Winny Kiepe
DS 710
12/1/2023

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.

*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). There were two further observations that struck me as noteworthy and potentially worth future exploration. First, students who didn't have high needs saw an increase in attendance rates between the 2019-2020 school year and the 2020-2021 school year (this was the year that many districts across the country used a virtual or hybrid model); they were the only group to see an increase in attendance rates in the 2020-2021 school year. Second, despite having the largest decrease in attendance rates the first year after the pandemic(at 7.29%), students experiencing homelessness had a notable increase in attendance between the 2020-2021 school year and the 2021-2022 school years (at 1.93%). Only two other groups saw an increase in attendance rates between these two years: English Language Learners (at 0.28%) and Black students (at .1%).

*Are these trends universal across school districts?*

While the trends found within the state are prevalent among school districts across Connecticut, they are not quite universal. 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. One observation I made is that while white students saw the lowest cumulative decrease in attendance rates, the only school district that saw an increase in attendance rates for white students was Old Saybrook School District which saw an increase in attendance in nearly all subgroups that they collected data on. A cursory glance of this data raises 3 questions for me.

1. Is the decrease in attendance rates across time statistically significant?
2. 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?
3. 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?

With only 3 years of data, I am skeptical that there is enough information to make any meaningful claim surrounding this first question. I would love to work towards answering questions 2 and 3 perhaps in a future class that incorporates more statistical analysis.

*How are districts that had higher than average attendance (top quartile, top 10%) in the year leading up to the pandemic faring currently?*

The top quartile of school districts consisted of 50 school districts, only about half of which were in the top quartile of attendance rates during the 2021-2022 school year. The top quartile of school districts in the 2019-2020 school year had a median and mean attendance rate ~1.5% higher than the top quartile of school districts in the 2021-2022 school year. Furthermore, the schools that made up the top quartile of attendance rates in the 2019-2020 school year saw their median attendance rate drop by about 2% and their mean attendance rate drop by about 2.5%. All of these are less than the 3.1% decrease in attendance rates throughout the state which seems to indicate that school districts in the top quartile of attendance rates in 2019-2020 saw their attendance decrease less significantly than other districts, though this claim needs more statistical evidence to support.

Of the schools in the top decile in the 2019-2020 school year, only 7 of the 21 schools remained in the top decile in the 2021-2022 school year, and only two more schools from the top decile remained in the top quartile in the 2021-2022 school year. The top decile saw a similar decrease in their attendance than the top quartile with the top decile of schools seeing their attendance rates decrease by ~1.4% while the schools that made up the top decile in the 2019-2020 saw a more severe decrease in attendance rates than the schools consisting of the top quartile from that year. The median attendance rate of schools from the top decile of the 2019-2020 school year decreased by 2.27% while the mean attendance rate decreased by 2.74%. Again, while

more in depth statistical analysis would be necessary to defend this claim, these results seem to indicate that schools in the top decile of attendance rates saw their attendance rates decrease less significantly than the average school, though not other schools within the top quartile.

*How are districts that had lower than average attendance (in the bottom quartile, bottom 10%) in the year leading up to the pandemic fairing currently?*

I next turn my attention to the districts that had below average attendance rates in the year leading up to the pandemic to see if their attendance rates were more significantly impacted than the average school district in Connecticut. It seems reasonable to predict that if school districts performing well leading up to the pandemic saw a smaller decrease in attendance rates than the average school district, then school districts that were performing poorly will see a larger decrease in attendance rates. I calculated the bottom quartile and decile for the 2019-2020 and 2021-2022 school years in order to create subsets of my dataframes to determine if this claim appears to be true.

The first thing I noticed was that while only about half of schools in the top quartile were in the top quartile both years, 36 of the 50 schools in the bottom quartile of attendance rates in the 2019-2020 school year were still in the bottom of attendance rates in the 2021-2022 school year. I then calculated the median and mean attendance rates of school districts in the bottom quartile of each school year. The bottom quartile of school districts saw a 4.2% decrease in median attendance rates and a 4.37% decrease in mean attendance rates between the 2019-2020 school year and the 2021-2022 school year. Furthermore, the schools that made up the bottom quartile of attendance rates in the 2019-2020 school year saw a 3.73% decrease in median attendance rates and a 3.9% decrease in their mean attendance rates between the 2019-2020 school year and the 2021-2022 school year.

*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%. At this point in my analysis I realized that some of my earlier claims may need to be revisited. 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. With this in mind, my plans for next steps to further answer my questions involve looking at the median and mean changes in attendance rates in order to compare districts with those numbers rather than the statewide numbers to avoid the skew of the data that these large districts are having. This also inspires me to create a scatterplot of district size and its decrease in attendance rates in order to get a clearer idea of the relation between district size and attendance rates.

The questions I am asking are important because ever since the COVID-19 pandemic districts across the country have noticed a decline in student attendance, and attendance is one of the most essential factors of student success. My goal as a data scientist is to find a job working with the state or a school district in order to identify trends and support decision making that supports student outcomes. With this particular set of questions I would hope that by finding districts who have significantly better attendance rates/changes in attendance rates than other districts, I could further analyze those schools to support statewide/district level policy that supports student attendance.