CALIFORNIA HIGH SCHOOL FUNDING EQUITY AND ITS IMPACT ON PERFORMANCE

An Exploratory Investigation

DS4A FALL COHORT
Revell Bell, Judith Ekwoge, Moyosoore Kemi-Rotimi, Gent Stone
Team 98

Contents

Introduction and Context

The K-12 public education system in California is one of the most populous in the country. However, the state ranks in the bottom half of the country in education performance and student funding in CA is lower than that of many states.[1][2] Thus, with this project, we set out to explore how/if there is a relationship between school funding and student performance. Specifically, we examine how school spending can impact student performance, vis-a-vis high school graduation rates and college acceptance rates as well as dropout rate. This analysis is limited to 3 high school cohorts that graduated from high schools in California, that are not in unified districts, between 2017-2019 academic years.

Given that there is sufficient data to show that contrary to human reasoning , low-income communities do receive less state and local funding than might be typical for school districts, the outcome of this project could serve as the starting point for revisiting very important policy reorientation regarding school funding. Beyond enriching policy/decision making, this project would ultimately highlight gaps that currently exist that the citizenry can work together to ensure closure on. With further insight into the discrepancies that currently exist, this project could unearth other areas of the California high school system where efficiencies can be gained with increased funding.

In a year that has been largely defined by the pandemic, 2020 showed greater indications for the necessity of having more people with post-secondary education to meet the rise in remote jobs.

---

[1] https://worldpopulationreview.com/state-rankings/public-school-rankings-by-state
[2] https://educationdata.org/public-education-spending-statistics

This makes our analysis even more important as we examine if increased funding can impact outcomes and ultimately serve as one of the ways to increase the employability of those in low-income communities.

Data Overview

In this section, we give an overview of the dataset used in this project and highlight some of their limitations and how that shaped our analysis and eventual direction of our project. We also share details of our exploratory and statistical analysis.

Datasets and Challenges

The California Department of Education has collected and stored a wide variety of data related to performance and expenditure, amongst others, through the years. For this project, we relied on the Adjusted Cohort Graduation Rate (ACGR) and Outcome Data[3] (hereafter referred to as Performance Data) and Current Expense of Education Data[4] (hereafter referred to as Expense data). We also sourced general information from the Public Schools and District Data Files.[5] This helped narrow down public schools and districts in California and helped in identifying those of interest as our analysis progressed.

Performance Data

This data is available per cohort over the duration of their high school career. For example, datasets for 2016-2017 capture data related to the high school cohort that started 9th grade in 2013-2014 and were slated to graduate 2016-2017. We focused on the performance data for 2016-2017, 2017-

---

[3] https://www.cde.ca.gov/ds/sd/sd/filesacgr.asp
[4] https://www.cde.ca.gov/ds/fd/ec/currentexpense.asp
[5] https://www.cde.ca.gov/ds/si/ds/pubschls.asp

2018, 2018-2019. This data is publicly available in .txt format at https://www.cde.ca.gov/ds/sd/sd/filesacgr.asp with a tab delimiter. Data for the respective years were downloaded to local destination from site using command line below (skipped checking for certificate since a secure connection could not be established and it was blocking access to data) and converted to csv file format, via command line prompts, for further analysis and exploration:

- wget--no-check-certificate https://www3.cde.ca.gov/demo-downloads/acgr/cohort1617.txt
- wget--no-check-certificate https://www3.cde.ca.gov/demo-downloads/acgr/cohort1718.txt
- wget--no-check-certificate https://www3.cde.ca.gov/demo-downloads/acgr/cohort1819.txt

Per data source, this data contains multiple aggregation and disaggregation at various levels of reporting. For example, given that this dataset reports for these cohorts at the District, County and School levels there were repetitive entries at these levels – details of the various layers of aggregation in this dataset can be found in the combined data dictionary in the appendix of this report. A cursory overview of the performance data showed very few null elements. There was consistency across which columns have null data as District Code, School Code, District Name and School Name seem to be the only fields with NULL data across the 3 datasets. Thus, a decision was made early on to leave as is and concatenate the datasets for all 3 years ahead of further analysis and cleaning. The underlying assumption was that the rows missing this data are those aggregated at the State "T" and County "C" level as it seems logical, they will not have either District or School code and/or name. This assumption changed slightly as we progressed with cleaning/further analysis as shown in the next section.

Upon merging, the combined performance dataset had **596,005** rows and **34** columns. This data contains information pertaining to the total cohort size, the number that met graduation

requirements for either University of California or the California State University, the number of dropouts, and the demographic of the cohort students among other indicators of performance which we explore further in our analysis. It is important to note that to protect student privacy, from the source, in instances where cohort size is 10 or less, performance data related to them was suppressed. There are also instances where a race/ethnicity group was suppressed due to aforementioned, thus data in that cluster is not reported for other race/ethnic groups.

Expense Data

The expense data shows current expenses on education per district level across California. The data corresponding to the years for performance were exported from the California Department of Education website. It quickly became apparent that this expense data is aggregated at the district level based on the District Ownership (DOC). Given that our project was focused on high school data, the subsets of this data that were relevant for our project were those with high school DOC type. This option inadvertently excluded high schools that fall under the Unified School District from the scope of our project as expense data was not independently available for these high schools. The expense data had **213** rows (indicating data for rows and districts over 3 years) and **8** columns.

District Data

The district data majorly served as the verification dataset of this project as it helped verify the classification of the various school District Ownership and the Districts' geodata including Longitude and Latitude.

Data Wrangling/Cleaning

Of all 3 datasets, only the performance data needed much wrangling as the other two datasets only required identification of common columns on which they all could be merged. As a

first step with the performance data, 11 calculated columns were dropped given they could eventually be inferred at a later point if necessary. Some structural issues such as leading or trailing spaces were stripped, and column names reformatted. We decided to drop the 3 rows that had a reporting category for "GX" given the very minute frequency of this category and the fact that all data associated with this category were suppressed. All rows reporting data at School, County and State aggregation level were dropped given the decision to base our analysis at the district level as this was the level at which funding information was made available. For our first level analysis, a subset of the performance data at the "TA" aggregation level was created; showing total data for each cohort in all **543** districts resulting in a data frame with **14,976** rows and **23** columns. We also created an adjacent dataset with the breakdown of the reporting categories - Gender, Race and Socio-economic category - and dropped the rows where data was suppressed because the cohort size in that category was less than 10.

