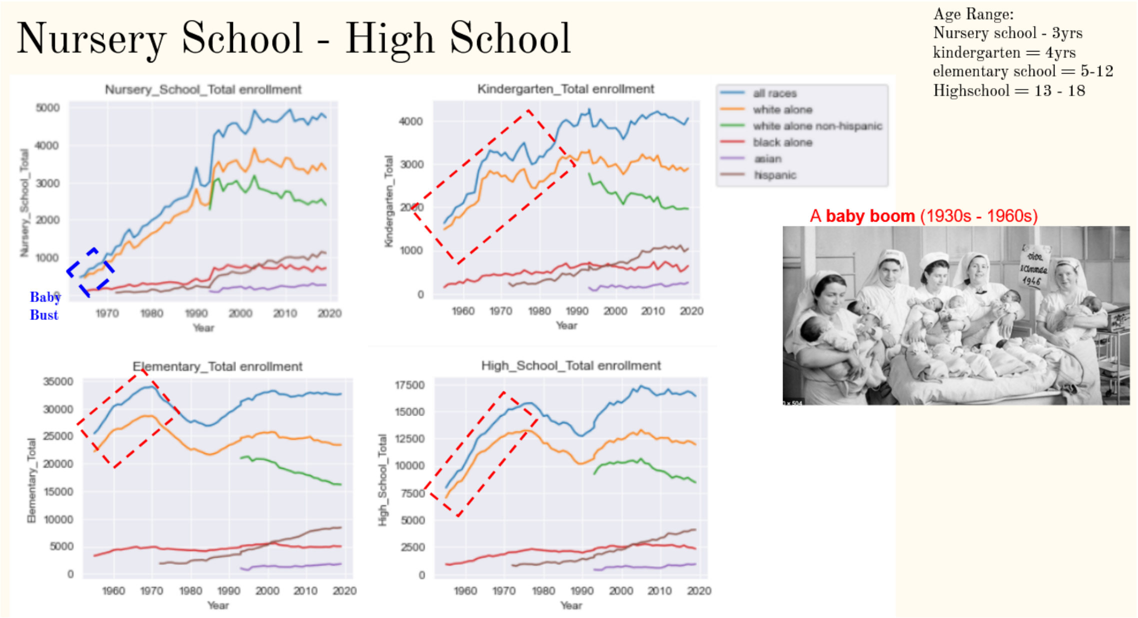
We were expecting to see a percentage increase in dropout lead to a percentage decrease in enrollment, and a percentage decrease in dropouts lead to a percentage increase in enrollment. We didn't necessarily see this to be a consistent pattern across the board. There were a few observations where it appeared that a relatively large percentage increase or decrease in dropouts vaguely coincided with what was expected to be observed for enrollment. No graphs are presented in the summary report but can be viewed in Jupyter notebook.

Question 5: Are there any historical events that may provide insight to the enrollment trends?

As we identified certain trends, we attempted to think about certain events that may have had a relation to the trends that we were observing. Our objective was not to suggest or imply causation, but to think about credible events that may have had a relation to the trends. This was an important part of our process because we believe that sometimes data scientist must look beyond the data to get a full picture of the story.

Below, we have provided a few examples that may explain the trend being observed.

Based on *Figure 7 (listed below)*, the red boxes show a high enrollment spike for elementary school, Kindergarten, and high school. Because the age demographics are between 4 and 18 years, we believe that is credible that there may be a connection to what's known as the “baby boom”; where after WW2, the birth rate of child births spiked exponentially during the 1930s and 1960s. The blue box under nursery school enrollment we associate with the “Baby-Bust” where the rate slowed down. Though co-relation does not mean causation, we can still assume that the Baby Boom had a significant relevance to this data trend.



