College Readiness of California High School Graduates

by Peter Bloomsburgh

Introduction, Motivation, and Goals

California has two public four-year university systems. The University of California system has ten campuses, and the California State University system has twenty-three campuses. In order for a California high school graduate to be eligible to apply to one of these universities, they must meet the "a-g" requirements. These requirements are based on courses they have successfully completed during high school. Data is available from the California Department of Education about the percentage of California high school graduates who meet these requirements, disaggregated along several dimensions.

I have two goals for doing this project. First, I want to encourage my local school district, the Berkeley Unified School District, to focus more on improving student achievement. There are currently too many students who are not eligible to apply to California's many excellent four-year public universities. I want the school district to set measurable goals, develop a specific action plan, and design a way to measure progress. I plan to encourage them to do this using information visualizations that tell a compelling story.

My second goal is to do an Exploratory Data Analysis to see what patterns and trends exist throughout California regarding preparing high school students for college admittance. Information visualizations will be important here as well.

Executive Summary

First, I located, loaded, explored, cleaned, and manipulated data about the college readiness of California high school graduates. This data was available for each of the past five years. It is available at the state, county, school district and school level. It is disaggregated by ethnicity, gender, socioeconomic status, and participation in two special programs.

Next, I developed six Python functions to more efficiently support analysis and visualizations. Finally, I used the data and Python functions to do Exploratory Data Analysis for the state as a whole and to compare my local school district with a neighboring school district.

A table of contents for the remainder of this report appears below. The body of the report shows sixteen visualizations with my comments below them.

Table of Contents

Project Methodology

The following six steps were completed first to facilitate the analyses and visualizations.

1. **Locate three data sources**

California public school graduation data

- for the five years from 2016-17 to 2020-21
- available from the California Department of Education website
- url: https://www.cde.ca.gov/ds/ad/filesacgr.asp
- format: tab-delimited text
- data includes the number of high school graduates and the number of graduates who met the "a-g" college readiness requirements
- data is available at four aggregation levels: state, county, school district, and school
- data is disaggregated by ethnicity, gender and participation in several special programs
- a total of about 1.1 million records

California public school funding data

- for the five years from 2016-17 to 2020-21
- available from the California Department of Education website.
- url: https://www.cde.ca.gov/ds/fd/ec/currentexpense.asp
- format: Excel
- data available for each school district
- data includes school district funding per student
- a total of about 4,700 records

Geojson data for United State counties

- "us_10m" dataset available from "vega_datasets"

2. **Load data**

- downloaded the five graduation data files from the California Department of Education website, converted them to Excel format, and then loaded them into Pandas dataframes
- downloaded the five school district funding data files and loaded them into Pandas dataframes
- accessed the geojson data for US counties as suggested by the Altair documentation

3. **Explore data**

- looked for missing or inaccurate data
- verified column labels were consistent across the five years

4. **Clean data**

- clean the data to address problems identified during the "Explore" phase, fixing bad data where possible
- converted some data to integers
- this step was automated and carefully documented to ensure it could be reproduced

5. Manipulate data

- dropped some dataframe rows
- combined some dataframe columns and dropped other columns
- added numerous calculated dataframe columns
- combined some "ReportingCategory" fields and modified all "ReportingCategory" labels to improve readability
- grouped the 58 California counties to create 6 regions
- merged graduation and school district funding data into one dataframe for each year
- combined dataframes for all five years into a single dataframe
- exported the single combined csv file for efficiency purposes [since it took five to ten minutes to run the Jupyter notebook which does data loading, cleaning, and manipulation]
- this step was also automated and carefully documented to ensure it could be reproduced

