DSC 465 – Data Visualization Class Project

Too Cool for School

‘Chicago Public Schools Dataset’

Group Members

Ashitha Yalavarthi

Lena Katterman

Yueting Zhao

Daniel O’Brien,

Kushal Varma Tatampudi

Group technical report

Introduction

The way we look at public schools may forever be changed due to the unexpected impact COVID-19 has had on public education. Chicago Public Schools is no exception. Throughout the past year we have heard teachers, administrators, parents, legislators and more contribute to the conversation of the direction of public schools, so we decided to take a closer look at a few key concepts regarding Chicago Public Schools:

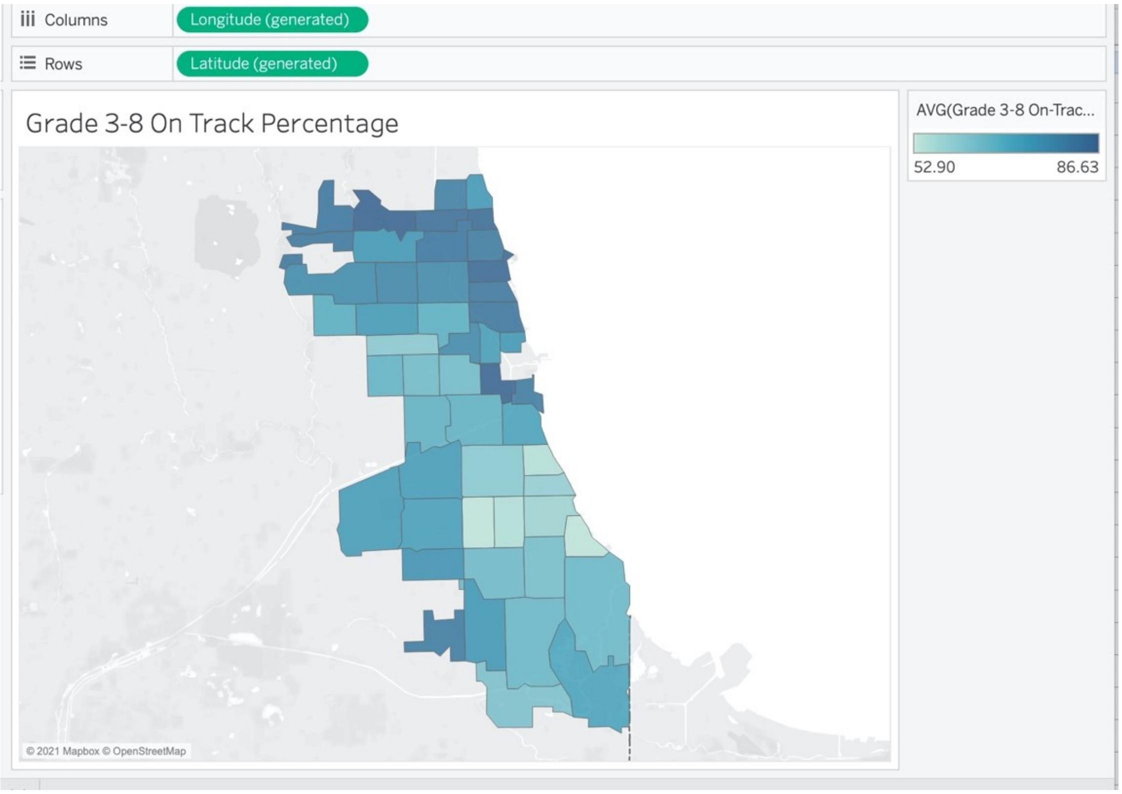
- What functions are schools serving besides education?
- What disparities exist within wealth and race?

With debates surrounding the necessity to reopen schools happening all throughout the country, the idea that schools offer more to communities that academics has been more apparent than ever. Schools serve as a meeting place for students, parents, and community members. Schools can also offer students a safe place to learn, eat, and grow. Schools throughout our country have continued to serve meals to students throughout the pandemic and will continue to do so. With all the different functions schools serve, it is only natural that we try to visualize these functions to some degree. Additionally, the conversations surrounding equity of race and socioeconomic status have continued to grow throughout 2020 and 2021. With the public consciousness shifting so much focus into equity, the ideas of inequity existing within one school district caught our attention. As it can be seen further in our analysis, the lens of racial and economic equity is placed on Chicago Public Schools to determine if and how equity impacts academics and other factors surrounding schools.

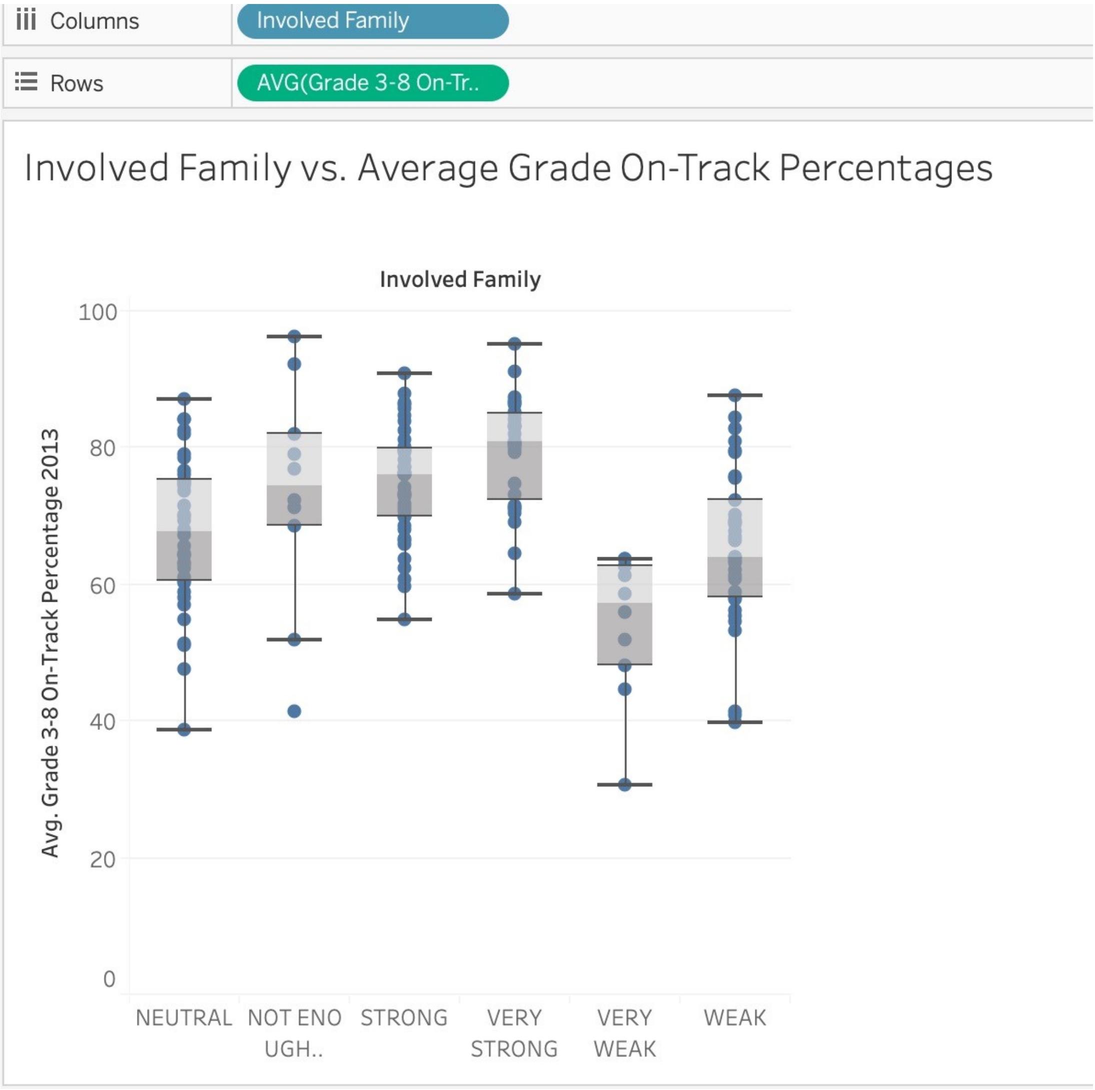
Our dataset consists of 483 rows. Each row in our dataset represents one school. The dataset contains 41 columns, containing 2 unique identifiers, 8 geographic variables including longitude, latitude, and zip code, among others. Our dataset also includes 14 categorical variables which include the results of surveys administered to staff and students, ratings, and certifications. And finally, our dataset has 17 continuous variables that include attainment and growth testing data and disciplinary and suspension data.

