

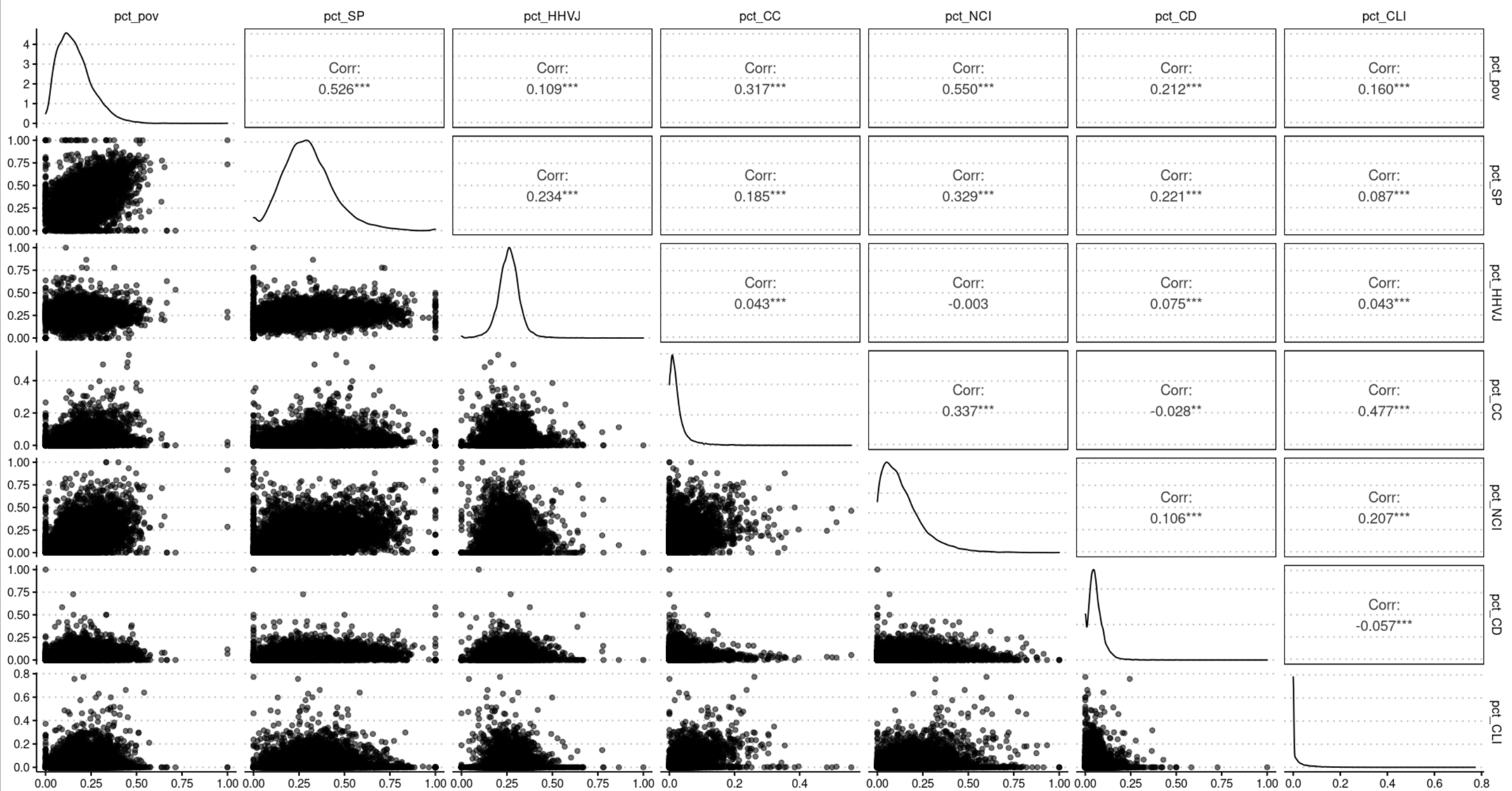


Household Conditions by Geographic School District

Jon Geiger, Noel Goodwin, Abigail Joppa

Data Story

Alabama School Districts: Socioeconomic and Demographic Data																
School ID	State	Geographic School District	Children 5-17 (SAIPE Estimate)	% Poverty (SAIPE Estimate)	% Single Parent Estimate	Single Parent Margin of Error	% HHs With Vulnerable Job Estimate	Vulnerable Job Margin of Error	% Crowded Conditions Estimate	HH With Crowded Conditions Margin of Error	% No Computer or Internet Estimate	No Computer or Internet Margin of Error	% Children with Disability	Children with Disability Margin of Error	% Linguistically Isolated Children	Linguistically Isolated Children Margin of Error
00001	Alabama	Fort Rucker School District	985	0.12442397	0.04897494	0%-10%	0.32755905	20%-46%	0.0280898884	0%-8%	0.02048947	0%-5%	0.032487310	0%-6%	0.0000000000	0%-3%
00003	Alabama	Maxwell AFB School District	292	0.15226337	0.10188679	3%-17%	0.30651340	4%-57%	0.0099667776	0%-10%	0.01452785	0%-6%	0.054794520	2%-9%	0.0000000000	0%-8%
00005	Alabama	Albertville City School District	4591	0.30053842	0.35292614	26%-44%	0.32653522	28%-37%	0.0476644412	3%-7%	0.22147135	13%-32%	0.018078851	0%-3%	0.1736005247	10%-24%
00006	Alabama	Marshall County School District	8299	0.26622182	0.29519674	24%-35%	0.23486671	21%-26%	0.0263831690	2%-3%	0.14248565	11%-18%	0.043499216	3%-6%	0.0386793576	1%-7%
00007	Alabama	Hoover City School District	15397	0.07259553	0.20831554	17%-25%	0.29216549	27%-31%	0.0154848723	1%-2%	0.03168848	1%-5%	0.038773786	2%-5%	0.0441644490	1%-8%
00008	Alabama	Madison City School District	9416	0.08298157	0.17839834	12%-23%	0.20930608	19%-23%	0.0160830505	1%-3%	0.10363837	6%-15%	0.057774000	4%-8%	0.0042480882	0%-1%
00011	Alabama	Leeds City School District	2324	0.16887608	0.27438369	14%-41%	0.30435523	25%-36%	0.0298975538	1%-5%	0.23763336	13%-35%	0.080034420	3%-13%	0.0000000000	0%-2%
00012	Alabama	Boaz City School District	1644	0.24957070	0.40097290	20%-60%	0.30367768	25%-36%	0.0126368999	0%-3%	0.11448503	3%-20%	0.064476885	2%-11%	0.0675182492	1%-12%
00013	Alabama	Trussville City School District	4476	0.06697699	0.22902320	15%-31%	0.25656864	22%-29%	0.0067690704	0%-2%	0.03920277	1%-7%	0.025245756	1%-4%	0.0000000000	0%-1%
00030	Alabama	Alexander City City School District	3064	0.37336412	0.50406188	39%-62%	0.23417529	19%-28%	0.0422178432	2%-6%	0.17756540	10%-26%	0.052219320	3%-8%	0.0287206266	0%-6%
00060	Alabama	Andalusia City School District	1425	0.31193927	0.57907951	50%-66%	0.27138391	21%-33%	0.0204906203	0%-4%	0.09485772	4%-15%	0.151578948	8%-23%	0.0000000000	0%-2%
00090	Alabama	Anniston City School District	3008	0.38064516	0.53800368	46%-61%	0.25753108	22%-30%	0.0120985014	0%-2%	0.26993594	20%-34%	0.094747342	6%-13%	0.0000000000	0%-1%
00100	Alabama	Arab City School District	1454	0.17518248	0.32910389	29%-37%	0.28344330	26%-31%	0.0059598493	0%-1%	0.09138656	6%-12%	0.056396149	4%-7%	0.0130674001	0%-3%
00120	Alabama	Athens City School District	3824	0.17829262	0.33690476	25%-42%	0.33215451	29%-37%	0.0160485804	1%-2%	0.14985251	9%-21%	0.092834726	4%-15%	0.0478556491	1%-8%
00180	Alabama	Attalla City School District	1144	0.26067907	0.45916334	22%-70%	0.31025642	21%-41%	0.0371672511	0%-8%	0.12264808	2%-22%	0.166083917	6%-27%	0.0000000000	0%-3%
00185	Alabama	Saraland City School District	2728	0.20179775	0.32961783	23%-43%	0.25502816	21%-30%	0.0215119850	1%-4%	0.07885906	3%-12%	0.039589442	1%-7%	0.0000000000	0%-1%