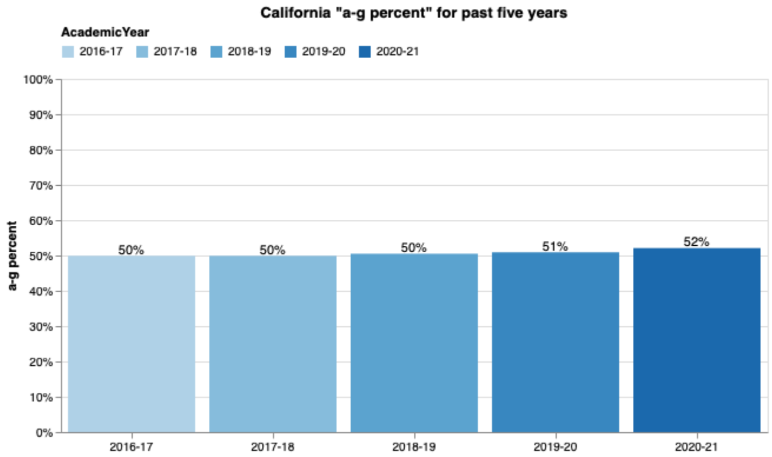
6. Develop Python functions for displaying visualizations

- five Python functions to create line, bar, scatter, pie and histogram charts using Altair
- a Python function named "filter_df" to select which dataframe rows to use for a particular visualization

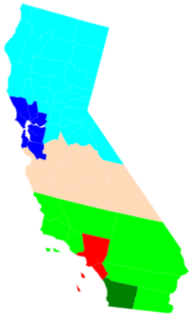
Two Jupyter notebooks document this process in detail:

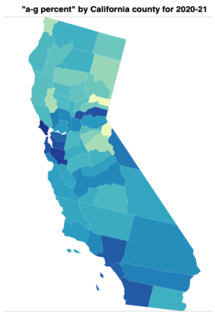
- "Notebook1.ipynb" – data loading, exploring, cleaning, and manipulation
- "Notebook2.ipynb" – a little more data manipulation, the six Python functions, and the visualization code