Exploratory Analysis and Visualizations

After finalizing a dataset and selecting the variables that we were interested in, we began our exploratory analysis. As a group, we used both R and Tableau to conduct our analysis and then model our findings. During this phase, there was not certain variables that we were invested in, so we started to look for patterns and trends among numerous variables. It was interesting because we had a good mix of categorial and geographical variables. Below is an example where we looked for patterns geographically, “On Track” is a measurement of student grades and attendance rates, this gives us a decent snapshot of attendance and grades across the city of Chicago.



In addition to exploring geographic patterns, we also explored the relationship between academic metrics and some of the other variables in our dataset. Below we can see a box and whisker plot shows the distribution of on track data with the different ordinal rating of involved families. The involved families' ratings are determined as a result of student and staff surveys and it shows that schools with strong and very strong involved families ratings have a higher on track percentage distribution. And schools with very weak involved families' ratings have a much lower distribution of on track percentages.

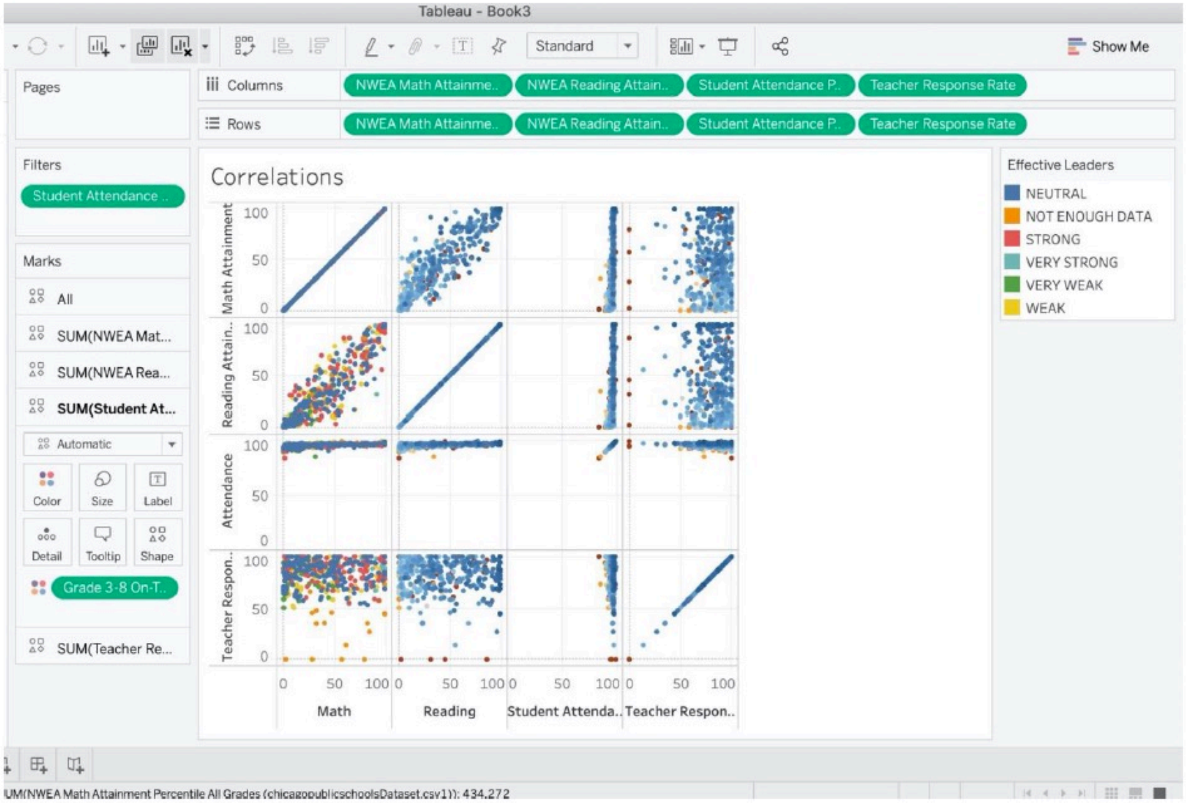


Columns Involved Family

Rows AVG(Grade 3-8 On-Tr..

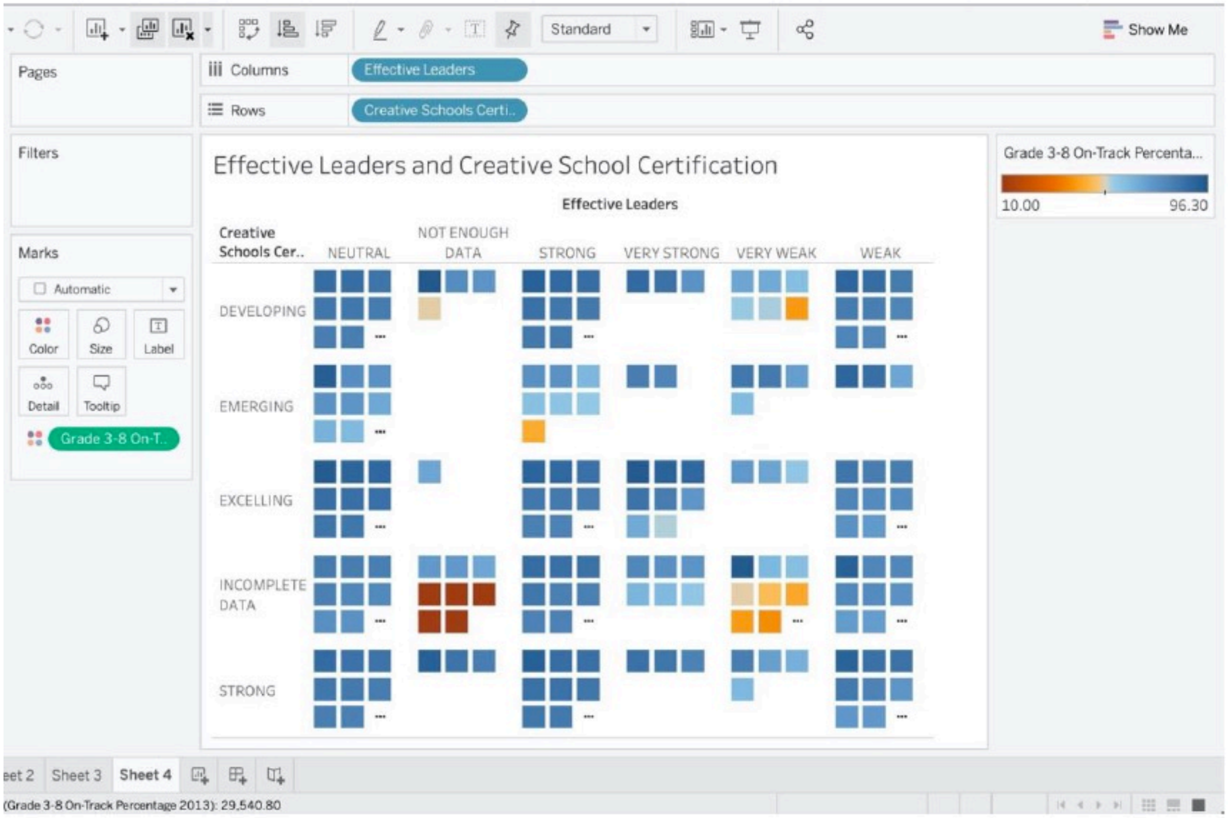
Involved Family vs. Average Grade On-Track Percentages

