Exam 2- Essay

For this project I used the kindergarten_CA dataset, which contains information about kindergarten students in California from the years 2001 to 2015. I chose it simply because it seemed to be the most interesting one; also, when I looked at it I saw that there was a good number of variables to work with, including quantitative and qualitative data. After I cleaned it up, there were nine variables left, being four of them categorical (school, county, district, and funding) and five quantitative (id, year, enrollments, immune, and proportion).

I did the cleaning with dplyr: used rename to change the name of some columns, mutate to create a new column for the proportion of immune kids, select to specify the order I wanted for the columns, arrange to reorder the rows by school id, and filter to remove missing values.

The first six visualizations are simple plots that I created to take a better look at some variables and choose what my last graph would be focused on. The bar chart for the year variable shows the school counts for each year. I was a bit surprised by it, since I was expecting that the number of schools would increase over the years and it turns out that there isn't a lot of variation.

The bar charts for county and district also don't show much variation on count, but they did help me realize that I had to be careful when including these two variables to my final visualization, because they have a lot of categories, which can be a problem when using aesthetics like colors and shapes.

I decided to create a pie chart for the funding variable, even though this kind of graph is usually avoided by statisticians. I did that because there are only two values, public or private, which makes it easy to compare quantities, even in a pie chart. The visualization basically shows that there are way more public than private schools.

The histograms for the immune and proportion variables show completely opposite types of distributions (the first one is right skewed and the second one left skewed), when they were supposed to be quantifying the same element: immune kids. This only confirms my thought that it's always better to work with proportion, or else you just have "out of context" numbers.

The more statistical part of this project is the boxplot I made for the enrollments variable, along with the five number summary of the distribution. The graph shows the overall behavior: with the exception of the two very high outliers, the data is concentrated below 100, and around 75, with not much spread. The summary statistics offer exact numbers and let us know that 25% of the data is below 35.5, 50% is below 73, and 75% is below 94. We also know that the school with least students has only 10 of them, and the one with most students has 316 (for this sample).

For my final R visualization, I wanted to explore the relationship between the year and the number of enrollments, but also include the funding and proportion variables to it. So I created a scatterplot for year x enrollments, using different colors for public and private schools, and varying the size of the points according to the proportion of immune kids. Here

we have a lot of information: there are not only more public schools, but they also have more students enrolled than the private ones; most schools seem to have an immune proportion close to 1, which is good news; the enrollments did increase a little bit over time, but not as much as I had expected.

I did the interactive part using Tableau. It is basically the same scatterplot as the R one, but with Tableau I could add a bit more information. When we mouse over the points, we can see the name of the school, which county and district it belongs to, and the actual value for the proportion and enrollments.

In short, I'm glad I included the simple exploration graphs to the project, because even though they're not very nice to look at, they did help me choose the variables that I wanted in my final visualization. I'm also satisfied with my scatterplot, since I was able to add all the variables I liked without making it confusing. In the interactive version, I included almost all of the variables in the dataset (7 out of 9). My only problem was with Tableau: I didn't know how to add a jitter element to my interactive graph. But besides that, everything worked out just fine and I was happy with the results I got.