Analysis and Visualizations – California Patterns and Trends

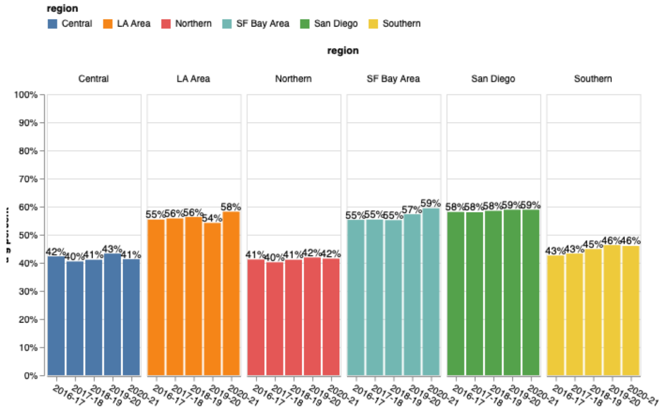


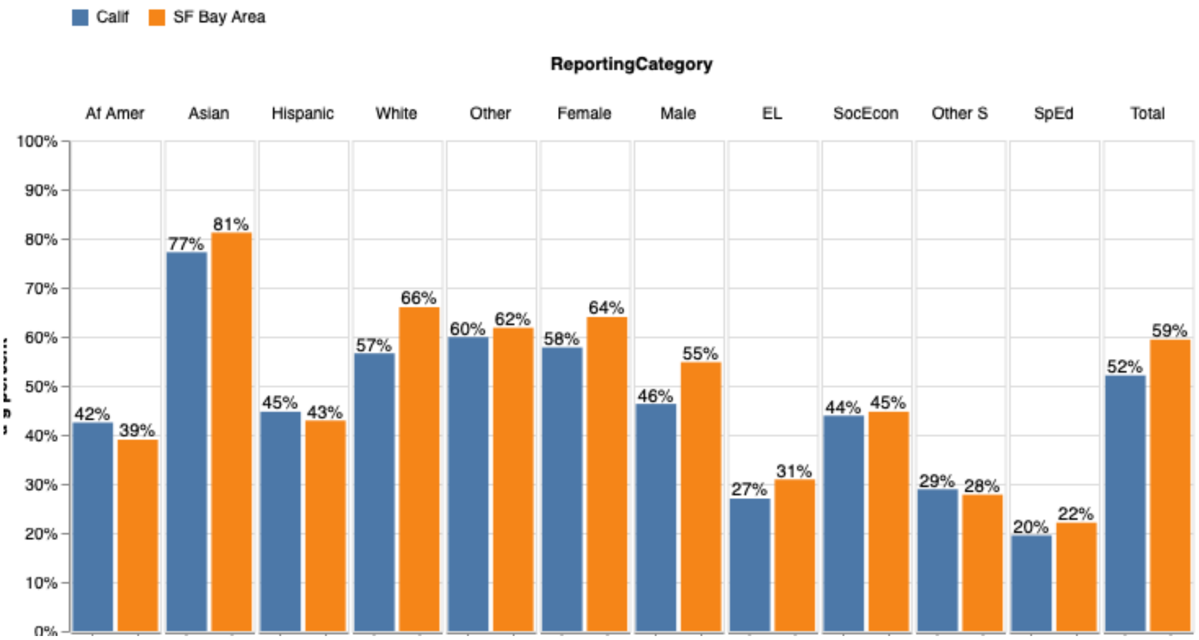
As shown above for the state as a whole, the overall % of graduating students meeting the "a-g" requirements remained very consistent over the past five years. It ranged from 50% for 2016-17 to 52% for 2020-21. This means that many students met them, but many did not. There were no apparent patterns involving COVID (2019-20 and 2020-21) compared to pre-COVID.





As shown above on the left, I assigned the 58 California counties to 6 regions, 3 urban (LA Area, SF Bay Area, and San Diego) and 3 rural (Northern, Central, and Southern). The urban regions tended to have higher "a-g" percentages than the rural regions, as shown in the charts above right and below.