We also made multiple scatterplots to see if any trends or correlations immediately stood out to us. It was clear and easy to observe that math and reading abilities had a connection. Meaning, if a student performed well in math, they usually also performed well in reading. Other than that, it was hard to see any clear patterns between the other variables that are included below.



UM(NWEA Math Attainment Percentile All Grades (chicagopublicschoolsDataset.csv1)): 434,272

Next, we created a heat map in Tableau to see if there was a connection between effective leaders and the creative school certification. We were interested in this to see if proper leadership would leave to the certification. After viewing the heat map below, it is not clear if there is a relationship between the two variables. However, we did notice that the “excelling” creative school and the “very strong” leadership do have higher on track percentages.



Columns: Effective Leaders

Rows: Creative Schools Certi..

Marks: Automatic

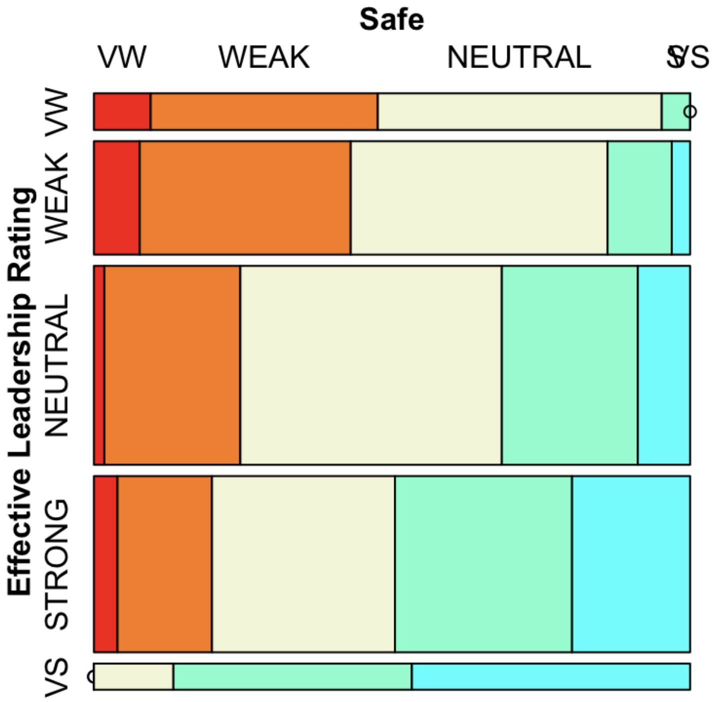
Color: Grade 3-8 On-T..

(Grade 3-8 On-Track Percentage 2013): 29,540.80

Exploratory analysis gave everyone the opportunity to survey the variables and pick what they wanted to investigate further. It was neat to see how we all took our variables in our own direction and then came back together.

Final Explanatory Visualizations

I) **Mosaic Plot – Effective Leadership and School Safety [O'Brien]**



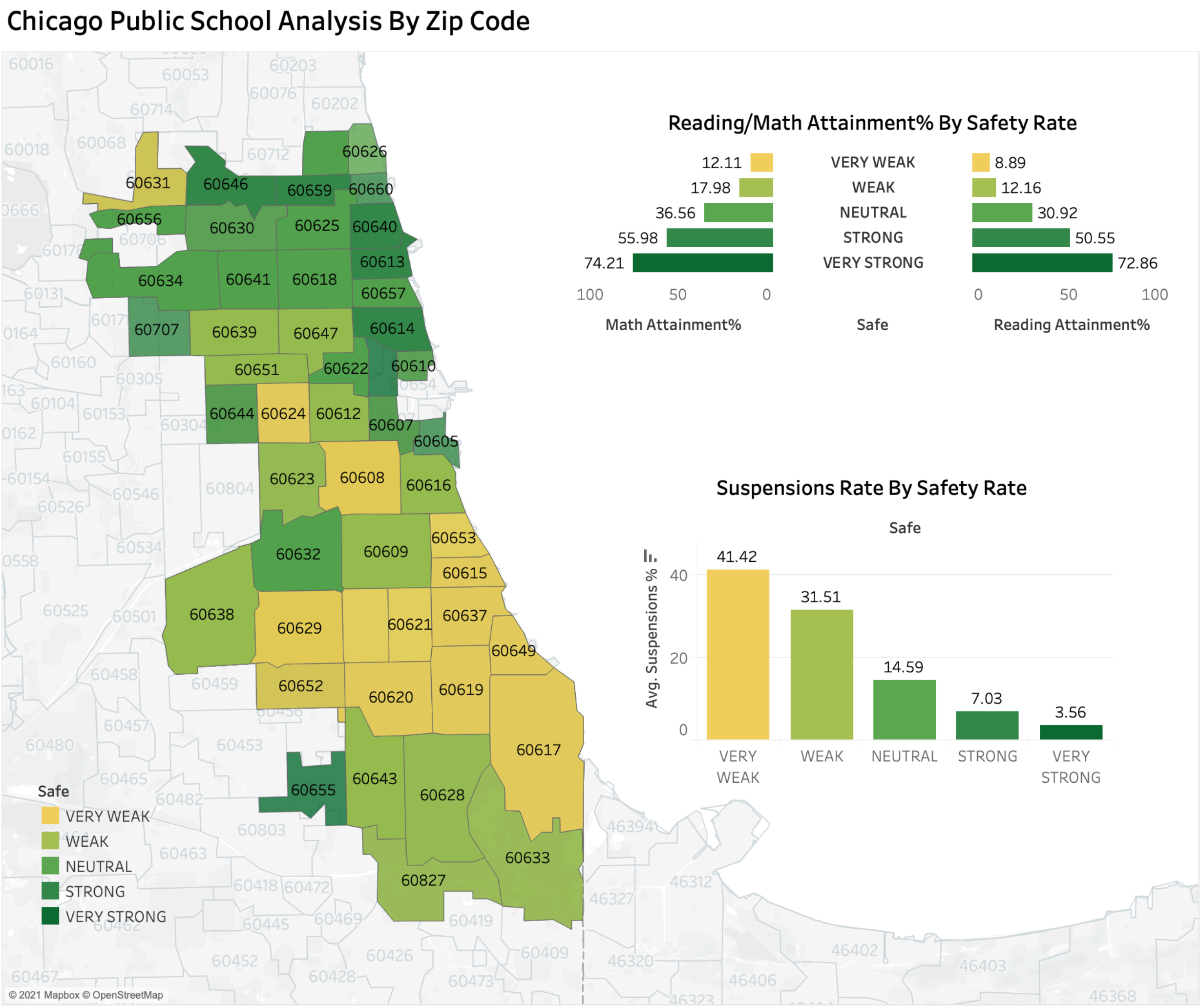
This mosaic plot that was created to compare the relationship between Effective Leadership and Safety. These ratings are assigned to schools as a result of an extensive survey given to students and staff every year. A variety of different questions are asked and depending on how they are answered collectively from the students and staff a rating is assigned to each school. The possible ratings are Very Strong, Strong, Neutral, Weak, Very Weak and Not Enough Data. The Not Enough Data ratings were removed to compare schools that submitted complete information.

