# COVID Cool for School?

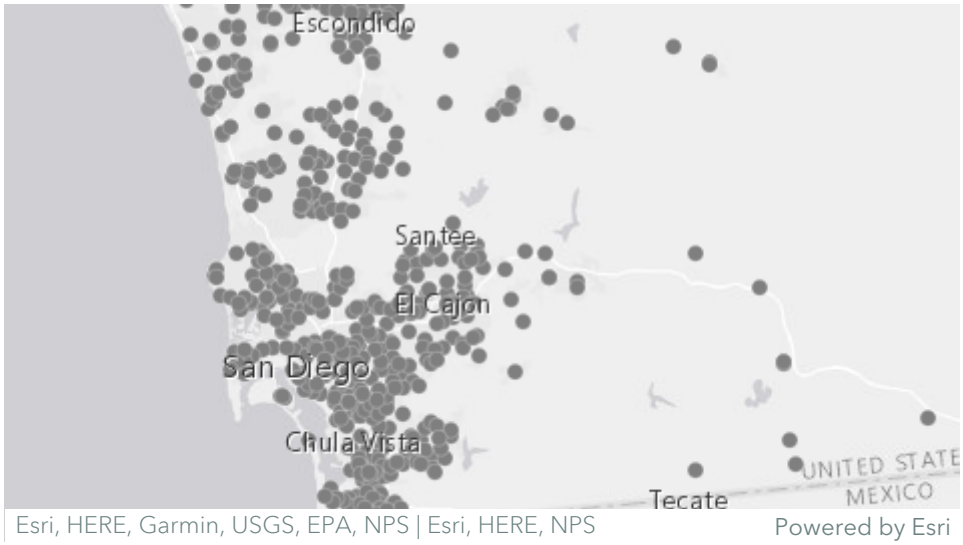
Should K-12 Schools in San Diego County open? Or should they stay virtual?

**September 27, 2020**

COVID-19 has impacted the school systems significantly in San Diego County. Since schools are finally opening back up, an important question arises regarding these schools: "Is it covid cool for school?". More formally, the question really means: "based on the conditions that we know today, is it safe enough for schools to go back to in person instructions?".

## Overview

Our project uses COVID-19 data to analyze the potential risk of opening elementary, middle, and high schools in San Diego. We identified 618 schools in San Diego county and built a classification model that determines the risk of each specific school based on the demographic information of its surrounding area. We used ArcGIS to build a web application to display the distribution of risky and non-risky schools.

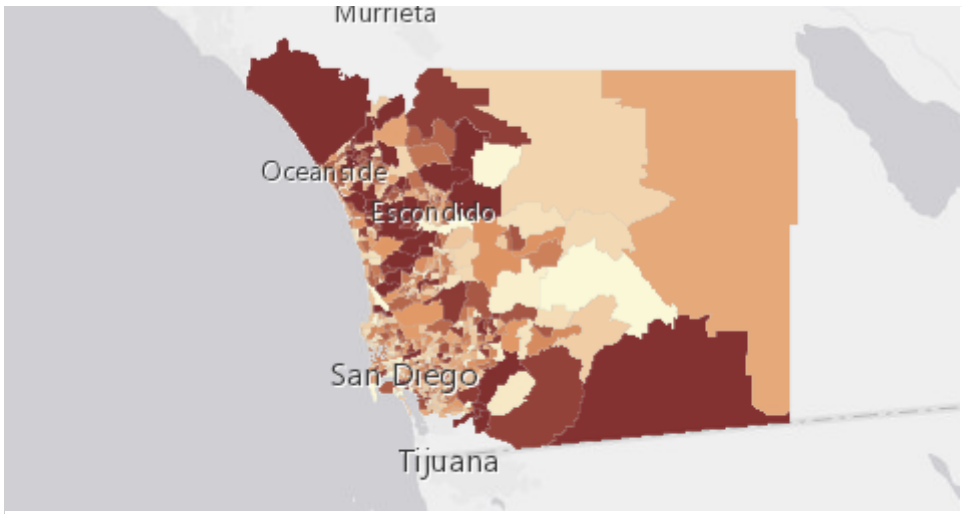



San Diego Schools Data Analysis

## Research Question

How can we determine whether or not it is safe to return back to school? This is a very complicated question because every school is different and its effects are multivariate. However, we realized by leveraging the data available we could identify the metrics that are statistically significant in relation to the COVID case count.

