Analysis of High School Graduate Data Using Database Analytics Tools

Ezana Ćeman*, Ajdin Salihović*, Samed Jukić*
*International Burch University,
Sarajevo, Bosnia and Herzegovina
ezana.ceman@stu.ibu.edu.ba
ajdin.salihovic@stu.ibu.edu.ba
samed.jukic@ibu.edu.ba

Original research

Abstract: *It can be confidently stated that access to education is one of the most prized possessions available to us today. Although there are underlying factors such as the discrepancies in the education being provided worldwide, it is imperative that data scientists and all those interested take advantage of the data publicly available to draw necessary insights into how to better the education sector in our respective countries. The purpose of this research is to showcase various analytical insights into the 2020 New York State (NYS) high school graduation rate data using various advanced database systems techniques, specifically using SQL. With these analyses, further studies and conclusions can be drawn for local governments to implement into their plans to increase the quality of the schooling system, to aim for equality for all without regard to cultural and ethnic background, and to find discrepancies within the current system.*

Keywords: **Database, data analysis, graduate, high school, New York State, SQL.**

1. Introduction

The realm of education is one topic always of discussion due to the worldwide discrepancies in the quality of education being provided. Although education is every human being's right, access to basic education is a duty that must be fulfilled within countries worldwide. “Governments are typically expected to ensure access to basic education, while citizens are often required by law to attain education up to a certain basic level [1].” Although every country has its education system, what the world lacks is one uniform system, ensuring equal access to education for all. What Figure 1 below shows us is government expenditure per student as a percentage of GDP per capita, in terms of secondary education by country [2]. We can see that the top four countries with the highest expenditure rates are: The United Kingdom, Japan, the United States, and India.

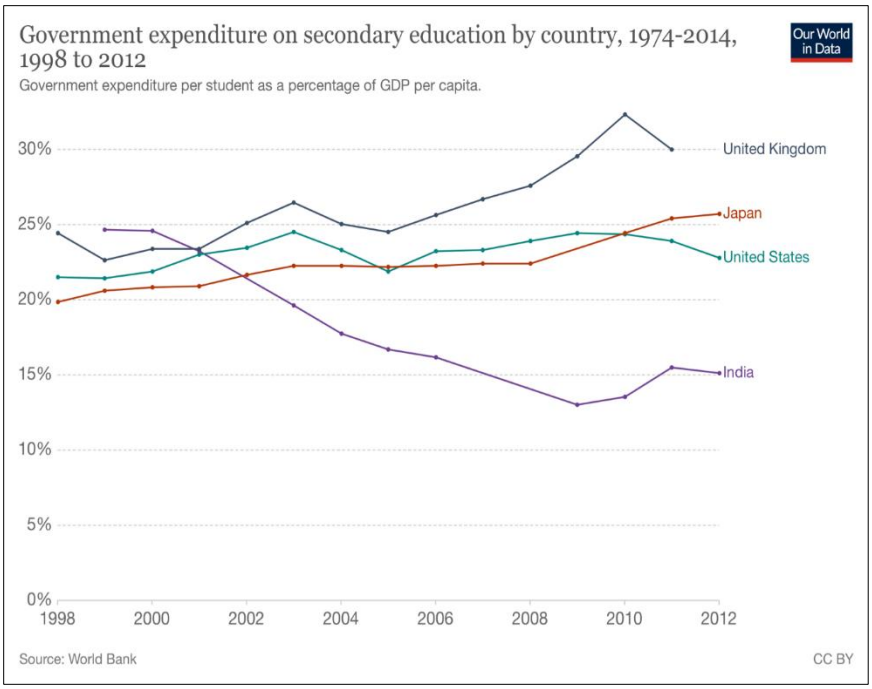

FIGURE 1. Government expenditure on secondary education by country, 1974-2014, 1998 to 2012 - *Source: Our World in Data*

Aside from government expenditures into the secondary education sector, gross enrollment ratios are imperative to analyze as enrollment of young men and women in secondary education, regardless of age statistics, can provide us with better insights into which countries are currently ranked with the highest values. Figure 2 below shows the world's leaders in gross enrollment in secondary education ranging from 1970 to 2014 with Portugal and Barbados ranking the highest with heavily indebted poor countries (HIPC) ranking the lowest [2].

We will admit that these results did come as a shock to us as the first instinct is almost always for our minds to look to world superpowers for the highest enrollment ratios. As we can see, this predisposition is false when looking at the data.