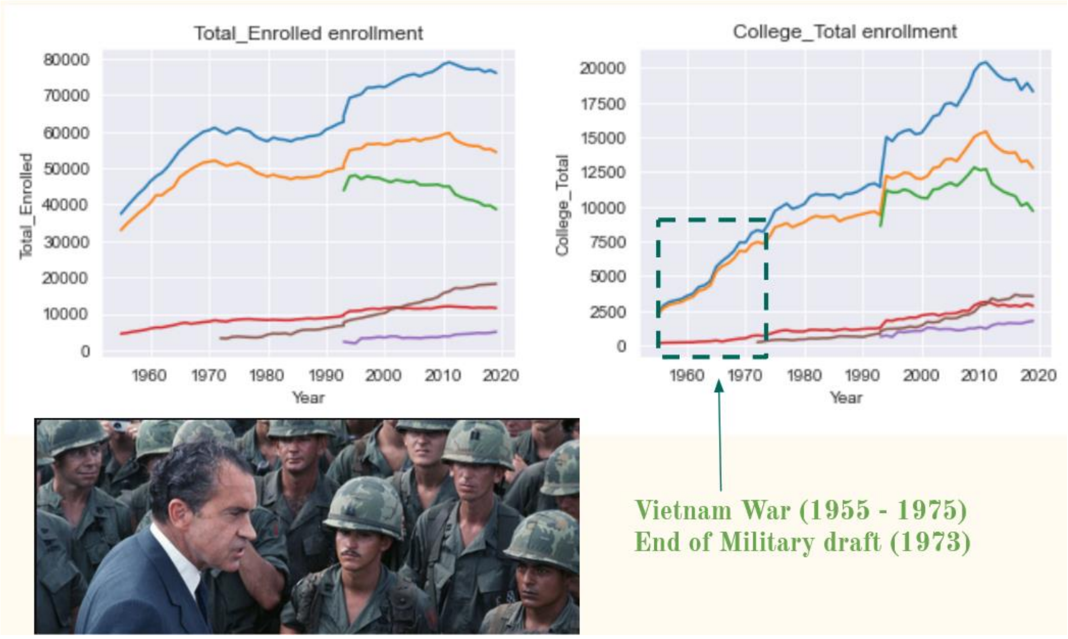
Nursery School - High School

Age Range:
Nursery school - 3yrs
kindergarten = 4yrs
elementary school = 5-12
Highschool = 13 - 18

A **baby boom** (1930s - 1960s)

Figure 7 The Baby Boom

Based on figure 8 (listed below), The Vietnam war lasted from 1955 to 1975, even though we see college enrollment steadily increasing during this time period, we believe that it is possible that the numbers may have been higher if it was not for the Vietnam war. We believe this is possible because during the Vietnam war millions of men were drafted who potentially could have enrolled in college.



Vietnam War (1955 - 1975)
End of Military draft (1973)

Figure 8 The Vietnam War

Lessons Learned:

One of our objectives was to find a data set that was interesting and valid. Once we were able to decide on a data set, we had to decide on which questions would help us to tell a compelling story. We learned that there are not always simple answers to questions. We constructed the code hoping that the answers would be clear and concise; but we realized that this was not always the case. There were times when we would have to refer to other graphs in our program to gain a better insight. There were also times when we had to look outside of the data to try and gain a better understanding of our results. Then, there were times when we could not gain any insight. We learned that data science will not always provide an easy and convenient answer to questions. We learned that in the search for answers, it easy to get carried away. Because of time constraints, we are unable to explore and dive deeper. Looking at enrollment data from 1955 through 2019 was like looking through a window into US history. If we were to continue this further, we would consider looking into other data sets that would help to provide more insight. We would look into other data that contains information about different regions, sex, and economic status.

References and Resources:

1. **Original data source**
   https://www.census.gov/data/tables/time-series/demo/school-enrollment/cps-historical-t ime-series.html

2. **Supplemental data**

   https://www.census.gov/data/tables/time-series/demo/school-enrollment/cps-historical-t ime-series.html

GitHub Repo (https://github.com/UC-Berkeley-I-School/Project2_Trinidad_Brown)