What metrics that are shared amongst different zip codes in the past year are the most correlated to the speed of COVID and can we use those signifiers to determine risk?



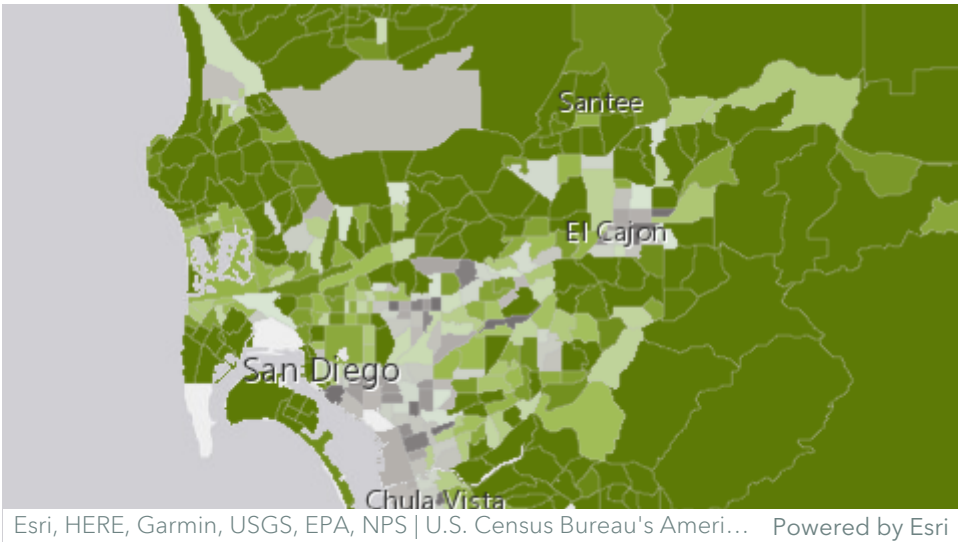

San Diego Demographic Analysis-Copy

## Background and Prior Work

COVID-19 has dramatically altered the lives of everyone, especially communities along the U.S.-Mexico border. While scientists and researchers try to understand the behavior and effects of this pandemic, they are able to mine and generate data that could be leveraged to a better understanding of what is happening.

We were specifically interested in how COVID will change the “back to school experience”. As the San Diego Unified School District has planned to begin schooling through online synchronous and asynchronous teaching, we are curious in determining the safety of returning students in different areas.

Our goal is to conduct an analysis of the relationship between COVID-19 case counts/increases and the demographics of the area surround the school. This will reveal any potential correlative relationships and give a better understanding of what potential factors increase the effects of COVID-19.



