Process Book (Final Version)
Prepared for CPSC 4030
Visualizations Dashboard Project

Created by:
Shail Patel
Joseph Geter
Evan Gaito

Overview and Motivation

Our group selected a dataset from the Stanford Enrollment Project, which began collecting data on enrollment in schools in the United States during the COVID-19 pandemic. The project aimed to track how enrollment may have been affected by the pandemic and shed light on potential implications and consequences for public schools due to enrollment changes.

The Stanford Enrollment Project recognized that school enrollment data was not tracked nationally, and different states or districts within states tracked their enrollment data differently. Therefore, the project not only worked to collect the data but to also standardize it into a common schema that could apply to all schools in the country.

Generally, our initial motivation as a project lined up with those of the Stanford Enrollment Project. That is, we wanted to look at how enrollment changed over time especially during 2020 and 2021 when COVID-19 made its impact felt. As you'll read in the following sections, that motivation did evolve over time, mostly for practical reasons.

Related Work

N/A

Questions

After initially considering our dataset, our first question was:

- How has COVID affected student school enrollments?

After discussion with Professor Iuricich following Checkpoint 2, we recognized that our first sets of visualizations were disjointed (more on this in Design Evolution). Essentially, our initial question was limiting and could only be answered by potentially over-summarized data, which would limit interaction opportunities. We decided that the demographic data may be more interesting for a visualization dashboard. That led us to focus our questions on Black and White students in the Washington state student body:

- How does the population of black students compare to the population of white students for each school?
- How does the ratio of black students to white students change across grade levels?
- Is the black student population higher in certain counties in Washington state?

Data

The Stanford Enrollment Project gathered data from schools across the entire United States. It contained records across all states, counties, school districts, down to the grade level at each school. This presented the group with an up-front decision on which data to select from the vast amount that had been collected. We recognized almost immediately that a nation-wide dataset would be too massive for our effort so we looked for a single state to focus on.

Choosing the right state was made somewhat easier for us because not all states contained the same amount of data. The project gathered enrollment data from 2015-2021 but not all states had data for all those years. Additionally, some states did not have as much demographic data on the students. Since our original motivation was to show enrollment changes due to COVID-19, we wanted to make sure we selected a state with all 6 years of data. The idea there was that we needed to establish a trend *before* COVID-19 in order to see how enrollment really changed. If we selected a state with only the years 2019, 2020, and 2021 then there would be less certainty on the true effects of COVID-19 on enrollment.

We ended up selecting Washington state because Washington had data for the maximum possible number of years. In addition, Washington had values for the most number of demographic features in the dataset.

Our dataset contains data about student enrollment in the state of Washington specifically. This data is spread across a large list of counties, with extra specifiers for grade level, race, and gender.

Exploratory Data Analysis

The first thing we looked at after selecting the Washington state dataset was the overwhelming number of records. In total, there were 140,371 records across the various counties, school districts, schools, and grades in Washington state. When we discussed dataset selection in class before the actual selection, we talked about an ideal size of somewhere in the 1000-5000 record range. At first, we thought perhaps we could limit the dataset to a single county within Washington. The major downside to that choice, though, is that we would not have been able to plot a county-level geographic map visualization.

A second option, and one we actually chose to move forward with during our first Checkpoint, was to reduce the size of the dataset by focusing our visualizations on a single grade level. After discussing our first Checkpoint, however, we decided not to reduce the number of records in our dataset simply to reduce the total size.

After digging further into the dataset, we did recognize an opportunity to reduce its size by removing the records with a “pk_12_total” grade value. We found that these records were actually summary records, where the data from all grades PK-12 for a given school were combined. We needed to drop these records from our dataset otherwise we would have shown significantly skewed data points that made all the individual grade-level records look “smaller” in comparison. If we did want to summarize data across all grade levels in a school, we knew we could do that ourselves by combining the individual records.

Next we looked into the attributes of the dataset. As mentioned in the previous section, we did know ahead of time that the Washington state dataset had a relatively low number of attributes with nothing but null values. Still, our first step was to identify specifically which attributes were useful to us, and which were not. After we had ignored our null-only attributes, we categorized the remaining ones as follows:

- General Information
    - State
    - County
    - Admin Level (District vs. School)
- District
    - District NCES ID - A national ID
    - District State ID - A state-level ID
    - District Name
    - CCD District Type - Numerical value corresponding to an NCES district category
    - CCD Charter - Charter Status
    - District Level - Elementary, Middle, High, etc. Mostly blank or “Other” in Washington
    - District Low Grade - The lowest grade level offered in the district
    - District High Grade - The highest grade level offered in the district
- School

    - School NCES ID - A national ID
    - School State ID - A state-level ID
    - School Name
    - CCD School Type - Categorization of the school
    - CCD Charter - Yes/No value
    - School Level - Numerical value corresponding to NCES category
    - School Low Grade - The lowest grade level offered by the school
    - School High Grade - The highest grade level offered by the school
- Enrollment
    - Year
    - Grade - Pre-K through 12
    - Total - total count of students
- Demographics (these are all counts of students)
    - White
    - Black
    - Hispanic
    - Native American/Alaskan Native
    - Asian
    - Hawaiian/Pacific Islander
    - Multiracial
    - Male
    - Female
    - Non-binary
    - English Language Learner - a student in the process of learning English
    - Homeless
    - Low Income
    - Section 504 - Students protected under US Code Section 504, related to disabilities

Design Evolution

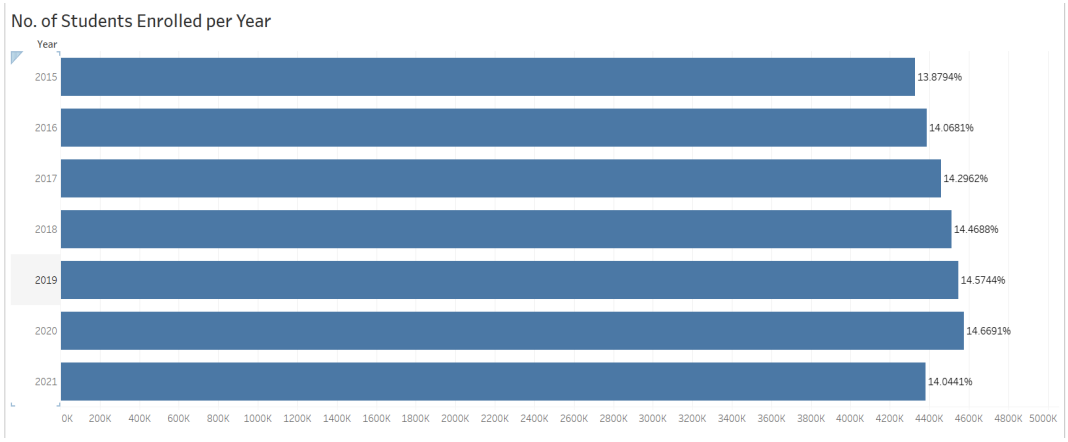
Our design changed significantly over the course of our project work, as we learned not to just create a collection of unconnected visualizations, but rather a group of visualizations that could build on one another to show interesting things.