00188	Alabama	Chickasaw City School District	1350	0.41289023	0.63652325	53%-74%	0.33821446	28%-39%	0.0390434340	2%-6%	0.29953918	19%-41%	0.039259259	1%-6%	0.0000000000	0%-3%
00189	Alabama	Satsuma City School District	922	0.21624266	0.22084367	9%-35%	0.19352290	13%-26%	0.0000000000	0%-2%	0.12757830	2%-24%	0.046637744	0%-9%	0.0000000000	0%-4%
00190	Alabama	Alabaster City School District	6469	0.11596363	0.21697639	16%-28%	0.30604193	28%-34%	0.0177251268	1%-3%	0.05473270	2%-9%	0.059823774	3%-9%	0.0347812660	1%-6%
00194	Alabama	Pelham City School District	4044	0.08137564	0.27000543	19%-35%	0.30181071	27%-34%	0.0126640042	0%-2%	0.10544980	5%-16%	0.040059347	2%-6%	0.0501978248	0%-10%
00195	Alabama	Pike Road City School District	1813	0.06903164	0.18050314	8%-29%	0.19800954	16%-24%	0.0009313878	0%-1%	0.03603977	0%-7%	0.040816326	1%-8%	0.0154440152	0%-4%
00210	Alabama	Auburn City School District	8277	0.11712439	0.28417316	22%-35%	0.28592813	26%-31%	0.0315098464	2%-4%	0.06876741	4%-10%	0.046756070	3%-7%	0.0421650372	2%-7%
00240	Alabama	Autauga County School District	10106	0.19297887	0.26168954	21%-31%	0.26019731	24%-28%	0.0141605493	1%-2%	0.08542150	6%-11%	0.119236097	9%-15%	0.0003958045	0%-1%
00270	Alabama	Baldwin County School District	33982	0.12897170	0.26784942	24%-30%	0.31789804	30%-33%	0.0126046147	1%-2%	0.09761074	8%-12%	0.045376964	3%-6%	0.0095344596	0%-2%
00300	Alabama	Barbour County School District	1698	0.38240576	0.45508981	36%-55%	0.21099067	17%-25%	0.0633918121	3%-10%	0.52163064	37%-67%	0.146054178	8%-21%	0.0477031805	0%-10%
00330	Alabama	Bessemer City School District	4276	0.31248683	0.69502950	62%-77%	0.34593496	31%-39%	0.0181190688	1%-3%	0.19586253	13%-26%	0.125818521	8%-17%	0.0116931712	0%-3%
00360	Alabama	Bibb County School District	3346	0.25439128	0.32525113	23%-43%	0.22743548	18%-27%	0.0076023391	0%-2%	0.17182428	9%-26%	0.053496711	2%-9%	0.0000000000	0%-1%
00390	Alabama	Birmingham City School District	29129	0.35049510	0.68142539	65%-71%	0.30610257	29%-32%	0.0143302176	1%-2%	0.22469211	20%-25%	0.083730988	7%-10%	0.0222802013	1%-3%
00420	Alabama	Blount County School District	8897	0.15214710	0.29853410	24%-36%	0.26388520	24%-29%	0.0157122519	1%-2%	0.18267849	13%-23%	0.032595258	2%-4%	0.0256266166	1%-4%
00450	Alabama	Brewton City School District	806	0.34977064	0.47727272	33%-63%	0.26670119	19%-34%	0.0055581289	0%-2%	0.29290205	15%-44%	0.064516127	2%-11%	0.0000000000	0%-4%
00480	Alabama	Bullock County School District	1554	0.65387118	0.70178044	63%-77%	0.19720812	15%-25%	0.0000000000	0%-1%	0.40093023	25%-55%	0.080437578	3%-13%	0.0000000000	0%-2%
00510	Alabama	Butler County School District	3361	0.34703895	0.47489393	40%-55%	0.27977172	24%-32%	0.0177400112	1%-3%	0.33245730	27%-40%	0.049985122	3%-7%	0.0000000000	0%-1%
00540	Alabama	Calhoun County School District	9446	0.18748659	0.31073233	26%-36%	0.26475266	24%-29%	0.0204309579	1%-3%	0.13825145	10%-18%	0.098666102	7%-13%	0.0000000000	0%-1%



Our Process

Household Conditions

Race Data

Graduation Rates!

Existing Models

Other indicators?

Can **we** build a model?



Data Explanation and Research Question Evolution

Data sources

Data Set	Source	Data File Name
NHGIS Household Conditions Data (cleaned)	ACS	hh.csv
Race/Ethnicity Data	ACS	race.csv
Graduation Rates	edfacts	grad.csv
Federal, State, and Local Funding Data (per child)	edfacts	finance.csv
Assessment Data	edfacts	assess.csv


Household Conditions

- Five-year estimates (between 2014-2018) of household conditions within a geographic school district calculated based on aggregate Census survey data.