I began with my first visualization with just a gray-scale color pattern. I added color to the mosaic plot to make the visualization more engaging and highlight the differences between the different ordinal values. Additionally, I did not have the ordinal values positioned correctly in my rough drafts, and so I ordered the values of the variables from very weak to very strong to show the progression of the values more clearly.

When examining the mosaic plot, the sections corresponding to a strong and very strong effective leadership have the largest proportion of strong and very strong safety ratings. This shows us that school with high effective leadership ratings are more likely to have a high safety rating than schools with neutral, weak, or very weak effective leadership ratings. Additionally, schools with weak and very weak effective leadership ratings have the largest proportions of weak and very weak safety ratings, showing that if a school has a weak or very weak effective leadership rating, they are more likely to have weak safety ratings.

II) Tableau Dashboard - SAFTY AND EDUCATION [Yueting]

Environment safety is a big concern of parents. In Chicago, students go to school based on their home location. Many parents are willing to relocate to a safer district where the schools also have a good reputation. To learn about the safe environment of the schools in Chicago, I decide to visualize the data into a map to check which schools are far away from the high crimes and which schools are near the center of the high crimes. Also, analysis how the school location effect on students' suspensions rate and attainment.



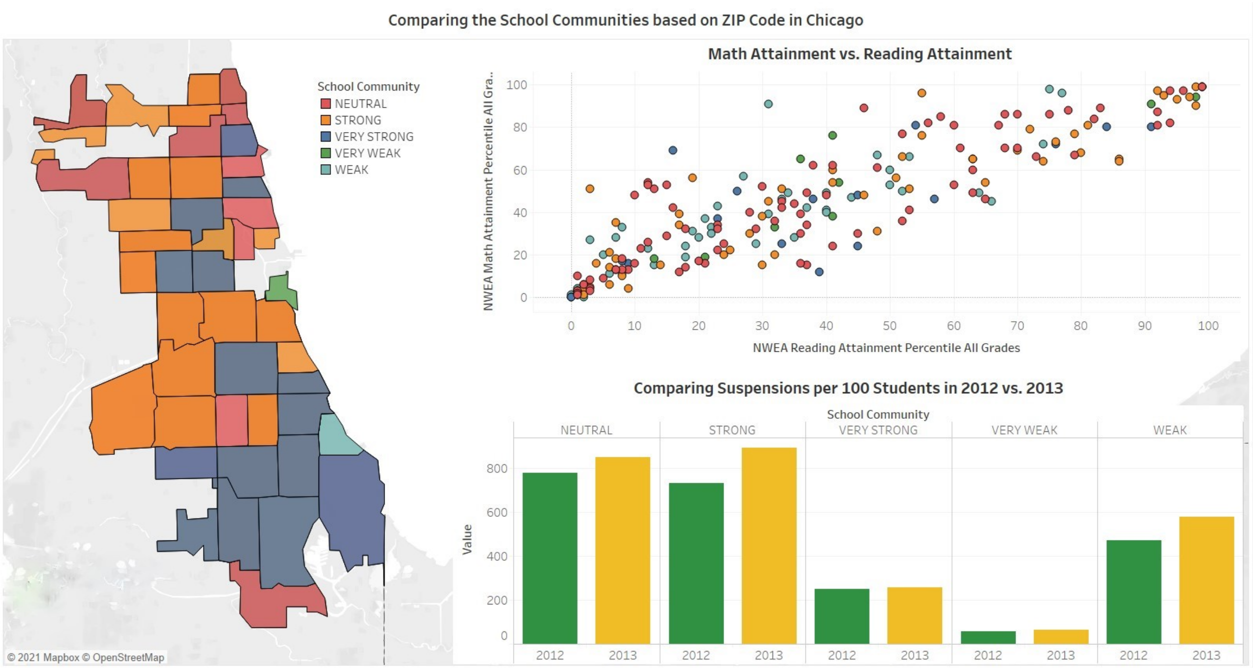
In this dashboard, I used the same color scale for all three graphs, as the Safe dimension is an ordinal variable, I order it from very weak to very strong, the lighter color indicates the weaker safety rate and the darker the stronger. The main purpose of this visualization dashboard is to present the underline relationship between area safety level and other features of education, in Chicago City area.

·The map graph by zip code on the left, shows within Chicago city area, north part of the city has the higher safety rate, while south party shows a lower rate in safety.

·The students Reading/Math attainment also shows a strong relationship with safety rate too. First there is a positive linear relationship between reading and math attainment. Second, the color of the bars shows the stronger safe area school have better student reading/math attainment. On the other hand, we can say, in the weaker safe area schools, students show worse attainment in both reading and math.

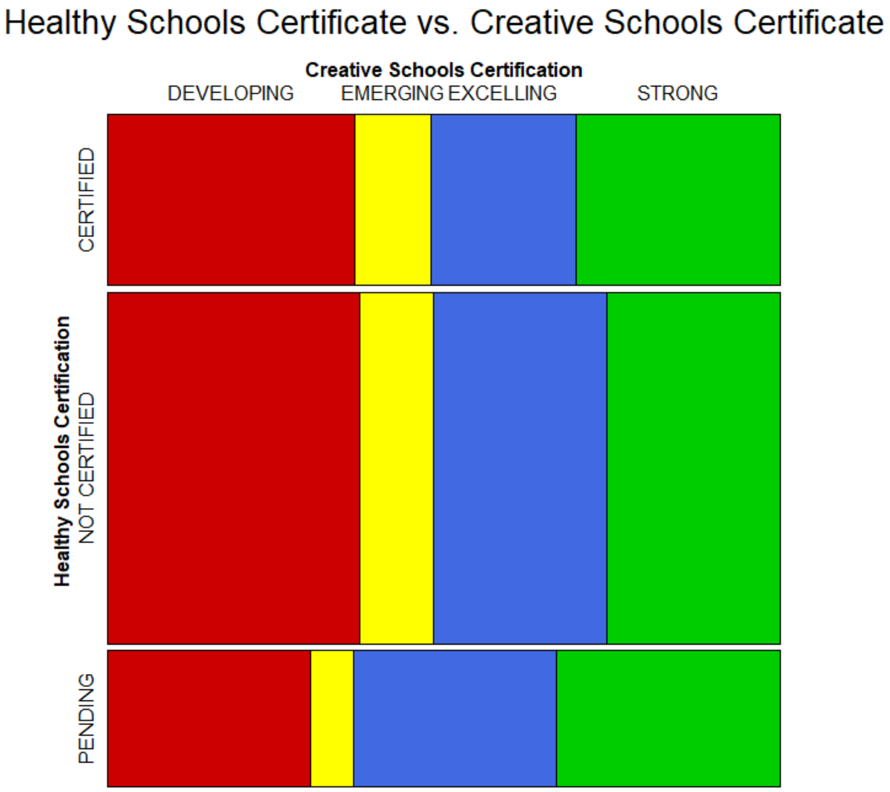
·The suspensions rate has a negative relationship with safety rate, by the bar chart, we can see the suspension rate significantly drop down when the safety rate is stronger.
In sum, from the visualized analysis above, I can conclude that districts' safety rate is an important factor for Chicago public schools. Students who study in the school located in a safer distinct might have a better Reading/Math attainment, also, the students might have lower chance from suspended from school.