For our first attempt at defining our visualizations, each group member worked on one or two visualizations independently. We then each worked on an alternative way to visualize the same information contained in each. This approach led to three visualizations and three alternative visualizations that were only loosely connected, following the theme of “student enrollments in the wake of COVID.”

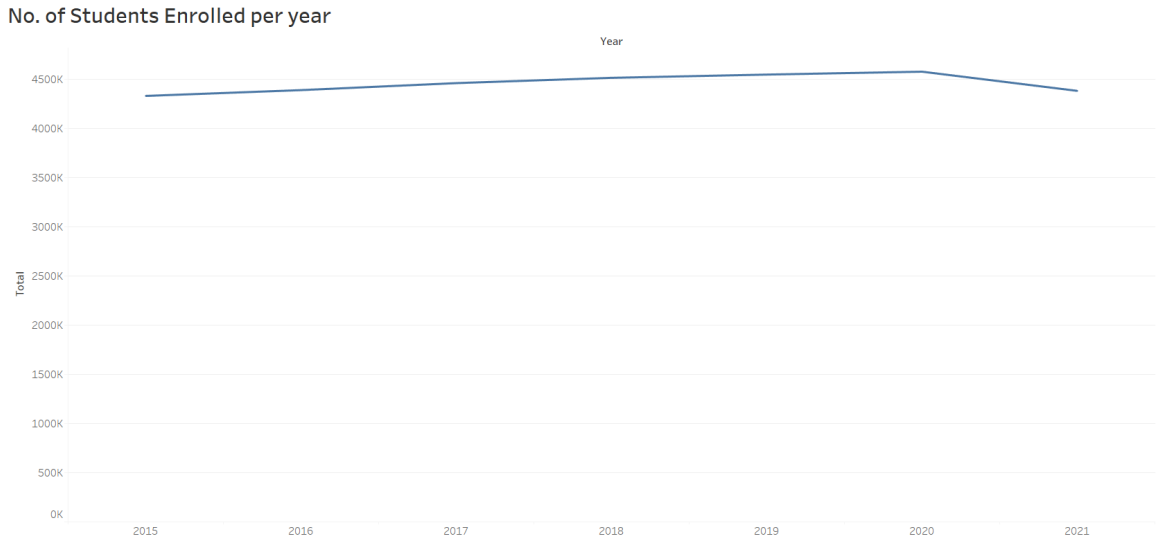
Visualization 1

The first visualization simply shows the total enrollment across all of Washington state per year. This data was visualized as a bar chart, and alternatively as a line chart.

Initial Visualization 1



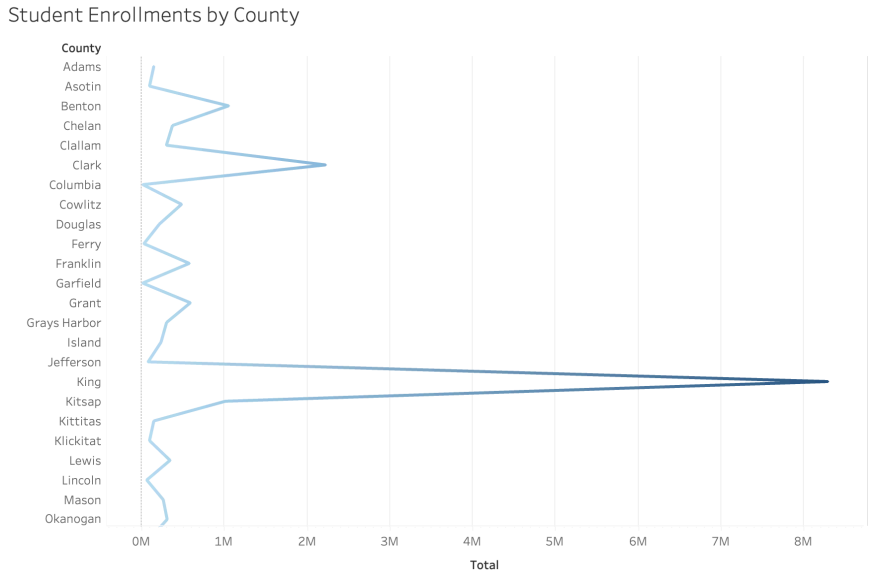
Alternate Visualization 1



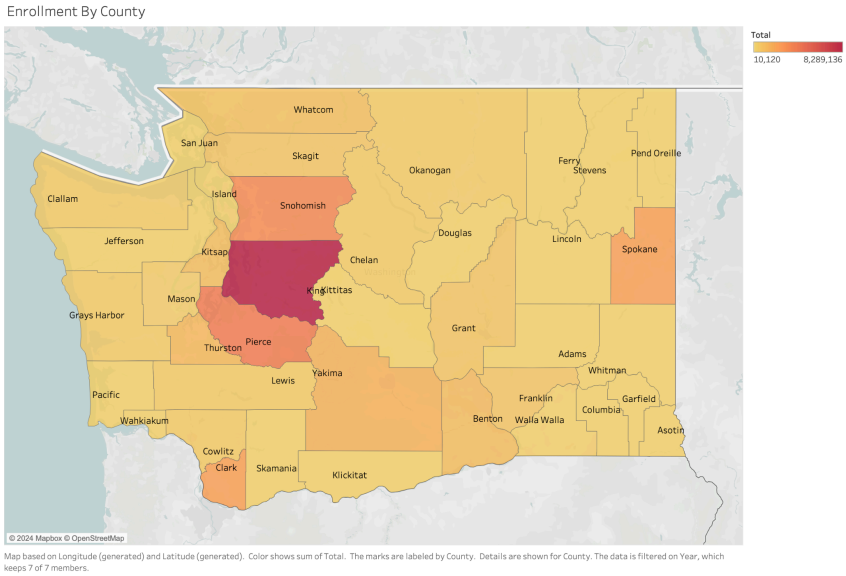
Visualization 2

Visualization 2 showed total school enrollment by county in Washington state. One option for this visualization used a line chart and the other used a map.