Margin of Errors

- Household condition is an estimate, so this dataset includes a margin of error variable for each estimate. *
- Transformed MOEs into one-sided margin of errors for easier use.

*No NA values, but many estimate that have MOEs of 50% which indicates that they are not accurate.



Household Condition Variables and Descriptions

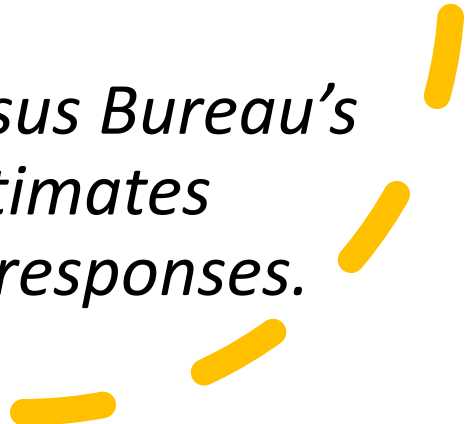
Variable	Description
Single Parents	students living in a household with only on father or one mother.
Linguistically Isolated	students who have cognitive, ambulatory, independent living, selfcare, vision, or hearing difficulties are considered to be children with disability.
Parents in Vulnerable Economic Sectors	parents fall into this category if they earn less than 800 dollars a week and works in industries that are most likely to be laid off.
Crowded Conditions	students fall into this category is there is less than one room per household member.
Lack of Computer or Broadband Internet	includes household with non-dial-up internet, also considers desktop computers, laptops, smartphones, and tablets as computers.

Household Conditions Variables and Descriptions Cont.

SAIPE estimates: *

Variables	Description
Children	an estimate of children between the ages of 5-17 who are enrolled in school within a certain geographic school district.
Poverty	a student is considered to be in poverty if their family's income is at or below 100 percent of the federal poverty level.

** yearly estimate made by US Census Bureau's Small Area Income and Poverty Estimates (SAIPE) Program based on Census responses.*

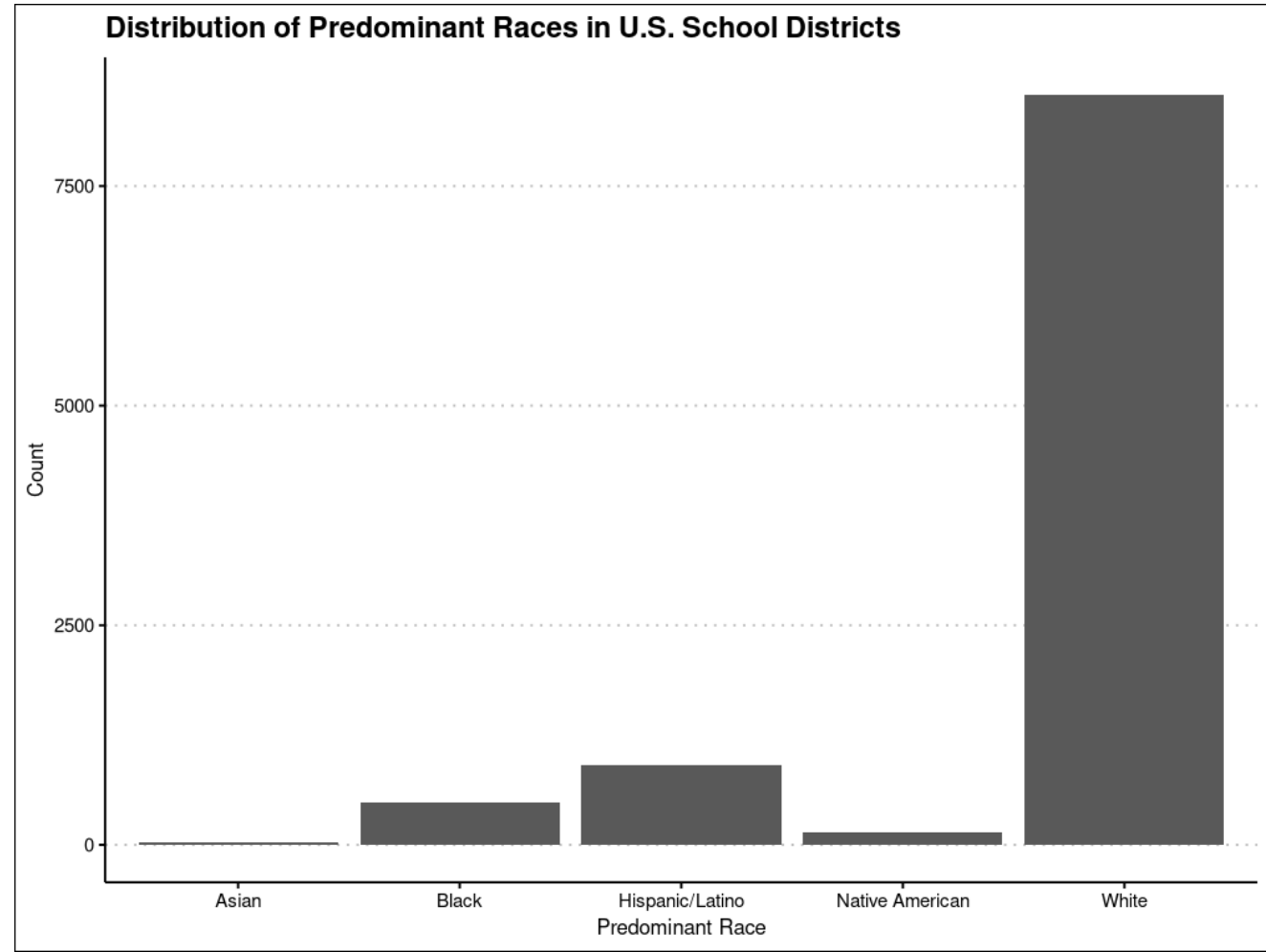



Working Research Questions (at this point)

- How much do racial/ethnic demographics correlate with household conditions across school districts?
- What are some of the biggest influences on *graduation rates* of school districts in the U.S.?
- How do household conditions compare to the funding each school receives, either independently or from the government?
- ~~What are some of the biggest influences on *year-to-year retention rate* of school districts in the U.S.?~~
- ~~To what extent did household conditions affect graduation rates before the spread of COVID compared to during the pandemic?~~

Race

- Entire dataset included 31 ethnic and racial percentage estimates by school district between 2014 and 2018.
- We narrowed it down to the following variables:
 - % Hispanic or Latino
 - % White
 - % Black
 - % Native
 - % Asian
 - % Pacific Islander/Hawaiian
- Added Predominant Race column



A large orange circle with a white outline, partially visible on the left side of the slide.