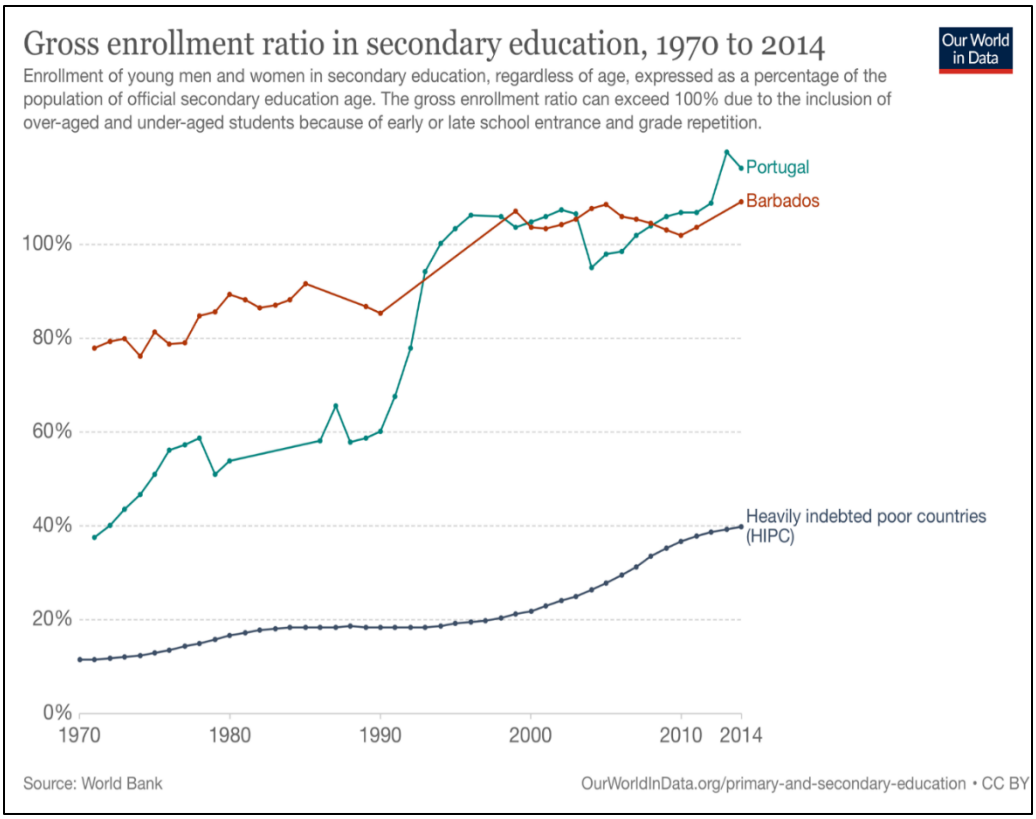

FIGURE 2. Gross enrollment ratio in secondary education, 1970 to 2014
- *Source: Our World in Data*

For this research, the region of the world that will be focused on pertains to the northeast United States, specifically the state of New York. NYS currently has 731 districts, 4,421 public schools, and 351 charter schools actively functioning at the moment [3]. Out of all the 4,421 public schools, there are a total of 2,053 high schools in NYS, including 1,520 public schools and 533 private schools [4]. As of June 30th, 2019, there were a total of 2,598,921 K-12 public school students in NYS [3]. From 2019-20, 815,707 students were recorded as high school students, ranging from grades 9, 10, 11, 12, and ungraded secondary which is the standard secondary school grades in NYS [5]. An additional grade level, denoted 'Ungraded Secondary' represents specialized high schools in NYS that do not follow the standard grading system. However, these students fall under the category of high school students nonetheless.

As ethnicity will play a significant role in this research, it is worth noting the latest available data regarding enrollment by ethnicity. Figure 3 below represents the 2019-20 enrollment count based on ethnicity [5]. White-identifying students made up the highest group, with an estimated 353,000 high school students. Hispanic or Latino high school students accounted for an estimated 219,000 students. Black or African American high school students totaled 141,000. Asian or Native Hawaiian/Other Pacific Islander high school students totaled 81,000. Multiracial high school students accounted for a total of 17,000 and lastly, there were a recorded 6,000 American Indian or Alaska Native high school students.

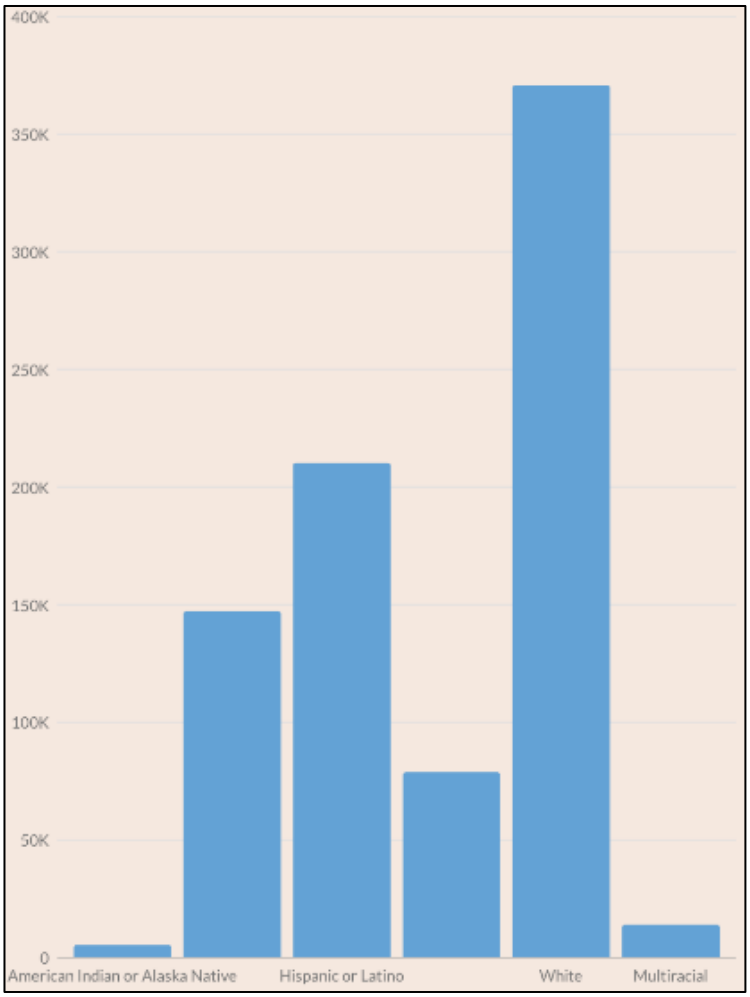

FIGURE 3. Enrollment by Ethnicity, 2019-20 - *Source: The New York State Education Department*

In addition, it is worth noting the latest data available in terms of enrollment by grade to provide us with an overview of how many students are in each of the high school grades relating to this study. Table 1 below represents the data available from 2019-20 in terms of enrollment by grade [5].

Above we can see that there is an about-even split between grades 9 – 12 regarding enrollment by grade.