Initial Visualization 2



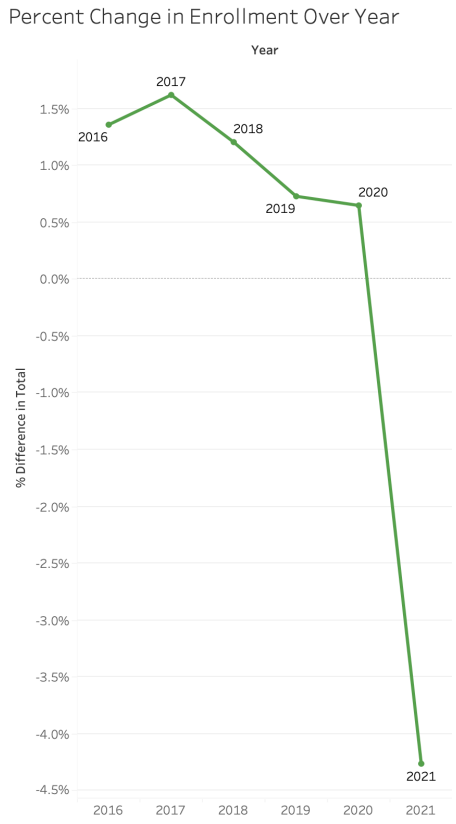
Alternative Visualization 2



Visualization 3

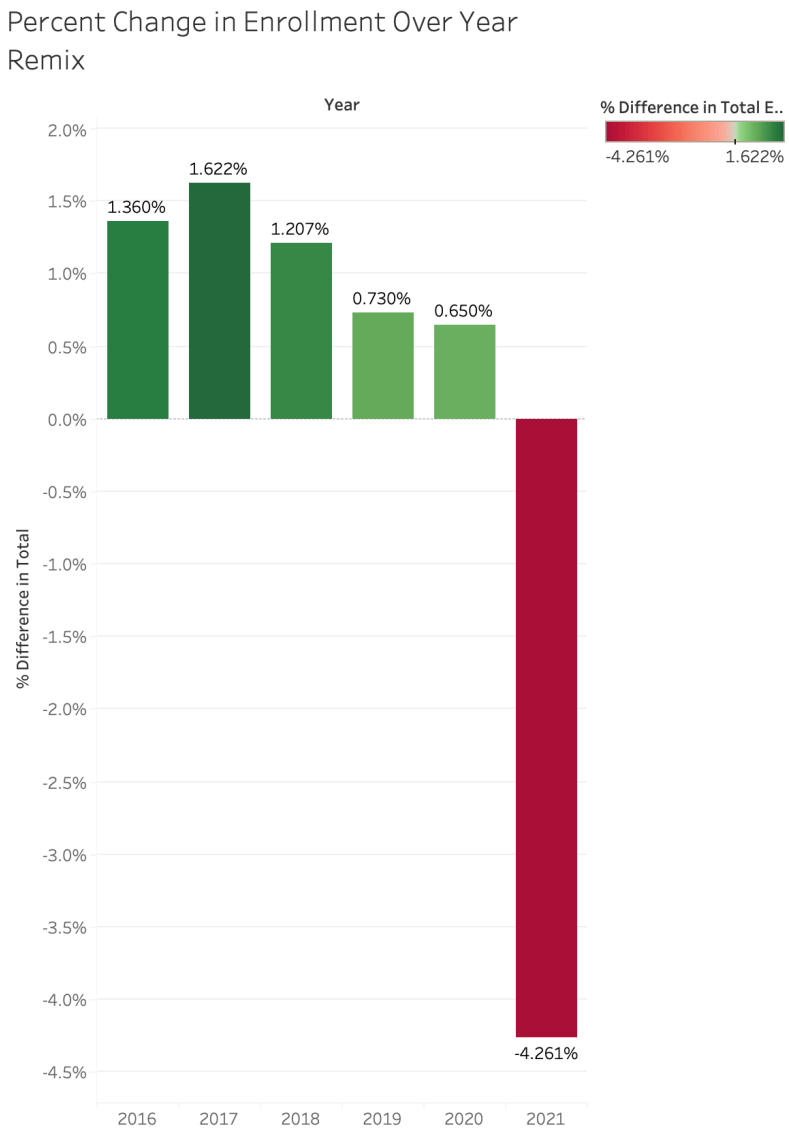
Visualization 3 was meant to directly address the question of how enrollment changed over time by showing a percentage change instead of raw enrollment numbers per year. The initial visualization was a line chart while the alternative was a bar chart with positive and negative values.

Initial Visualization 3



The trend of % Difference in Total for Year. The marks are labeled by Year. The view is filtered on % Difference in Total, which keeps non-Null values only.