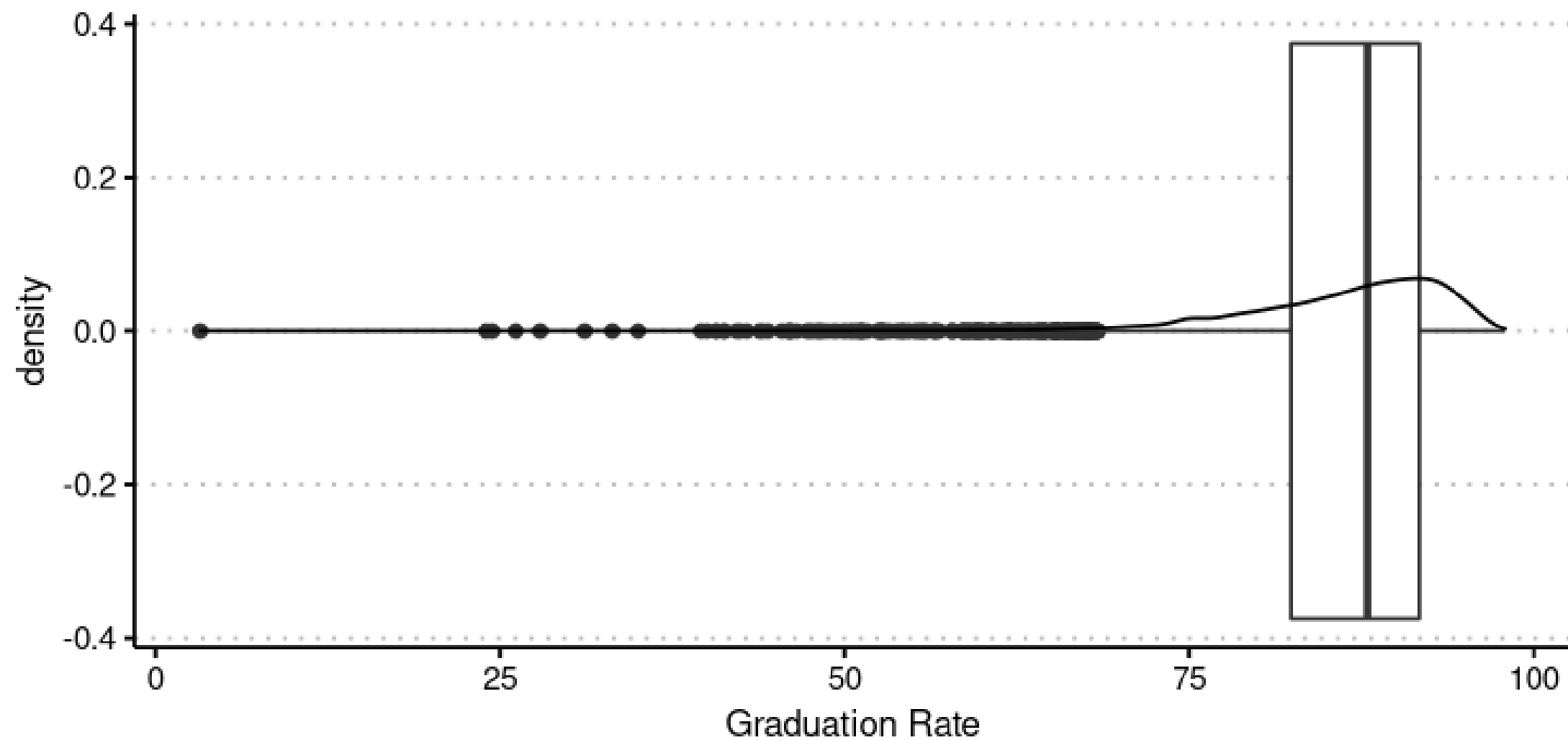
Updated Working
Research Question
Influenced by the
Inclusion of Race Data

How do the metropolitan school districts in the US compare in demographic, social, and economic household conditions, and what factors play a leading role in those discrepancies?

