Project Instructions:

Link to GitHub Repository: https://github.com/mekhali/Teacher_Quality_Research.git

In my GitHub repository, all my data and code is stored in the “Data_studio_proj” file.

Part 1: Raw Data (Optional)

***I strongly recommend skipping this section and going straight to Part 2 since I have all the data sets in the GitHub repository ready for use.

While the data sets I used are in my GitHub repository, if you want to get the raw data for yourself from the source, here are the links and instructions to downloading it:

Staffing and Vacancy Data:
https://www.doe.virginia.gov/teaching-learning-assessment/teaching-in-virginia/education-workforce-data-reports

- Opens up to the Virginia Department of Education website
- Click “Staffing and Vacancy Report” where it takes you to the “Build Your Own Report” page
- I used 2023-2024 Data
- Report level is “School”, on “All Divisions” and “All Schools”
- Category: “Personnel”
- Position Type: “Teachers”
- English Learners: “All Positions”
- Adult Education: “No” because we are only looking at K-12 students
- Special Education: “All Positions”
- Submit, and then download the data as a .csv

SOL Test Data:
https://www.doe.virginia.gov/data-policy-funding/data-reports/statistics-reports/sol-test-pass-rates-other-results

- Opens up to the Virginia Department of Education website
- Click “SOL Test Results” where it takes you to the “Build Your Own Report” page
- I used 2023-2024 Data
- Report level is “School”, on “All Divisions” and “All Schools”
- I used “All Races,” “All Genders," and “All Grades”

- “All Students” should be selected for Disadvantaged, English Learner, Migrant, Homeless, Military Connected, Foster Care, Disabled (this ensures all these populations are included in the data set)
- Test Level: “All Levels”
- Test Source: “SOL”
- Subject Area → Control Shift to select all
- Test: “All Tests”
- Statistic: “Pass Rate”
- Submit, and then download the data as a .csv

Median Income Data:
https://hdpulse.nimhd.nih.gov/data-portal/social/table?age=001&age_options=ageall_1&demo=00011&demo_options=income_3&race=00&race_options=race_7&sex=0&sex_options=sexboth_1&socialtopic=030&socialtopic_options=social_6&statefips=51&statefips_options=area_states

- Click download on the right and click “Export Data (CSV)”

All Other Data Used in this Project:
https://schoolquality.virginia.gov/download-data

For all following data sets, select:

- Reporting Level: “School”
- Division: “All Divisions”
- Schools: “All Schools”

Then for each individual data set follow the following instructions (these will populate with separate spread sheets):

- *Post Secondary Enrollment:*
  - Data Type: “College and Career Readiness”
  - Select Indicators: “Postsecondary Enrollment”
  - School Year: Here I selected “Most Recent Year Available” because as of July 2025 that is the 2023-2024 school year, however, soon the 2024-2025 school year will be available, and if you choose to download the raw data instead of use my spreadsheet, you will have to select “All Available Years”
  - Hit “Download Spreadsheet”
- *Provisionally Licensed:*
  - Data Type: “Teacher Quality”
  - Select Indicators: “Provisionally Licensed"
  - School Year: Here I selected “Most Recent Year Available” because as of July 2025 that is the 2023-2024 school year, however, soon the 2024-2025 school year will be available, and if you choose to download the raw data instead of use my spreadsheet, you will have to select “All Available Years” and filter for the year you want to use in the R code.
  - Hit “Download Spreadsheet”
- *Teacher Educational Attainment:*

    - Data Type: “Educational Attainment”
    - Select Indicators: “Teacher Quality"
    - School Year: Here I selected “Most Recent Year Available” because as of July 2025 that is the 2023-2024 school year, however, soon the 2024-2025 school year will be available, and if you choose to download the raw data instead of use my spreadsheet, you will have to select “All Available Years” and filter for the year you want to use in the R code.
    - Hit “Download Spreadsheet”
- *Teacher Quality:*
    - Data Type: “Teacher Quality”
    - Select Indicators: “Teacher Quality"
    - School Year: Here I selected “Most Recent Year Available” because as of July 2025 that is the 2023-2024 school year, however, soon the 2024-2025 school year will be available, and if you choose to download the raw data instead of use my spreadsheet, you will have to select “All Available Years” and filter for the year you want to use in the R code.
    - Hit “Download Spreadsheet”

Part 2: Combined R Code

For Part 2 either download the entire “Data_studio_proj” file folder from the GitHub repository and open the “MP_Combined.R” code from there and run it (as recommended), or use your own data sets that you downloaded from Part 1.

You need to combine all the data sets you just downloaded from GitHub or by yourself.
The code for this is in the R script file titled “MP_Combined.R” which is in the “Data_studio_proj” folder.

