Winny Kiepe
DS 710
12/1/2023

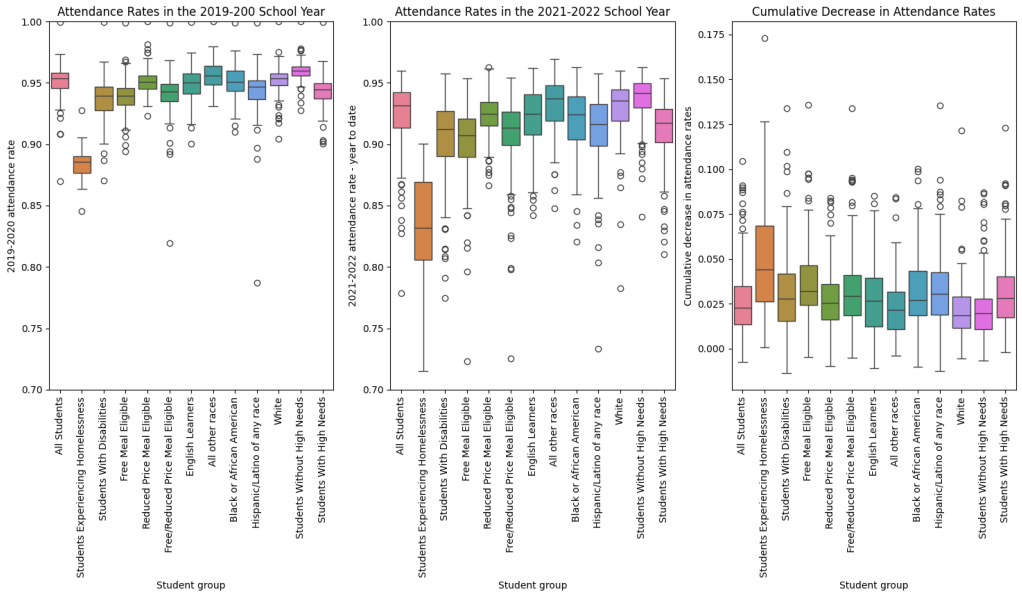
I chose to analyze data surrounding attendance rates in Connecticut's public schools. I chose this dataset because I knew I wanted to work with data related to public education in the United States since my ideal job will be working with school districts/states to analyze district data. I chose this data because I wanted to analyze a trend that I had seen happening in my district and that I hear many other educators talking about: decreases in attendance rates since the pandemic. I found the data on data.gov on 11/29/2023.

The goals of my analysis were to identify trends in attendance rates from the start of the COVID-19 pandemic to the school year two years later. I was motivated to look into these trends to see if other school districts were seeing the same trends I was seeing when I was a teacher. Furthermore, the dataset I found includes attendance rates for various demographics which allows me to look at possible inequities. My goal is that this analysis can be used to help identify districts that have successfully mitigated attendance loss in order to facilitate further study to determine methods that districts are using to promote attendance in their schools. For my final analysis I am focussing on trends across student subgroups as well as the relationship between district size and the change in attendance rates.

What are the trends in student attendance since the start of the COVID-19 pandemic?
Are these trends arising in all subgroups within school districts? If so, are different groups impacted more or less than others?

I was initially wondering what trends in school attendance rates could be seen in the years since the COVID 19 pandemic. When I was working in Sun Prairie my colleagues and I had noticed a decrease in student attendance over the past few years, which is a pattern I have heard many other educators reporting on. Using data from the state of Connecticut I wanted to determine whether they were seeing the same patterns I had noticed in my own school district. Furthermore, because the dataset has data for varying subgroups I wanted to see how different groups were impacted. I started out by looking at trends throughout the state as a whole. Because the dataset already had aggregate data for the state of Connecticut, this required minimal data wrangling. In order to compare attendance rates I created a function that takes in as inputs a list of student groups and a district to analyze with the default inputs being a list of all student groups and the district being the entire state. The function creates a subset of the csv that contains only the entries from the chosen district and then calculates the change in attendance rates between the 2019-2020 and 2020-2021 school years, the change in attendance rates between the 2020-2021 and 2021-2022 school year, and the cumulative change in attendance for each of the groups from the input. Because not every district collected data for each of the groups, if the district does not have data for a particular group, the value for that calculation is set to NaN. The function then returns a data frame of the changes in attendance rate for each group to allow them to easily be compared. From this analysis I saw that all students saw a decrease in attendance from the 2019-2020 school year (when the pandemic first hit) to the 2021-2022 school year and each subgroup of students saw a decrease in attendance rates with students experiencing homelessness being the most impacted (seeing a cumulative 5.36% decrease in attendance) and white students and students without high needs being the least impacted groups (seeing a cumulative 2.05% and 2.08% decrease in attendance respectively).