There is a multitude of database analytics tools available for use with the main purpose of utilization being for data analysis. Not every tool fits every single need however, great insights can be found through researching which tool(s) work the best for the respective field of research. The field of data analytics has been growing day by day as new insights are being pulled from data in endless ways imaginable. The purpose of this research is to showcase various data analyses through the use of database analytics tools by utilizing comparative study techniques in terms of graduation percentages. After finding the right dataset to work with, various steps were taken in the preprocessing of the dataset to ensure that all of the data is being analyzed as accurately as possible. A variety of exploratory and description data analyses were conducted on the 2020 NYS graduation rates dataset to find various comparisons that can be implemented further in additional research.

TABLE 1. Enrollment by Grade, 2019-20
Source: The New York State Education Department

9th Grade	
211,978	26%
10th Grade	
203,562	25%
11th Grade	
191,168	23%
12th Grade	
189,493	23%
Ungraded Secondary	
19,506	2%

2. Methods and Materials

Dataset Overview

The New York State Education Department (NYSED) frequently publishes data in regards to a multitude of aspects of the education sector in the state for public use, which is where we found the dataset to be used in this research. Titled "GRAD_RATE_AND_OUTCOMES_2020", this dataset features a total of 227,451 rows and a total of 37 columns [6]. This gives us an estimated 8,415,687 numerical and categorical values to work with. This dataset showcases "outcomes of designated subgroups are reported by the total public school (aggregated data for all districts and charter schools), county (aggregated data for all districts and charter schools in the county), Needs-to-Resource-Capacity (N/RC) group, district, and public schools [6]."

One of the most significant indicators that will be the main priority of this research pertains to 'SUBGROUP_NAME', which consists of a text value and 25 differentiating characteristics. NYS has a standard definition of the subgroups used within their documentation which is shown in Table 2 below.

TABLE 2. Subgroups (NYS)

1. 'All Students'	13. 'English Language Learner'
2. 'Female'	14. 'Formerly English Language Learner'
3. 'Male'	15. 'Economically Disadvantaged'
4. 'American Indian/Alaska Native'	16. 'Not Economically Disadvantaged'
5. 'Black'	17. 'Migrant'
6. 'Hispanic'	18. 'Not Migrant'
7. 'Asian/Pacific Islander'	19. 'Homeless'
8. 'White'	20. 'Not Homeless'
9. 'Multiracial'	21. 'In Foster Care'
10. 'General Education Students'	22. 'Not in Foster Care'
11. 'Students with Disabilities'	23. 'Parent in Armed Forces'
12. 'Not English Language Learner'	24. 'Parent Not in Armed Forces'.

Each subgroup has its corresponding 'SUBGROUP_CODE', consisting of a specific 2-digit code identifying the demographic at hand.

Utilized Technologies

For the purpose of this research, due to the large size of the dataset being analyzed, we found that through the use of the programming language SQL, we would be able to acquire effective results while using various advanced database system techniques. SQL stands for Structured Query Language and it is "a database computer language designed for the retrieval and management of data in a relational database [7]." We have found through trial and error by testing out other database analytics tools, that we can acquire the greatest results by using the SQL language. By creating our tailored queries, we were able to retrieve various insights into the 2020 graduate rate dataset which in turn will lead us to our conclusions which are noted towards the end of this work. In addition, the Microsoft Excel program was utilized to create the visuals relating to the results of this research.

Data Cleaning and Preprocessing

For this research, cleaning is the process of detection and correcting records from records set, table, or database. For example, if there are incomplete, incorrect, inaccurate, or irrelevant, we would modify or delete the faulty data. In the case of discrepancies in the education being provided worldwide, in this case for the New York State (NYS) high school graduation rate. We might see a discrepancy between students enrolled versus students who graduated in case of seeing more graduates than the enrolled students of that class. We would check the number of students that repeated the year with the number of enrolled students for an example of class 2020. If a certain number of students of class 2019 have repeated the year they would also be graduating at the same time year after. This filtering is used when using WHERE when using SQL in which we can filter records and add more using AND until we are ready to run the query.

3. Results

As with every dataset, significant conclusions can be drawn through the use of statistical analysis techniques which was our first step in the data analysis process. Using both the 'SUBGROUP_CODE' and 'SUBGROUP_NAME' indicators as a basis for our queries, various statistical values were found which in turn helped bring light to the current demographic relationship to graduate rates. Based on these indicators, what was found was: the sum of the total enrolled based on the subgroup, the sum of total graduates based on the subgroup, the sum of total locals based on the subgroup, the sum of total registered based on the subgroup, the sum of total advanced registered based on the subgroup, the sum of total diploma credentials based on the subgroup, the sum of the total still enrolled based on the subgroup, the sum of total dropouts based on the subgroup, the average of enrolled based on the subgroup, the average of graduates based on the subgroup, the average of locals based on the subgroup, the average of registered based on the subgroup, the average of advanced registered based on the subgroup, the average of diploma credentials based on the subgroup, the average of still enrolled based on the subgroup, and the average enrolled based on the subgroup.

From these analyses, the most relevant results found are summarized below.

Figure 4 shows that students in the Non-Migrant, Not in Foster Care, Parent Not in Armed Forces, Not Homeless, and Not English Language Learners categories are the highest in terms of enrolled students.