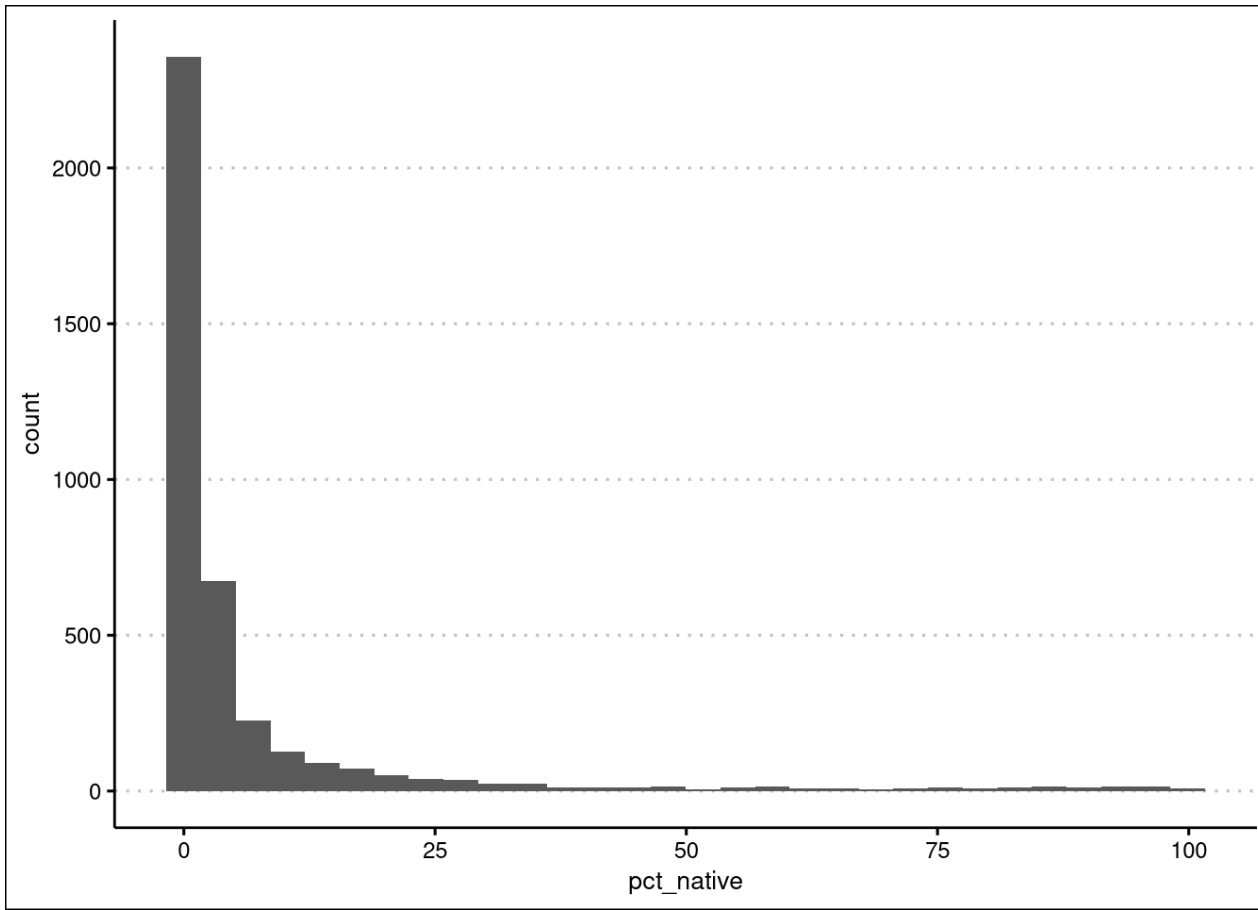
Graduation Rates

- Accessed through Urban Institute R Package
- Includes graduation rates by cohort
- Transformed the data to include graduation rates by district rather than cohort.
- We narrowed it down to the following variables:
 - leaid
 - grad_rate_midpoint – average graduation rate

Distribution of Graduation Rates



Native Population



- We did not want to disregard the native population, so we turned our percent native variable into a dummy variable
- If native students comprise most of the district population, we might assume that the district is on a reservation (proxy)

Big Pivot in Research Question with addition of Grad Rates

One of the only measures being used to predict graduation rates are GPA, tests, disciplinary issues, attendance etc. using a tracking system called EWS (early warning system).

However, we know that household conditions significantly impact these indicators. This leads us to the question:

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

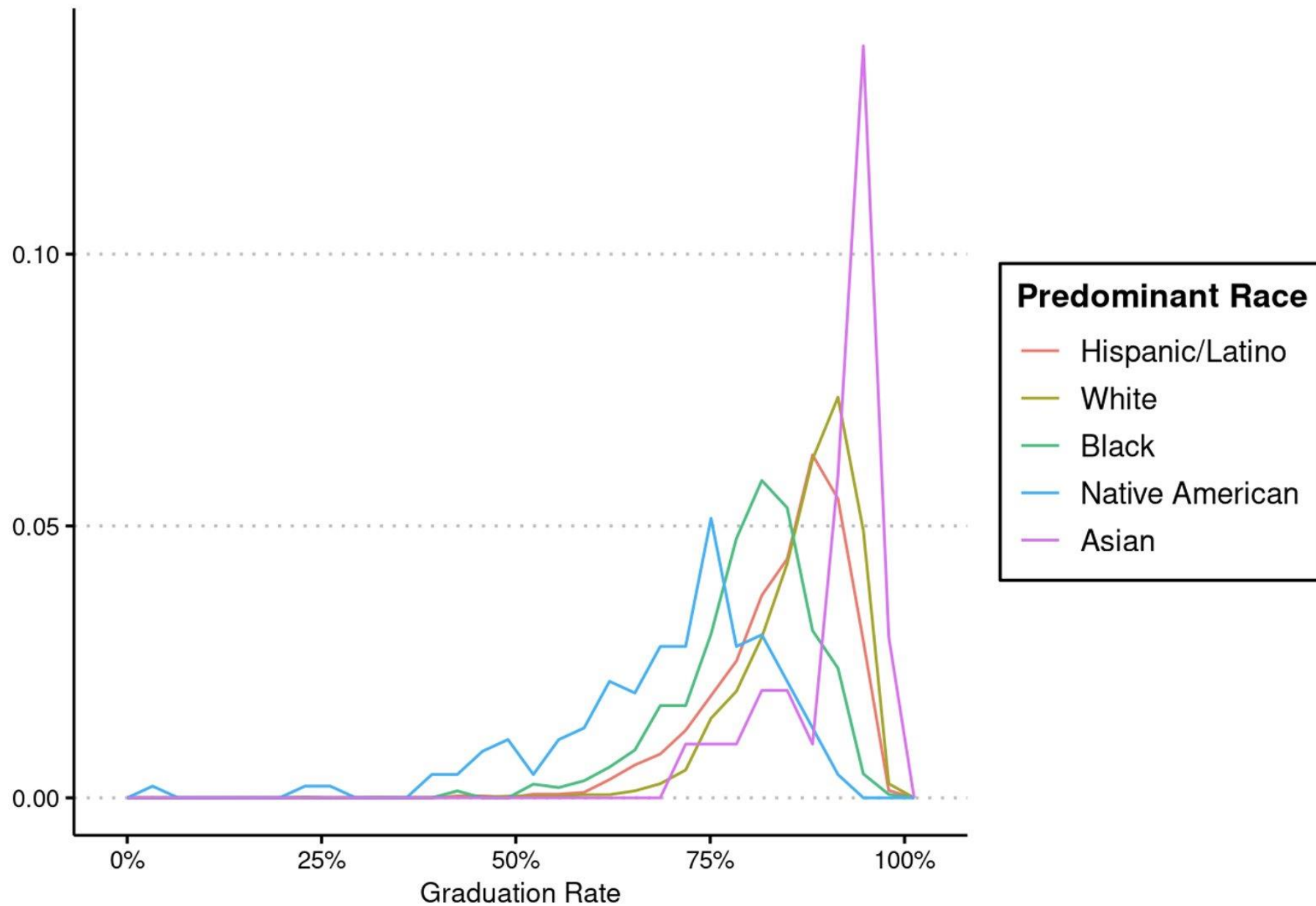
How does this differ across:

Racial and ethnic groups?

Regions?

Household income groups?

Graduation Rate Distribution per Race













This led us to wonder...

How do the current predictors of graduations rates (GPA, tests, disciplinary issues, attendance, etc.) compare to using household conditions as predictors of graduation rates?

(Research on the top indicators in predicting graduation:
https://ccrscenter.org/sites/default/files/EvidenceBasedPractices_EarlyWarningIndicators.pdf)

Table 1. Frequently Used Ninth-Grade Indicators and Thresholds

Ninth-Grade Indicators	Description	Threshold	Evidence-Based Rating
 First 20- or 30-day attendance (Allensworth & Easton, 2007)	The number of absences within the first 20 or 30 days of each grading period is the biggest risk factor for failing ninth grade (Neild & Balfanz, 2006). This indicator is particularly important because the data are available early in the school year, allowing for timely intervention.	Missed 10% or more of instructional time	 PROMISING EVIDENCE
 Attendance (Allensworth & Easton, 2005, 2007)	Attendance is a frequently used indicator because attendance during the first year of high school is directly related to high school completion rates (Heppen & Theriault, 2008).	Missed 10% or more of instructional time	 STRONG EVIDENCE
 Course failures (Allensworth & Easton, 2007)	This indicator applies to failures in any subject, not just the core content areas (English, mathematics, science, or social studies). This indicator is strongly related to the next indicator, grade point average (GPA).	Failed one or more courses per grading period	 STRONG EVIDENCE
 GPA (Allensworth & Easton, 2007)	If using a 4.00 GPA (rather than a weighted average) as the norm, students are considered off-track if they have a GPA of 2.00 or lower following each grading period and at the end of the year. This calculation includes all credit-bearing classes.	2.00 (less than half the maximum attainable GPA)	 STRONG EVIDENCE
 On-track indicator (Allensworth & Easton, 2007)	This composite indicator is the minimal expected level of student performance (Allensworth & Easton, 2007). To be considered on-track, a student must have accumulated enough credits for promotion to the next grade and have no more than one failing grade in a core subject (English, mathematics, science, or social studies) by the end of the school year.	Either two or more core course failures or a failure to earn enough credits to be promoted to	 STRONG EVIDENCE

New 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?
- Tiers of school district funding?

