LAUSD School District Zoning: Predictive Analysis

Group 3: Taya Stewart, Raymond Delgado, Adnan Elahi

**ABSTRACT**

School district zoning has substantial historical implications that still have an effect on today’s education. The current state of school district zoning in Los Angeles County and surrounding counties have an inevitable tie to Ruby Bridges being the first African American child to desegregate in Alabama Elementary school in 1960 and the redlining process in Los Angeles County of the 1940s through 1970s. Although these political and social ramifications still exist, we must attempt to use an unbiased balance of machine learning, mathematical algorithms, and geographical information system mapping to outline a new approach to school districting. This report analyzes the variables considered in the aforementioned model. The K-Means Analysis has determined that it is possible to find similarities between the schools based on the population of students who are disenfranchised. We were able to determine that most schools had a manageable student population, and the vast majority of students were not disenfranchised. Using the Gini Index and school capacity, we affirmed that both income inequality and educational inequality exists in Los Angeles County. Los Angeles County’s gini index will continue to increase toward perfect income inequality unless changes are made to the school district zoning process.

Keywords: Education, Predictive Analytics, K-Means Clustering, Forecasting, Geographical Information Systems Mapping

**INTRODUCTION**

Historically, school districts have received substantial scrutiny for zoning policies that may have been inherently discriminatory. Although explicit discrimination is illegal, the effects of redlining and zoning still reside. In most California counties, diversity is expected and is evident when compared to the rest of the United States. However, there are still pockets of similar demographics in one area due to social, economic, and environmental factors. In the current schooling model, schools are divided into districts that are determined by a local district governing board1. This board is elected during the November General Elections and holds a fixed length term of 4 years or more. Given that school district boundaries are determined by a board of potentially 6 individuals, human bias is inevitable.

With the advent of machine learning, geographical information system mapping, and complex data analytics models, bias can be minimized. Mathematical algorithms may help determine which factors contribute to a successful school district. Moreover, this analysis will use K-Means Clustering to identify the factors that lead to underperforming schools. The analysis will also conduct a forecasting analysis to determine how inequality will expand in LA county if nothing is done.

**REVIEW OF LITERATURE**

A Report on Absenteeism in the Nation’s Public Schools determined three effective categories that define reasons for for student absenteeism (Balfanz, 2012)

* Cannot Attend: housing instability, juvenile justice, family responsibility
* Will Not Attend: bullying & harassment
* Do Not Attend: no value, no discipline

From these categories, the report determined that most absenteeism had a direct correlation with a students low income status. Students were forced into low academic performance based on the lack of resources in their household and school district.

Education Inequalities at the School Starting Gate further studies this issue as the study follows 2 cohorts, a kindergarten class of 1998 and 2010(Garcia, 2017). The study uncovered 2 major conclusions:

* Children who start behind stay behind
* The solution is for districts to have equitable access to nutrition, health, academic support to offset income inequality

After following each group of children over the course of ten years, it was determined that those who had more equitable access to district resources, were able to overcome the disadvantages of their low income status. Ultimately, school district funding and zoning play a key role in overall academic performance and some districts have greater resources than others.

John Mackenzie conducted first-hand experimentation to determine if the following hypothesis are correct (Mackenzie, 2009):

1. Is school performance independent from the school’s funding

2. Do local funding have no impact on school funding

3. Will taxes for school funding hinder property values of the local area

The metrics used for the analysis are:

* School funding in terms of property taxes.
* Student performance indicators defined by the National Assessment of Educational Progress.

The results of the analysis determined a positive correlation between school funding and the school’s average NAEP score. Moreover, the report shows that property values increased when school funding school funding increases—thus proving a positive correlation. These findings are significant because it shows that a lack of academic success is not a result of a “bad student”, but it is a direct result of a lack of funding and resources. This report shows that increasing school funding through taxes provides benefits for schools and the local communities.

Johanna Wald and Daniel Losen state that the disparities of school performance runs along racial and social-economic lines. This means that schools with a high minority population will have less resources (funding, teachers, and advanced level-courses) than their white counterparts. Wald and Losen also highlights the fact that regressive policies, which include ending school desegregation plans and high stakes testing. Moreover, zero-tolerance policies disproportionality impact students of color(Wald & Losen, 2003). For instance, black students are 2.6 times more likely to be suspended than white students. These findings prove that our educational system has failed to achieve racial and social equality. These troubling developments also emphasize the desperate need for reform on how our nation funds and organizes the public-school system.

**METHODOLOGY AND PROCEDURE**

The methodology of the project consists of a linear process that is divided into nine phases. These phases are:

* Group Forming
* Topic Selection
* Data Sourcing & Selection
* Data Preparation & Processing
* Algorithm selection
* Data Modeling
* Record Insights
* Process Insights into actionable solutions

During the Group Forming phases consists of the professor randomly dividing the class into six groups. The groups were randomly created to reflect the fact that in a corporate setting people will not have privilege to select their co-workers or project members. Moreover, this challenge will force the students to organize and divide the work in a fair manner.

Our group topic selection phase began with us just listing all the topics that we care about or want to learn more about. Then, we removed possible topics based on our group decision. It must be noted that the topics that were not relevant to LA county were the first to be removed. After about an hour of discussion, we have selected school zoning as our topic. The major reason for the decision was because we all are familiar with school districts and how zoning works.

The professor imposed the rule that we must get our data from the LA County’s Open Data web page. The professor also informed the class that we can supplement our data sets with other agencies that are within LA county. It must be noted that our project has an emphasis on the Los Angeles Unified School District(LAUSD), so the professor allowed our group to get more datasets from the official LAUSD website. Therefore, we have selected a main dataset from the LA county website and smaller datasets from the LAUSD website.

We did the vast majority of data preparation through excel. The data preparation mainly consists of removing missing values and combining our datasets into multiple usable files. Our group then uploaded these files into SAP Predictive Analysis and some minor data preparation. Those smaller data preparation mainly consists of renaming the fields so that they are easier to understand. We also used SAP Predictive Analysis to create calculated features.

During the data analysis phase, we used SAP Predictive Analysis to perform some basic analysis on our data. The goal of the data analysis is to develop a profile of the students and schools in LA county. Moreover, these data analysis give us a better understanding of what relevant questions we might want to answer. Please note that the graphs generated for our analysis can be found in the tables section.

Based on the result of our data analysis our group has selected three algorithms: K-Means Clustering, Forecasting, and multivariate regression. The K-Means Clustering will be used to group the schools within LA County into several unique clusters. Forecasting will allow our group to predict how the current trends in economic inequalities will affect the quality of the school. The multivariate regression will identify what features predict academic performance.

During the data modeling, we configure the algorithms to work with our dataset. We have provided very detailed explanations on how our groups configured these algorithms and research procedures.