III) Tableau Dashboard - Suspensions and grades of school communities based on the zip codes in Chicago [Ashitha]


Comparing the School Communities based on ZIP Code in Chicago
School Community
NEUTRAL
STRONG
VERY STRONG
VERY WEAK
WEAK
© 2021 Mapbox © OpenStreetMap
Math Attainment vs. Reading Attainment
NWEA Math Attainment Percentile All Gra..
0
20
40
60
80
100
0
10
20
30
40
50
60
70
80
90
100
NWEA Reading Attainment Percentile All Grades
Comparing Suspensions per 100 Students in 2012 vs. 2013
School Community
NEUTRAL
STRONG
VERY STRONG
VERY WEAK
WEAK
Value
0
200
400
600
800
2012
2013
2012
2013
2012
2013
2012
2013
2012
2013

This dashboard is created in Tableau, which compares the school communities based on the ZIP codes in Chicago. From the Choropleth we can see that most of the very strong school community is in the south part of Chicago and the strong community is spread over above the south part. The scatterplot in the upper right side, shows that when the NWEA Reading attainment percentile is increasing the NWEA Math attainment percentile is also increasing it shows a strong positive correlation. Most of them are from the strong and Neutral community. The Bar graph in the right bottom shows the Suspensions per 100 students in 2012 and 2013 based on the school community. In this we can see that the Strong and Neutral Community suspensions have been increased in 2013 compared to 2012. And the very weak community have low number of suspensions in both the years.

IV) Mosaic plot - Healthy Schools Certificate vs. Creative Schools Certificate [Kushal]

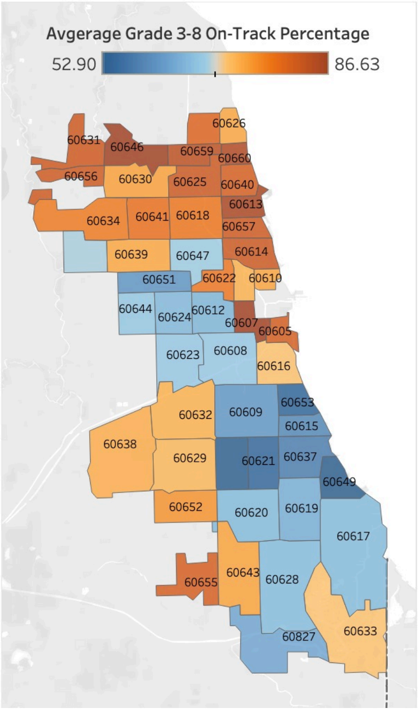


The above graph is made in RStudio with the Healthy Schools Certificate and the Creative Schools certificate variables using the ChicagoPublicSchoolsDataset. The Healthy Schools Certificate has 3 factors – Healthy Schools Certified, Not Certified and Pending Certification. The Creative Schools Certificate has 4 factors – Developing, Emerging, Excellent and Strong. From the above mosaic plot –

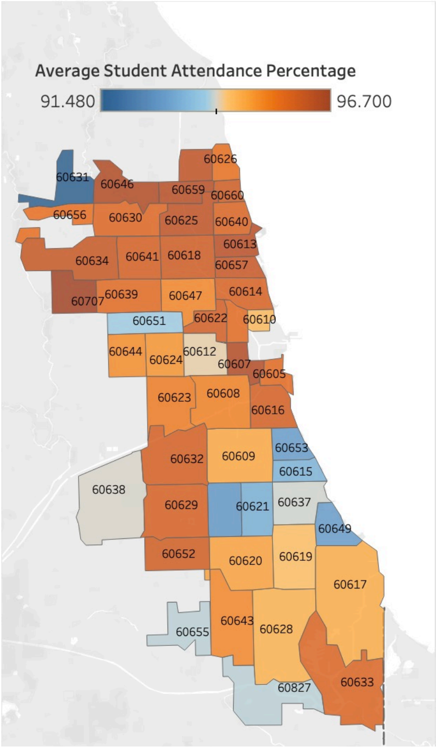
- We can see that majority of the schools (about 2/3rd) are not certified/pending certification from the Healthy Schools Certificate.
- Majority of the schools that are certified as Developing by the Creative Schools are not certified from the Healthy Schools.
- Schools that are certified as strong and Developing by the Creative schools represent most of the schools that are certified by the Healthy Schools.

V) Choropleth- Relationships between Attendance, On-Track Percentage, and Wealth [Lena Katterman]

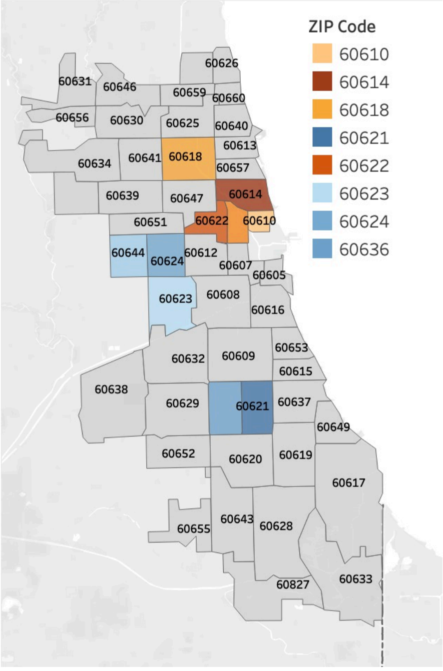
Grade 3-8 On-Track Percentage Average



Student Attendance Average



Top 5 Wealthiest and Least Wealthy
Zip Codes in Chicago, IL



After performing exploratory analysis and receiving feedback from Professor Brown, I decided to focus on the impact of wealth and student performance and attendance on Chicago Public Schools (CPS). The wealth gap continues to grow in the United States, and it is becoming very apparent in the public-school systems. I began by identifying the top 5 most wealth and top 5 least wealthy zip codes in Chicago that are in our dataset. Considering this was geographical data, I thought a choropleth would be the best technique for displaying the information I wanted to convey. On the choropleth on the right, it can be observed that the wealthiest zip codes have a lower on-track percentage, there is clear evidence that the less wealthy zip codes have lower on-track percentages as well. The average student attendance is also higher in areas of wealth, and much lower in less wealthy areas. The darker red areas have higher on track percentages and high attendance. The darker the blue, the lower the on-track percentages and attendance averages are. I used orange-blue diverging colors on the legends because it would call attention to the large differences in the data.

