Thesis - Draft

Nathan Yang

Project Background

The goal of this project is to model academic performance in school districts across the United States through various demographic and socioeconomic factors. The data sources include the American Community Survey (ACS), the Educational Opportunity Project, the Longitudinal School Demographic Dataset (LSDD), the Common Core of Data (CCD), and the Census. The data will be joined by school district or county to create a comprehensive dataset for analysis.

I first explored the relationship between a few socioeconomic factors and academic performance in school districts. I created a dashboard to visualize my findings and developed some simple linear regression models to explore the relationship between these factors and academic performance.

Next, I plan to expand on this project by incorporating more data sources and more advanced modeling techniques. I also plan to overhaul the joining process of datasets to preserve more records and develop a more comprehensive dataset for analysis.

Methodology

The first step of my project was to identify the datasets that I would be using for this project. I started with only looking at datasets aggregated by school district as that would be the most granular level relevant to my research question. I started with the American Community Survey (ACS), the Educational Opportunity Project, the Common Core of Data (CCD), and the Census. I reviewed the data dictionaries for each dataset to understand the variable encoding and the extent of the data. I also reviewed the data sources to understand the data collection process and the limitations of the data.

I first joined the datasets solely by school district name. This was a simple join that matched the exact names of the school districts. However, this method had issues as many school districts had different names in different datasets due to no common naming convention. This would result in many records not being matched and a loss of data.

I created some preliminary simple linear regression models within an interactive dashboard to explore the relationship between academic performance and various socioeconomic factors. I used the shiny (Chang et al. 2023) package to create the dashboard and the ggplot2 (Wickham 2016) package to create visualizations for the data. I used the rsconnect (Atkins et al. 2024) package to deploy the dashboard onto shinyapps.com where it can be publicly accessible.

Next, I reviewed the data sources I initially picked out and identified additional data sources that could be useful for this project. These came primarily from reading various research papers and reports that studied similar topics. I used an excel spreadsheet to track all of the data sources and variables of interest.

After identifying an exhaustive list of datasets and variables, I began the process of downloading and cleaning the data. I used the tidyverse (Wickham et al. 2019) package to clean and manipulate the data to prepare the datasets to be joined. For my joining process, I used fuzzy matching techniques to join records that were similar but not exact matches. The metrics I used for fuzzy matching were string distance and Jaccard difference.

String distance is a metric that calculates the number of character changes needed to transform one string into another while Jaccard difference is a metric that compares how many 2-letter pairs are shared between two strings. I used the stringdist (Loo 2014) package to calculate both of these metrics and determined thresholds from examining the distributions and matching strength for each metric. I then joined the datasets purely by matching state and calculated the metrics for every pair of school district names within a state. Once I have this dataset with all the potential matches, I start an extensive filtering to ensure I am getting the most accurate matches possible.

1. Filter for matches that both begin with the same letter: This prevents matches names containing North/South and East/West at the beginning are not accidentally mapped together due to the characters in these cardinal directions being similar
2. Filter for matches that end with the same three letters: This prevents matches such as "Abcdefgh county" and "Abcdefgh city" where the school districts may have the same name but are clearly different entities. This also resolves matching names that have numbers at the end such as "Abcdefgh 231" and "Abcdefgh 562" that clearly represent different school districts
3. For each school district in the academic performance dataset, I find its best match based on string distance with ties broken by Jaccard difference (and ties at this stage decided randomly).

This is an example of a dataset joined between my academic performance data and a dataset from the CCD. Using these string comparison metrics, I was able to preserve many records that would have been unmatched if I performed a direct name join. It is especially noticeable with abbreviated words that these metrics help to identify matches like with "Heights" being reduced to "Hts." or "Community" being abbreviated to "Com" as shown below. Additional common abbreviations found in the school district names are "Saint" written as "St." and cardinal directions only represented by the first letter.

Table 1: **Example of School District Matching**. This table shows the similarity between `seda_district` and `ccd_district` using a distance measure and Jaccard index.

seda_district	ccd_district	dist	jaccard
Beaverton Rural Schools	Beaverton Schools	6	0.2727273
North Daviess Community Schools	North Daviess Com Schools	6	0.2580645
Southern Wells Community Schools	Southern Wells Com Schools	6	0.2580645
North Lawrence Community Schools	North Lawrence Com Schools	6	0.2500000
South Harrison Community Schools	South Harrison Com Schools	6	0.2500000
Greenfield-Central Community Schools	Greenfield-Central Com Schools	6	0.2285714
Minnetonka Public School District	Minneapolis Public School District	5	0.2857143
Morris Area Public Schools	Moorhead Area Public Schools	5	0.2758621
West St. Paul-Mendota Hts.-Eagan	West St. Paul-Mendota Heights-Eagan	5	0.2432432
Minnesota Public School District	Minneapolis Public School District	5	0.2424242
North Branch Public Schools	North Branch Area Public Schools	5	0.1724138
Ridgefield Park School District	Ridgefield School District	5	0.1666667
Hamilton County CUSD 10	Hamilton Co CUSD 10	4	0.2727273
West Washington County CUD 10	West Washington Co CUD 10	4	0.2142857
Rising Sun-Ohio County Com	Rising Sun-Ohio Co Com	4	0.1818182

By joining datasets by exact district name, I would have only had 4441 records with the CCD data. However, using the fuzzy matching techniques, I was able to match 4576 records. This is a 3% increase in the number of records that were matched.

Through district name joining I was only able to match about 250 school districts with ACS income data. However, using the fuzzy matching techniques, I was able to match 281 school

districts. This is a 12% increase in the number of records that were matched.

Early on in my project, I only selected datasets that were aggregated by school district and it unfortunately did not prove fruitful as many of the ACS datasets I investigated had very limited data on school districts. This resulted in poor record retention for future dataset merging in addition to reduced modeling data as demonstrated by the ACS income dataset. I transitioned to identifying the counties for school districts in the academic performance dataset and then joining the datasets by county. This proved to be much more successful as I was able to use many ACS Data Profiles (DP) datasets which are a selection of curated features from various ACS datasets that have greater consistency in data. This allowed me to retain more records and have a more comprehensive dataset for analysis.

The new comprehensive dataset is much larger and contains more features than the previous dataset. However, the loss of granularity from school district to county may have an impact on the accuracy of the models. I will need to investigate this further in my modeling phase and keep this in mind when interpreting the results.

References

Atkins, Aron, Toph Allen, Hadley Wickham, Jonathan McPherson, and JJ Allaire. 2024. "Rsconnect: Deploy Docs, Apps, and APIs to 'Posit Connect', 'Shinyapps.io', and 'RPubs'." https://CRAN.R-project.org/package=rsconnect.

Chang, Winston, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges. 2023. "Shiny: Web Application Framework for r." https://CRAN.R-project.org/package=shiny.

Loo, M. P. J. van der. 2014. "The Stringdist Package for Approximate String Matching" 6: 111–22. https://CRAN.R-project.org/package=stringdist.

Wickham, Hadley. 2016. "Ggplot2: Elegant Graphics for Data Analysis." https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the {Tidyverse}" 4: 1686. https://doi.org/10.21105/joss.01686.