After we run these algorithms, we will record our findings. These findings will include the direct output from SAP Predictive Analysis and summarization of the output. Then, we will explain the importance of the results.

The last phase of the project methodology is to convert our insight into actionable solutions. These solutions provide a course of action for both LA county and the individuals. We will also explain any potential drawbacks and hurdles in employing our solutions .

Covid-19 has highlighted the educational disparities created by the current school zoning practices. With students struggling to find equitable access to food, internet, safe study environments, and medical treatment, you must question why this is happening. While some students are receiving free loaner laptops, take home lunches, and on campus medical services, other students in lower funded districts are struggling to receive the same resources. Research questions were developed to better determine and analyze the cause of this disparity.

The factors used to determine whether or not a school district was equitably zoned were as follows:

**Figure 4.1**

|  |  |
| --- | --- |
| Class Size >=15 | Performance |
| # Of students in free & reduced lunch program | # Of students experiencing homelessness |
| # Of English Learners | #Of students who are classified as migrants |
| # Of Students in a school | Compound effect of numerous indicators |
| #Of Teachers compared to students | #Of Suspensions |
| #Of Foster Youth | Racial Demographics |
| Median Income | District Size |
| Gini Index | School Level Distribution |
| Poverty Index | Distance to school |
| Attendance | #Of Full Time Staff |

These considerations were narrowed and queried into considerable clusters and groups to determine which factors were evident in the same area. These clusters were supported by K- means analysis feature of the R Studio Package of SAP Predictive Analytics.

The goal of the K-means clustering is to discover if it is possible to group the 87 school districts based on the total of students who are experiencing hardship. The California Department of Education defines these hardships as (California Department of Education, n.d.):

* Lack of english fluency
* Being part of the foster system
* Being a migrant
* Being homeless
* Experiencing nutrition insecurity

The dataset represents these hardships by the classification of students and enrollment of the free or reduced lunch programThe independent variable for the K-Means clustering are:

* Total Enrollment: How many students are in a certain school
* Free & Reduced Meal Program: How many students are part of the free & reduced meal program due to their family being unable to afford lunch.
* Foster: Those students who are part of the foster system.
* Homeless: Those students who are homeless
* Migrant Program: Those students who are a consider migrant
* English Learner (EL): Those students who have net yet achieve fluency in English

In the sake of transparency and minimizing potential biases, we must first explain the limitations of these labels. For instance, the dataset does not define the severity of a student’s homelessness. The California Department of Education defines the homeless population as “individuals who lack a fixed, regular, and adequate nighttime residence.”

Based on the initial data analysis, we have discovered that there are a total of 87 school districts in LA county. The equation to find the optimum number of clusters for K-means analysis is:

For our analysis n equals 87, which is the total number of school districts. Therefore, our equation will give us the following outcome:

= = 6.59 or ~7

Similarly, SAP Expert Analytics was used to forecast future outcomes based on factors listed in Figure 4.1. These forecasts were predicted over five years using triple exponential smoothing and time series analysis. These features are enabled by the R Studio package which allows for additional customization within a visualization. Triple exponential smoothing was chosen because the data may have been repetitive over the provided data period given the large population size.

Multiple line charts can be overlapped within SAP Expert Analytics to show a comparison between data points. Given that there are numerous variables to consider across multiple counties in California, it is important to display the relative position. However, over a short period of time these charts may be misleading. In the case of school districts and county zoning, the number of years and the sample size is large enough to be represented by multiple line graphs on a single x axis. In this analysis sample size is denoted by “n”.

Due to the limitations of SAP Predictive Analytics, bar and column charts were employed to display the distribution and relationship between geographic regions that were experiencing the compound effect of key indicators of a series of years. An alternative method would be to use calculated fields to further define segmented data.

**RESEARCH QUESTIONS AND HYPOTHESIS**

This report aims to address the following hypothesis:

H1: Is it possible to find similarities between the schools based on the population of students who are disenfranchised?

H2: Is there a relationship between income inequality and school capacity?

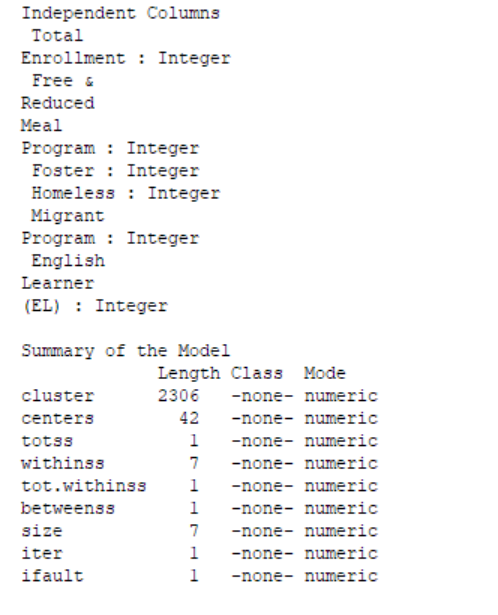
As well as, the following research questions:

1. What are the median incomes in the differentLAUSD Zones?
2. What is average class size in the different LAUSD Zones?

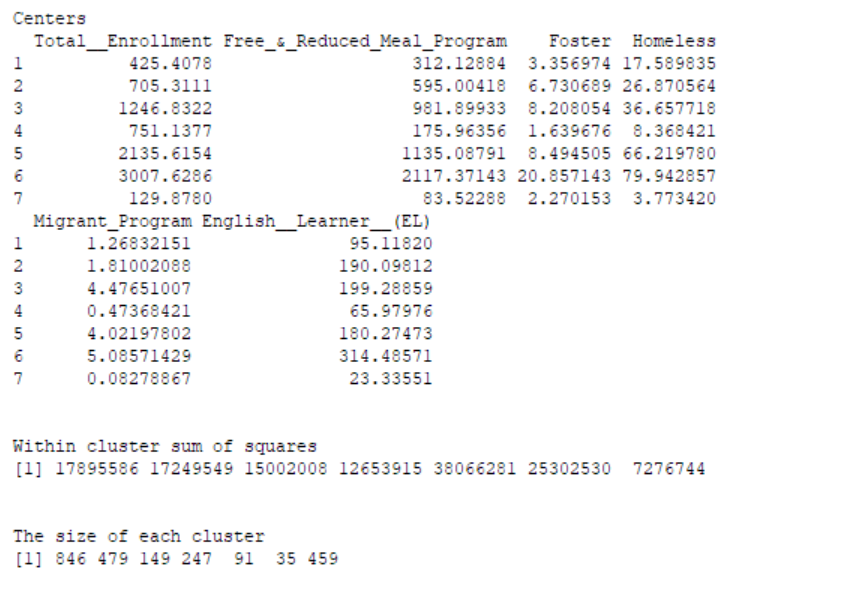
These hypotheses and research questions will help determine if there is a correlation between the aforementioned factors and overall school district functionality. Based on the analysis, we have affirmed hypothesis 1, it is possible to determine similarities between schools based on their population. We deemed hypothesis 2 to be positive, the gini index for Los Angeles affirms this to be true. Research question 1 and 2 are further concluded in the data analysis and results.