We ultimately merged all performance, expense and district datasets referenced above to proceed with our data analysis after working through a few challenges highlighted below.

Challenges

There were a few data specific challenges that influenced the direction of this projects, few of these are highlighted below:

1. **Data Availability:** the funding allocation data we had initially hoped to find was not publicly available online, hence our resort to use the Current Expense of Education data for this project. This limitation of this was that we were not able to determine at a policy level what is currently being earmarked per student in California High School district. Having to work with data of what was spent as opposed to what was planned to be spent limits the kind of recommendations that can be made from the project without overreaching.

2.   **Data Parity:** Another early challenge we encountered was that the expense data lacked the level of granularity that could help distinguish high schools in unified schools' districts, thus we only had data for 71 standalone high school districts. However, given that the performance data has the school aggregate level, we had high school performance data across all districts with a high school regardless of whether they were unified or not. To overcome this challenge, we created two separate sets of unique district codes from the expense data and the funding data and then created an intersection resulting in dataframes with equal shapes – **213** rows and **23** columns. Relevant rows from these two data frames were then merged with Longitude and Latitude data from the District data.

Another data parity challenge we encountered was how to avoid aggregating the expense data at the racial and socio-economic reporting category level after merging performance with funding. To navigate this, we used a groupby to determine the count of reporting categories per district, per year. Expense data per district, per year was then divided by the resulting count from the groupby. Thus, ensuring accuracy when aggregation is applied to the expense columns per reporting category.

Exploratory Data Analysis

Interested in the relationships between school expenditure in these districts and performance, we examined which expense data have correlation with the performance parameters. The result of this analysis is the correlation matrix below which shows a high correlation between EDP 365 (expense data) and some performance data parameters. We observe expenditure has the strongest correlation with those who met UC/CSU admissions requirements (for the purpose of this project, this is defined as College Readiness), those who received regular high School diploma, Cohort Students (and early indication that cohort size determines the total expense) and those who

received special education certificates. There is also a marginal correlation observed with dropout rate.
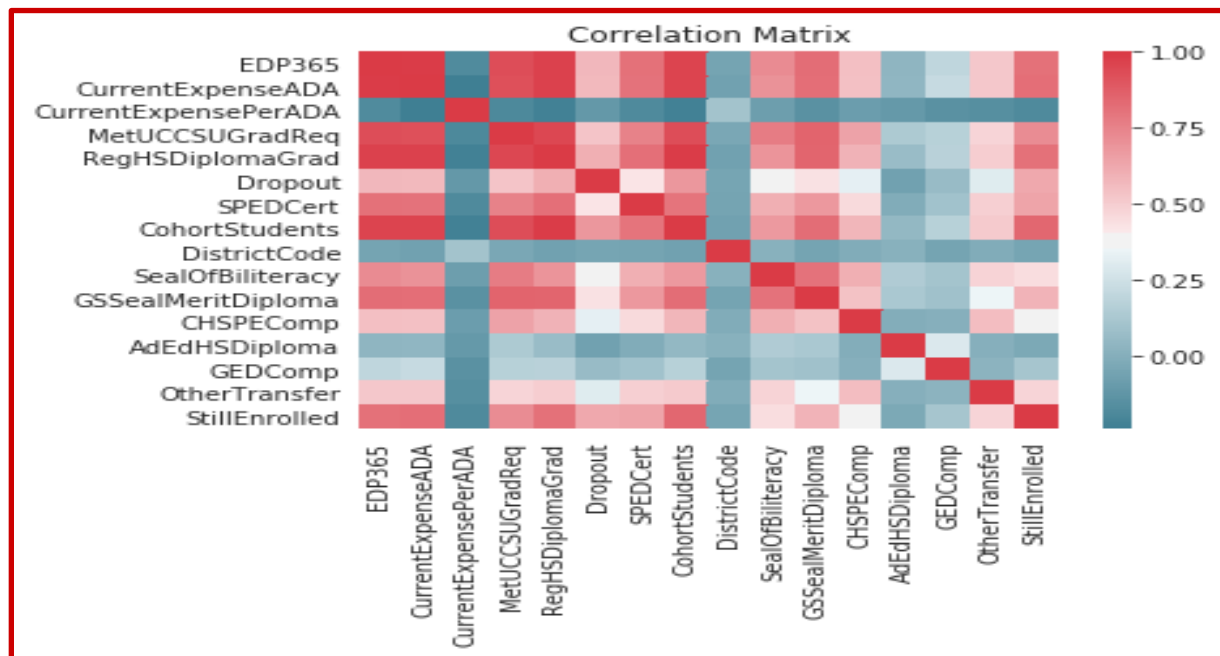
The correlation observed above is further investigated with these subplots below. Except for EDP 365 v. the number of dropouts, we see the correlations are positive for the most part.
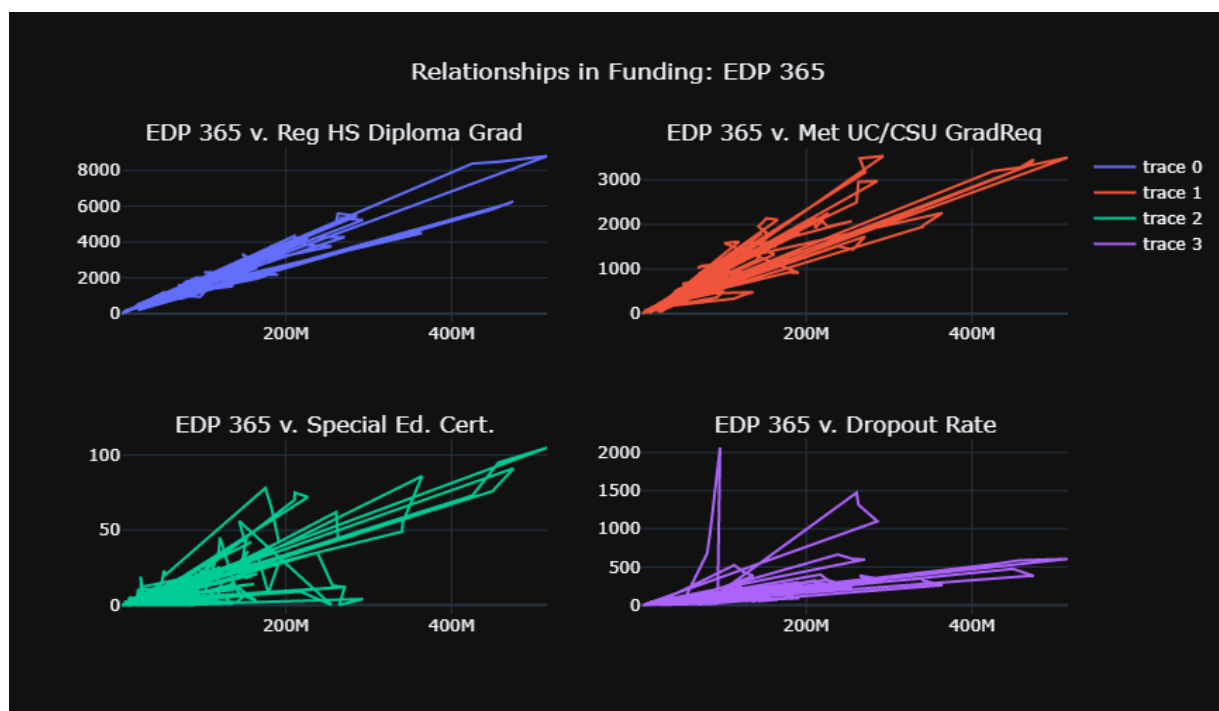
We further explored how funding was being allocated across the state and visualized below the spread out of funding per district.
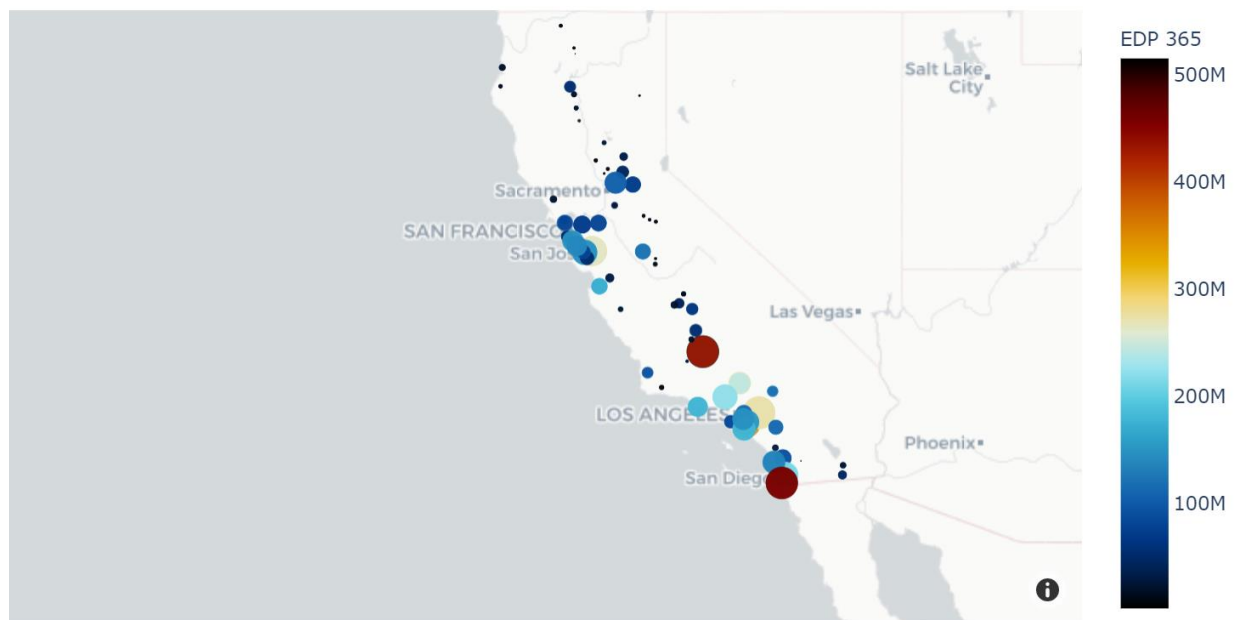

