Joining Methodology

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.

This project is continuing off the independent research project from last semester, where I explored the relationship between a few socioecnomic 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.

This semester, 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

Dataset Descriptions

...

Prior Work

Previously, I had 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.

...

Current Work

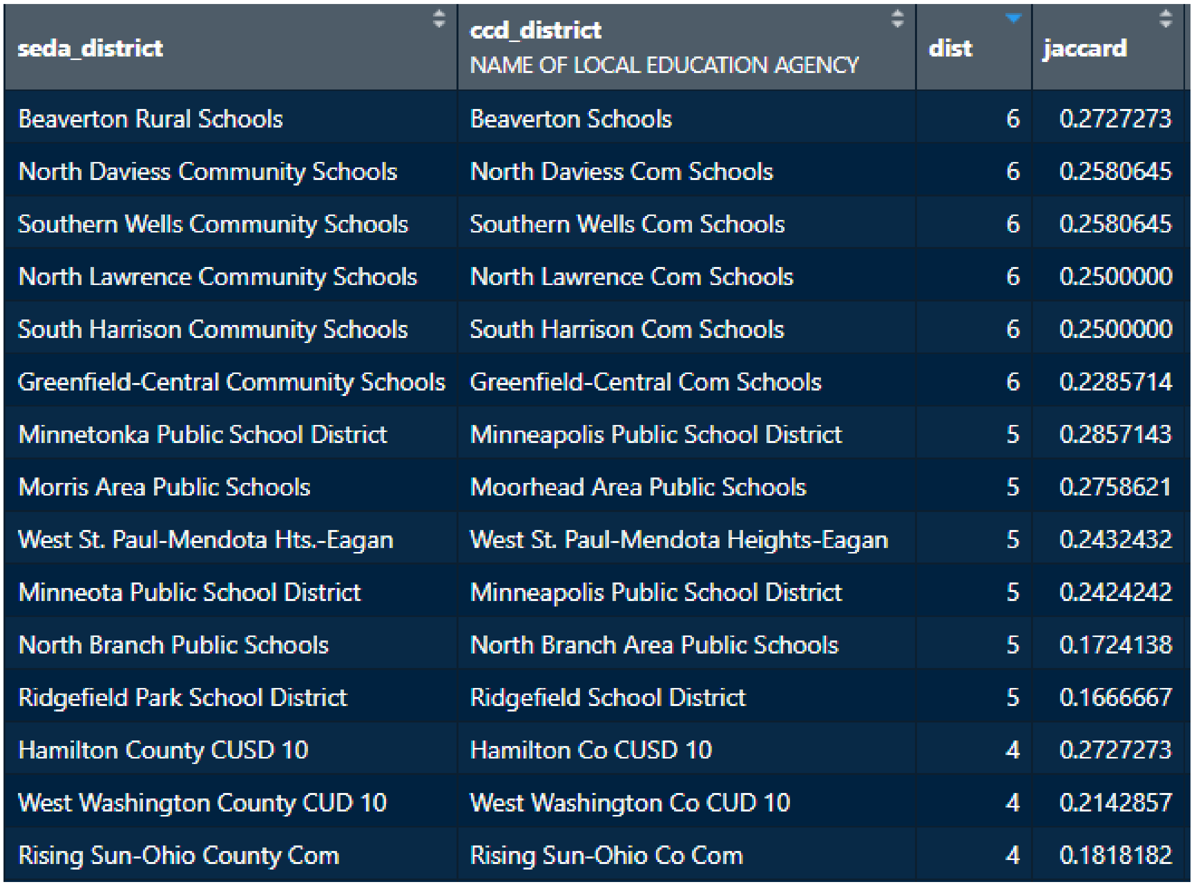
I first reviewed the data sources I utilized in the previous semester and identified additional data sources that could be useful for this project. These came primarily from reading various research papers and reports that studied similar topics. I then reviewed the data dictionaries for each dataset to understand the variables and extent of the data. 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 (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?**) 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 both of these metrics for every pair of school district names. Once I have this dataset with all the potential matches, I start an extensive filtering to ensure I am getting the more 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 found 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.



seda_district	ccd_district NAME OF LOCAL EDUCATION AGENCY	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
Minneota 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

Had I only performed a direct name join, 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.

```r
test_dataset <- inner_join(
    test_mth_all_admindist_gys,
    trim_ccd_financial,
    by = c("seda_district" = "ccd_district", "stateabb" = "stateabb")
  )
test_dataset |> count()
```

In my original project 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.

I plan to continuously refine my matching process and explore additional techniques to improve the accuracy of my matches. I will also explore additional data sources and variables to incorporate into my analysis. I will then begin to develop more advanced models to predict academic performance in school districts across the United States.

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.