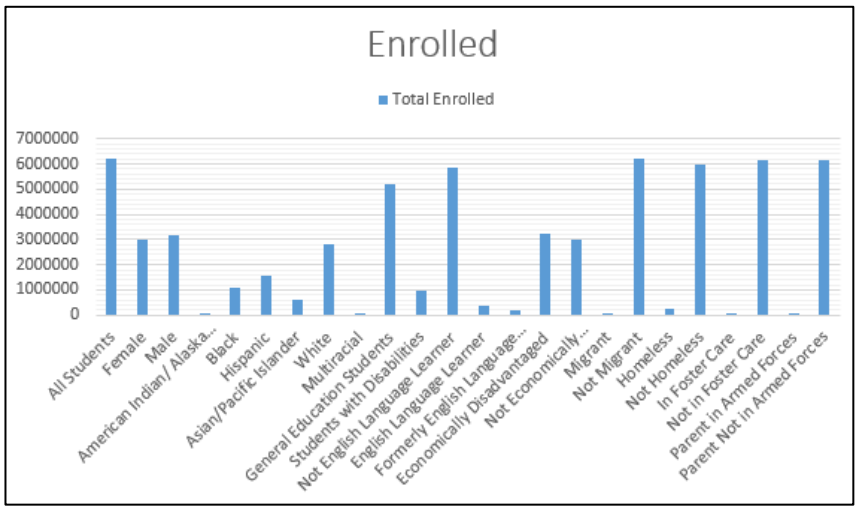

FIGURE 4. Calculates Sum of total Enrolled based on Subgroup.

Figure 5 shows that students in the Non-Migrant, Not in Foster Care, Parent Not in Armed Forces, Not Homeless, and Not English Language Learners categories are the highest in terms of graduated students.

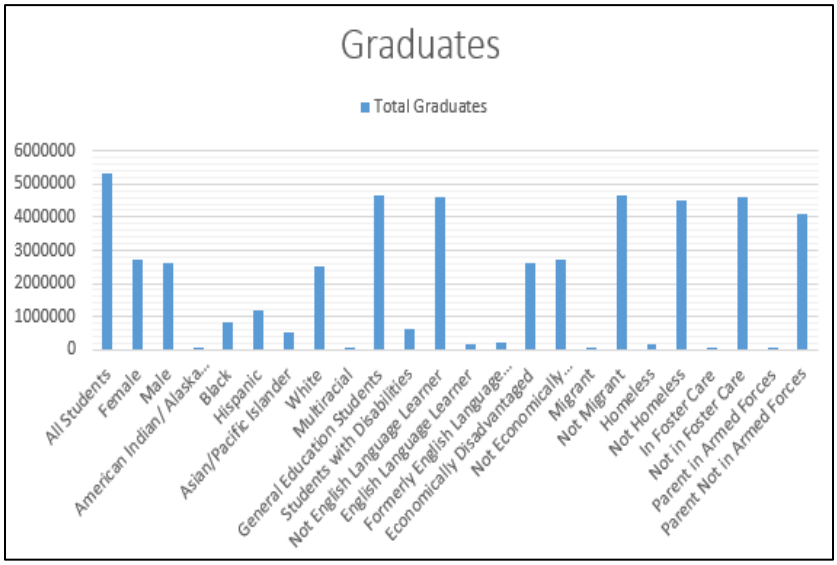

FIGURE 5. Calculates Sum of total Graduates based on Subgroup.

Figure 6 shows us that students in the Non-Migrant, Not in Foster Care, Parent Not in Armed Forces, Not Homeless, and Not English Language Learners categories are the highest in terms of local students. As can be seen, there is a recurring trend here thus far.

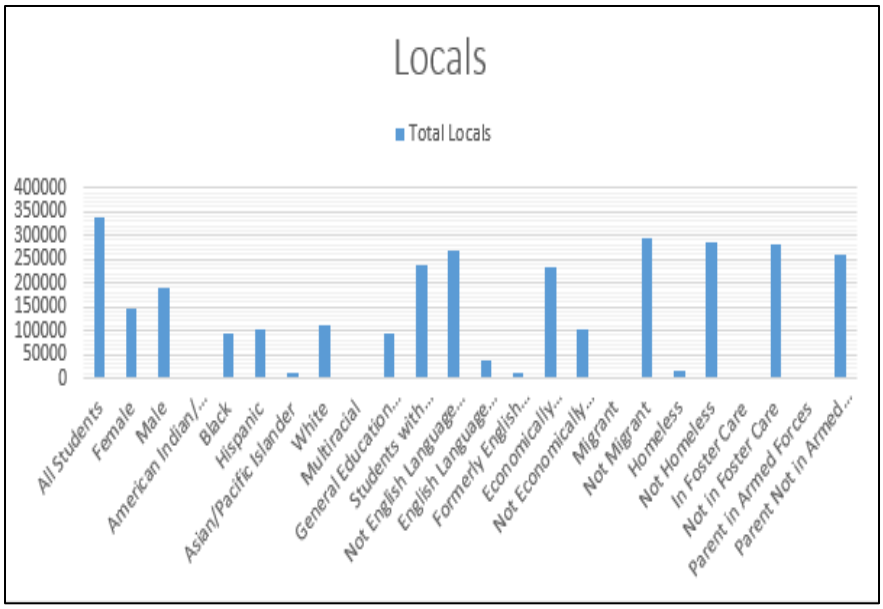

FIGURE 6. Calculates Sum of total Locals based on Subgroup.

Figure 7 shows us a representation of the sum of total registered students based on their respective subgroups. Similar to Figures 4 – 6, there is a recurring theme present.

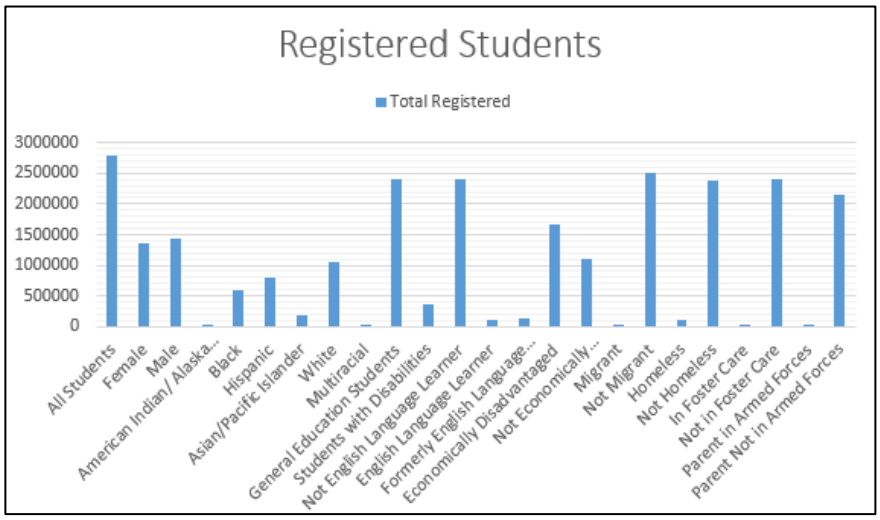

FIGURE 7. Calculates Sum of total Registered based on Subgroup.