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

District funding

- Accessed through Urban Institute API
- Sourced from edfacts
- Focused in on total local, state and federal funding by district
- Calculated funding per child between 2014-2017

leaid	fed_per_child	state_per_child	local_per_child
100005	1541.9299	6573.459	3037.7369
100006	892.2461	4340.764	1620.4964
100007	415.3407	4625.787	6419.9519

Assessment Dataset

- Accessed through the urban institute R package API
- Sourced from edfacts
- We narrowed it down to the following variables:
 - leaid
 - read_score – 5-year average reading assessment score (2014-2018)
 - math_score – 5-year average math assessment score (2014-2018)
 - total_score – 5-year average total (reading+math) assessment score (2014-2018)



Final Research Question

To what extent can we use household conditions, race, assessment scores, and school funding to predict school district graduation rates in the United States?

Modeling

Predicting Graduation Rates from
our data

Models



Tuned parameters with 10-fold cross-validation and random grid search

Linear Regression {lm}

Lasso Regression {glmnet}

Multivariate Adaptive Regression Spline {earth}

Support Vector Regression {kernlab}

Decision Tree {rpart}

Random Forest {ranger}

Gradient Boosted Trees {xgboost}

Data Filtering

NA rows omitted at each step

All data, with no filtering

School Districts with no MOEs of 50% and > 100 children

High School Districts with no MOEs of 50% and > 100 children

Data Processing

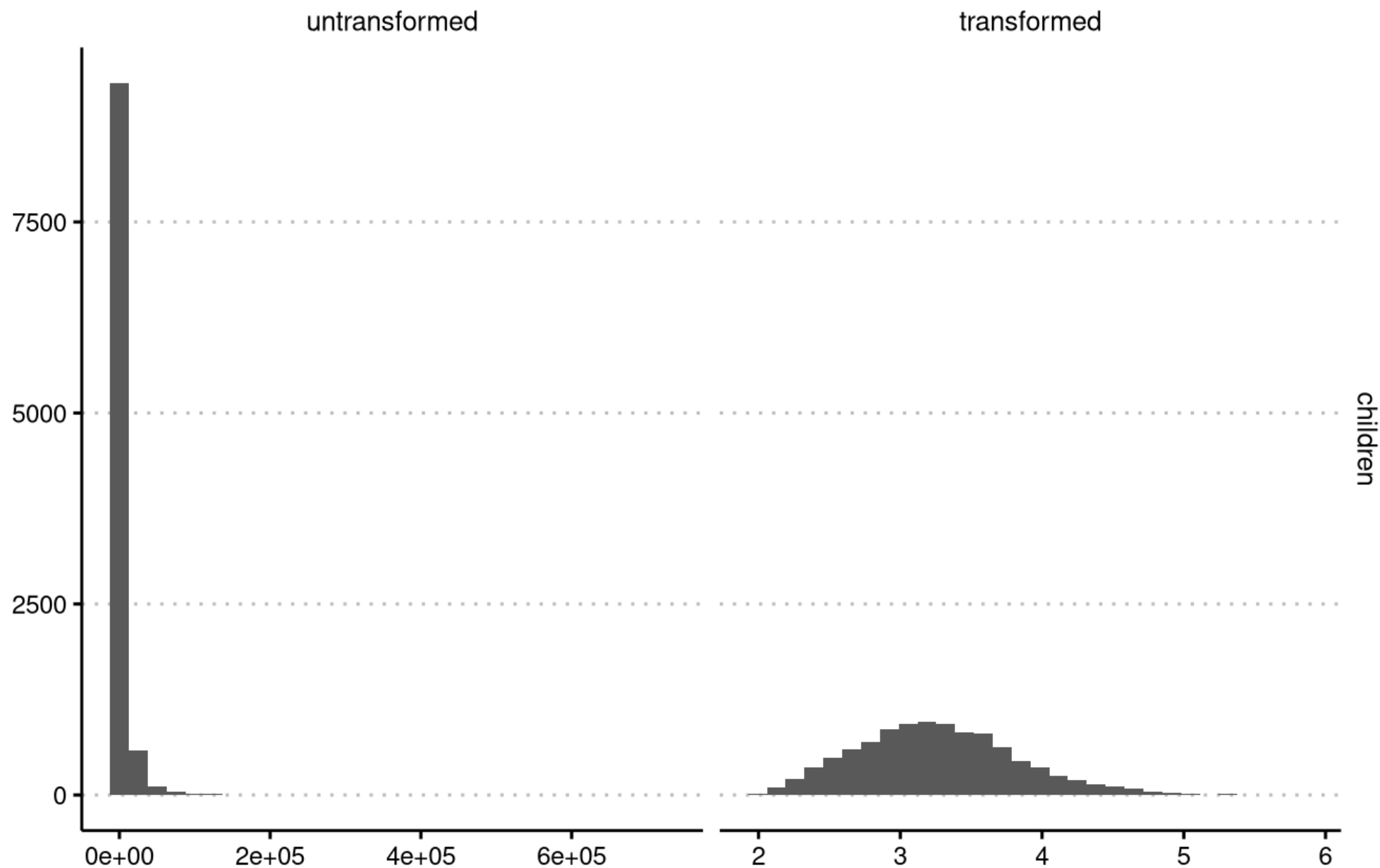
Interactions!