The top 5 wealthy zip codes perform significantly better in both areas, highlighting the problem that Chicago Public schools are facing. Wealth should not be an indicator of academic success and every child should experience comparable educations, regardless of their wealth and race. This data identifies the geographical areas that need extra help and resources in order to start performing as well as the wealthy areas. The racial wealth gap must be addressed in order to improve CPS, these wealthy areas are also the whitest areas. There have been practices and policies for many years to prevent closing this gap and that contribute to white families continuing to have an advantage.

Zip code data collected from:
https://www.zipdatamaps.com/economics/income/agi/state/poorest-zipcodes-in-illinois

d) Analysis and Discussion

We set off to uncover and analyze functions schools served outside of academics and examine inequities that exist within the Chicago Public Schools district. Throughout our analysis we discovered that many connections exist between several factors outside of academics, although at times the cause and effect of the relationship remained unclear. For example, looking at the first mosaic plot comparing effective leadership and safety ratings, we know that schools that have very strong and strong effective leadership ratings are more likely to have a very strong or strong safety rating than schools with neutral, weak, and very weak effective leadership ratings. However, this does not mean that effective leadership causes schools to be safer. It is possible that schools with higher safety ratings allow leaders to focus on areas outside of safety, such as additional student activities, professional development for teachers, and building stronger school communities. So, although we see that schools with very strong and strong effective leadership ratings have a higher proportion of strong and very strong safety ratings and schools with weak and very weak effective leadership ratings have a larger proportion of weak and very weak safety ratings the cause of this pattern remains unknown.

e) Appendix
Section 1: Personal Writeups

A. Daniel O’Brien:

After forming a group by responding to each other in the discussion thread, we agreed that we would all do a little research and find a dataset for consideration. I suggested the Chicago Public Schools dataset, and that ended up being the popular choice the group chose. Shortly thereafter I volunteered myself to serve as the group liaison to submit our work and ask for any needed clarification from the professor if necessary.

When it came to milestone 2, I volunteered to take on the description of the variables and dataset, while other group members contributed by performing an exploratory analysis and researching similar visualizations connected with the same topic and/or dataset.

Prior to the milestone 4 submission, we discussed the directions for our research and settled on the two research questions: What functions are schools serving besides education? What disparities exist within wealth and race? For the direction of my visualization, I focused on the first question regarding what schools offer other than academics. I created a mosaic plot comparing the ratings of effective leadership and school safety ratings. I started with a black and white mosaic plot comparing the different ratings of Very Strong, Strong, Neutral, Weak, Very Weak and Not Enough Data. I excluded Not Enough Data and compared the other ratings. I added color to make the differences in category more noticeable.

When it came to the group presentation, in addition to the slides with my visualization, I wrote the introduction and conclusion slides. During the presentation itself, I introduced the topic and read the introduction slide in addition to the slide including my visualization. For our final submission, I transferred and added information to the introduction section. After our presentation, I review the feedback left for us in the VoiceThread. I reordered my ordinal variable, made slight changes to the color scheme of my mosaic plot, and altered the way I described my visualization, removing all mention of correlation as suggested.

I feel that this project has helped me learn a lot about creating visualizations, but more importantly, understanding and communicating the meaning of the visualizations. I began this course with a basic knowledge of visualizations primarily used for exploratory analysis. From visual encoding and Gestalt psychology to learning about creating and explaining interactive and complex visualizations, I feel much better equipped moving forward. Having a deeper understanding of visual encoding and what types of visuals are easier to interpret, I feel that I am now able to create, present and explain visualizations proficiently and I also know the dangers of chart junk and 3D visualizations.

Overall, I am happy with my contribution and I feel grateful to be a part of a hardworking, collaborative, and dedicated group. I am thankful for the opportunity to work with Too Cool for School, and I am pleased with the final product that we have put together.

B. Yueting:

I played the team number in this project. After I joined this team, I started going through the dataset, detailed learnt each of the variables and the possible relationship among them. And checked the different data type of each variables. I did research on all the variables in this data set and record the results to prepare for the further data analysis.

When time came to Milestone 2, I contributed two parts of the milestone 2 report. The first is I did part of the work on exploratory analysis and provided three graphs. A map graph based on longitude and latitude, color filled by safe variable, zip information shown for Zip code. As result, we can see in Chicago city area, the north part has better safety rate since the color on the map show that information. And two stacked bar charts, which have the same color scale with safe variable, one for Count of Collaborative Teachers and CPS Performance Policy Level (level1 the best, level 3 the worst) as result, we can see lower CPS schools most located in weaker safe distinct, while the CPS level2 has the highest count of Collaborative Teachers. I also did another stacked bar chart, which presented the Avg. Grade 3-8 On-Track Percentage 2013 and Quality of Facilities. The same with the other graphs, the color filled with safety rate. From this visualization, I can see the safety rate does not affect the Avg. Grade 3-8 On-Track Percentage 2013very significantly. The second section which I worked on Milestone 2 was I did the research online and find the similar visualizations which related to our project topic and did analysis for each of them. As result, all the type of visualizations are used in our final presentations. I have also tried with other variables, each of them, so this step provided me a better understanding of the dataset and gave me some ideas that which direction I would like to do further analysis.

Time came to Milestone4, after team meeting, I created my first version of visualization, the topic is focus on safe and education. The reason that I decided to choose this direction is I suppose my audience is the normal parents and they plan to send their kids to Chicago public school. I want to use my data analysis skill to find the underlying relationship between area safety rate and other important education factors which parents might extremely concerned and then visualized it into a reading friendly graph to let my audience understand the result from the data analysis. As result, I created a dashboard, which included three graphs, a map by zip code, a scatter plot to present students' academic study attainment, and the last one is a bar graph which illustrated the suspension rate. The main point is I used a light yellow to dark green color scale to present the safety rate, the lighter the weaker safety the darker the strong, and applying it to all the three graphs. So, this dashboard is very user friendly, because my audience and easily use the ordinal color scale to immediately tell which area is good for their kids' education performance.

Before the presentation, I created my second version of visualization. In which I used a butterfly bar chart replaced the scatter plot. As we know, in scatter plot there are two numerical variables, and I used the safe variable to fill the color the all the plots on the plot, it showed a noticeably clear positive linear trend but each of the spot's color is not noticeably clear. And after the research I found the butterfly bar chart can satisfy my request of this set of data (two numerical and one categorical/ordinal variables), the new butterfly bar char not only presented the positive linear correlation but also showed a noticeably bright color with different safety level. After the updated, I used the second version for the presentation. And I received very good commands from the audience (classmates from other team), also some recommendation from professor. My unalterable version came out after I made the last update according to professor and other classmates' recommendations. I changed the title size and color to make them clearer and better located. Change the scale in to ordered from light to dark color. So, in this final report, it is my last version of visualization.

In sum, in my visualization, I used the dashboard to combine three different graphs into one page. Used statistical knowledge to find the linear correlation between numerical variables and make them into a better visualization rather than scatter plot only. I used the generated latitudes and attitude variables to generate the map graph and added different layers with zip code, states boundaries, landscapes. Fill color into one common variable (safe) which presents the same information on all three graphs on this dashboard. I majorly used three numerical variables, two categorical variables and two geographical variables in this dashboard. As result, I have 4 major findings which all related to Chicago city safety level of each distinct. It will provide a clear picture for my audiences, the public-school students' parents, to realize where to relocate could provide their kids a better education.