Figure 8 shows us that General Education Students followed by Not Migrant, Not Homeless, and Not in Foster Care make up the highest groups for advanced registered students. In NYS, advanced registered stands for students who are taking college-level classes at a high school level.

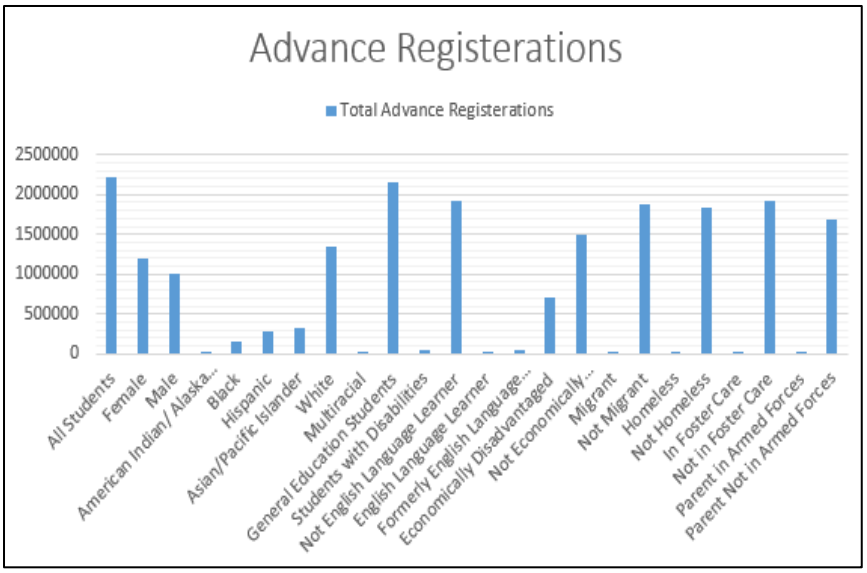

FIGURE 8. Calculates Sum of total Advance Registered based on Subgroup.

Figure 9 shows us that Students with Disabilities followed by Not Migrant, Not Homeless, and Not in Foster Care make up the highest groups for total diploma credentials.

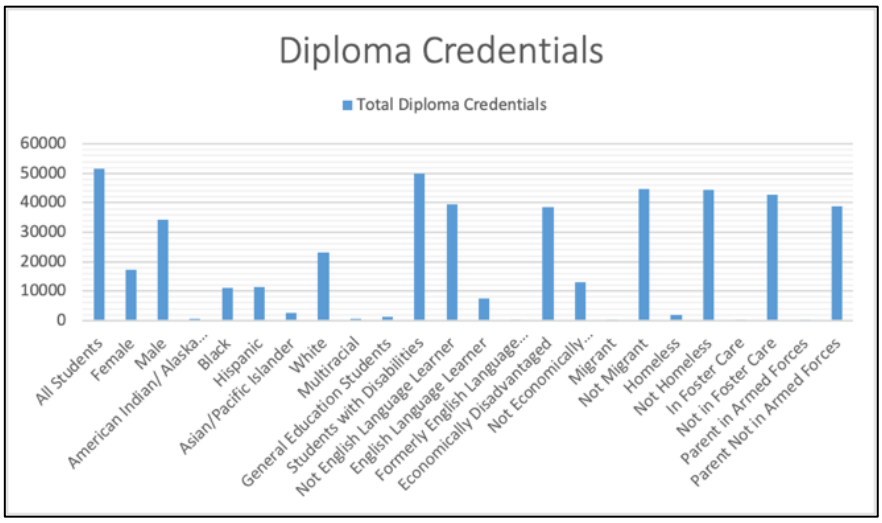

FIGURE 9. Calculates Sum of total Diploma Credentials based on Subgroup

Figure 10 shows us a graphical representation of the sum of the total still enrolled students based on the subgroup. A recurring theme to the previous figures remains.

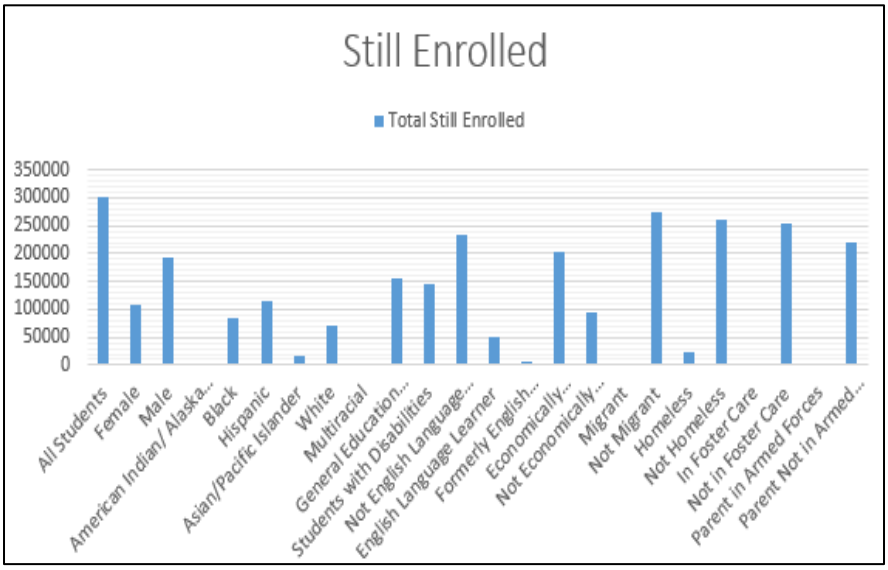

FIGURE 10. Calculates Sum of total Still Enrolled based on Subgroup.