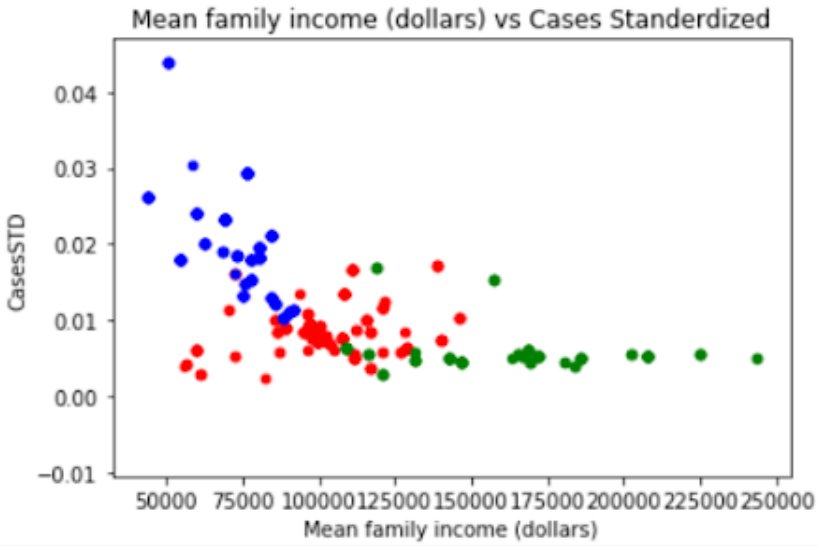

San Diego Schools Data Analysis

## Hypothesis

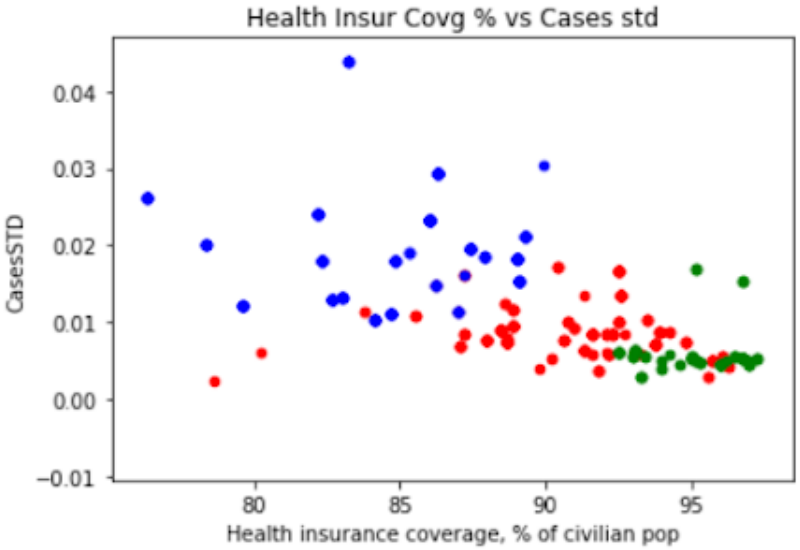
We hypothesize that the effects of COVID will be the most severe for underprivileged areas and schools. We predict that certain demographic metrics such as income, insurance coverage, and the number of remote workers greatly affect the case count and increases in that specific area. Because of the inequities of access to resources, areas with less will have fewer luxuries and preventative methods against COVID. We also hypothesize that areas with greater concentrations of people such as National City will be more susceptible to COVID. Furthermore, we think that the location closer to the border will also be more susceptive to COVID. We make this prediction because we understand that COVID is an airborne virus and is easily transmitted through close proximity.

## Data Analysis & Results

While analyzing the increases in COVID from July to August, the most correlative indicators are reaffirmed with our initial suspicions. Not having health insurance had a direct correlation with the increases of COVID cases. Furthermore, median family income is another factor that is strongly related to the increases in COVID.