Alternative Visualization 3



Focused Visualizations

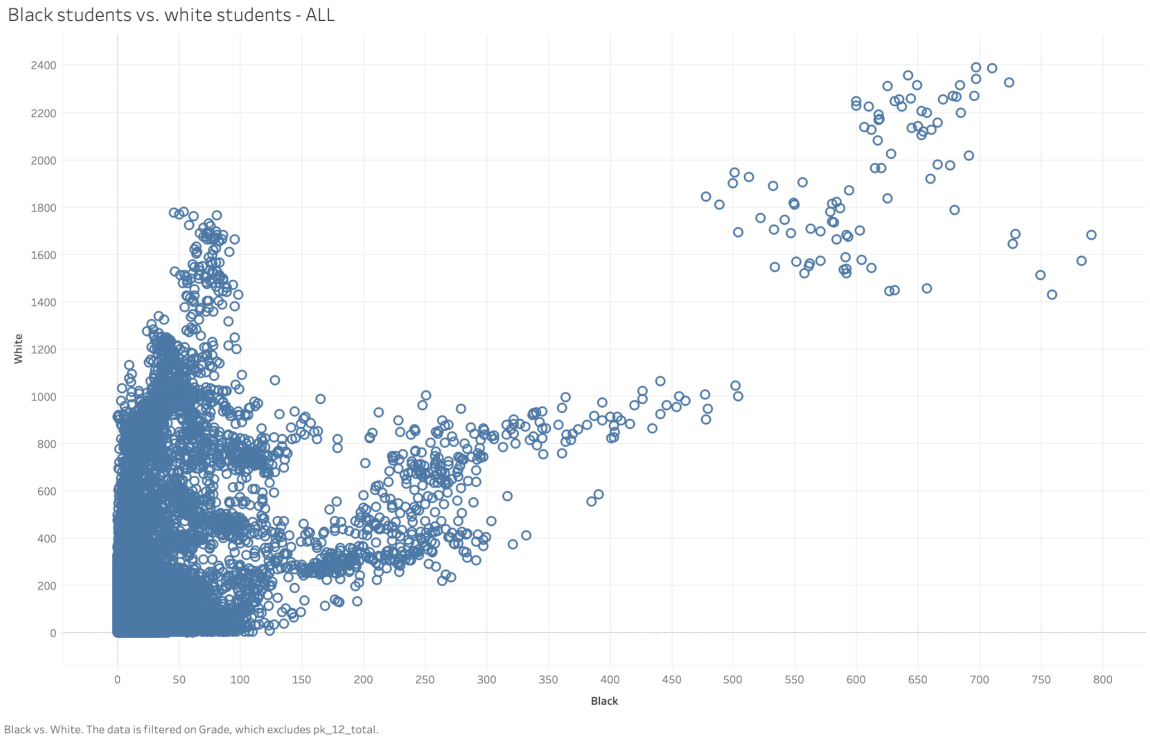
Naturally, the lack of a specific focus point made the visualizations look similar to each other. Visualizations 1 and 3 were effectively telling the same story, just with different values. Additionally, all three visualizations required a significant amount of summarizing. That is, the number of data points shown in each was minimal. Both visualizations 1 and 3 summarized enrollment numbers across all schools, in all counties, to a single number per year. Visualization 2 contains more data points but doesn't meaningfully connect to the other two.

Our design began to evolve after we discussed a more concrete theme to focus on, deciding on Black and White enrollment, which allowed us to show more data points in our visualizations and think of various ways to visualize the data while still centralizing their “point” on Black and White enrollment.

We decided to essentially redo Checkpoint 2 and make new visualizations in Tableau that fit this new theme.

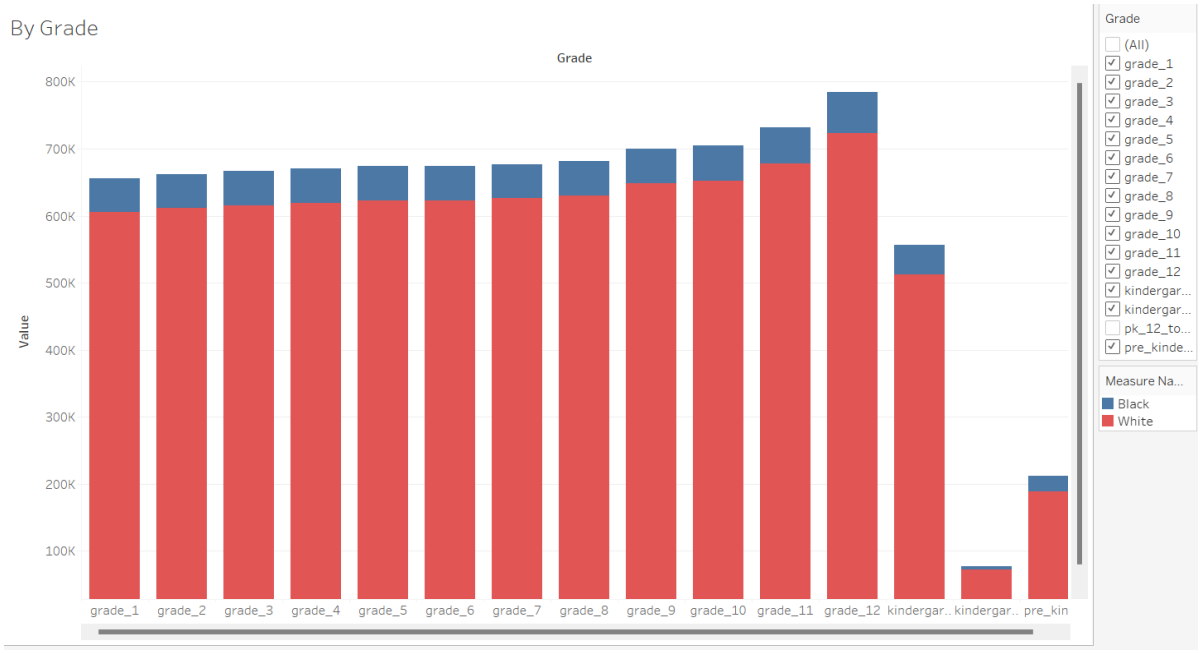
Focused Visualization 1

This visualization shows a mark for every grade in every school in Washington state and plots the number of Black students and the number of White students for each. This visualization lends itself to further filtering.