**DATA ANALYSIS & RESULTS**

**Figure 6.1: Summary of K-Means Analysis for Hypothesis 1**

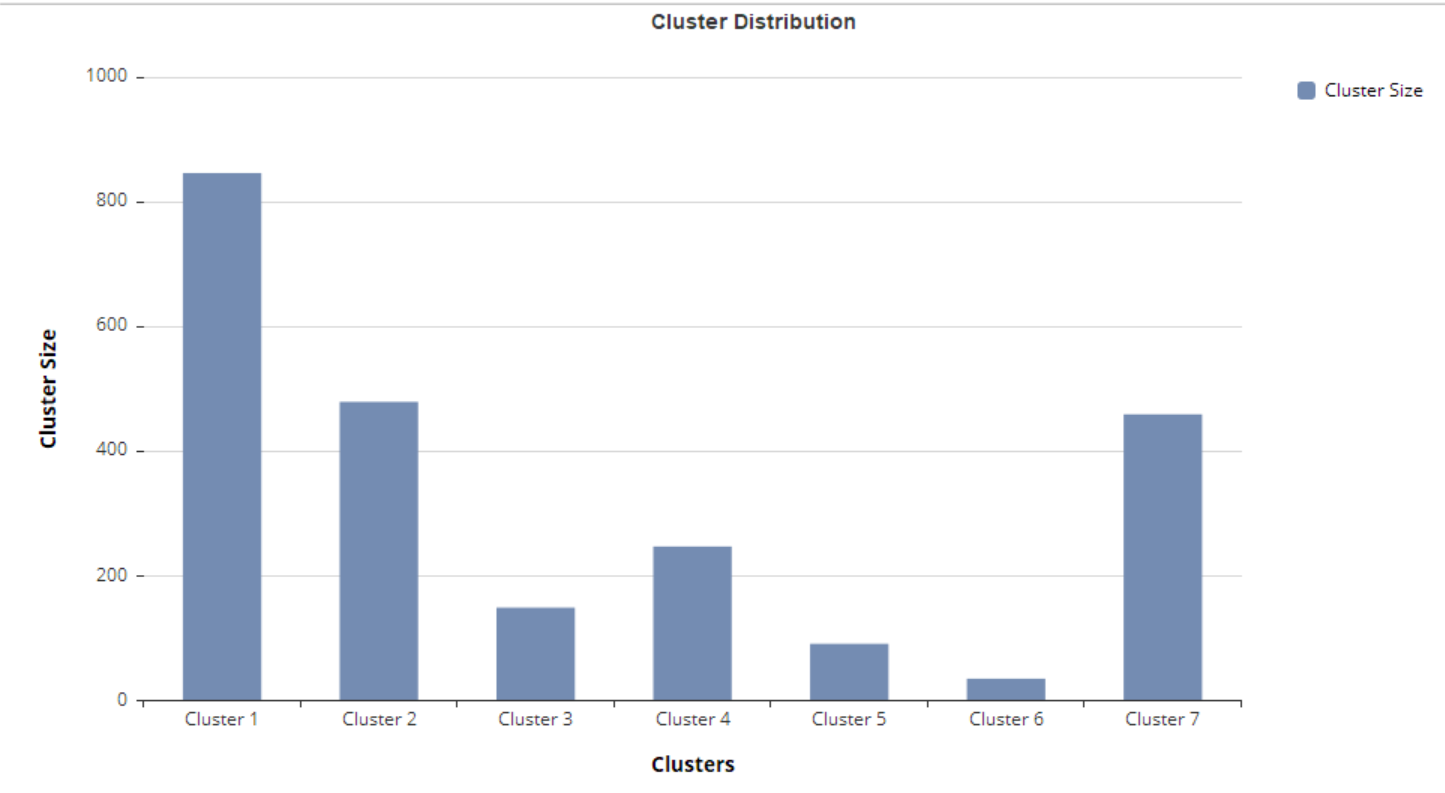


**Figure 6.2: Summary of K-Means Analysis (cont.) for Hypothesis 1**



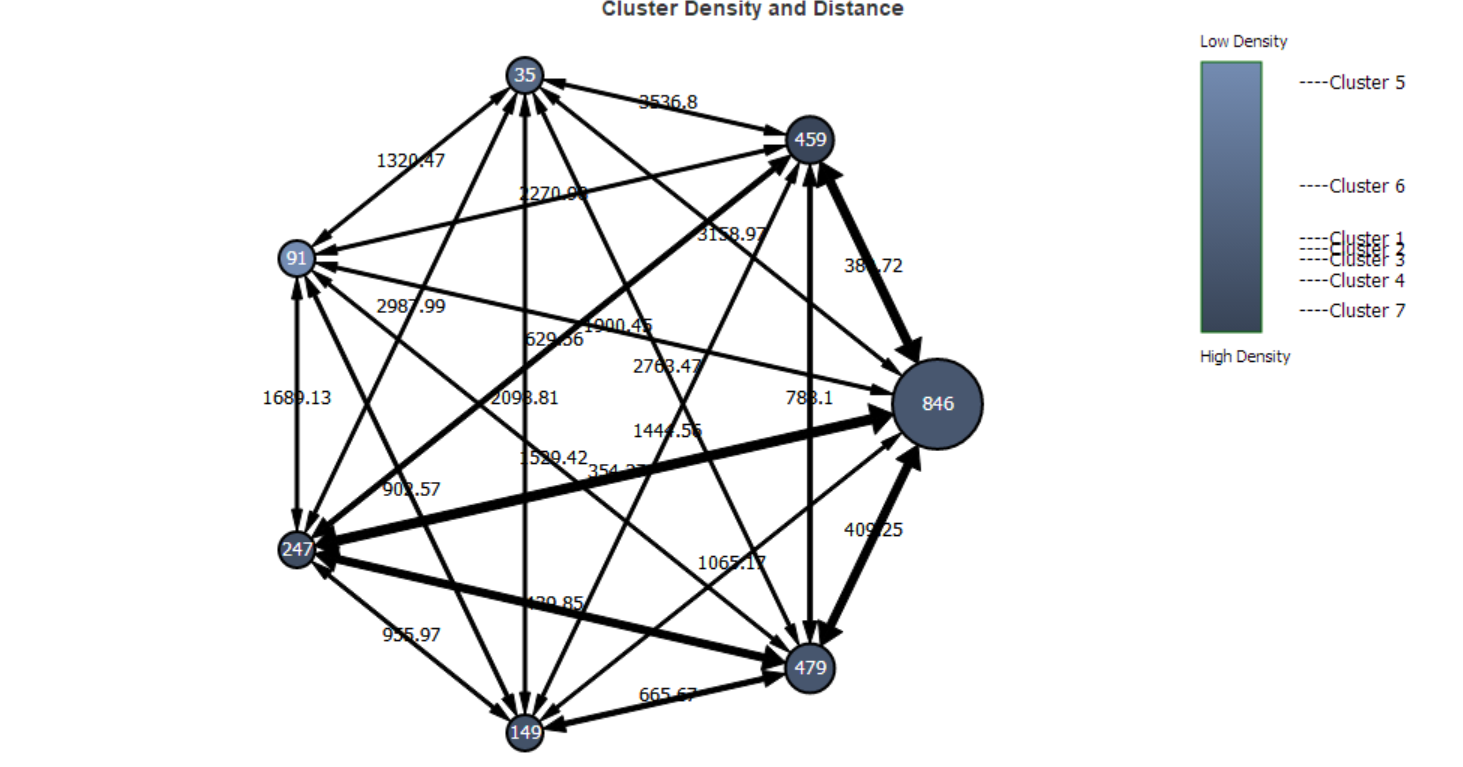
Based on the result of the K-means clustering it has been determined that Cluster 06 has the biggest centers. Conversely, Cluster 07 has the smallest centers.

