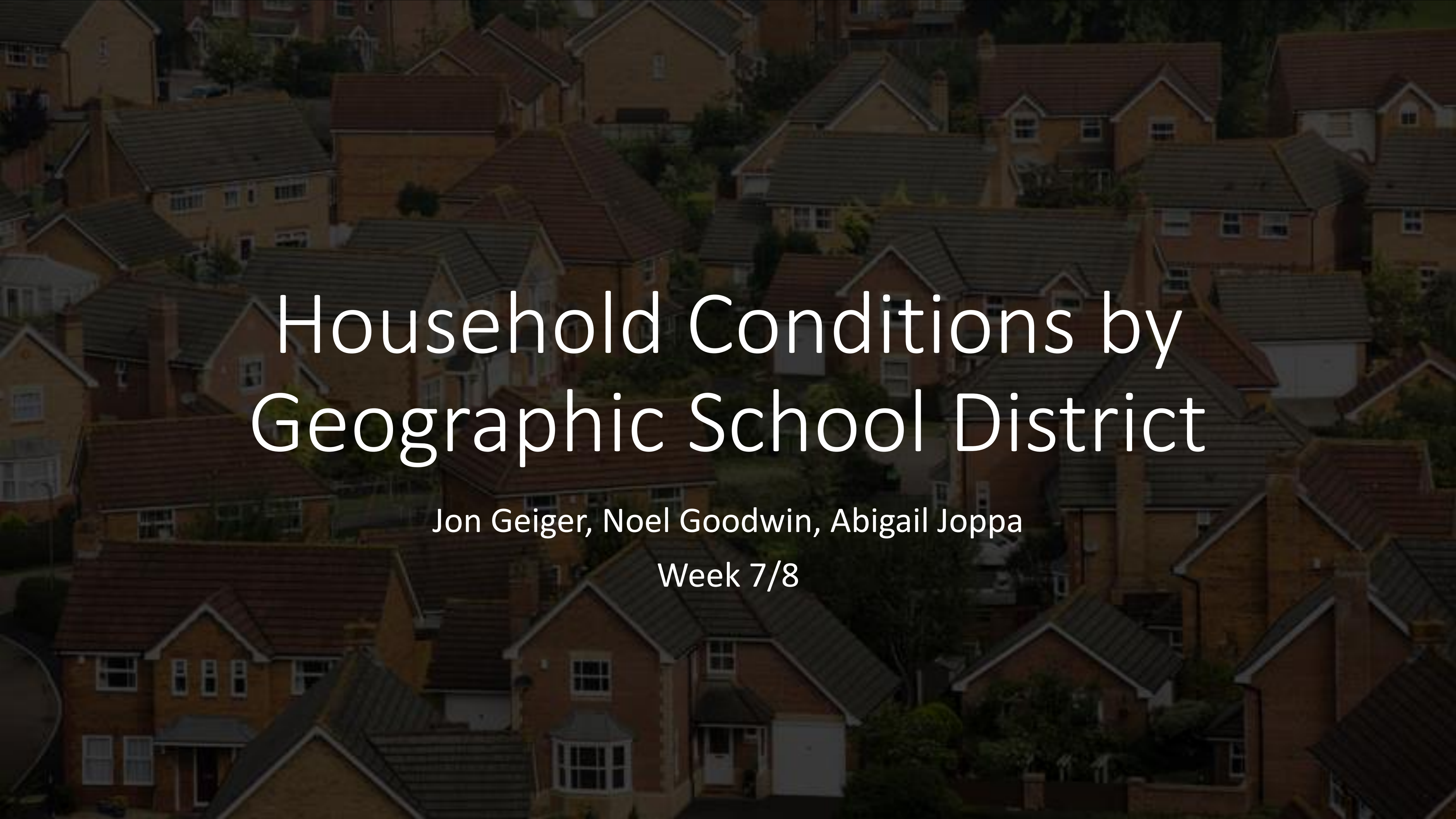


Household Conditions by Geographic School District

Jon Geiger, Noel Goodwin, Abigail Joppa

Week 7/8

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?

Last Week's Steps..

