Household Conditions by Geographic School District

Jon Geiger, Noel Goodwin, Abigail Joppa

Week 9/10

Research Question

We know that correlations between graduation rate and household/demographic conditions vary by region.

What household conditions are the biggest indicators for graduation rates across school districts?

How does this differ across Regions in the U.S.?

Does district-wide assessment data provide a significant improvement to a regression model?

Last Week's Next Steps

- Include assessment data in the model
- Decide how to split our analysis by region
- Interpret variable importance
- Evaluate modeling techniques

Model Variables

Current Variables:

- Household Conditions
- Racial/Ethnic Distribution
- Finance Data
- Assessment Data

Reproducibility!

Our Repo

- analyses
- data
- raw
- scripts
- slides
- .gitignore
- README.md
- _config.yml
- index.md

Our New Data Directories

- NHGIS_District_data.xlsx
- Public_Schools.csv
- README.md
- race_raw.txt

Our raw data sets

- Original data (NHGIS_District_data.xslx)
- Info about Schools (Public_Schools.csv)
- Original downloaded race data (race_raw.txt)

Our cleaned data sets :

- HH Conditions (hh.csv)
- Grad Rates (grad.csv)
- Race Distribution (race.csv)
- Assessments (assess.csv)
- Financial (finance.csv)
- High School IDs (hs_dist.csv)

- README.md
- assess.csv
- finance.csv
- grad.csv
- hh.csv
- hs_dist.csv
- race.csv

Scripts

- create_hh.R
- create_hs_dist.R
- create_race.R
- download_assess_data.R
- download_finance_data.R
- download_grad_data.R
- names_list.R
- prune_race_variables.R
- to_moe.R

Data Dictionary

Household conditions (hh.csv)

Cleaned and pruned version of the original household conditions dataset. Changes include:

- Rename variables for easier use (see [names_list.R] (https://github.com/jonmgeiger/household-conditions/blob/main/scripts/names_list.R))
- Transform the original "Margin of Error" (MOE) variable into a single sided MOE for future anlayses
- Include LEAID (state fips code + district ID) for easy joining
- Filter out New York Department of Education "School District" -- This is an extreme outlier, as it includes the children for all 32 NY districts. See create_hh.R for more details.

Variables	Description
`state`	State where each district resides
`leaid`	Local education agency identification number
`dist`	District name
`children`	An estimate of children between the ages 5-17 who are enrolled in school within a certain geographic school district.
`pct_pov`	Percent of students within each geographical district boundary estimated to be living in poverty
`pct_SP`	Percent of households within each geographical district boundary estimated to be living in a household with only one father or one mother.

GitHub Pages

https://jonmgeiger.github.io/household-conditions/index.html

Household Conditions by Geographic School District

Data and Society Capstone Project
Seattle Pacific University
By: Jon Geiger, Noel Goodwin, and Abigail Joppa

Home | Data | Analysis

Data Dictionary

Household Conditions Dataset (hh.csv)

- Provided by the Urban Institute

Variables	Description
state	State where each district resides
leaid	Local education agency identification number

Modeling!

Investigating Relationships

untransformed
transformed
children
count
7500
5000
2500
0
0e+00
2e+05
4e+05
6e+05
2
3
4
5
6
value

children
log_children
grad_rate_midpt
100
75
50
25
0
0e+00
2e+05
4e+05
6e+05
6
8
10
12
value

untransformed
quadratic
arcsin
exponential
count
value
1200
800
400
0
0
25
50
75
100
600
400
200
0
0
2500
5000
7500
10000
600
400
200
0
0.0
0.5
1.0
1.5
7500
5000
2500
0
0e+00
1e+43
2e+43

This week's results

Last week's results

Result Specifics

```
## # A tibble: 7 × 4
##   wflow_id       rank  rmse   rsq
##   <chr>         <int> <dbl> <dbl>
## 1 main_xgboost      1  5.50 0.505
## 2 main_rf           2  5.59 0.488
## 3 main_mars         3  5.95 0.421
## 4 main_lasso        4  6.11 0.385
## 5 main_lm           5  6.16 0.380
## 6 main_svm          6  6.16 0.387
## 7 main_dtree        7  6.31 0.368
```

Model Results:

- Test RMSE: 5.98 --> 5.40
- Test R^2: 0.451 --> 0.494

Variable Importance

Last week's variable importance: Random Forest

This week's variable importance: Boosted Tree

Next Steps

Work on Final Presentation & Deliverables (Our Data Story)

Include only districts with high schools (interpreting grad rates?)

Compare model accuracy with four regional models OR state-wide **dummy variables instead of region**