**Figure 6.3: The Cluster Distribution for Hypothesis 1**



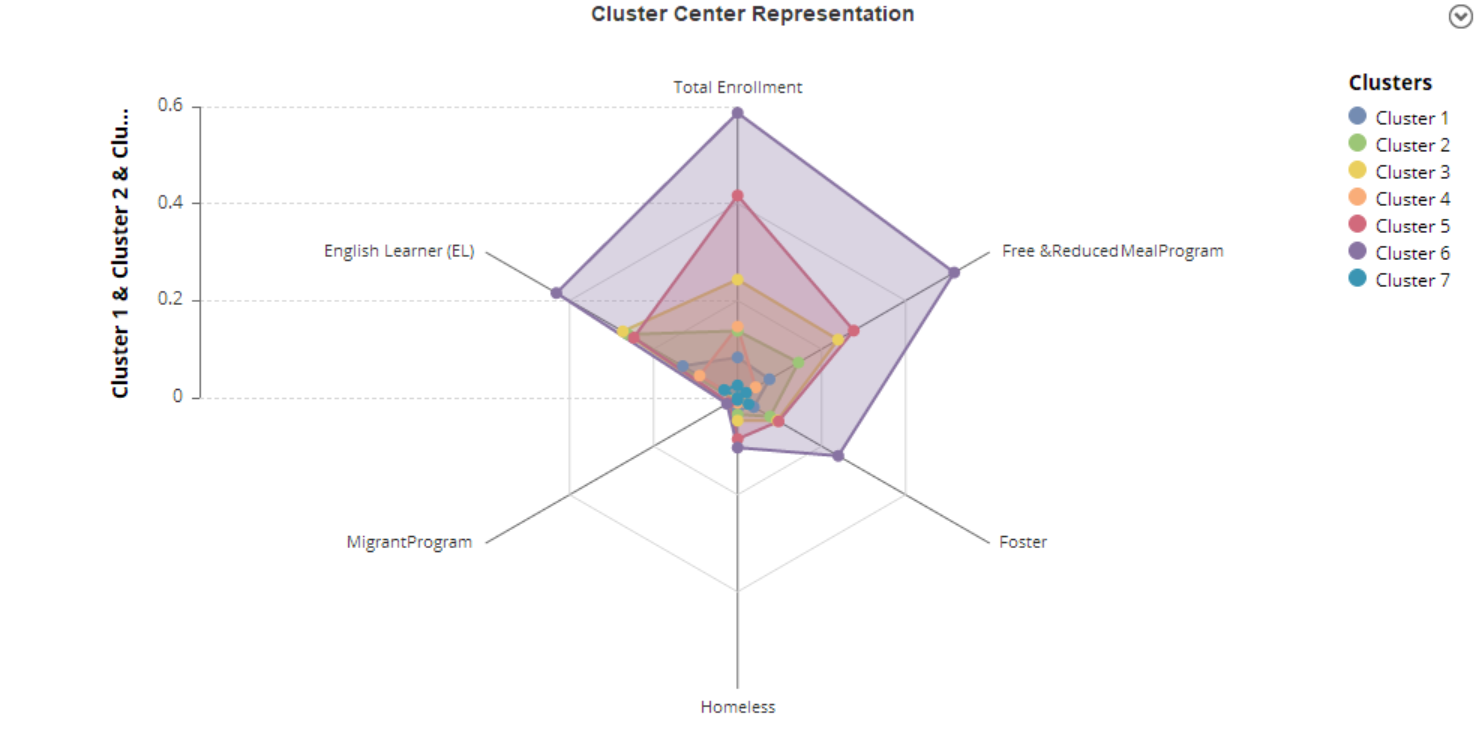
Most schools were classified under Cluster 01. The least amount of schools were classified under Cluster 06.

**Figure 6.4: Cluster Density and Distance for Hypothesis 1**



Cluster 07 is the most dense cluster and Cluster 05 is least dense. Moreover, Cluster 07 is the nearest to the other clusters. Cluster 04 is the farthest away from the other clusters.

**Figure 6.5: Cluster Center Representation for Hypothesis 1**



Cluster 06 has the largest concentration of total students and students who are underprivileged. Cluster 07 has the smallest concentration of total students and students who are underprivileged.

Major insights gained from the K-means clustering are:

1. The K-means analysis has highlighted the schools with the largest concentration of underserved students.
   1. The clusters with the highest concentration: 6, 5, 3
2. Highlights the most common issues faced by students: class sizes, food insecurity, and lack of proficiency in English.