Are these trends universal across school districts?
As a whole, school districts across the state of Connecticut saw a decrease in attendance rates for all groups of students as seen in figure one below.



The first 2 boxplots show the attendance rates in the 2019-2020 school year and the 2021-2022 school year. As can be seen, the median value of each group saw a decrease. We also see higher variance in attendance rates after the pandemic with each group of students having a larger interquartile range and all groups except for students experiencing homelessness containing more outliers (I think it is worth noting that while 2019-2020 had a handful of outlier districts perform better than average, all the outliers in 2021-2022 had performed below average). This seems to indicate that while most districts are struggling with decreased attendance rates, some districts are adapting more successfully than others.

The third boxplot shows cumulative decreases in attendance rates. We see that each group saw a median 2.5% - 4.5% decrease in attendance rates. By looking at the bottom of the whiskers on the graph, we see that several groups of students had at least one school where there was an increase in attendance rates for that group. Upon looking more closely at the data I found two school districts saw an overall increase in attendance rates during the time period of this data collection, and 23 school districts saw at least one subgroup of students increase their attendance rate. My hope is that analysis of these schools can be used to identify and implement policies and practices that can be applied across the state to better support attendance for all students.

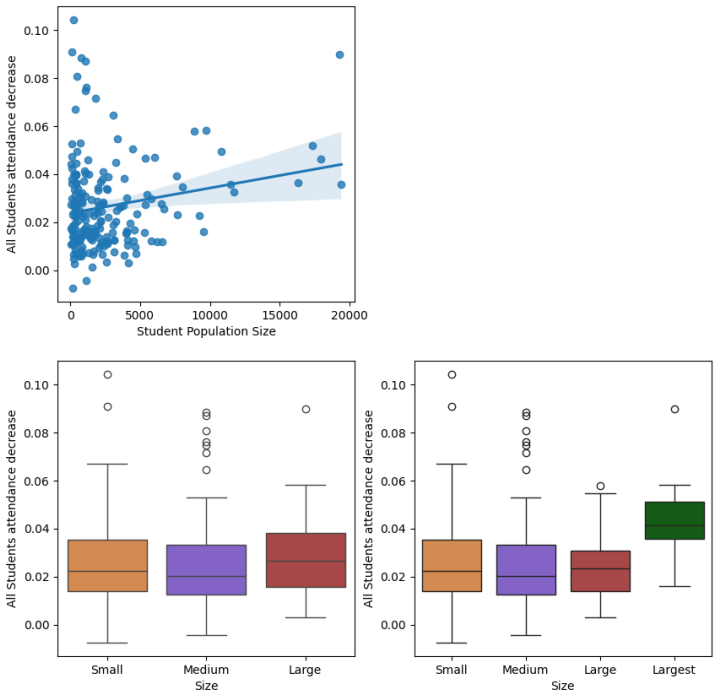
I do consider it to be worth noting that I have not done any statistical analysis on the probability of the varying changes in attendance rates. That is to say, the current analysis cannot answer the question “Are the different changes in attendance rates of varying subgroups from separate districts statistically significant?(E.g. is Bolton School District’s cumulative change in attendance rates for students with disabilities significantly different than Connecticut’s change in attendance rates for students with

disabilities or could the differences in attendance rates be attributed to chance)?" Nor does it answer the question "Are the different changes in attendance rates among different subgroups within a school district statistically significant? (E.g. is the change in attendance rates for Bolton School District's white students significantly different than the change in attendance rates for students of other races or could the differences in attendance rates be attributed to chance?)". These two questions should be answered prior to any further in depth analysis of individual district's.