Note: if you are not downloading the entire “Data_studio_proj” folder from my GitHub and running the script from that R Project file, then you need to add your own local file path for all of the data sets under the “Import data” section of the code.

Run all the code in that R.script file. This code will import the data sets, clean them, combine all of them into one data frame, and then export that data frame into the “Data_studio_proj” folder as a .csv file titled “combined_data_set.csv”. You will use this data set in the next two R Script files. There are notes throughout this code with instructions on what each line does.

In the MP_Combined.R code, I left more data sets than I ended up using in the final project. All those variables were tested prior to deciding on the final variables used. I left these in the combined data set that this code creates because I think all these variables are useful, and I wanted to make them accessible for anyone who wanted to test them further.

However, for those who only want to replicate what was in the final article, the datasets used for that were:

- staffing_and_vacancy_report_statistics.csv
- assessment_statistics.csv
- Postsecondary Enrollment.csv
- Provisionally Licensed.csv (had school poverty level data)
- Teacher Educational Attainment.csv
- Teacher Quality.csv
- HDPulse_data_export_copy.csv

Part 3: Regression R Code

Next, open the R Script file titled “MP_Regressions.R” from the “Data_studio_proj” file. Run the entire code and it will give you the regression output as well as check for multicollinearity. I screen shotted the regression results after the code line: summary(ps_enrollment_model). There are notes throughout this code with instructions on what each line does.

I made decisions about what schools I would keep in the data set and what should be removed in order to keep the data standardized. More instructions are in the “MP_Regressions.R” code on what was being kept and why. However, a full list of the schools that were removed along with reasons for each removal can be found in the GitHub repository titled “Schools Omitted from Analysis.”

Part 4: Graphs R Code

Next, open the R Script file titled “MP_Graphs.R” from the “Data_studio_proj” file. Run the entire code. This will provide you with all the graphs I used in the Medium Article.

Note: While these are the base graphs I used in the article, I edited them on google slides to make them more aesthetically pleasing. So font, and title sizing will look different on the R-Studio graphs.

This code includes other graphs that I considered using in the final Medium article but were ultimately scraped. They are there to play around with.

The graphs that were used in the final Medium article were labeled as follows in the code (in order of article appearance…):

- Under **#Scatterplots** (the first ggplot graph was used, that was the one with the Post Secondary Enrollment dependant variable)
- Under **#By Poverty Level - #Academic Outcomes Stats: Post-Secondary Enrollment**
- Under **#Bined Graphs - #Pass Rate by Teacher Quality**

- Under **#By Poverty Level - #Resource Stats: Out-of-Filed Teachers**
- Under **#By Poverty Level - #Academic Outcomes Stats: Pass Rate**

At the end of the code, it creates a “datawrapper_df.csv” file that is saved into the “Data_studio_proj” folder. The .csv file is cleaned and adjusted to be uploaded to Datawrapper to create the maps. It averages Post Secondary Enrollment by district, so it can be used at district-level on the map of Virginia. It also fixes some name inconsistencies to match Datawrapper’s database. However some changes still need to be made on Datawrapper itself…

Part 5: Data Wrapper Instructions

Now that you have the “datawrapper_df.csv” file, you can go to https://www.datawrapper.de/ and click the “Create New” button in the top right hand corner, and select “Map.”

Then select “Chloropleth Map.”

You will then need to select your map. Search for “Virginia School District” and select “USA >> Virginia >> School Districts” and click “Proceed.”

Then you need to upload the “datawrapper_df.csv” file.

When you upload the data set there will be 4 errors due to the names we used in our data set not matching Datawrapper’s database.

Go to the spreadsheet on the right, and for the 4 Division Name errors highlighted in red, click the the drop down arrow to change the following:

- Alleghany Highlands Public Schools → Alleghany County Public Schools
- Colonial Beach Public Schools → Colonial Beach Town Public Schools
- West Point Public Schools → West Point Town Public Schools
- Williamsburg–James City County Public Schools → Williamsburg City Public Schools

I did this process twice, once to create the map for Median Income and once to create the map for Master’s Percentage.

Once you proceed to the "Visualization" section I imported colors from Colorbrewer for my map.

Under colors, “Type” was “continuous.”

Under “Annotate” and “Customize Tool Tips” this is what I had written to get the proper labels on the maps:

- Median Income Map: ${{ FORMAT(median_income, "0,0.[00]") }}
- Master’s Average Map: {{ FORMAT(district_masters_avg, "0.0%") }}

Most everything else was the default settings. I added Title, Data Source, and Links in the “Annotate” section.

Part 6: How I Made the Graphic

I basically made this graphic on canva. Didn’t use any data, it was just an idea that I thought would be better visualized than simply explained in words.