The Gini Index is a ratio between 0 and 1 that represents the level of income inequality in a given area. The Gini Index is different from the poverty index in that it contrasts low and high earners to reveal an income gap. The Gini Index can be interpreted as follows (County of Los Angeles Open Data, December 2019):

* Index = 1; Perfect Inequality
* Index=0; Perfect Income Equality

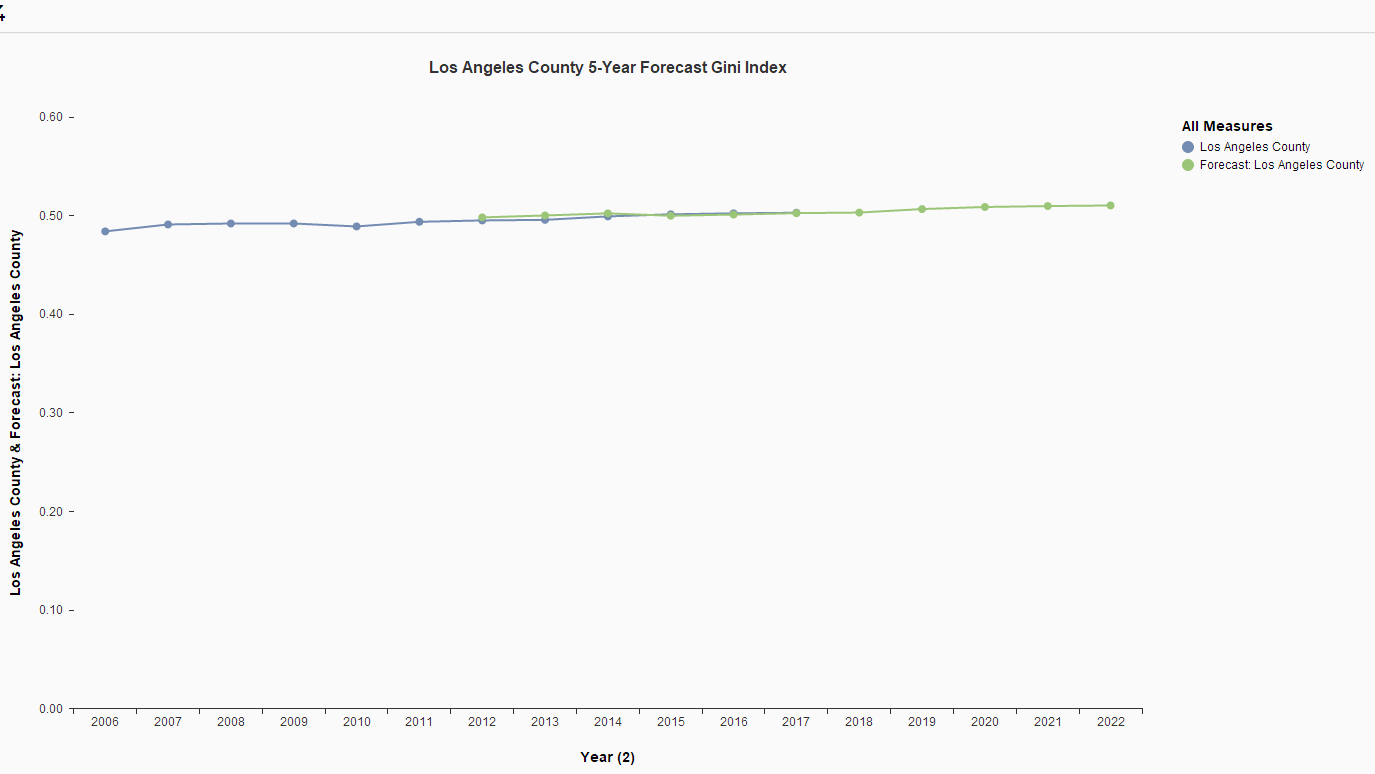
The Gini Coefficient in Figure 6.6 measures within a 95 percent confidence interval an index of 0.52 in Los Angeles County in 2014. 0.52 is a relatively high index denoting that there is income inequality in Los Angeles County. It is forecasted that there will be a slight increase in income inequality over the next 3 years.

The above cluster examines the relationship between underprivileged youth and the current educational system. While the Gini Index can be used to relate the economics, it is also applied to education. A work paper conducted by the International Monetary Fund, an offshoot of the United Nations, uses the following formula to explore the relationship between income and educational inequality (Coady, May 2017):



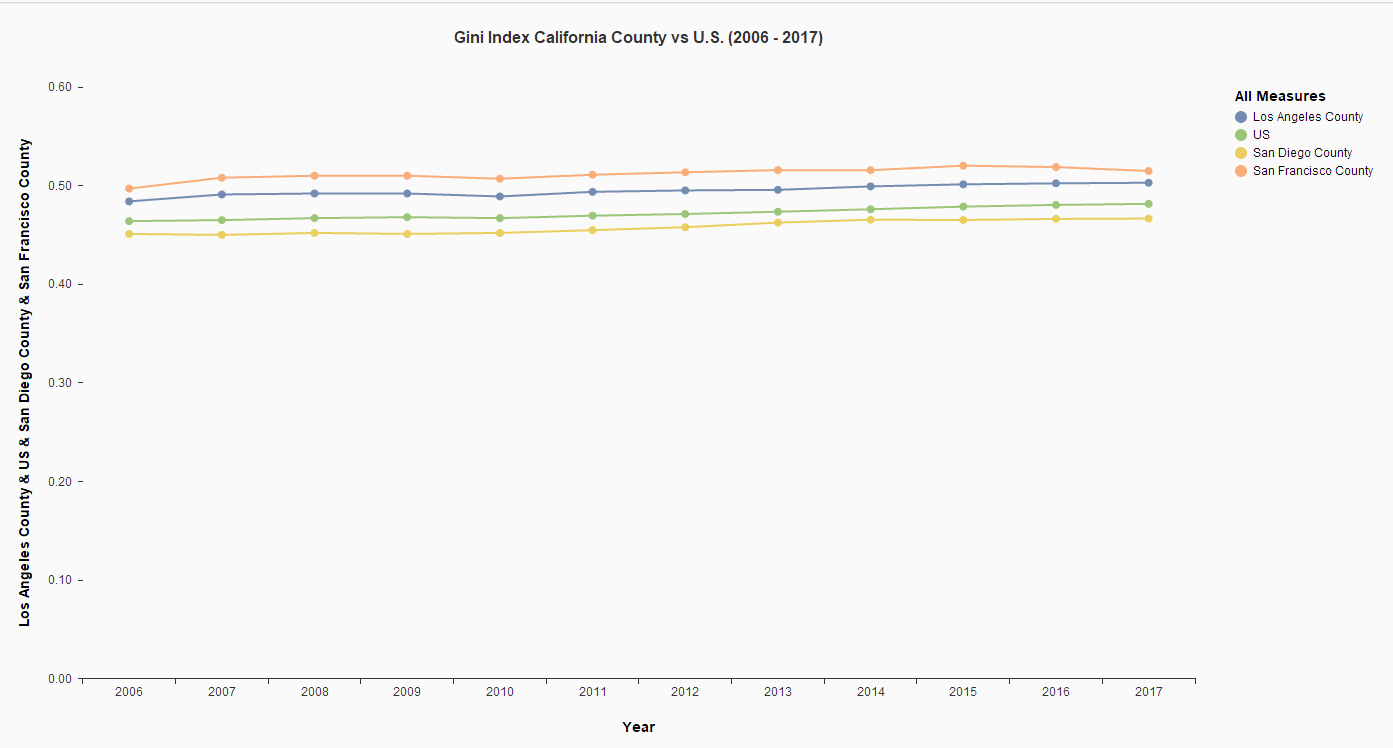
With I representing Income Inequality, E as Average Education and Alpha as Education Inequality. This equation makes income inequality dependent on the level of education and existing educational inequality. An idealized school zone would take into account the Gini Index when zoning. A computerized mathematical model can determine a best fit model that limits factors that result in high income and educational equality in a disproportionate amount of school districts.