Near-zero variance
predictors

Log-transformations

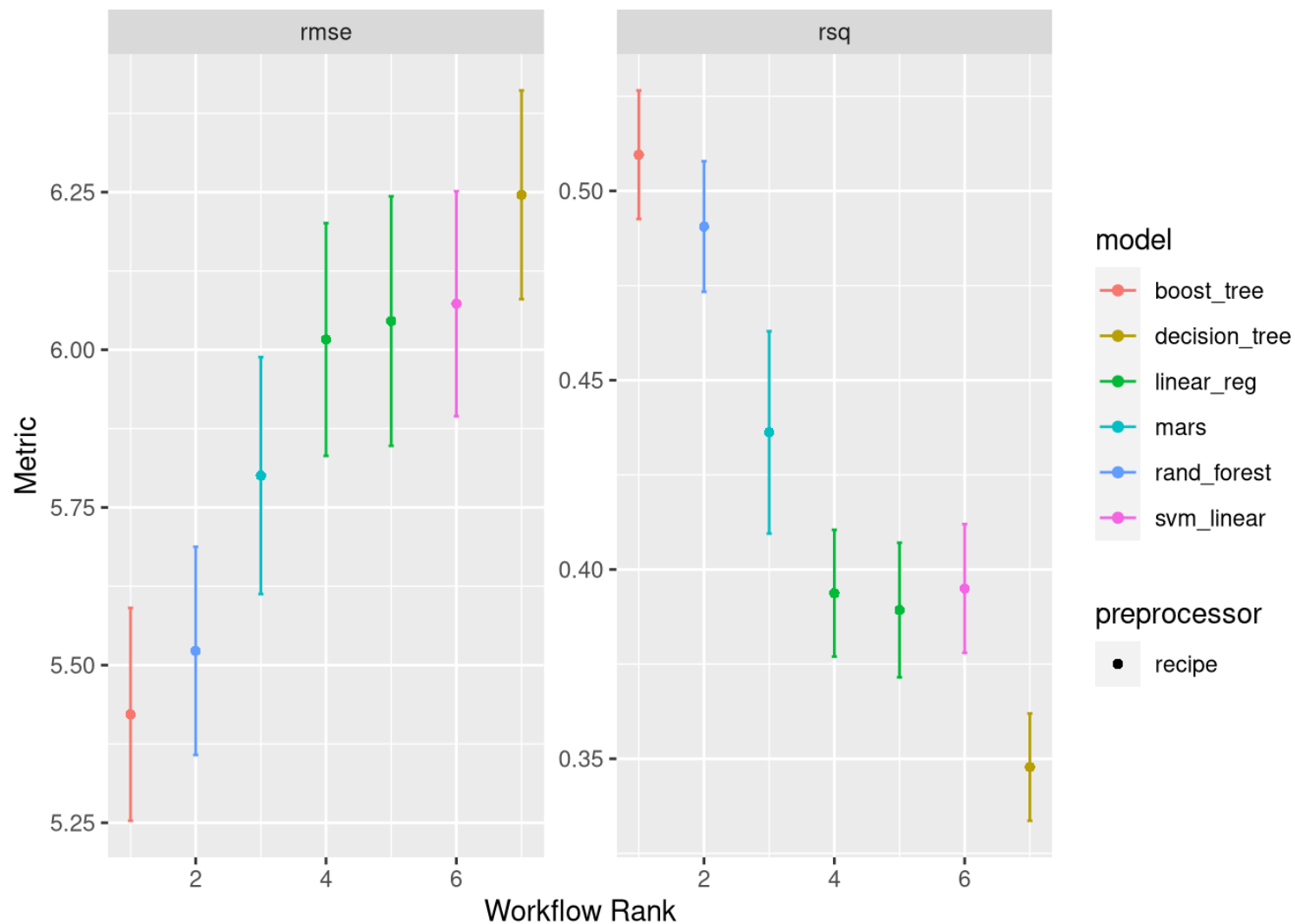
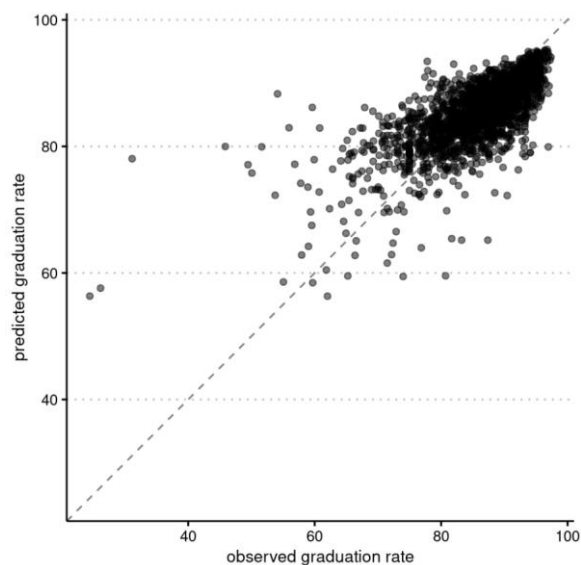
Original vs. Log-transformed Children

(Log transformations are base-10)