Mean Income in relation to Standardized Cases




Health insurance coverage in relation to Standardized Case

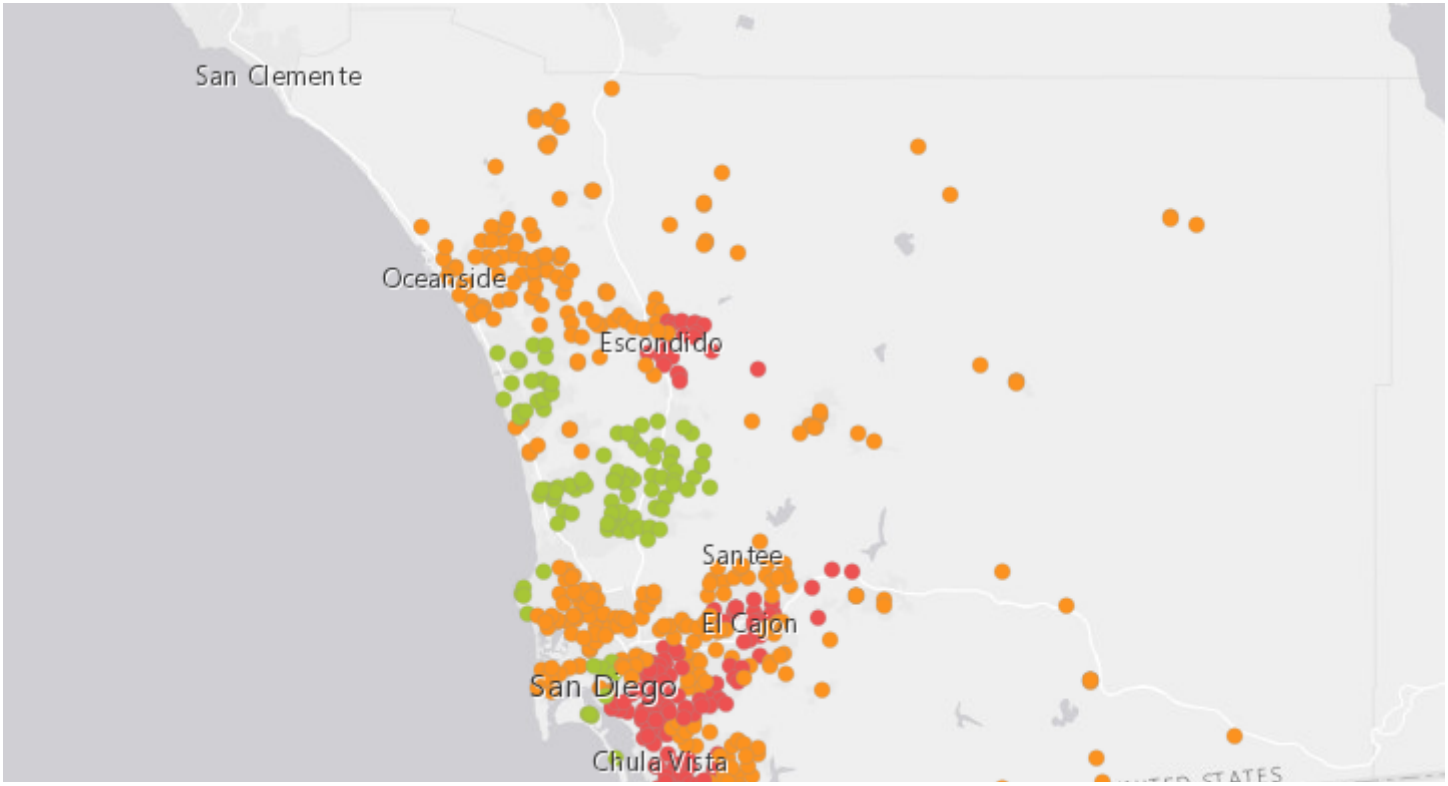
## Feature Selection

One of our datasets titled “income” contained many of the features that we would later use in our clustering model, including Health insurance coverages, people that worked at home, children of the households, and various household incomes all grouped by zip codes in San Diego County. Initially, there were about eight or nine divisions of income levels, so for simplicity, we grouped these into three sections 0-35k, 35k-100k, and 100k plus. From our other datasets, we were able to include up-to-date information on both the flat number of cases and the absolute increase of cases over the month of August 2020. All of these features were then standardized to fit the population. At this point we had many features to distinguish between zip codes, but not as many for distinguishing between schools within a zip code, so we used cumulative enrollment of each school, as well as the percentage of Hispanic students for each school because we suspected that COVID affects minorities more heavily, and standardized those as well. Ideally, we would have liked to have geo-codes for the boundaries and size of each school and even classroom within, but finding information like that on the internet proved to be exceedingly difficult without proper

communication and exchange with California Department of Education officials, so we decided to create our model using these eleven features.

## Modeling

None of the available data was labeled, so we had to use an unsupervised machine learning approach. Also, our goal was to cluster the data by the similarity of the selected features, so we selected a K-means algorithm from Scikit learn. We decided on grouping the schools into three clusters and ran the algorithm for a thousand iterations and selected the grouping with the lowest inertia.




San Diego Schools Data Analysis

## Evaluation

The K-means algorithm grouped the schools into three separate clusters, but we needed to decipher the groupings. At