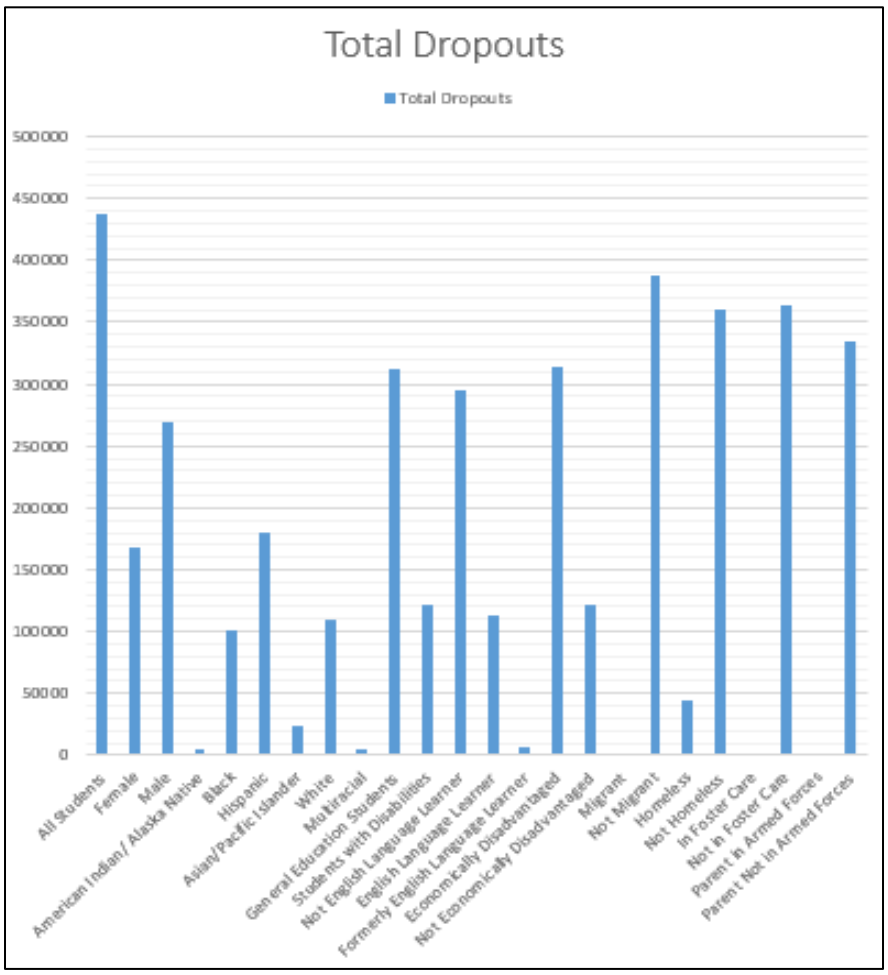

FIGURE 11. Calculates Sum of total Dropouts based on Subgroup.

Following these relevant sums and averages computed, we wanted to showcase the statistics of total dropped out students county-wise for those counties which have more than a total of 10,000 dropout students. For reference, NYS is broken up into 62 different counties which are all included within the dataset as well.
Figure 12 below was created which represents the desired output.

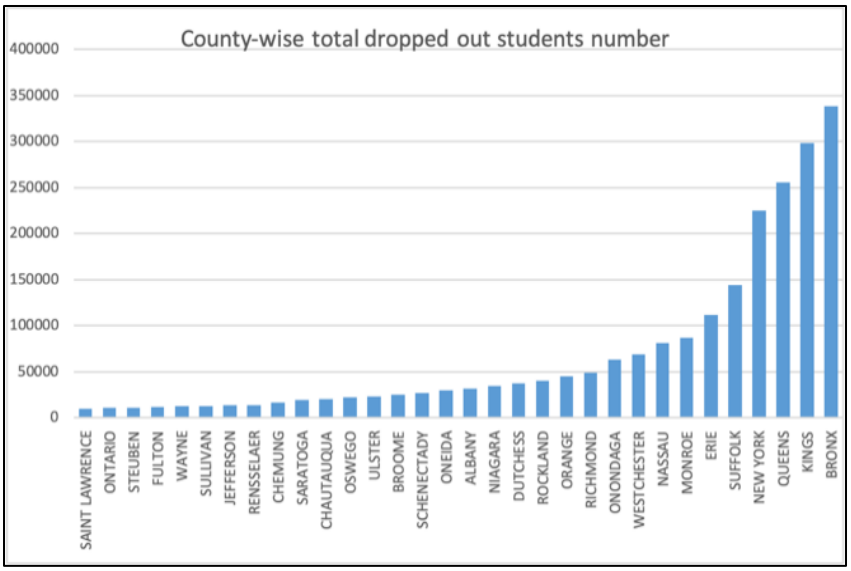

FIGURE 12. The county-wise total dropped out students.

The graph clearly shows that Bronx county has the maximum number of dropout students whereas Saint Lawrence county has the minimum among the counties that have more than 10,000 dropout students.
Following the computation of county-wise data, what follows is the statistics of graduation percentages statewide over the various subgroups for the '2014 Total Cohort - 6 Year Outcome' membership.
Figure 13 below was created which represents the desired output.