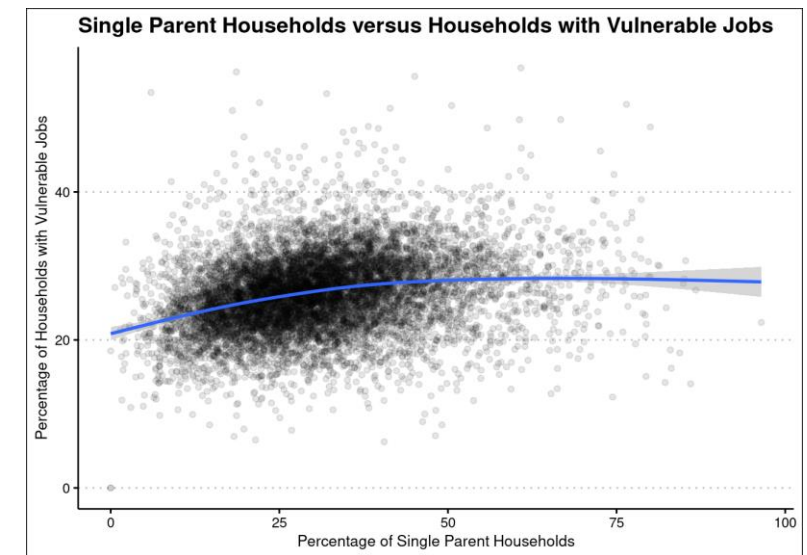
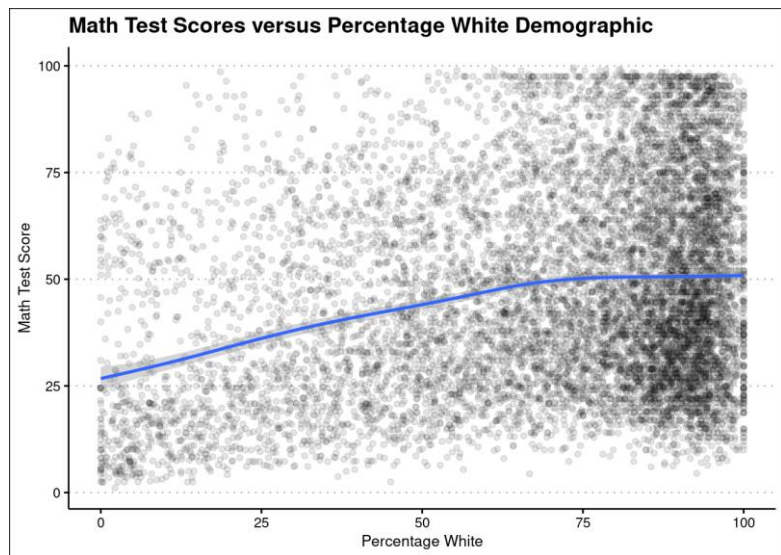
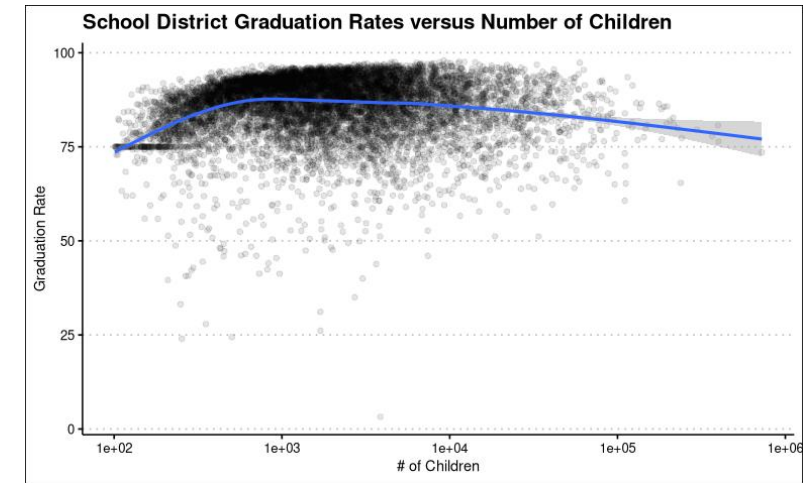
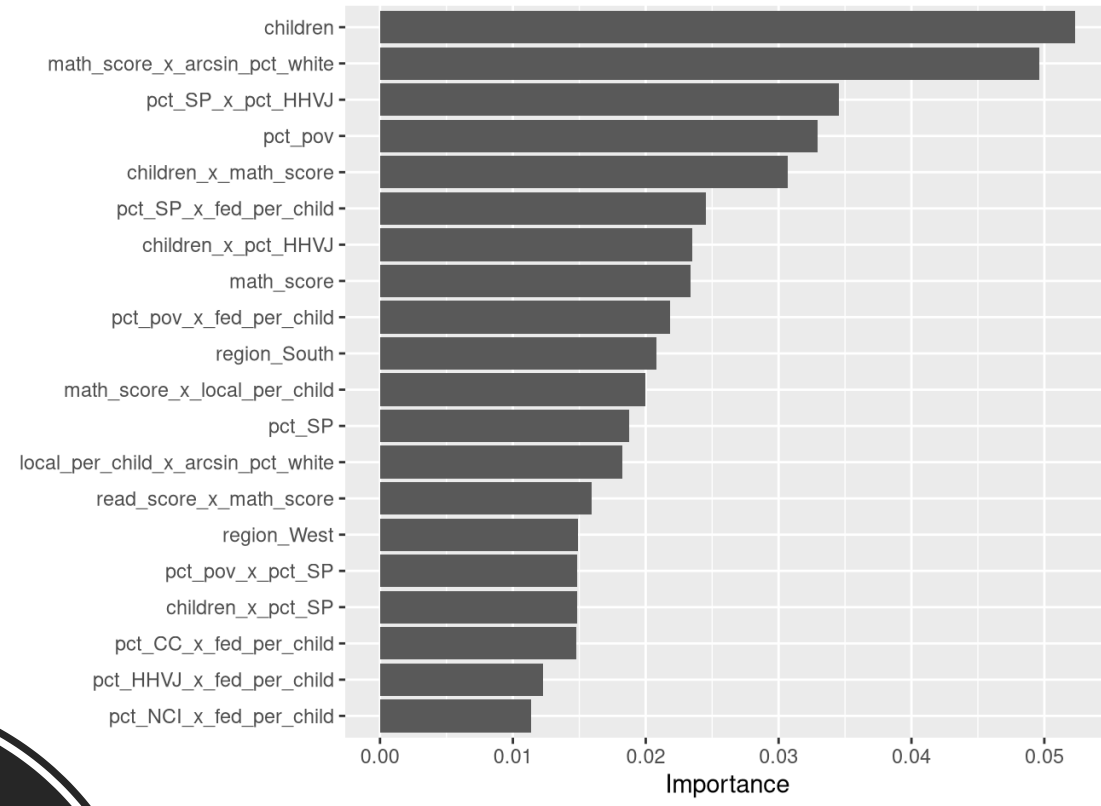


Final Model Results

##	wflow_id	rank	rmse	rsq
##	<chr>	<int>	<dbl>	<dbl>
## 1	main_xgboost	1	5.42	0.510
## 2	main_rf	2	5.52	0.491
## 3	main_mars	3	5.80	0.436
## 4	main_lasso	4	6.02	0.394
## 5	main_lm	5	6.05	0.389
## 6	main_svm	6	6.07	0.395
## 7	main_dtree	7	6.25	0.348



Interpretation



Discussion



Societal Implications

- Very difficult to predict graduation rates with *just* household conditions
- An individual's identity is lost at the school district level
 - We see this at a university level as well... 🙄
- Each of the household conditions impacts the model in some meaningful way

Strengths & Limitations

Strengths

- Only known model to predict district-level graduation rates
- Utilizes multiple sources of data
- Fully reproducible (mostly)

Limitations

- Predicting graduation rates by districts vs at school level
- The data: All estimates (accurate?)
- Models aren't *super* useful and are hard to interpret

What we would have done differently

Workflow

- GitHub
- Branching



Reproducibility

```
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 Management

📄 NHGIS_District_data.xlsx
📄 README.md
📄 assess.csv
📄 clean_graduation_data.csv
📄 finance.csv
📄 finance_data.csv
📄 grad.csv
📄 grad_predom-raceP_household.csv
📄 grad_raceP_household.csv
📄 grad_raceP_household_rev.csv
📄 grad_race_household.csv
📄 hh.csv
📄 public_schools.csv
📄 race.csv
📄 raceP_household.csv
📄 race_household.csv
📄 school_assess.csv
📄 school_grad_race_hh.csv

Better data
management

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

Future Research?

- Correlation between funding for predominate native schools?
- Graduation rates per school (instead of district)?
- Four separate models for regions?
- States instead of regions?





Questions?