Statistical Analysis

Our main focal point was in examining whether expenses correlate with the size of the student population. We wanted to investigate if there were other factors that potentially can boost funding for schools. Listed below we will have three different figures that show:
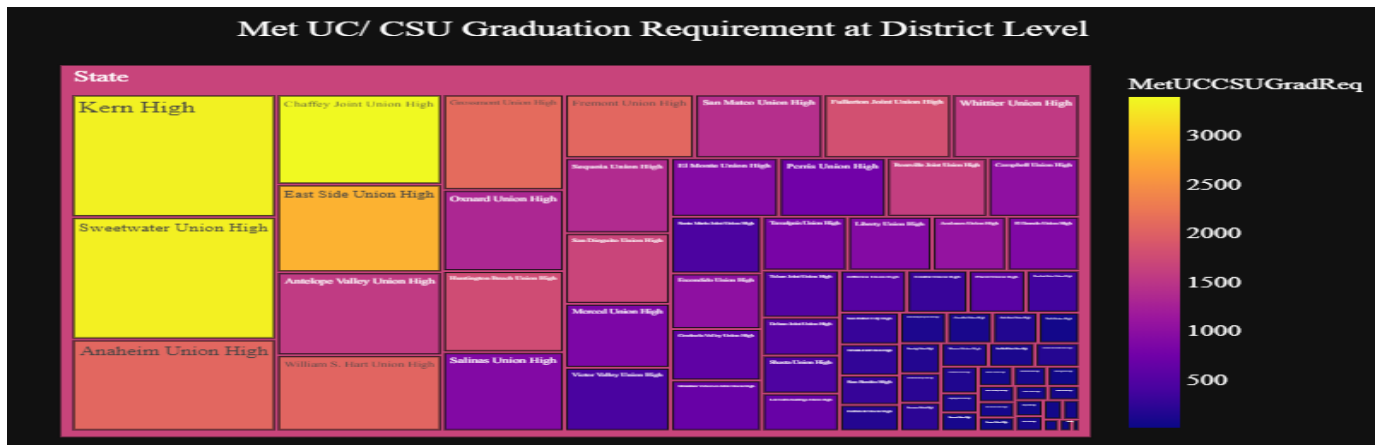
1. The distribution layout of funding from county to county.
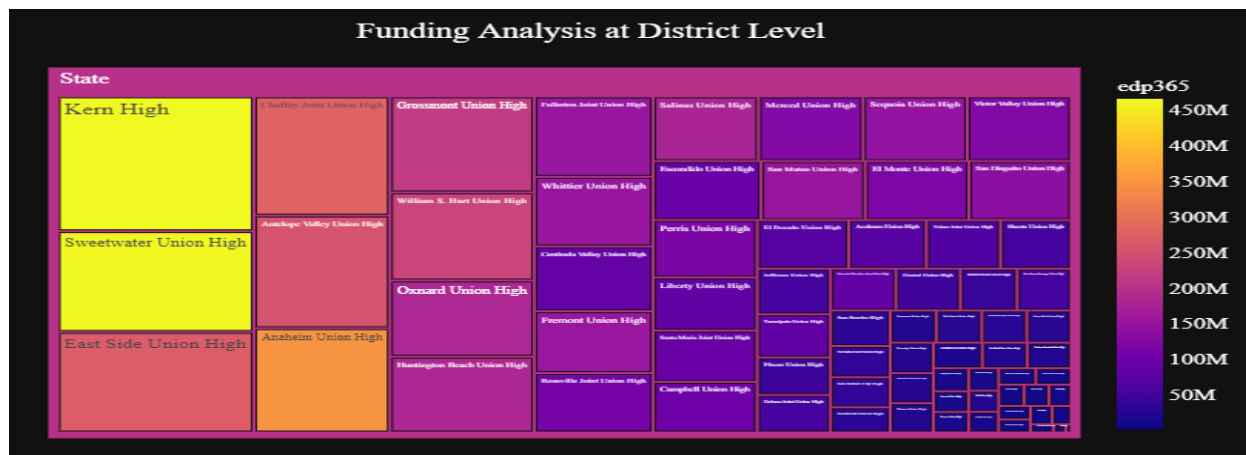
2. The amount of students who received Special Education Certificates by county.