**Figure 6.6: Los Angeles County 5 Year Forecast of Gini Index for Hypothesis 2**



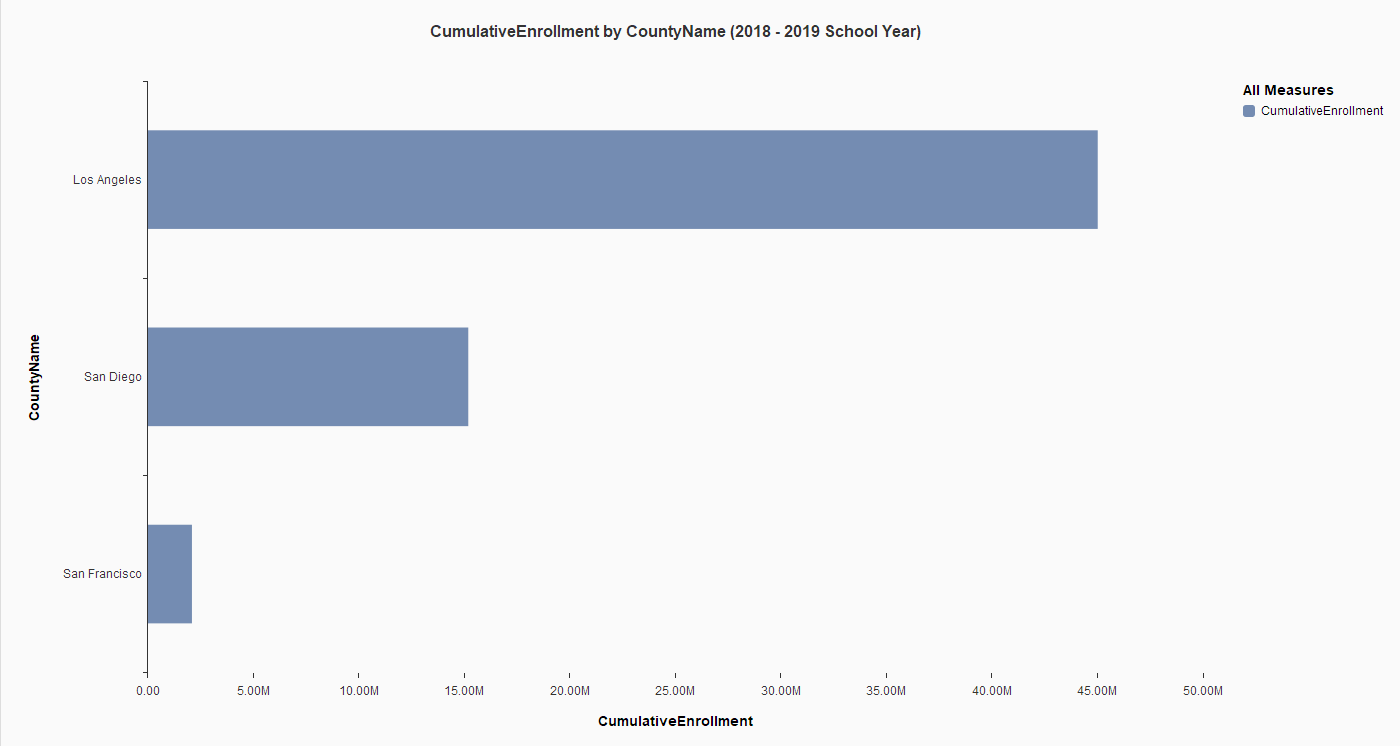
Los Angeles County is one of the largest counties in California and the United States as a whole. The county consists of 3.8% or $710.29 billion USD of the U.S. total GDP, yet there is a notable income disparity in comparison to the rest of the nation (Tartar, December 2019). The county also struggles with a substantial homeless population of close to 60,000 individuals. With this in mind, it is evident that the Gini Index found in Figure 6.7 is an accurate reflection of income distribution in Los Angeles County. Los Angeles County rests above the U.S. Gini Index consistently for 9 years between 2006 and 2017. This means that despite the massive economy in Los Angeles County, there is significant income inequality. This income disparity is more evident in certain geographic counties than others. Figure 6.7 reveals that Los Angeles and San Diego County have comparably high indexes compared to the United States average and San Francisco County. Given the Income Inequality formula, educational inequality should be happening most in school districts located in this county.

**Figure 6.7: Gini Index in California Counties Versus the U.S. from 2006 to 2017 for Hypothesis 2**



An adverse effect of income inequality is larger class sizes. Larger class sizes have been proven to result in increased distractions, thinly spread resources, and lower academic performance. LA County consists of approximately 45 million students in the 2018 - 2019 school year, compared to San Diego County’s 15 million and San Francisco’s 2.5 million. While there were a whopping 45 million students, there were only 73, 737 teachers in LA County from 2018 to 2019 and 25,443 in San Diego County (Learning Policy Institute, December 2019). Although the number of teachers is significantly lower than the cumulative student enrollment shown in Figure 6.8, the student to faculty ratio is lower because teachers are being spread thin by teaching multiple courses in a single semester.

**Figure 6.8: Cumulative Enrollment by County Name for the 2018 -2019 School Year for Hypothesis 2 and Research Question 2**