Additionally, this analysis looks solely at attendance rates for the varying student groups but does not take into account the population size of those groups. Perhaps machine learning algorithms could be used to generate a model that predicts attendance rates for school districts based on their current attendance data which could be used by districts at risk of seeing excessively high attendance decreases to take proactive steps to ensure students are making it to class.

Does the size of the school district appear to have an impact on the change in attendance rates?
To answer this question I created three more dataframes: one for small school districts (bottom quartile of district size), one for large school districts (top quartile of district size), and one for average sized school districts (the remaining districts). Small school districts saw a median 2.2245% decrease in attendance and a mean 2.643% decrease in attendance. Average sized schools saw the smallest decrease in attendance rates with a median decrease of 2.03% and a mean decrease of 2.51%. Lastly large school districts saw the largest decrease in attendance rates with a median decrease in attendance of 2.675% and a mean decrease in attendance of 2.916%. I recognized that all of these were below the statewide attendance decrease of 3.1% and was initially confused as to how this could be possible. I realized that the largest school district must be having a significant impact on the statewide attendance rates since none of our average changes in attendance rates were at or above the 3.1% decrease in attendance that Connecticut saw as a whole. This led me to create one more dataframe of the largest school districts consisting of the 10 largest districts in the state. Looking at the data frame of the 10 largest districts, I saw that all but the 10th largest district had a decrease in attendance higher than 3.1%. The median decrease in attendance rates for these 10 largest districts is 4.1% with a mean decrease in attendance of 4.5%. This seems to indicate that the largest school districts are having a significant impact on the state's data, especially since the largest 10 districts make up nearly 30% of the state's student population.

I then created a scatterplot with a line of best fit to visualize the relationship between district size and attendance rates which can be viewed in figure 2.



The scatterplot shows a slight positive correlation between a district's size and its decrease in attendance rates though I am skeptical that it is a statistically relevant correlation given how much variance exists in the smaller school districts. Skepticism of a linear correlation between district size and change in attendance seems further justified when looking at the first set of boxplots which shows small, medium, and large districts all have comparable changes in attendance rates. It is not until we separate the largest districts out that we see any significant change with the largest districts showing a median decrease in attendance rates approximately twice that of other districts. This fact alongside the shape of the scatterplot leads me to wonder if a quadratic model (or perhaps some other type of nonlinear model) may appropriately model the relationship between district size and attendance rates. What I think is more likely, though, is that there is currently a hidden variable that is more prominent in the largest school districts than smaller school districts. Perhaps they have a higher proportion of student groups that saw a large decrease in attendance (such as students with high needs) than the smaller districts, or perhaps being in more populated areas means that they were more likely to have students exposed to COVID. This would require more in depth analysis of student demographics in each of the school districts than this report goes into, but it is a possible next step in understanding the decrease in attendance rates across Connecticut. I would also like to reanalyze this data changing the criteria for qualifying as a small, medium, and large district. In this analysis they were determined by the smallest and largest quartile of student populations. Looking at the distribution of district sizes though, this seems like a flawed method for determining district size given that a vast majority of districts have under 5,000 students while only about a dozen districts have more than 10,000. Perhaps instead small, medium, and large districts should be defined by total student population size rather than relative student population size.

Next Steps

This report has already mentioned a few potential next steps: first, running a more in depth statistical analysis of the change in attendance rates within and across school districts; second, running a more thorough analysis focusing on the largest school districts to try to identify causes for their decreases in attendance rates. Furthermore, the original data has several factors that this analysis did not dive into including demographic information about school districts. As was suggested earlier in the report, a more complete analysis could potentially utilize machine learning algorithms that utilizes that demographic information to create a more accurate model to predict districts at high risk of seeing above average declines in their attendance rates. My goal of such a model is to support those high risk districts in taking proactive measures to support their students in attending class.