Special Education Certificates at District Level
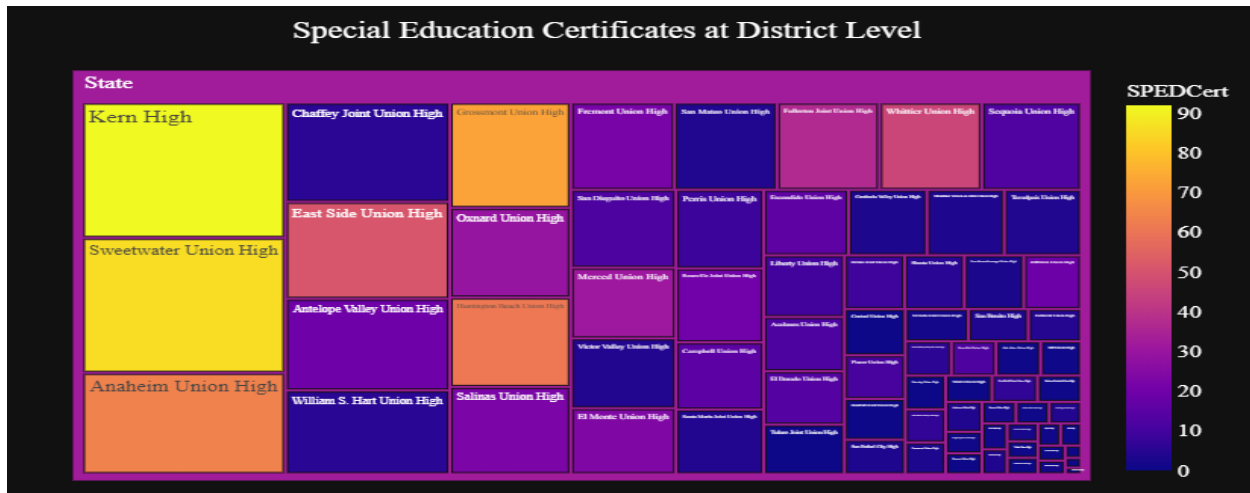
3. The number of students who graduated and met the UC/CS graduation requirements.

Met UC/ CSU Graduation Requirement at District Level

Findings:

1. As expected, larger districts spend the most money, award most special education certificates and a higher count of graduates who have met UC /CSU graduation requirements

2. Some high-spending districts do not produce a large amount of special education certificates, namely Anaheim Union and Chaffey Joint Union.

3. Lastly, we see that larger districts produce more graduates who met UC and CSU admission requirements, but at what rate? We explore this further by exploring averages based on the percentage of total graduates who met that requirement and cohort size.