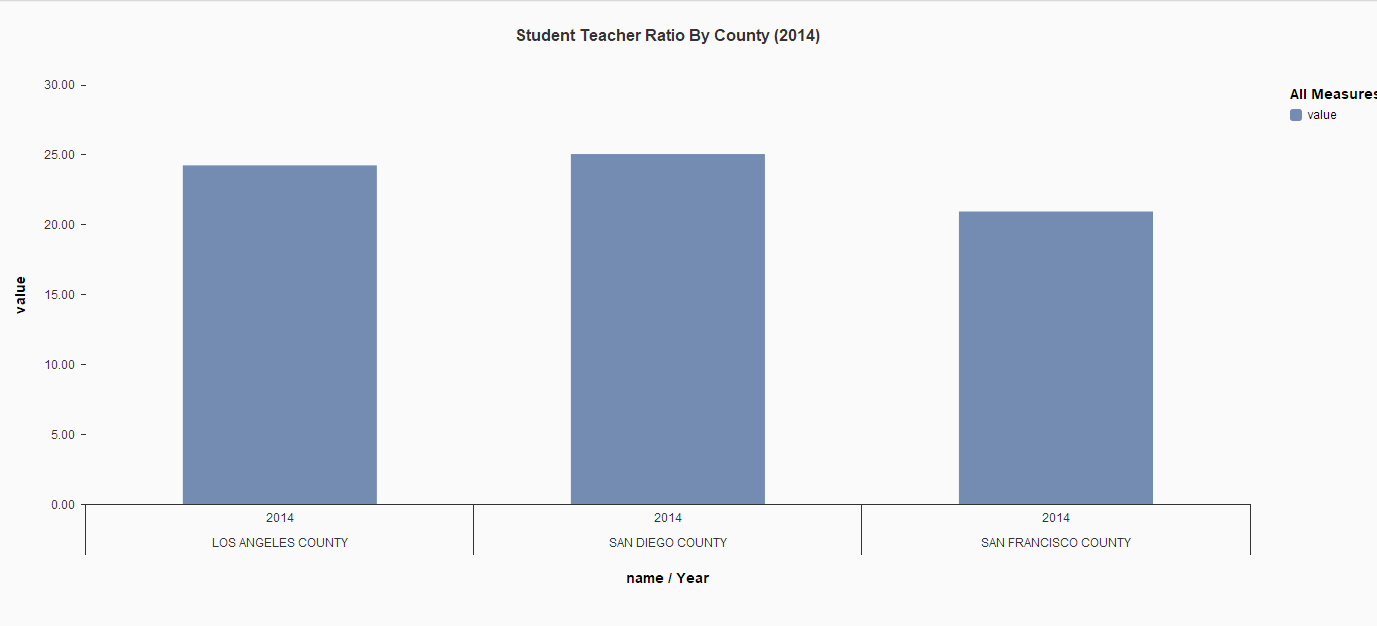


If the student to teacher ratio was calculated using the number of teachers and students total, it would be:

* Los Angeles County; 610 students per teacher
* San Diego County; 590 students per teacher

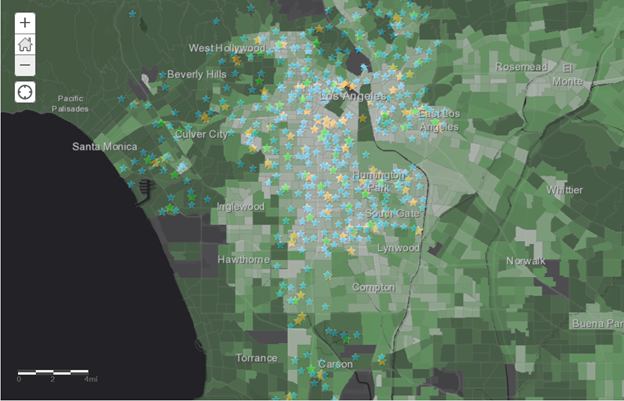
This figure is also offset by the school's ability to petition to receive no penalty for having larger class sizes. In a model that resulted in an equitable school zone, class sizes would be automated based off of a best fit model. Petitions and overworked teachers is a norm in the current model and needs to be removed by having an automated process.

**Figure 6.9: Student to Teacher Ratio By County in 2014 for Hypothesis 2**



California schools receive the majority of their funding from the State (58%), primarily from income and sales tax revenues, but also from local (32%) property taxes and other sources. Federal government only covers 9% of the school funding. Because income and sales taxes are more volatile revenue sources than property taxes, California school districts face dramatic cyclical funding variations as the economy rises and falls.

**Figure 6.10: LAUSD Elementary (blue), Middle (green) and High school (orange) and Median household Income for Research Question 1**



Educational inequalities begin from the funding of the government. Low income schools are more likely to receive less funding than high income schools. Los Angeles Unified School District has among the highest concentrations of low-income students in the state, with more than 80% living below the poverty line (Spectrum, January 2019). Most public’s schools get paid depending on their location. There is a larger disparity between low income and high-income cities.

**Figure 6.11: Earning by LAUSD School City for Research Question 1**

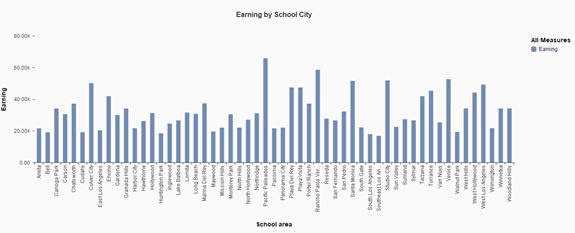
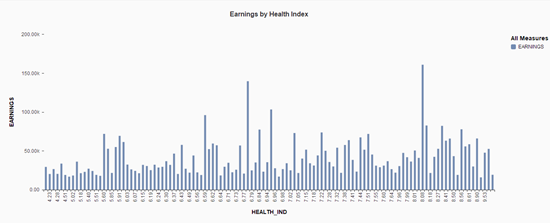


Figure 6.10 shows Los Angeles Unified School District median household income. Dark green areas with the highest income are mostly white neighborhoods and schools in those areas receive the most funding for the school. Light green and gray areas on the geo map are low income neighborhoods. Low income areas are consistent of Latino/Hispanic and Black/African American neighborhoods. Figure 6.11 shows that top median household income is over $50,000 and bottom median household income is less than $20,000.

**Figure 6.12: Earning by Los Angeles County Health Index for Research Question 1**



Most low poverty schools struggled with economic and social stressors, including unstable housing, hunger and lack of medical treatment as indicated by figure 6.12. Low poverty schools are mainly under pay students in low income communities leaving them with less resources. Teachers in low poverty schools were more likely to report problems in school that might create less teaching time, classroom lock downs, lack of access to technology, insufficient substitutes teachers and interruptions to the class.

These disruptions can take up to half an hour a day of teacher time. Teachers at low poverty schools also spent more class time counseling students with emotional and social problems, advising them on colleges and careers preparedness and discussing community problems and societal inequities. Inequality in learning time is hurting number of California’s most vulnerable students.Also lower income students don’t get the same opportunities to learn and achieve as their more affluent peers.

Fewer than half of the 70% of students who graduate from high school in the LAUSD complete the courses required for admission to University of California and California State University schools. students should have access to the school prep curriculum, nonetheless (UCLA, October 2015). Students should be provided tutoring and uninterrupted learning that would take to perform well in the classes. Teachers need to make sure that the learning time is rich, engaging and uninterrupted. Students in low poverty communities also lose more after school learning time than their more affluent peers.

**Figure 6.13: Top 50 Earning by High School & Bachelors for Research Question 1**

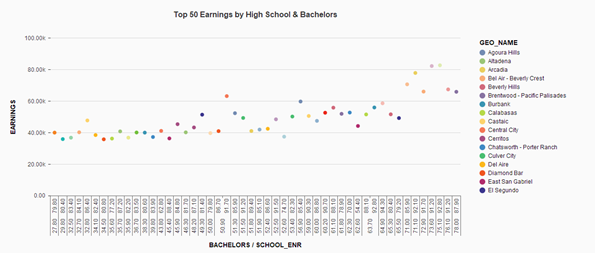


Figure 6.13 shows people living in high income areas are better educated. Students in high income areas have smaller classes, dedicated teachers, secure environments and ample resources. They also have many after school programs and parents are more involved with those programs.