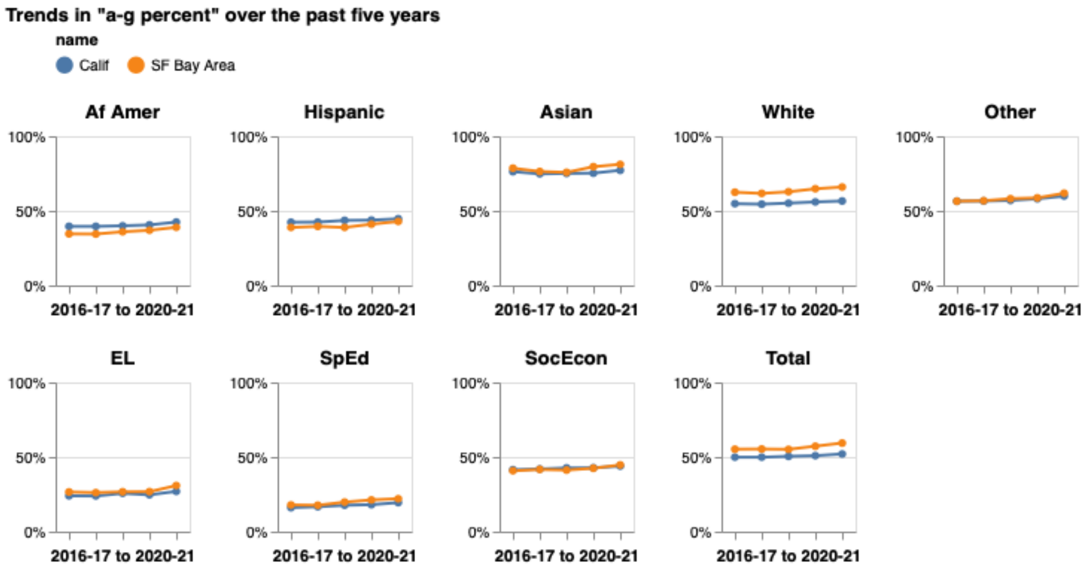




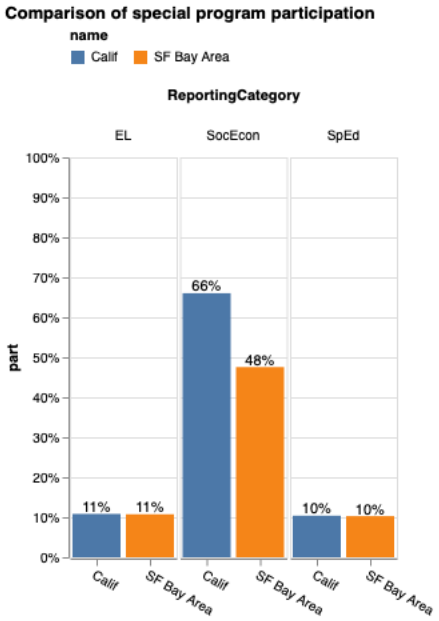
As shown above, there was significant variation among different ethnicities, genders, socioeconomic status, and participation in special programs. This was true for both the state as a whole and for the San Francisco Bay Area region:

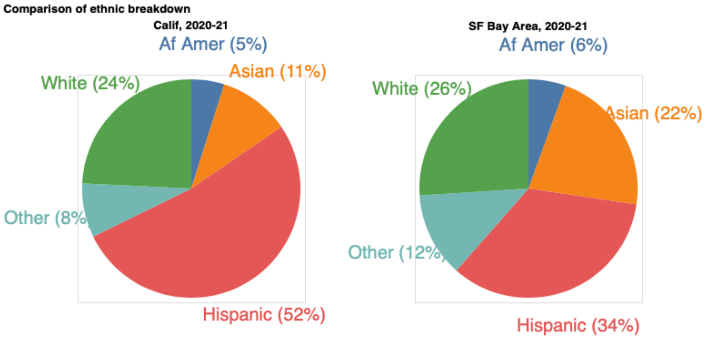
- Asian students were highest and African American and Hispanic were much lower
- females were higher than males
- students with low socioeconomic status were much lower
- English Learner and Special Education students were lower

None of these findings were surprising to me.

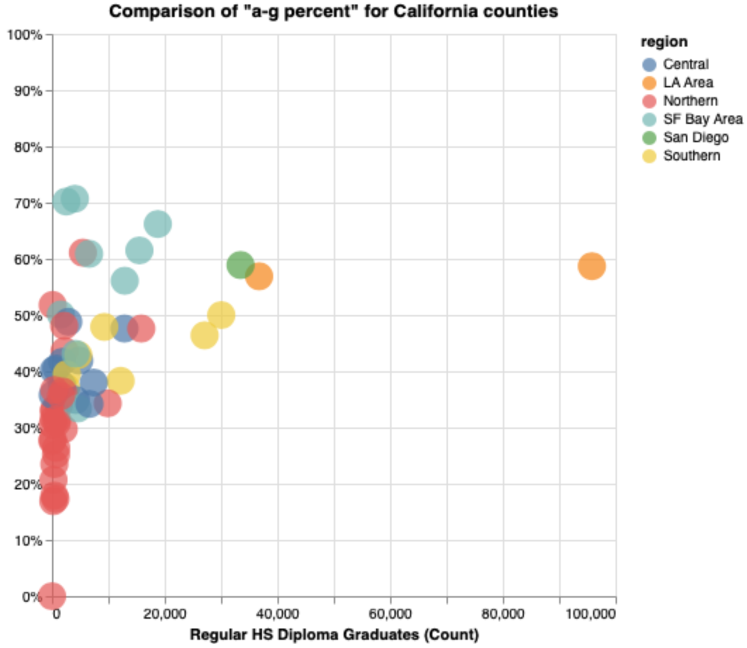


The patterns described above remained fairly consistent throughout the past five years.