Leading with the results observed above, we further extracted the districts with the least and most funding from 2017-2019 and explored the relationship between funding and performance in these districts.
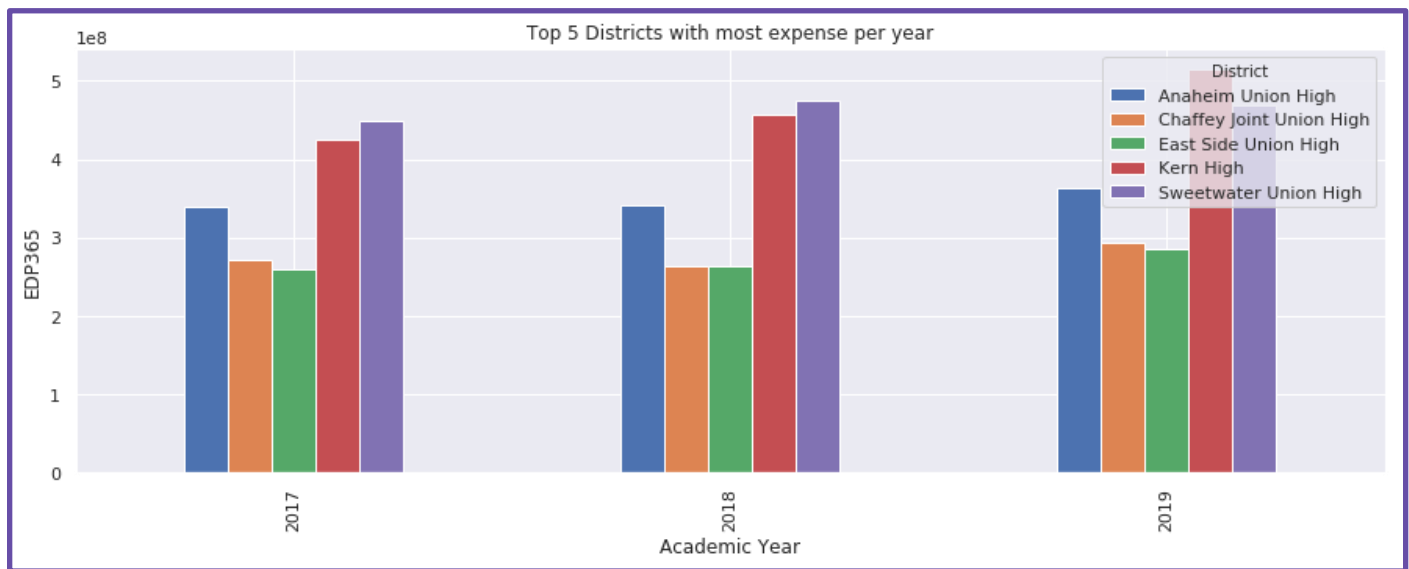
Above we see the top 5 districts with the most expenses throughout the 3 years we are examining, and we see below those with the least expenses.
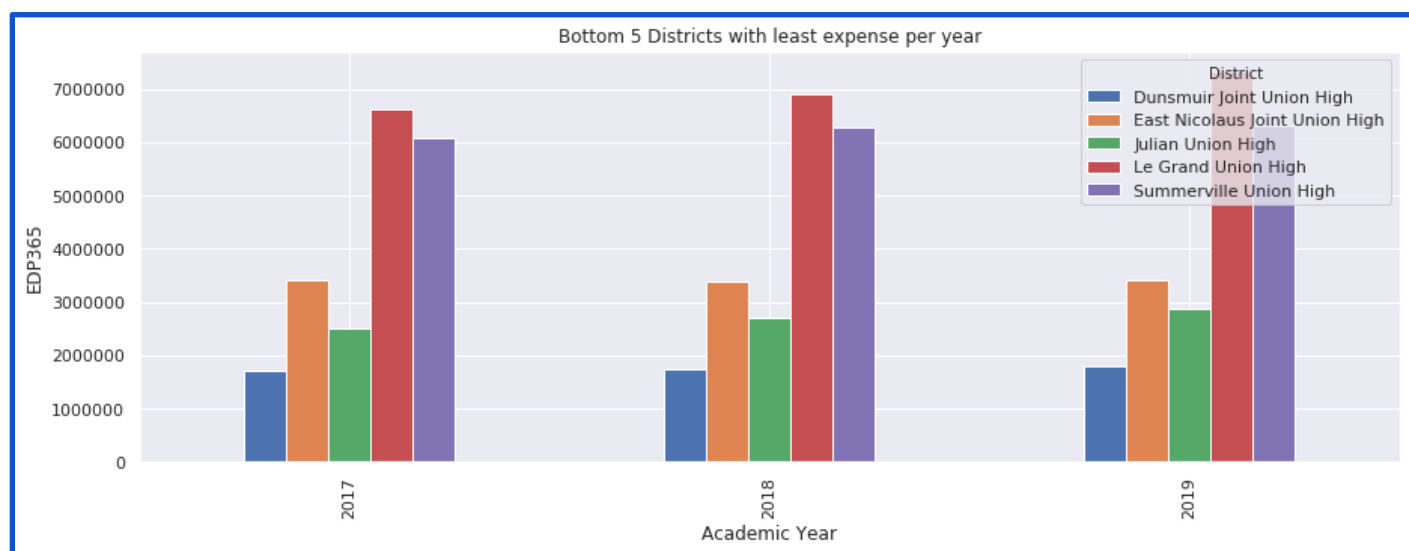
A quick look at their average college readiness rate for all 3 years below, we observe comparable college readiness between districts in the low expense categories and those in high expense categories. As is the case with Dunsmuir High School District's college readiness rate with a median that is visibly higher than all schools in the top category except Chaffey Joint Union High.
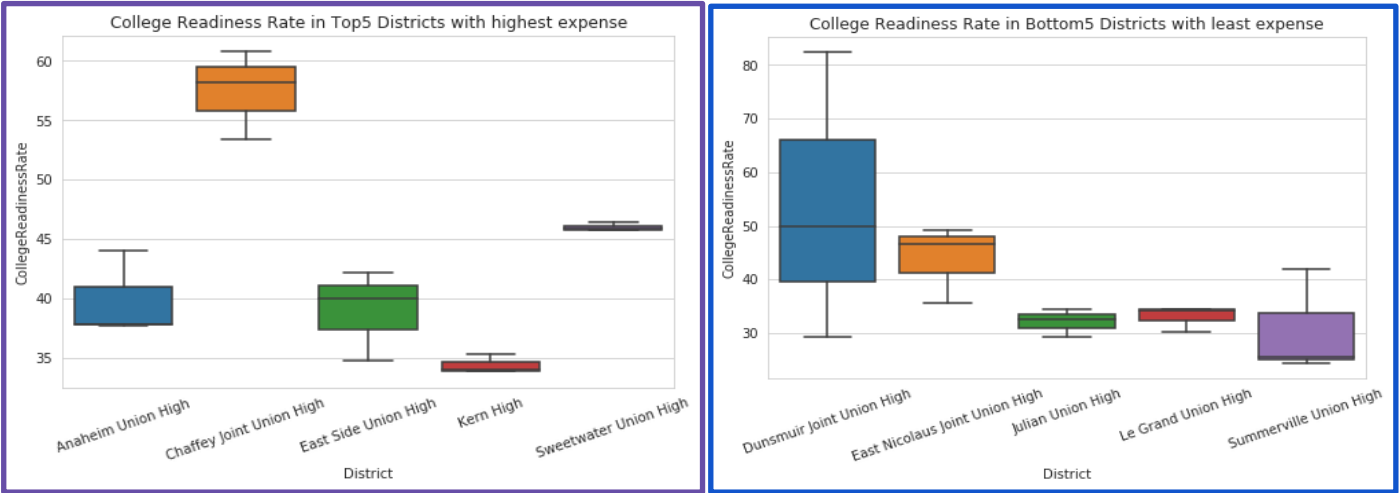
We also examined college readiness by the various reporting categories i.e socio-economic composition and race. It is evident below that all students that declared their socio-economic categories, have a trailing median. While there are outliers, we observe that students who identify as socio-economically disadvantaged show the highest promise of meeting UC/CSU admissions requirement among the categories reported. These are closely followed by the migrant student
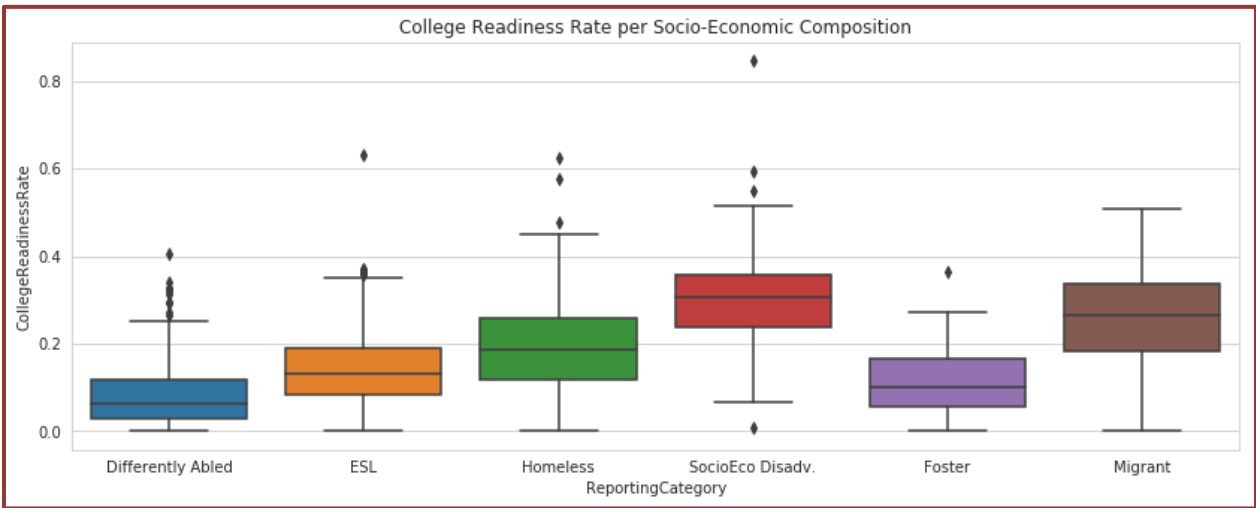
population. Students with disabilities and those in the Foster care system are on the lower spectrum of college readiness among this population.