When students feel the support of their teachers or they learn confidence from positive interactions in the class, they are more determined to pursue continuing education and approach college with admiration. Teachers need to make sure that students are developing emotional, physical, and social skills before they learn core classes. School needs to Incorporate parent involvement programs to help build a relationship with academics and personal life.

1. **DISCUSSION**

Based on the K-Means Cluster analysis it has been determined that the majority of schools have a manageable student population. Moreover, most schools are not overburden by students who face additional hardship. However, there our analysis has highlighted the huge disparities that exist between schools.

An actionable solution is to put a bigger emphasis on schools that have a large student population and a high concentration of students who are members of underserved communities. In a macro perspective, LA county must consider more investment or possible real location to better assist the schools that are labeled as high risk. The major obstacle to these reforms is the fact that funding will most likely come from an increase in taxes. However, the defeat of Proposition 15 shows that the possibility of levying new taxes is unlikely. Therefore, we should construct our action plan around the individual citizen. Here is a list of what a single person can do:

1. Donate to underprivileged schools
2. Volunteer your time to those underprivileged schools

The last suggestion is that schools should transition from mailing out paperwork to electronic documentation. This will help the schools save money and better assist students who are homeless or part of the foster system. Important documentation can be emailed to the parent, which will bypass the lack of a permanent mailing address.

1. **CONCLUSION**

Our recommended model attempts to remove the human bias associated with school districts and county borders. In exchange, a more equitable school district is adopted that reduces income and educational inequality. While this model may require extensive and accurate data collection, the mathematical approach helps to undo some of the system problems that have had a historical effect on the educational system.

The K-means analysis was able to divide the 87 school districts into 7 distinct clusters. Most schools were classified under cluster 01and cluster 06 had the smallest number of schools. It must be noted that most clusters were similar to cluster 01 than cluster 06. This finding is important because it shows that most schools have a manageable student. Moreover, the analysis shows that most schools do not have a high concentration of underprivileged students. However, the analysis also highlights the disparities that exist between the different schools. Schools that are part of clusters 6, 5, and 3 have the larger student population and concentration of underprivileged students. The Gini Index calculation provides sufficient evidence that there is income and educational inequality in Los Angeles compared to the United States and surrounding counties. There is a relationship between class size and educational inequality, as districts with a high student to teacher ratio are underperforming and underfunded. The median income in Los Angeles correlates with the Gini Index, in that in high income school districts there is a lower student to teacher ratio and greater resources on campus. The K-Means analysis, forecasting, and income calculations provide substantial evidence that there is an educational, financial, and environmental prejudice existing in Los Angeles County school districts. The K-Means clustering also affirms that with increased equitable school zoning there is increased performance.

1. **REFERENCES**

Balfanz, R. (May,2012). The Importance of Being In School: A Report on Absenteeism in the Nation’s Public Schools. Retrieved from <https://coseboc.org/sites/coseboc.org/files/assets/ChronicAbsenteeismReport.pdf>.

Ballotpedia. (2020). California Proposition 15, Tax on Commercial and Industrial Properties for Education and Local Government Funding Initiative (2020). Retrieved from Ballotpedia: https://ballotpedia.org/California\_Proposition\_15,\_Tax\_on\_Commercial\_and\_Industrial\_Properties\_for\_Education\_and\_Local\_Government\_Funding\_Initiative\_(2020)

California Department of Education. (n.d.). District/School Boundaries. Retrieved November 30, 2020, from <https://www.cde.ca.gov/re/di/fq/distschlboundaries.asp>.

California Department of Education. (2020, June 29). Retrieved from California Department of Education: <https://www.cde.ca.gov/ds/sd/sd/fscupc1617.asp>

California Department of Education. (2020, July 10). Definition of Homeless. Retrieved from California Department of Education: https://www.cde.ca.gov/sp/hs/homelessdef.asp#:~:text=The%20McKinney%2DVento%20Act%20defines,regular%2C%20and%20adequate%20nighttime%20residence.&text=Children%20and%20youths%20who%20have,sleeping%20accommodation%20for%20human%20beings

County of Los Angeles Open Data. (December 6, 2019). Income Inequality. Retrieved November 30, 2020 from <https://data.lacounty.gov/Sustainability/Income-Inequality/jh5d-y5fe>.

Coady, D., Dizoli, A.. (May 2017). Income Inequality and Education Revisited: Persistence, Endogeneity, and Heterogeneity. (WP/17/126). Retrieved November 30, 2020 from <https://www.imf.org/en/Publications/WP/Issues/2017/05/26/Income-Inequality-and-Education-Revisited-Persistence-Endogeneity-and-Heterogeneity-44854>.

Garcia, E., Weiss, E., (September 27, 2017). Education inequalities at the school starting gate. Retrieved from <https://www.epi.org/publication/education-inequalities-at-the-school-starting-gate/>.

Learning Policy Institute. (December 5, 2019). Interactive Map: Understanding Teacher Shortages in California. Retrieved November 30, 2020 from <https://learningpolicyinstitute.org/product/interactive-map-understanding-teacher-shortages-california>.

Mackenzie, J. (2009). *Public School Funding and Performance*. Retrieved from Semantic Scholar: https://www.semanticscholar.org/paper/Public-School-Funding-and-Performance-Mackenzie/55c148dfacc3502bfd1298eb7429e9d1ea60658b?p2df

Spectrum News 1 (January 09, 2019) Looming Strike Has Some Parents Worried Over Future of LAUSD. Retrieved November 30, 2020 from

<https://spectrumnews1.com/ca/la-west/news/2019/01/09/lausd-parent-worries-school-district-is--toxic->

Tartar, A. (December 19, 2019). One-third of the U.S. economy is jammed into just 31 counties. L.A. is the biggest. Retrieved November 30, 2020 from <https://www.latimes.com/business/story/2019-12-19/los-angeles-largest-economy#:~:text=L.A.%20is%20the%20biggest,of%20the%20U.S.%20total%20economy.&text=While%20the%20nation's%20economy%20has,1%25%20of%20the%20nation's%20counties>.

UCLA Blue Print (October 28, 2015) Unequal Schools, Generations of Poverty Retrieved November 30, 2020 from <https://blueprint.ucla.edu/feature/unequal-schools-generations-of-poverty/>

Wald, J., & Losen, D. J. (2003). Defining and redirecting a school-to-prison pipeline. *New directions for youth development*, 9-15.