As shown above for the state as a whole, 52% of graduates were Hispanic, 24% White, 11% Asian and 5% African American. A total of 66% were of low socioeconomic status, about 10% were English Learner students and a similar number were Special Education students. For the San Francisco Bay Area region, there were twice as many Asians, fewer Hispanics, and fewer of low socioeconomic status. These statistics help explain why the San Francisco Bay Area region had a higher percentage of graduates meeting the ”a-g” requirements.

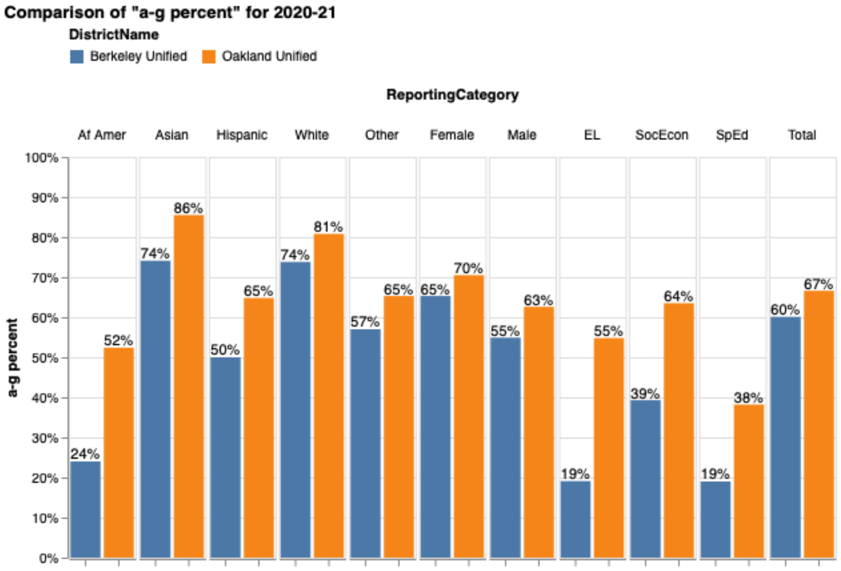


As the above chart indicates, the number of high school graduates varied greatly among the state’s 58 counties, from a few hundred to about 96,000 (for Los Angeles County). There was also tremendous variation in the % meeting the “a-g” requirements, ranging from 18% to 71%.

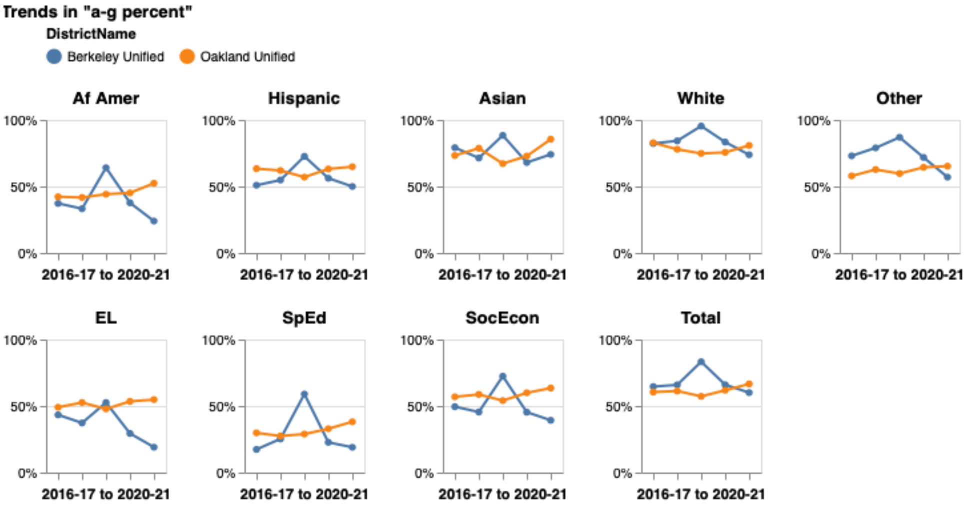
Analysis and Visualizations – Berkeley vs. Oakland

I used the saved dataframe and the six Python functions to compare two school districts, Berkeley and Oakland. Berkeley is a city of 123,000 people, and Oakland is a city of 423,000 people. They border each other and are across the water from San Francisco.