We cannot determine the relationship between socio-economic and race composition – if any exists at all – as this data was not accessible. Judging by the totals though, it is evident that socio-economic data was only reported for a select population of the student cohort. Thus, we proceeded to examine performance by Race alone. In examining this, as shown below, we see that the Asian student population has a very high college readiness rate with a median of about 70%. They are closely followed by the Filipino student population with a median of about 65% while the Native American student population have the lowest college readiness rate.
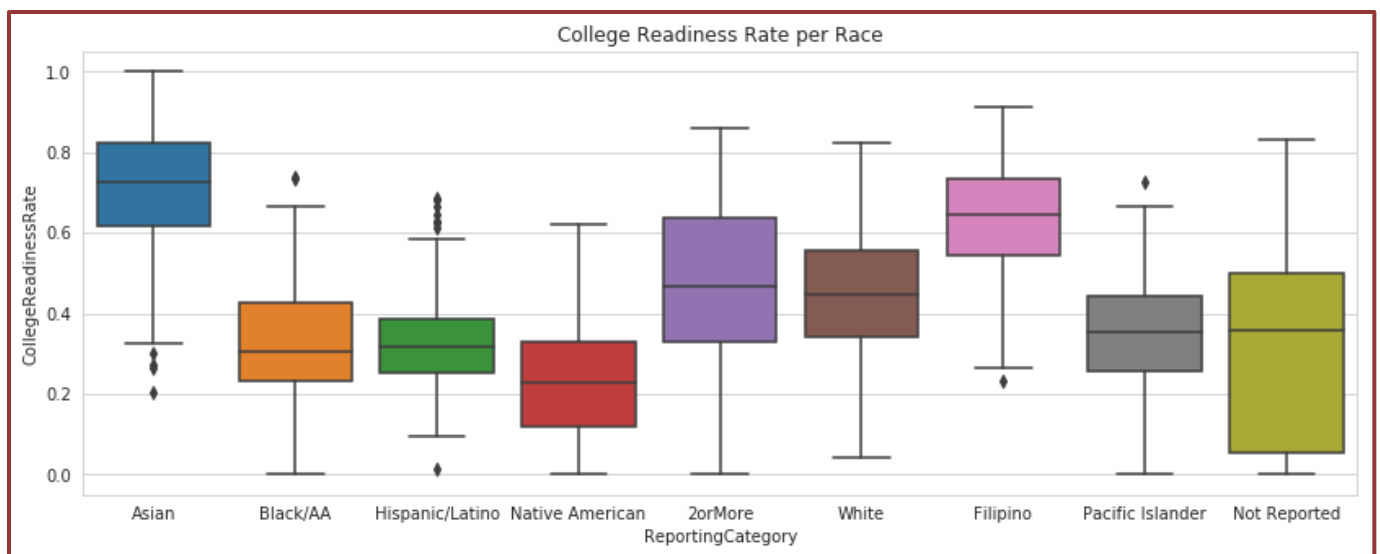
Description of Dashboard - Google Data Studio

Use Case: Earmarking key factors of performance and their relationship to high school district expenditure

The Google Data Studio dashboard we used for this project is a 8-page dashboard full of charts and plots showing various combinations of data for 71 high school districts in California.

This dashboard can be used to examine, at a granular level, the different interactions between expenses and performance levels of these school districts. There are various insights that can be gained by examining the relationship between race and performance, socio-economic composition and performance, expenses – at a collective level and per Average Daily Attendance among others.

Data Engineering: Google Data Studio

Charts used: scorecards, geomap, scatter/bubble chart, pivot chart, bar chart, donut chart, animated barplot.

We used scorecards in our overview page for totals we wanted to highlight and a geomap of the total school districts by cohort size(circle size) and expense(color shades). On the second page, we used a scatter/bubble chart to investigate the tri-dimensional relationship between performance (college readiness rate=the percentage of students who met UC/CSU graduation requirements), high school spending and cohort size. The chart can be zoomed in to by using the metric sliders at the upper right corner of the page. The next page pulled out the five highest and lowest spenders for a more detailed look at the types of spending done by each school district. The parameters that interacted in this page for a more rounded outlook of each school district were the performance levels(college readiness, graduation, and dropout rates), the school year, three types of expenses, and cohort size. The possibility to observe all of this was possible with a combined use of a bar chart, a pivot table, and scorecards.