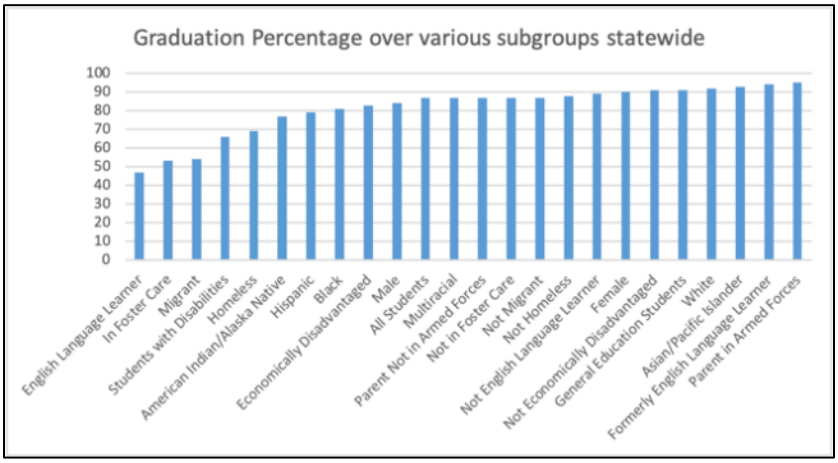

FIGURE 13. Graduation percentages over various subgroups.

By analyzing the output data, it was found that English Language Learners have the lowest percentage of graduation whereas students whose parents are in the Armed Forces have

the highest percentage of graduation. It's tough for a nonnative English Language Learner to get used to an entirely new language and way of life so it may be a reason for the lowest percentage. The students whose parents are in the Armed Forces may have better government-provided facilities which is a catalyst for their highest percentage. The graph also shows that female students have a higher percentage of graduating compared to male students. English Language Learners, In Foster Care, and Migrant subgroups students should be given special care because their graduation percentage is the lowest and below 60. Following the aforementioned analysis, we wanted to list the counties and number of schools in each of them where no students graduated in various subgroups over the cohort '2014 Total Cohort - 6 Year Outcome'. Only those county schools are to be considered where the Need to Resource Capacity category is 'Urban-Suburban High N/RC Districts'.

The results have been visualized in Figure 14 below.

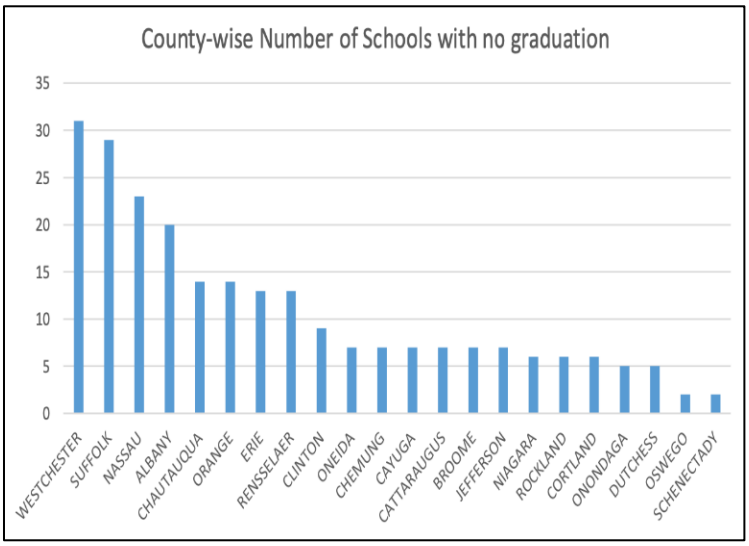

FIGURE 14. County-wise Number of Schools with no graduation.

By analyzing the above data, we see that Westchester County has the highest number of schools where no students in various categories graduated in the last six years.

Next, a comparative study was created between the graduation percentage of all students vs. black students over various counties where at least 100 black students were enrolled during the '2014 Total Cohort - 6 Year Outcome' period.

The results have been visualized in Figure 15 below.

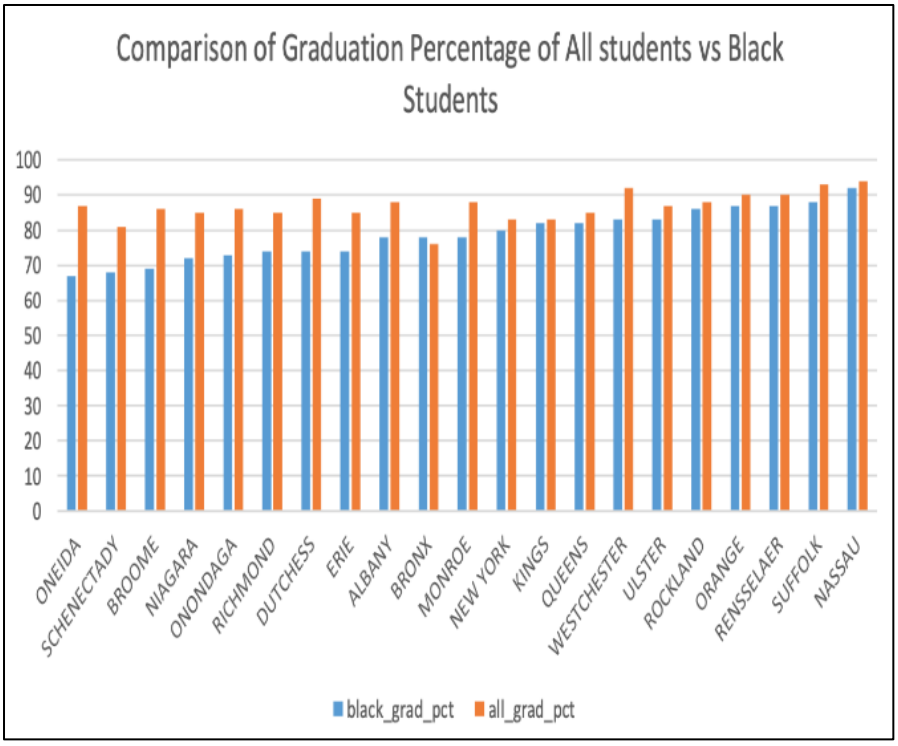

FIGURE 15. Comparison of Graduation Percentages of All Students vs. Black Students.