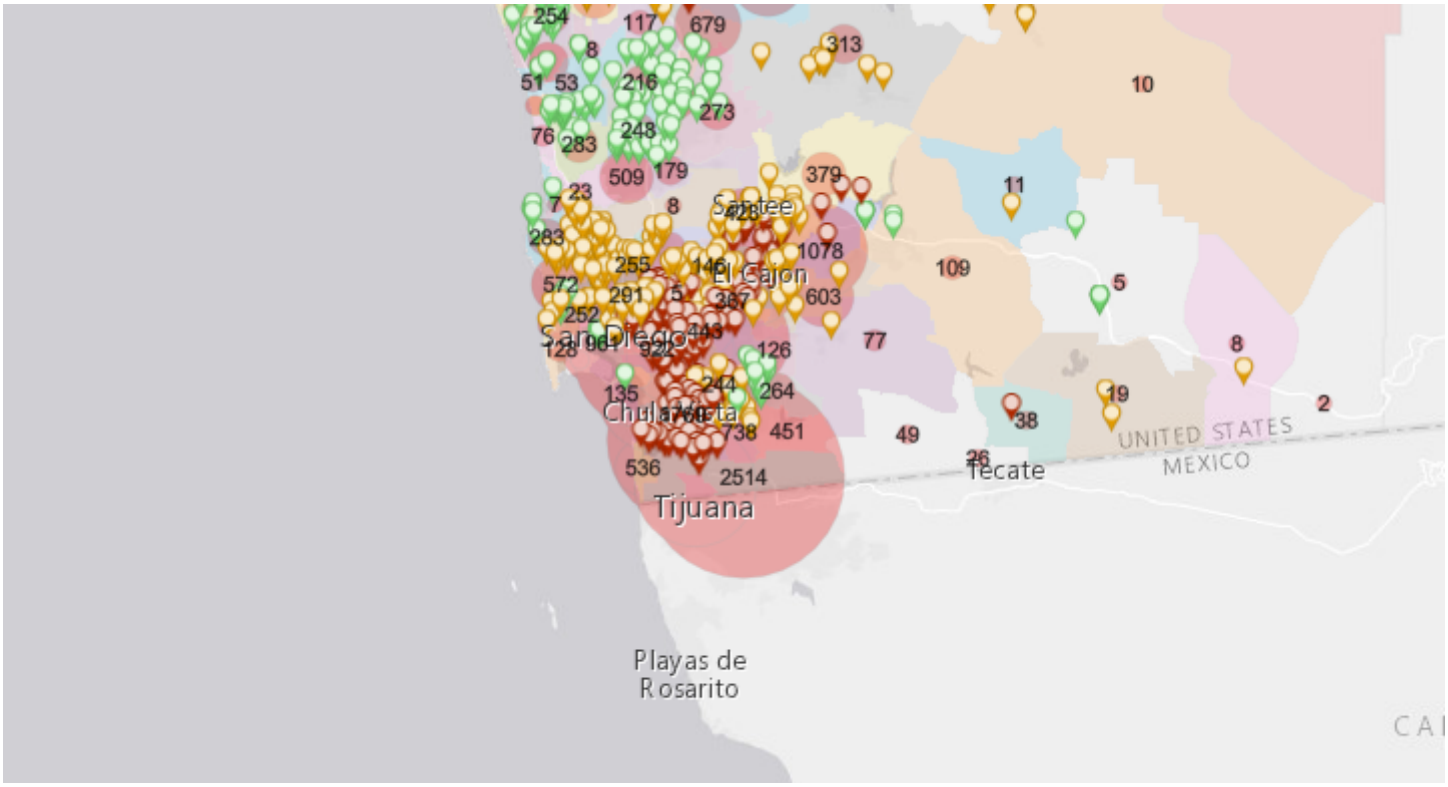
first, we tried to take the mean value of each feature in each cluster to get a better understanding of the clusters. However, the mean value told us nothing about the distribution of each feature, so we decided to graph two features on a scatter plot and color code them to their cluster classification. This allowed us to have a manual look at the data and then decide with a team consensus on how to rate each cluster relative to one another. From the data exploration, we made sure not to have a dominant feature that could single-handedly predict a school's cluster.

## Ethics & Privacy

As responsible data science students that have taken COGS 9, we realized that we do not want to perpetuate a biased garbage-in garbage-out model.

## Conclusion & Discussion

A majority of the features we used to classify risk were based on the zip codes in which a school was located in. Although COVID affects groups regionally, we aimed to have more granularity for each specific school. However, we weren't able to find a limited amount of data for specific schools in San Diego and because of this couldn't include it as features of our model.




School Map Risk Analysis with Zip Code Cases