The next page enables a study of all high school district performance rates by year in a bar chart, with each school's racial and socio-economic composition represented by donut charts. You can also explore this information at the county level by using buttons at the upper left corner of the page. The next two pages use donut charts and stacked bar charts to show a better visual representation of performance levels by year or racial and socio-economic categories. The last

chart is an animated bar chart that shows each school district's total expenses accumulating each year.

Conclusion

While there are valuable insights derived in the cause of exploring the datasets we started out with, most of which have been represented above and on the dashboard, there are gaps in the data available that preclude us from drawing conclusions about relationship between funding and performance. Neither can we conclusively determine the adjustment needed to per pupil spending that could positively or negatively impact funding. However, there are useful insights we have been able to derive from this project.

The first thing we were able to recognize is that it is clear, larger school districts have to receive more money at an operational level and that much of that funding goes beyond the students, just the school being open is expensive. Also, we observed that smaller districts had better performance and that could be because of better student to teacher ratios, where a lot of the cohorts were less than 100 students large. However, there are indications that the cost incurred per Average Daily Attendance for these smaller cohorts adds up quickly. While at first glance it seemed obvious that Dunsmuir High School District has the lowest expense (EDP365), when compared with their cohort size of 50 students, the low number is justified. However, further investigation is needed to determine why their Current Expense per ADA runs so high at total of ~$91,000 over the 3 years examined.

In addition, our investigation showed that there is comparable, and sometime, better performance shown by cohorts in smaller school districts as opposed to larger school districts as is the case with Dunsmuir Joint Union High District in comparison to Sweetwater Union School

District and Kern High School District, respectively. It is insights like these that raise these two questions – how much of the reported expenditure is dedicated for students' academics (performance) and what logic can be used to determine how California will meet or exceed metrics of performance as set by the state department of education? For example, will an increase or decrease in expenditure by a certain percentage also impact performance – either positively or negatively by a certain percentage?

Lastly, we also observed that race is a clear marker of performance with clear distinctions between the performance of Asian as opposed to Native American student population. While it would have been helpful to explore the socio-economic composition of all students, we find that this data is only reported for a subset of student who all seem to belong to marginalized categories. The trends observed in performance and dropout among these population raises the question about what the state is currently doing to support these students and whether more can be/should be done?

Although these questions remain unanswered, this study has been eye-opening as it has revealed to us that high level analysis is long, tedious and most of the time consist of multiple layers of investigation. There is also a tediousness brought on by the way education related data is coded and made publicly available. One thing remains certain, for future work to make more meaningful progress in examining the relationship between funding and performance, the department of education needs to make publicly available the amount of money earmarked for each school district in California. We opine that this source, much more that the expenditure data, is needed to make inferences that might be noteworthy for policymaking decisions. In the interim, it will be helpful if the publicly available expenditure data had clearer definitions about what they

encompass i.e. expense broken down by students, staff, equipment, upkeep, innovation among others.

Future Work

Some future work that would be of extreme use would be to create a real time dashboard that exists in real-time connecting all districts measuring both performance and funding. There is a strong need for more transparency regarding these two parameter and real time accessibility such as dashboard provides will allow for quicker recovery and better results. This also makes the information open to the public, officials and parents making for a great start to accountability from all angles.

Appendix

Performance Data Dictionary

Below is a definition of the columns starting with the categorical variables that are basis of aggregation and disaggregation/subset for this dataset and must be carefully selected during analysis in order to obtain an unduplicated count:

Categorical variables

1. **AggregateLevel**: This data is aggregated at 4 categorical levels. *In order to obtain and unduplicated count, one of these levels must be chosen in analyzing this data*:
    - **T** = State
    - **C** = County
    - **D** = District
    - **S** = School

2. **CharterSchool**: This column is indicative of whether data in that category encompasses all schools, only charter schools or non-charter schools. *In order to obtain and unduplicated count at the selected level, one of these values must be chosen in analyzing this dataset*:
    - **All** = Includes data for **all schools** regardless of charter status. For AggregateLevel=S, all schools are included. For Aggregate Levels T, C, and D, data for all schools are included in the aggregate.
    - **Yes** = Includes **only data for charter schools**. For Aggregate Level=S, **only charter schools** are included. For Aggregate Levels T, C, and D, **only data for charter schools** are included in the aggregate.
    - **No** = Includes **only data for non-charter schools**. For Aggregate Level=S, **only non-charter schools** are included. For Aggregate Levels T, C, and D, **only datafor non-charter schools** are included in the aggregate.

3. **DASS**: An indicator of whether data for all schools, only schools participating in theDashboard Alternative School Status Program (DASS), or only non-DASS schools are included. *In order to obtain and unduplicated count at the selected level, one of these values must be chosen in analyzing this dataset*:
    - **All** = Includes data for **all schools regardless of DASS status**. For AggregateLevel=S, all schools are included. For Aggregate Levels T, C, and D, data for all schools are included in the aggregate.
    - **Yes** = Includes **only data for schools participating in DASS**. For AggregateLevel=S, **only DASS schools** are included. For Aggregate Levels T, C, and D, **only data for DASS schools** are included in the aggregate.
    - **No** = Includes **only data for non-DASS schools**. For Aggregate Level=S, **only data for non-DASS schools** are included. For Aggregate Levels T, C, and D, **only data for non-DASS schools** are included in the aggregate.