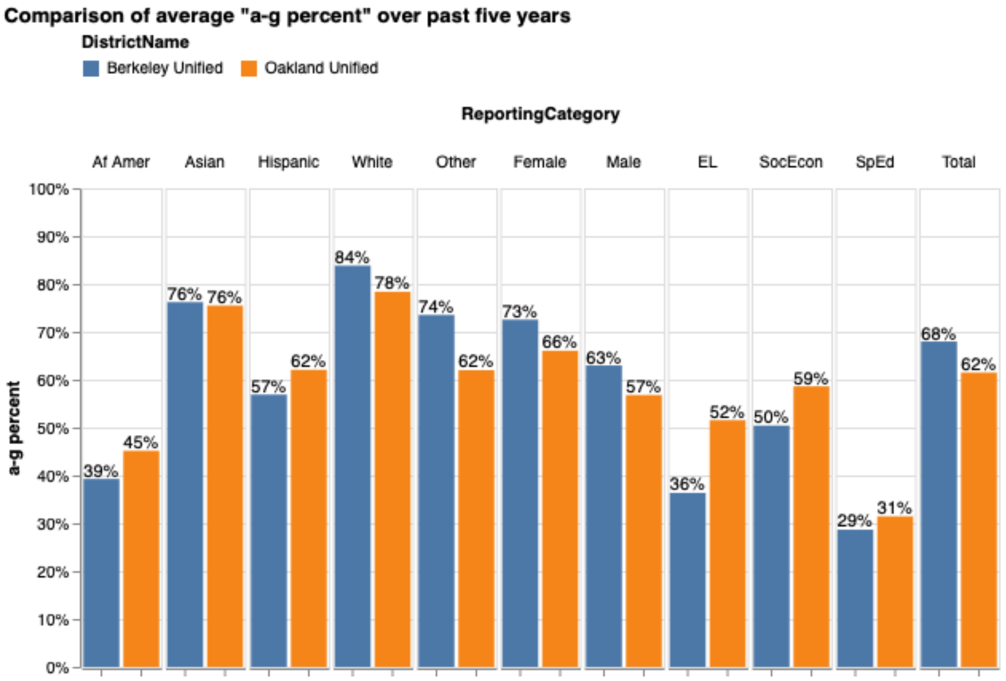
The results were very surprising to me. About 25 years ago, I analyzed standardized test scores for White, Asian, African American and Hispanic students from several school districts in the county, including Berkeley and Oakland. I found that Berkeley had some of the best scores for every ethnic group. That was not true this time, twenty-five years later.



As shown above, for the most recent school year, Berkeley had lower results than Oakland for every ethnic group, every gender, low socioeconomic students, and both special programs.

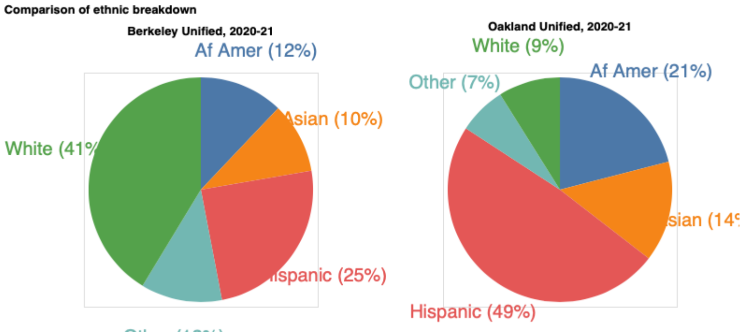


As indicated above, these patterns mostly held for four of the past five school years. More research is needed to better understand what happened in the 2018-19 school year.