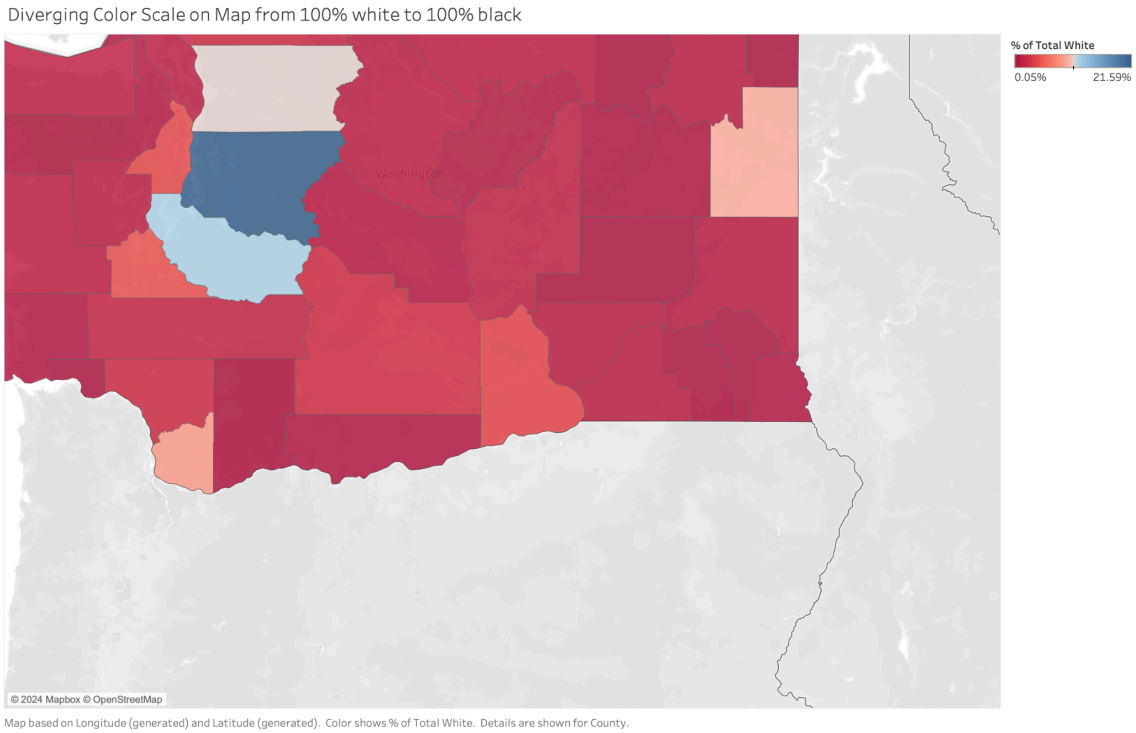
Focused Visualization 2

This visualization shows the total number of Black students and the total number of White students for each grade level across the entire state of Washington.



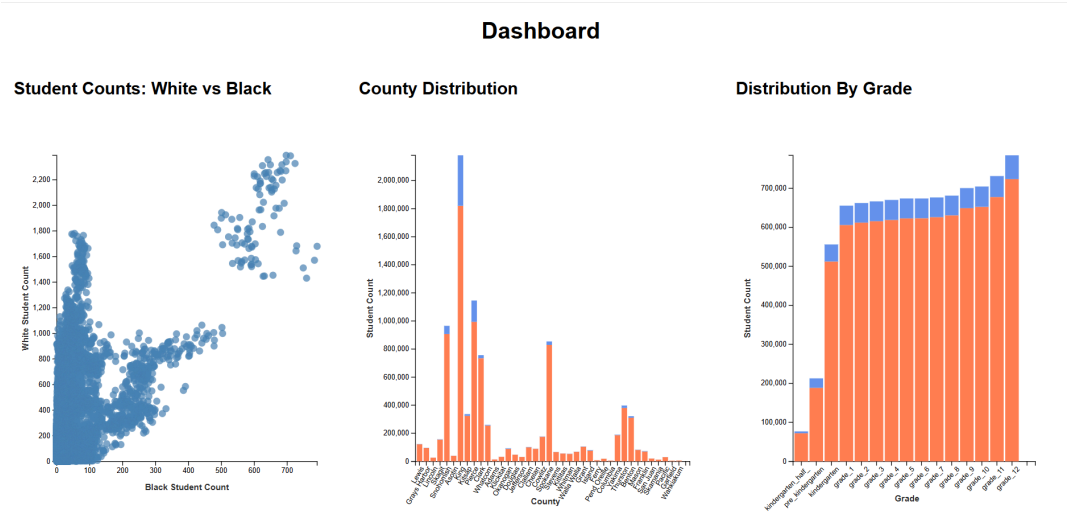
Focused Visualization 3

This visualization shows a diverging scale where a higher percentage of White students is blue and a lower percentage of White students is red. It needs some more refinement for the final project as it is not very clear how to interpret the color scale.



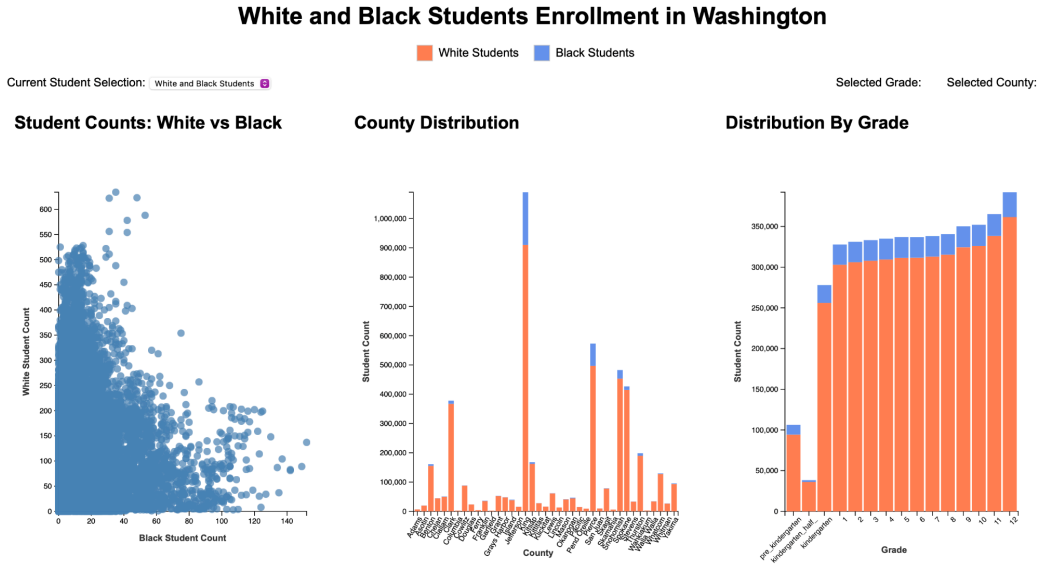
Dashboard Prototype

Lastly, after considering the best ways to show our visualizations with the d3 Javascript plugin, we produced our first dashboard prototype in HTML/JS:



Further Evolution

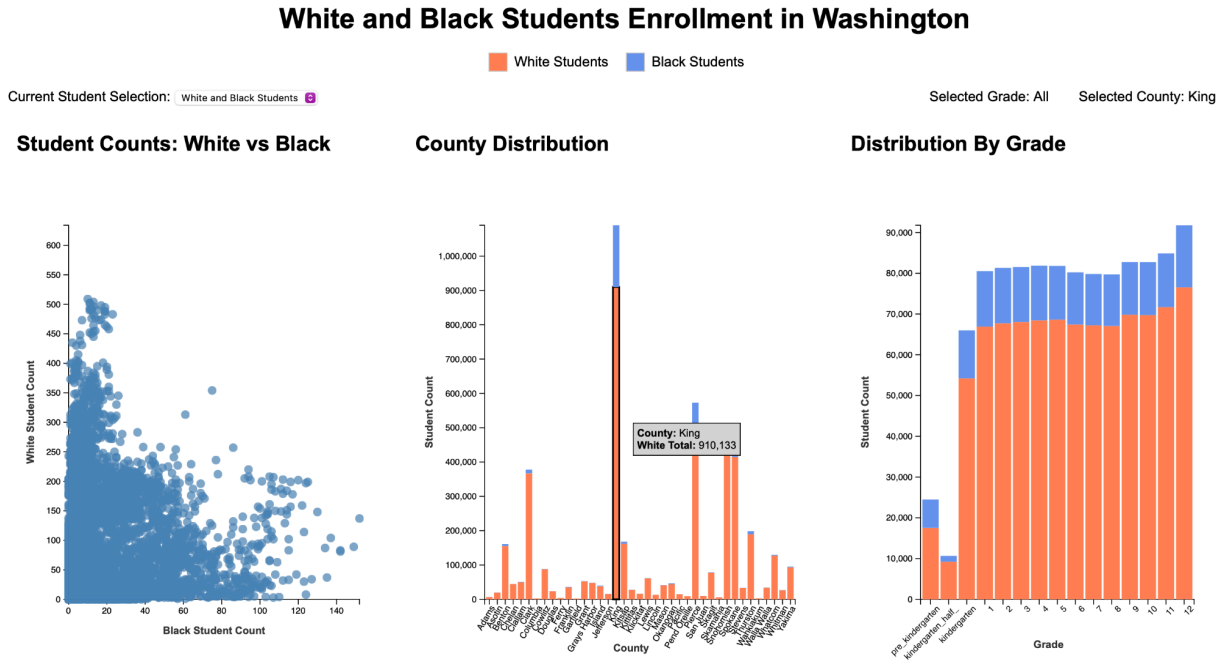
Our dashboard design evolved following various group meetings, as well as the final in-class presentation. To start with, our prototype became more interactive once we established how it would be formatted on the HTML webpage.



We added a key that shows the colors for white and black students, as well as a dropdown menu that allows the user to separate the graphs by race. The graphs can also change when hovering over a specific bar or dot. However, there were also certain functionalities that we

couldn't immediately implement and bugs we were unable to fix; this initial dashboard was shown during the final in-class presentation.

We met up for one final group meeting following the presentation, where we made our final changes to the design. Functionality for showing data when hovering over a bar or dot was added, along with bug fixes for the scatterplot, which experienced issues in our initial dashboard.

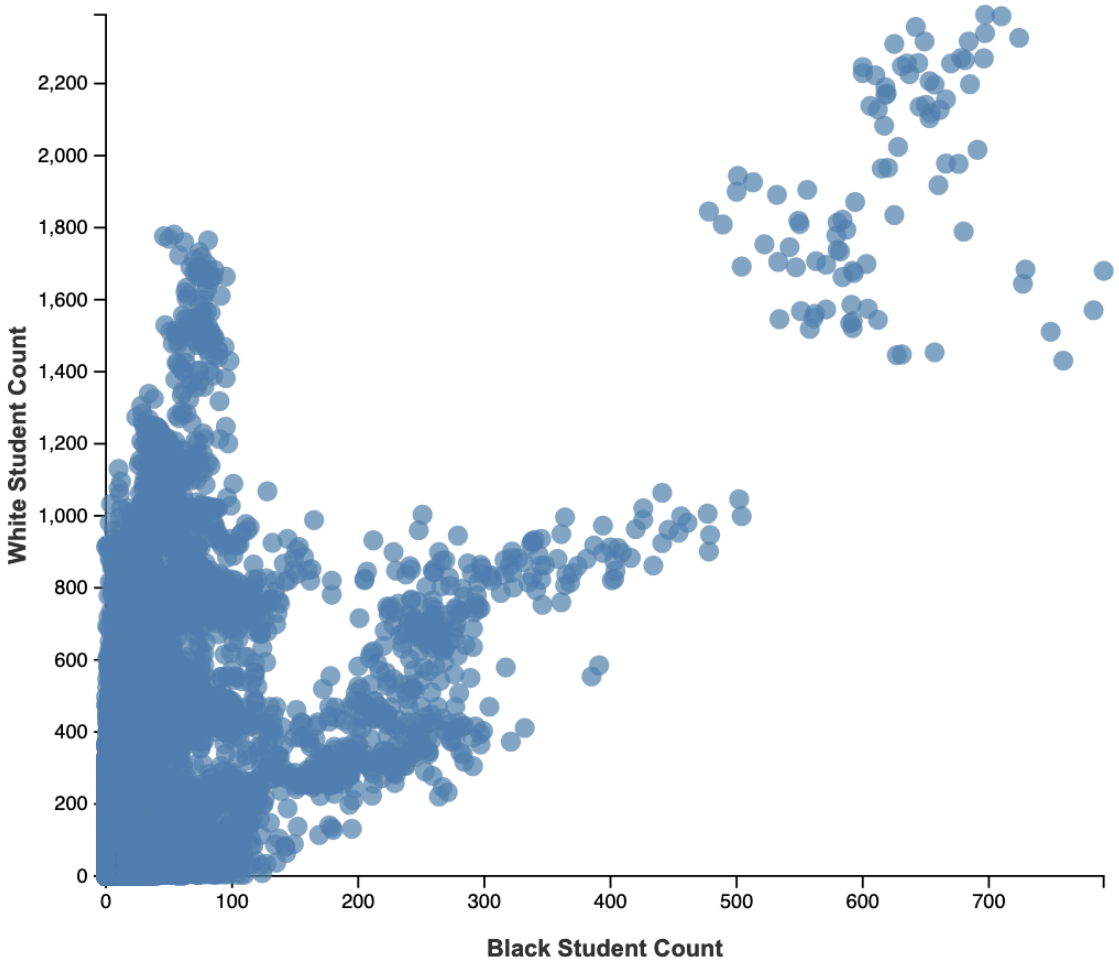


(This is an example of hovering over a bar and showing the data accordingly.)