The graph shows that almost every county graduation percentage for Black students is smaller than the percentage of All Students. Only Bronx County is different where the percentage is greater for Black students than that of All Students. The demographic breakdown for Bronx County is 43.6% Black or African American [8].

Lastly, the final analysis conducted in this study was to find the dropout percent of students based on the 'Need to Resource Capacity' category for the 'All Students' subgroup and the cohort '2016 Total Cohort - 4 Year Outcome - August 2020'.

The tabular output is depicted in Table 3 below.

TABLE 3. Dropout percentages based on selected indicators.

aggregation_name	dropout_pct
NRC: Low Needs	1%
NRC: Charters	3%
NRC: Average Needs	4%
NRC: NYC	6%
NRC: Rural High Needs	7%
NRC: Buffalo, Rochester, Yonkers, Syracuse	10%
NRC: Urban-Suburban High Needs	11%

The tabular output shows that the NRC: Low Needs category has the minimum dropout percentage and the NRC: Urban- Suburban High Needs category has the maximum.

4. Discussion and Conclusions

Throughout this research journey, various advanced database systems techniques were utilized to be able to provide insights into the graduation rate dataset provided to the public by the New York State Department of Education. The application of various data analytic processes was successfully done here through various steps along the way: data collection, data cleaning, data preprocessing, data analysis, and both data and results interpretation. There were a few minor obstacles that came in the way when it came to our implementation of the aforementioned steps yet results were found, analyzed, and interpreted.

Here is a summary of the results gathered from this research study:

1. Students in the Non-Migrant, Not in Foster Care, Parent Not in Armed Forces, Not Homeless, and Not English Language Learners categories are the highest in terms of enrolled students.
2. Students in the Non-Migrant, Not in Foster Care, Parent Not in Armed Forces, Not Homeless, and Not English Language Learners categories are the highest in terms of graduated students.
3. Bronx county has the maximum number of dropout students whereas Saint Lawrence county has the minimum among the counties that have more than 10,000 dropout students.
4. English Language Learners have the lowest percentage of graduation whereas students whose parents are in the Armed Forces have the highest percentage of graduation.
5. Westchester County has the highest number of schools where no students in various categories graduated in the last six years.
6. Almost every county graduation percentage for Black students is smaller than the percentage of All Students. Only Bronx County is different where the percentage is greater for Black students than that of All Students (due to current demographics).

This topic was an interesting one that has peaked our internet for potential further research opportunities. There are various ways to go about interpreting data for usable results and this research paper just shows a few methods using the tools available to us to do so. The possibilities are endless with data science, and hopefully, through the use of this work, conclusions will be drawn to ensure differences are made within local, state-wide, and national governments to push for equality for all in terms of accessibility and quality of information.

7. References

[1] Roser, M. (2016, August 31). Global Education. Our World in Data. https://ourworldindata.org/global-education

[2] Roser, M. (2013, July 17). *Primary and Secondary Education*. Our World in Data. https://ourworldindata.org/primary-and-secondary-education

[3] NYSED Data Site. (n.d.). https://data.nysed.gov

[4] High Schools (h.s.). New York High Schools. https://high-schools.com/directory/ny/

[5] 2018 | NY STATE - Enrollment Data | NYSED Data Site. (n.d.). NYSED Data Site. Retrieved June 1, 2021, from https://data.nysed.gov/enrollment.php?state=yes&year=2018&grades%5B%5D=09&grades%5B%5D=10&grades%5B%5D=11&grades%5B%5D=12&grades%5B%5D=14

[6] https://data.nysed.gov/files/gradrate/19-20/gradrate.zip

[7] SQL Tutorial - Tutorialspoint. (n.d.). Tutorialspoint. Retrieved June 1, 2021, from https://www.tutorialspoint.com/sql/index.htm

[8] Bronx County, NY. (n.d.). Data USA. Retrieved June 1, 2021, from https://datausa.io/profile/geo/bronx-county-ny

[9] National Research Council and National Academy of Education. (2011). High School Dropout, Graduation, and Completion Rates: Better Data, Better Measures, Better Decisions. Committee for Improved Measurement of High School Dropout and Completion Rates: Experuidance on Next Steps for Research and Policy Workshop. R.M. Hauser and J.A. Koenig, Editors. Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies PrPress.

[10] Anderson, K.P., and Ritter, G.W. (2017). Disparate use of exclusionary discipline: Evidence on inequities in school discipline from a U.S. state. Education Policy Analysis Archives, 25(49).

[11] Balfanz, R., Herzog, L., and MacIver, D.J. (2007). Preventing student disengagement and keeping students on the graduation path in urban middle-grades schools: Early identification and effective interventions. Educational Psychologist, 42(4), 223-235.

[12] American Educational Research Association. (2015). AERA statement on the use of value-added models (VAM) for the evaluation of educators and educator preparation programs. Educational Researcher 44(8), pp. 448-452. Available: https://journals.sagepub.com/doi/10.3102/0013189X15618385 [August 2019].

[13] Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. Educational Researcher, 36(5), 258-267.

[14] National Academies of Sciences, Engineering, and Medicine. 2019. Monitoring Educational Equity. Washington, DC: The National Academies Press. doi: 10.17226/25389.×

[15] Aucejo, E., and Romano, T.F. (2016). Assessing the effect of school days and absences on test score performance. Economics of Education Review, 55, 70-87. Available: https://doi.org/10.1016/j.econedurev.2016.08.007 [March 2019]