4. **ReportingCategory**: This category helps disaggregate this data into select Race, Ethnicity, gender or program subgroups spanning 18 categories. *If not interested in exploring any of the underlisted categories, Reporting Category should be set to TA*:
    - **Black/AA** = African American
    - **Native American** = American Indian or Alaska Native
    - **Asian** = Asian
    - **Filipino** = Filipino

- **Hispanic/Latino** = Hispanic or Latino
- Not Reported
- Pacific Islander
- 2orMore = Two or More Races
- White
- **GM** = Male
- **GF** = Female
- **ESL** = English Learners
- **Differently Abled** = Students with Disabilities
- **SocioEco Disadv.** = Socioeconomically Disadvantaged
- **Migrant**
- **Foster**
- **Homeless**
- **TA** = Total

Other columns

5. **CountyCode**: A unique two-digit code corresponding to the county.
6. **DistrictCode**: A unique five-digit code corresponding to the district.
7. **SchoolCode**: A unique seven-digit code corresponding to the school.
8. **CountyName**: County name.
9. **DistrictName**: District or Administrative authority name.
10. **SchoolName**: School name.
11. **AcademicYear**: expected year of on-time, four-year graduation from high school
12. **CohortStudents**: The number of students who enter grade 9 for the first time, plus any students who transfer in later during grade 9 or the next three years. This does not include any students from the cohort who transfers out, emigrates from the country, transfers to a juvenile facility or prison, or dies during that same period.
13. **Regular HS Diploma Graduates (Count)** : Total number of cohort students who received the standard high school diploma from a school aligned with State education standards. *This is a reference columns for a lot of the other columns in this dataset*
14. **Regular HS Diploma Graduates (Rate)**: The percentage of cohort students who received the standard high school diploma.
15. **Met UC/CSU Grad Req's (Count)**: Total number of cohort graduates who met all a-g requirements for admission into a UC or CSU school.
16. **College Readiness Rate**: The percentage of cohort graduates who met admission requirements for a UC or CSU school.
17. **Seal of Biliteracy (Count)**: Total number of cohort graduates who attained a high level of proficiency in one or more languages in addition to English and earned the State Seal of Biliteracy.
18. **Seal of Biliteracy (Rate)**: The percentage of cohort graduates who earned the State Seal of Biliteracy.
19. **Golden State Seal Merit Diploma (Count)**: Total number of cohort graduates who earned the Golden State Seal Merit Diploma by earning a high school diploma and demonstrating mastery in at least six subject matter areas, including mathematics, English language arts, science,U.S. history, and other areas.
20. **Golden State Seal Merit Diploma (Rate)**: The percentage of cohort graduates who earned the Golden State Seal Merit Diploma.
21. **CHSPE Completer (Count)**: Total number of cohort students who withdrew from regular high school without receiving a regular high school diploma after passing the CA High School Proficiency Exam.
22. **CHSPE Completer (Rate)**: The percentage of cohort students who took and passed the CHSPE.

23.   **Adult Ed. HS Diploma (Count)**: Total number of cohort students who withdrew from regular high school without receiving a regular high school diploma, and later received documentation for completing an adult education program.

24.   **Adult Ed. HS Diploma (Rate)**: The percentage of cohort students who received documentation for completing an adult education program.

25.   **SPED Certificate (Count)**: Total number of cohort students with exceptional needs (having an IEP) who received a certificate or document of education achievement or completion.

26.   **SPED Certificate (Rate)**: The percentage of cohort students with exceptional needs who received a certificate or document of education achievement or completion.

27.   **GED Completer (Count)**: Total number of cohort students who withdrew from regular high school without receiving a regular high school diploma and have received a High School Equivalency Certificate by passing the GED exam, the TASC exam, and/or the HiSET.

28.   **GED Completer (Rate)**: The percentage of cohort students who received a High School Equivalency Certificate.

29.   **Other Transfer (Count)**: Total number of cohort students who withdrew from regular high school without receiving a regular high school diploma and transfer to an adult education program or to community college during the cohort period.

30.   **Other Transfer (Rate)**: The percentage of cohort students who transfer to an adult education program orto community college during the cohort period.

31.   **Dropout (Count)**: Total number of cohort students who do not graduate with a regular high school diploma, do not complete high school, and are not still enrolled as a "fifth year senior".

32.   **Dropout (Rate)**: The percentage of cohort students who do not graduate with a regular high school diploma, do not complete high school, and are not still enrolled as a "fifth year senior".

33.   **Still Enrolled (Count)**: Total number of cohort students who did not graduate high school by the end of their cohort outcome period who: are re-enrolled as a "fifth year senior" at the end of their cohort period, or who completed high school no later than mid-September of the subsequent academic year.

34.   **Still Enrolled (Rate)**: The percentage of cohort students who did not graduate high school at the end of their cohort outcome period who either re-enrolled or completed high school no later than mid-September of the next academic year.

The following additional information is also shared by CDE as added note about using the ACGR and Outcome Data Files:

This file contains different subsets of data based on the values of different variables. If these sub-setting variables are not used properly, it is easy to end up with incorrect or duplicative data. Please refer to the following examples below as a guide for properly using these data:

- Get school-level data totals with one record per school with school totals (without student group totals)
    - Aggregate Level = S (School)
    - Charter School (Y/N) = All
    - DASS (Y/N) = All
    - Reporting Category = TA
- Get school-level data totals by student group (Reporting Category) for all non-charter schools
    - Aggregate Level = S (School)
    - Charter School (Y/N) = N
    - DASS (Y/N) = All
    - Reporting Category Not equal to TA
- Get school-level data totals by student group (Reporting Category) for all DASS schools

- o   Aggregate Level = S (School)
- o   Charter School (Y/N) = All
- o   DASS (Y/N) = Y
- o   Reporting Category Not equal to TA
- Get district-level data totals by student group (Reporting Category) for non-charter schools
  - o   Aggregate Level = D (District)
  - o   Charter School (Y/N) = N
  - o   DASS (Y/N) = All
  - o   Reporting Category Not equal to TA
- Get statewide-level data totals for charter schools that are DASS schools
  - o   Aggregate Level = T (State)
  - o   Charter School (Y/N) = Y
  - o   DASS (Y/N) = Y
  - o   Reporting Category = TA

Expense Data Dictionary

Below are the definitions for the columns

1. **CO Code:** County Code
2. **LEA (Type):** Local educational agency
3. **EDP 365:** Expenditures for Current Expense of Education. includes (1) classified salaries, (2) employee benefits(including state payments to retirement systems on behalf of districts as from 2014–15), (3) books and supplies, (4) equipment replacement, (5) services and indirect costs. Excludes (1) non-agency activities; (2) community services; (3) food services; (4) fringe benefits for retired persons; and (5) facilities acquisition and construction.
4. **District Code:** Numerical District Codes
5. **District:** District Name
6. **Current Expense ADA:** Current Expenses for Average Daily Attendance
7. **Current Expense Per ADA:** Current Expenses Per Average Daily Attendance
8. **Year:** Expense Year
9. **ADA:** Average Daily Attendance