Implementation

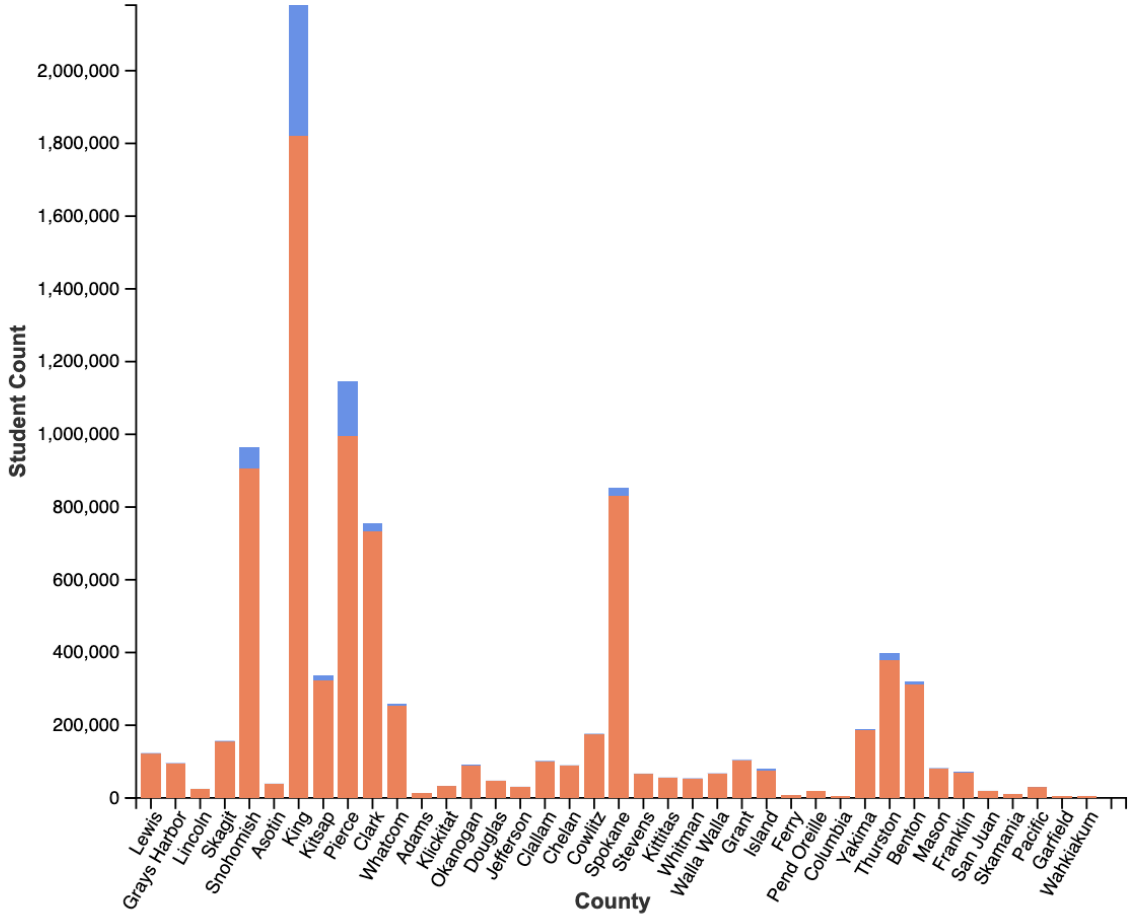
Our first prototype of the dashboard lines up with our second round of focused visualization proposals. Our second prototype of the dashboard lines up with the focused visualization proposals and contains more interaction for further exploration of the data. For implementation, we further trimmed the data set to just include the required columns. As mentioned in previous sections, we also excluded the data from the "pk_12_total" summary grade so that we were not showing misleading data points in the visualizations. The first visualization in the dashboard matches our first focused visualization:

Student Counts: White vs Black

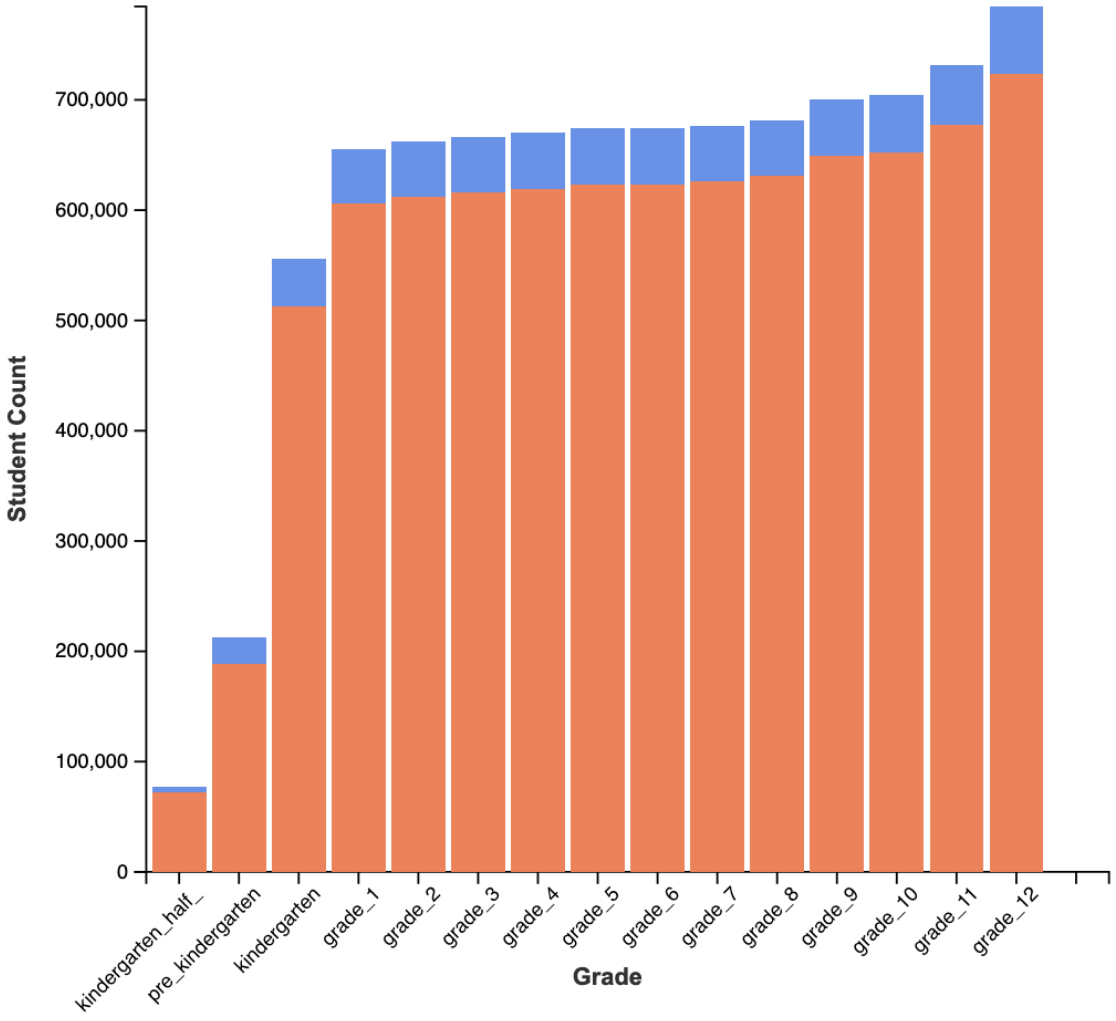


This visualization shows a mark for every single grade at every single school in Washington state. While at first this visualization contains a lot of overlapping marks, making it difficult to pick out individual items, the hope is that by filtering the data set using interactions that it will help the user understand how the filter(s) are being applied (see Further Implementation). The intent of the next two visualizations is to show the number of black and white students enrolled in each county and also by grade.

County Distribution



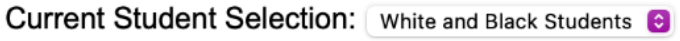
Distribution By Grade



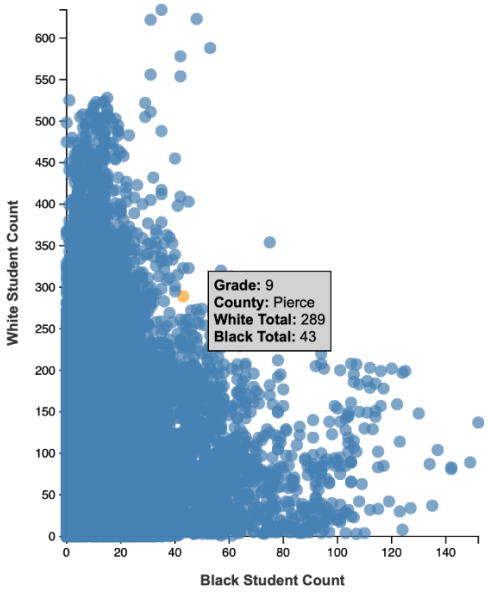
Further Implementation

Our second prototype of the dashboard lines up with the focused visualization proposals and contains more interaction for further exploration of the data. For implementation, we followed the same procedure as the first prototype, further trimming the data set to just include the required columns and excluding the data from the “pk_12_total” summary grade in order to not show misleading data points in the visualizations. As proposed in the prior chapter, we also implemented interactive functionalities to help the user understand how the data changes under **County** and **Grade Level**.

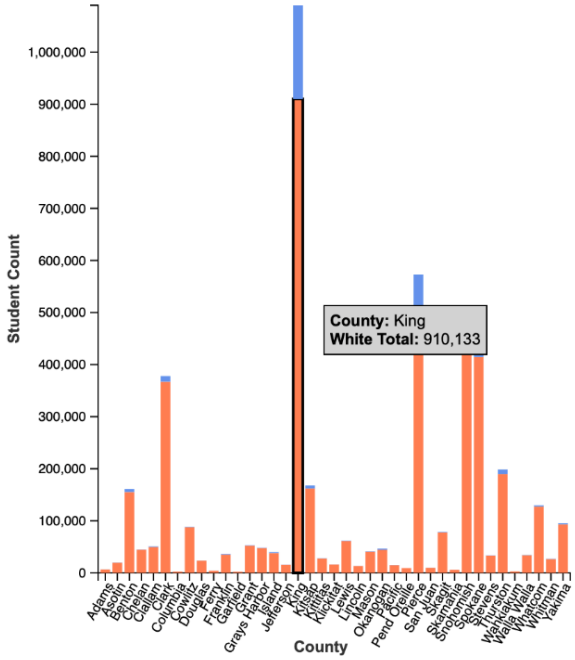
A dropdown menu allows the user to filter the graphs by race (black or white students):



The user can also hover over a bar or dot to make the graphs change and to show the individual data values:



(hovering over a dot)



(hovering over a bar)

Final Evaluation

Overall, we believe these visualizations do a good job of showing how Black student enrollment and White student enrollment relate to one another. We can picture how the interactions with each plot will affect the others.

At this final Checkpoint, we have made a considerable number of improvements to our final dashboard, resulting in a final product that represents our main goal. We believe it is clear from our graphs and interactions that there is a difference between black and white students with regards to school enrollment following the COVID-19 global pandemic. In addition, the selected visualizations, which show data related to county and grade-level, provide further insight into the nature of this difference.

However, we also recognize that our visualization can be improved even more if we chose to do so outside of this class. With more space, we could have potentially experimented with making the “by county” visualization a map, similar to the initial proposal visualization. There may have also been an opportunity to adjust the scale of the scatter plot even further, to make it more obvious how much difference there is between the student races. If we were feeling ambitious,

we might have even included a fourth visualization that shows some change over time, given the large size of our dataset.