Comparison of average "a-g percent" over past five years

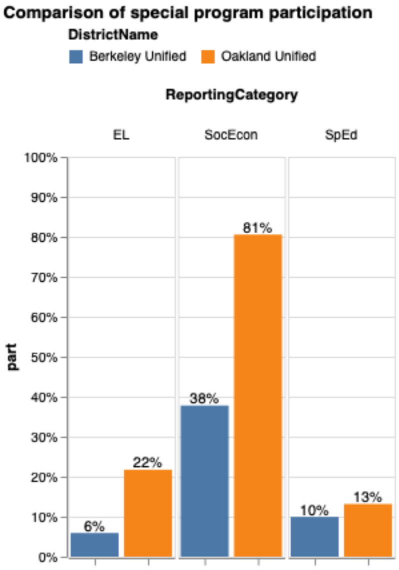
I averaged the past five years, including the unusual 2018-19 results. As shown above, Berkeley was still significantly lower for African American, Hispanic, English Learner, and low socioeconomic students. However, Berkeley was higher overall, 68% to 62%. A comparison of student populations probably explains this, as shown below.



Comparison of ethnic breakdown

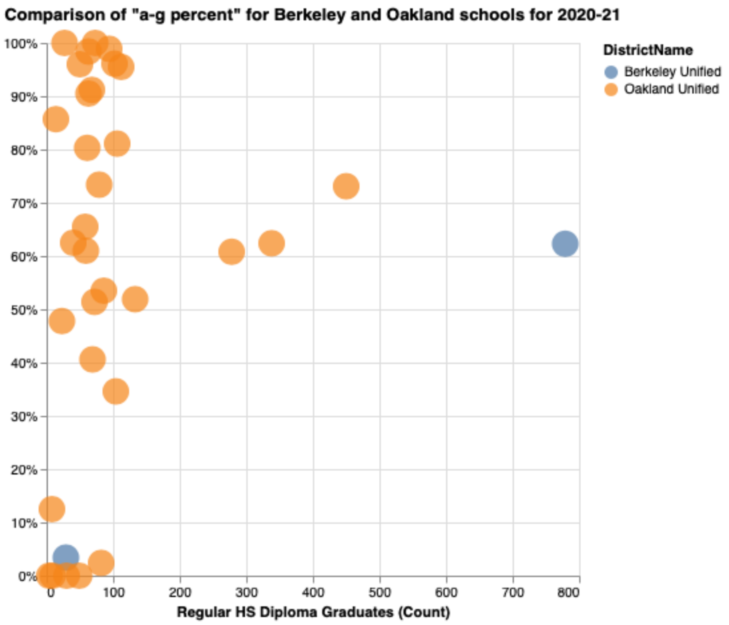
Berkeley Unified, 2020-21

Oakland Unified, 2020-21

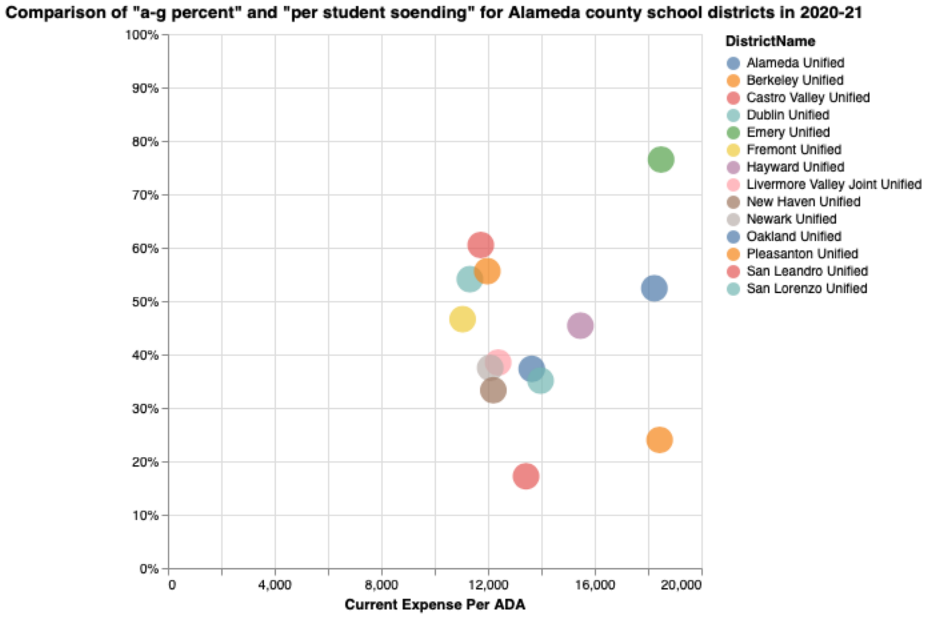


Comparison of special program participation

As indicated above, Berkeley had more White students and fewer Hispanic and African American students. Berkeley also had fewer English Learner and low socioeconomic students. In other words, Berkeley had more of the types of students who tend to perform better and fewer of those who do not.


Comparison of "a-g percent" for Berkeley and Oakland schools for 2020-21

As shown above, both Berkeley and Oakland schools varied greatly in terms of both size and the percentage of graduates meeting the "a-g" requirements. Berkeley has two high schools, a small high school with 3% meeting "a-g" and a very large high school with 62% meeting. Oakland has 31 high schools, with the percentage meeting "a-g" ranging from 0% to 100%.


Comparison of "a-g percent" and "per student soending" for Alameda county school districts in 2020-21

As indicated above, a comparison of the fourteen school districts in Alameda county shows some interesting information. Berkeley has one of the three highest funding levels per student because of an extra local tax. But Berkeley has the second lowest percentages of African American students meeting the "a-g" college readiness requirements.

Conclusions and Next Steps

Generalizable tools: One of the goals of this project was to build generalizable tools to facilitate comparison of California regions, counties, districts and schools. I believe that I have accomplished this, yielding a variety of powerful visualizations that allow the user to see patterns and trends "at a glance".