In this project, I used what I have learnt from the Data Visualization course. I will make visualization for the audience's needs. So, I supposed my audience were students' parents who are looking for public school for their kids. Second, chose the most useful variables which would highly contribute my topic. Third I used the color scale to catch my audience eyes. In this dashboard, every part with lighter yellow color means weak safety area, and they can easily find out how this weak safety level will affect other important education factors. In other words, the darker color area is where they should highly consider relocating.

If I have more time, I would love to do a more detailed analysis for science background audience. And connect all the visualizations into a story. For example, I can used box plots to analysis the outliers, and do further analysis on these outliers. Since in this dataset there are a lot of categorical variables, a mosaic plot and categorical scatterplot might be an appropriate choice for the further analysis and visualization too.

C. Ashitha

I have joined team and we decided a platform to meet that is Microsoft teams. We created a separate channel for our project. After that we had our first meeting, we started searching for datasets. Daniel brought up a dataset which was good for our project. We tried different visualizations with that dataset.

In our second meeting we discussed about the milestone 1, in which we have submitted the group members names along with the group name as 'Too cool for school' and described briefly about the dataset.

Coming to milestone 2, we have built different visualizations using the techniques learned from the tutorials. In this milestone we have described the dataset in detailed like how many rows and columns are present, type of the variable (categorical, numerical, ordinal). We have performed cursory analysis on the data including producing some exploratory visualizations and have explained about the visualizations like how we performed and what do we get to know from it.

I have contributed around three visualizations, which I have built using Tableau. In these visualizations I have used the variables School community and performance level, School community geographic map, Scatterplot with math attainment ad math growth. In this milestone we also included the visualizations done by other people in different ways using our dataset in the past by performing research online.

In the homework 4, I have performed visualization (Mosaic Plot) using School community versus the school certificates in R studio. From this we can get know which community of the school has been certified, not certified, or is pending. And the other visualization was done in Tableau using Safe and Zip code variables to build a geographic map. Which describes the safe school community in Chicago based on the zip codes.

In milestone 4 we have created drafts of the exploratory visualizations. Each of us in the group have performed each visualization using different variables in the dataset. I have performed a dashboard in Tableau which compares the school communities based on the ZIP codes in Chicago. The Choropleth shows that most of the very strong school community is in the south part of Chicago and the strong community is spread over above the south part. The scatterplot in the dashboard, shows that when the NWEA Reading attainment percentile is increasing the NWEA Math attainment percentile is also increasing it shows a strong positive correlation. Most of them are from the strong and Neutral community. The Bar graph shows the Suspensions per 100 students in 2012 and 2013 based on the school community. From the dashboard we saw that the Strong and Neutral Community suspensions have been increased in 2013 compared to 2012. And the very weak community have low number of suspensions in both the years. Coming to the final presentation, we had a meeting on Teams, where we decided to do the presentation and recorded it. We then posted the recording in voice thread. Overall, it was pleasure working with this team. From this project I have learned how to use the skills learned in the class. Using Tableau and R studio on the dataset we selected was interesting and have learned to perform visualizations. I have performed visualization keeping in mind of Data, audience, and message.

A. Lena

After completing milestone 0, which was the discussion board used to connect with classmates, I was able to quickly find a group to work with. We used Microsoft Teams to communicate with each other and then later started to utilize WhatsApp as well. After establishing our group, we all set out to find and suggest a dataset that could be used. I was looking at Kaggle and dataset on Chicago Housing prices, but it was messy and did have many variables. Daniel was able to share a dataset he found with Chicago Public School data. It had a lot of variables and I thought it was useful that it had geographical data included. In one of our first Microsoft Team's meeting, we officially picked the CPS dataset. Then Daniel submitted our survey on D2L, and we started on Milestone 2-4.

We began by dividing roles and sharing the responsibilities for completing Milestone 2. At this point, we got the chance to pick roles to split up sections A-D. Daniel manipulated our dataset to exclude some variables that we collectively decided not to use. Then we had a finalized and manipulated data set to start our exploratory analysis. We all participated in exploratory analysis to get a feel for what interested us individually. We used many different techniques for these visualizations. For this process, R and Tableau were used by different group members. I personally used Tableau for my exploratory analysis and ended up contributing 8 different visualizations.

After completing 4 and receiving feedback, we were able to focus in on a story we wanted to tell by focusing on 2 general questions. They were: What functions are schools serving besides education? What disparities exist within wealth and race?
Considering my personal interest in CPS and wealth, I thought a choropleth would be the appropriate way to display this information. Tableau is a great tool for making choropleths so that is what I used. I focused in the variables on-track percentage and student attendance. In my visualization I used the averages. It quickly became obvious that wealth had a strong correlation to higher attendance and better grade percentages.

In order to properly display this data, I went through a couple different drafts. At first, I only had 2 choropleths that were focused on wealth and on-track percentage. After discovering in exploratory analysis that there is a relationship between on-track percentages and student attendance, I thought it made sense to include a choropleth for student attendance average as well. This data can help show CPS that there are areas that require extra attention in order to be successful. Beyond the wealth data that I used, it is very apparent that the wealthier the areas are, the whiter the areas were as well. The racial wealth gap and inequality amongst race in Chicago is very problematic and leads to only certain children getting a proper education.

After we all had visualizations that we were happy with, we added them to a word document on Microsoft Teams and provided feedback for each other. Then we went on a Microsoft Teams call and Daniel was able to share his screen of our PowerPoint and record our voices. I presented the exploratory analysis section and included some of my initial visualizations. We all shared our individual visualizations and after Daniel posted it to VoiceThread, that wrapped up the presentation.

This class has been a great experience for me. I am used to taking classes and having to apply statistic to datasets, however I never had to focus on the best way to visualize the data. More often, we would just have to summarize the results and explain the meaning. This was a great opportunity to model my findings and think about all of the decisions when choosing a visualization. From color choices to technique, it was fun to create visualizations and learn Tableau/R along the way. I am glad I got to be a part of “Too Cool for School” and very thankful for each member who made this project go smoothly and were all very engaging.

B. Kushal

As a part of milestone 0, I found this group with similar interests on the discussion board. We discussed about interesting topics from each of the group member and agreed to see if we could find a dataset that could suit the class project. We choose Microsoft teams as the main platform for communication and collaboration. Daniel suggested that we could use the Chicago public schools' dataset as it has enough variables that we could investigate on. I tried to explore the different variables in the dataset to find any interesting relationships. I played the role of a team member in this group. I was team representative in a previous class project, I wanted to see how different it is to be a team player. This was a nice experience.

As a part of milestone 1, we decided to name our group as 'Too cool for School' as suggested by Daniel along with the dataset we will use for the project and a brief description. In our second group meeting, we discussed about the milestone 2 and divided the work accordingly. In our meeting we agreed to exclude some variables that we will not be using in the future. We finalized the dataset with the most important variables and compiled the milestone 2 with dataset description and included our initial exploratory visualizations we got from R and Tableau. Each of the team member brought all kinds of interesting visualizations – bar graphs, scatterplots, geographic visualizations. We found some interesting that we could explore more in the next milestones.