Statewide patterns and trends: Another goal was to explore patterns and trends for California over the past five years. I was not surprised by the results here. College readiness statistics at the state level have been fairly stable over the past five years. The patterns involving ethnicity, gender, socioeconomic status, and special program participation were not surprising to me.

Berkeley vs Oakland comparison: These results were very surprising to me. About 25 years ago, I analyzed standardized test scores for White, Asian, African American and Hispanic students from several school districts in the county, including Berkeley and Oakland. I found that Berkeley had some of the best scores for every ethnic group. That was not true this time, twenty-five years later.

I believe that Berkeley can do much better. I plan to communicate these results to school leaders. Hopefully, they will feel an urgency to develop and implement a plan with measurable objectives, specific actions, and a system to monitor progress.

There is something that is worth pointing out here. It is based on President Dwight Eisenhower's Decision Principle, which was popularized in Stephen Covey's 1989 book, The 7 Habits of Highly Effective People. Basically, the concept is that people should prioritize important tasks and not get distracted by urgent but less important tasks. Applied to high school students, the challenge is how to motivate them to spend time on academics instead of being distracted by their cellphones and other activities. From my experience, this is a huge challenge that high schools face.

Possible areas of further research

- issues of "data provenance": How are the "a-g" statistics calculated? What explains the unusual Berkeley data for 2018-19? Are lower curriculum standards or more lenient grading a factor explaining why some school districts appear to be doing better?
- incorporation of other interesting datasets, many of which are available from the California Department of Education and the US Census Bureau
- machine learning, such as clustering, to find and compare similar school districts
- development of an interactive website to help others analyze this college readiness data

Statement of Work

All work was done by Peter Bloomsburgh.