Following the feedback on our milestone 2, we discussed the directions that we wanted to take for the milestone 4. We agreed to stick to the two research questions: "What functions are schools serving besides education?" and "What disparities exist within wealth and race?." As most of the team members focused on the Safety, Attendance-Suspension rates, leadership, and Scores. I focused on the two certifications – Healthy schools' certificate and Creative schools' certificate. I decided to do a mosaic plot using these two variables as these were categorical variables. The Healthy Schools certificate has three factors – Healthy schools certified, Not Certified and Pending Certification. The Creative Schools Certificate has 4 factors – Developing, Emerging, Excellent and Strong. I made basic mosaic plot with greyscale for the milestone 4. From the mosaic plot, we were able to know that majority of the schools are not certified or pending certifications from the Healthy schools' certificate. And most of the schools that are certified by the Healthy school's certificate are either certified as strong or developing by the Creative schools' certificate.

For the group presentation, we compiled the visualization made by each of us along with introduction about the dataset and our directions with the summary of our findings. During the presentation, I spoke about the mosaic plot about the two certifications and the summary of our findings in the VoiceThread. The presentation recording was quick and easy. We all joined the meeting in Microsoft teams and Daniel took control of changing the slides and recording the meeting. We all were happy with the first recording and decided to post it on the VoiceThread. It took us some time to figure how to do it, but we were able to submit our presentation before it was due. For the final report, I changed the axis of the variables and added color scale to the mosaic plot as per the feedback from the team members and professor.

This project helped me to learn a lot about creating visualizations in Tableau and R studio. I began this course with basic skills of data analysis. I was able to learn how to create visualizations in a way that they communicate the message to its audience. Having a better understanding of the visual encodings and color schemes, and working with Table and different packages in R. I feel I can use the skills that I learnt in this course in my future career. I am glad that I got the chance to work in this group, and I thank each of the group member that made this project a success.

Section 2: Code for R Visualizations

A. R code for mosaic plot- Effective leadership vs Safety and drafts. [O'Brien]

```
chi <- chicagopublicschoolsDataset

library(plyr)

head(chi)

names(chi)

count(chi, 'ZIP.Code')

count(chi, 'Wards')

count(chi, 'CPS.Performance.Policy.Level')

count(chi, 'CPS.Performance.Policy.Status')

count(chi, 'Safe')

count(chi, 'Effective.Leaders')

chiTable <- matrix(chi=c())

mosaicplot(chi, main = 'Safe Schools Vs Policy Level',
         color = TRUE)
mosaicplot(~Safe + CPS.Performance.Policy.Level, data = chi,
         color = TRUE)

chi1 <- chi[!(chi$Safe=='NOT ENOUGH DATA'),]

count(chi1, 'Safe')

chi1 <- chi1[!(chi1$CPS.Performance.Policy.Level=='NOT ENOUGH DATA'),]

chi1 <- chi1[!(chi1$Effective.Leaders=='NOT ENOUGH DATA'),]

chi1 <- chi1[!(chi1$School.Community=='NOT ENOUGH DATA'),]

 head(chi1)

count(chi1, 'CPS.Performance.Policy.Level')

chi1$Safe <- as.character(chi1$Safe)
```

```
chi1$Safe[chi1$Safe == 'VERY STRONG'] <- 'VS'

chi1$Safe <- as.character(chi1$Safe)
chi1$Safe[chi1$Safe == 'VERY WEAK'] <- 'VW'

chi1$Safe <- as.character(chi1$Safe)
chi1$Safe[chi1$Safe == 'STRONG'] <- 'S'

chi1$Effective.Leaders <- as.character(chi1$Effective.Leaders)
chi1$Effective.Leaders[chi1$Effective.Leaders == 'VERY STRONG'] <- 'VS'

chi1$Effective.Leaders <- as.character(chi1$Effective.Leaders)
chi1$Effective.Leaders[chi1$Effective.Leaders == 'VERY WEAK'] <- 'VW'

count(chi1, 'Safe')

mosaicplot(chi1, main = 'Safe Schools Vs Policy Level',
          color = TRUE)
mosaicplot(~Safe +CPS.Performance.Policy.Level, data = chi1,
          main = 'Safety and Performance Level',
          xlab = "Safety",
          ylab = "Performance Level",
          color = TRUE)

library(ggplot2)install.packages("devtools")
devtools::install_github("haleyjeppson/ggmosaic")

library(ggmosaic)
mosaic_examp <- ggplot(data = chi1) +
  geom_mosaic(aes(x = product(Safe), fill = Effective.Leaders)) +
  labs(y="Safety", x="Effective Leaders", title = "Safety of Schools and Effective
Leadership")

mosaic_examp

chi1$Effective.Leaders <- factor(chi1$Effective.Leaders, levels = c('VW', 'WEAK',
'NEUTRAL', 'STRONG', 'VS'))
chi1$Safe <- factor(chi1$Safe, levels = c('VW', 'WEAK', 'NEUTRAL', 'S', 'VS'))

cTable = table(chi1$Effective.Leaders, chi1$Safe)
cTable

library(expss)
chi1 = apply_labels(chi1,
                   Effective.Leaders = "Effective Leadership Rating",
                   Safe = "Safety Rating")
```

```
library(vcd)

mosaicPlot = mosaic( ~ Effective.Leaders + Safe, data = chi1,
            highlighting = 'Safe', highlighting_fill = c('red', 'darkorange', 'beige',
'aquamarine', 'cyan'),
            labeling_args = list(set_varnames = c(Effective.Leaders = 'Effective
Leadership Rating')
            ))
mosaicPlot
```

B. R code for mosaic plot - Healthy Schools Certificate vs Creative Schools certificate [Kushal]

```
library(plyr)

head(chicagopublicschoolsDataset)

names(chicagopublicschoolsDataset)

count(chicagopublicschoolsDataset, "Healthy.Schools.Certification")

count(chicagopublicschoolsDataset, "Creative.Schools.Certification")

CPS <- chicagopublicschoolsDataset

CPS <- CPS[!(CPS$Creative.Schools.Certification=='INCOMPLETE DATA'),]

count(CPS, "Creative.Schools.Certification")

count(CPS, "Healthy.Schools.Certification")

CPS$Healthy.Schools.Certification<- as.character(CPS$Healthy.Schools.Certification)

CPS$Healthy.Schools.Certification[CPS$Healthy.Schools.Certification == 'HEALTHY
SCHOOLS CERTIFIED'] <- 'CERTIFIED'

CPS$Healthy.Schools.Certification[CPS$Healthy.Schools.Certification == 'PENDING
CERTIFICATION'] <- 'PENDING'

count(CPS, "Creative.Schools.Certification")

count(CPS, "Healthy.Schools.Certification")

library(ggplot2)

library(ggmosaic)
```

```
Table <- table(CPS$Healthy.Schools.Certification, CPS$Creative.Schools.Certification)

Table

HealthySchoolsCertificatevsCreativeSchoolsCertificate <-
table(CPS$Creative.Schools.Certification, CPS$Healthy.Schools.Certification)

mosaicplot(HealthySchoolsCertificatevsCreativeSchoolsCertificate , main = "Healthy
Schools Certificate vs. Creative Schools Certificate", xlab = "Creative Schools
Certificate", ylab = "Healthy Schools Certificate", color = TRUE)

library(vcd)

mosaicPlot <- mosaic( ~ Healthy.Schools.Certification + Creative.Schools.Certification,
data = CPS, main = "Healthy Schools Certificate vs. Creative Schools Certificate",
highlighting = 'Creative.Schools.Certification', highlighting_fill = c('red3', 'Yellow1',
'royalblue', 'green3'), labeling_args = list(set_varnames = c(Creative.Schools.Certification
= 'Creative Schools Certification', Healthy.Schools.Certification = 'Healthy Schools
Certification')))

mosaicPlot
```