Visualize predictor distributions for regression

Improve regression model predicting graduation rates using HH conditions and race

Continue to look for prior research that looks at graduation rates and indicators

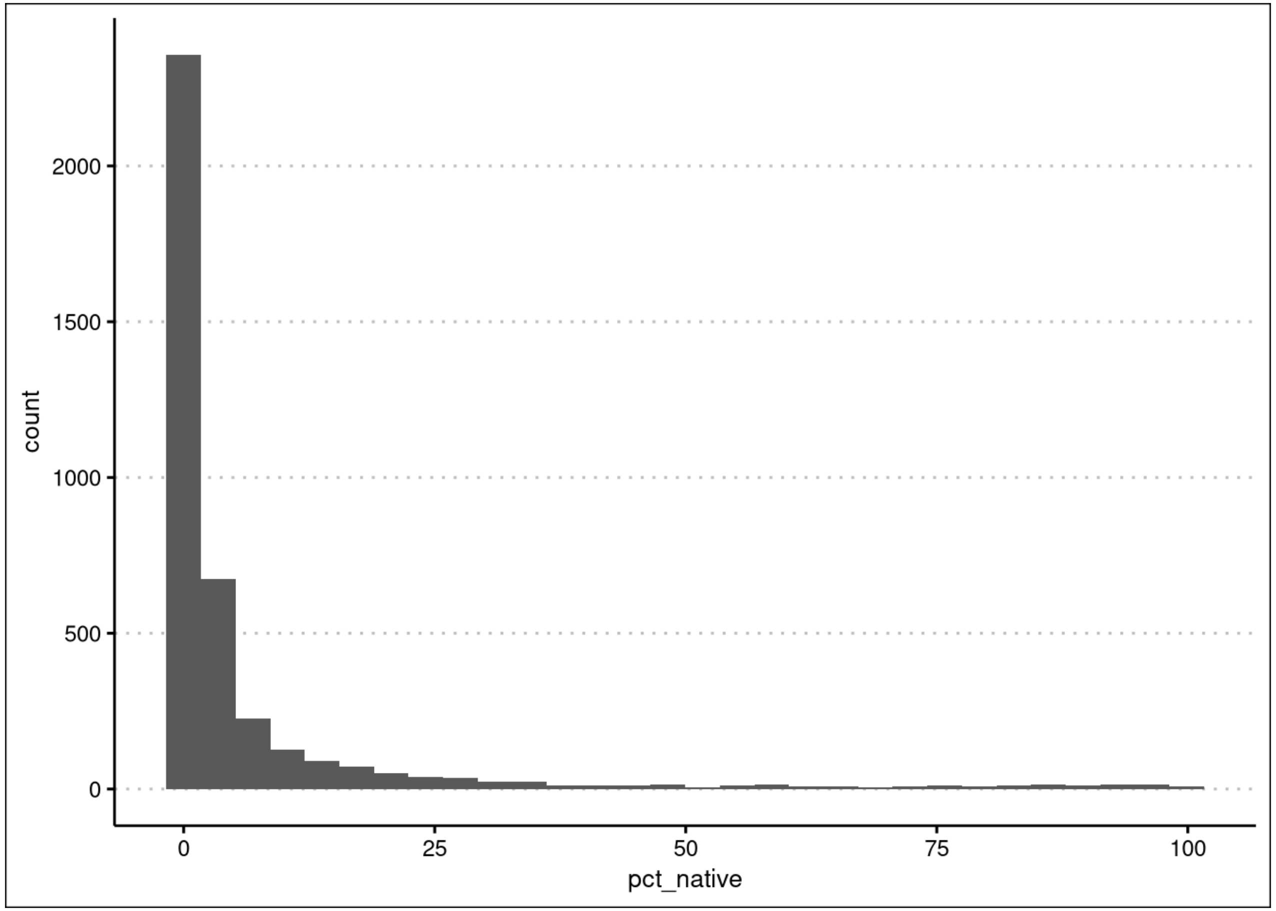
Last Week's Difficulties:

- Difficult to "group by" racial/ethnic groups
- Defining regions
- Household income group data
  - Proxy with school district funding?
  - Look not at income groups, but ranked groups of amount of funding n

Difficult to "group by" racial/ethnic groups

- Got rid of PI
  - Just over 800 districts with PI students
  - Of those 800 districts, the majority have under 5 percent
- Native racial group

Native Population

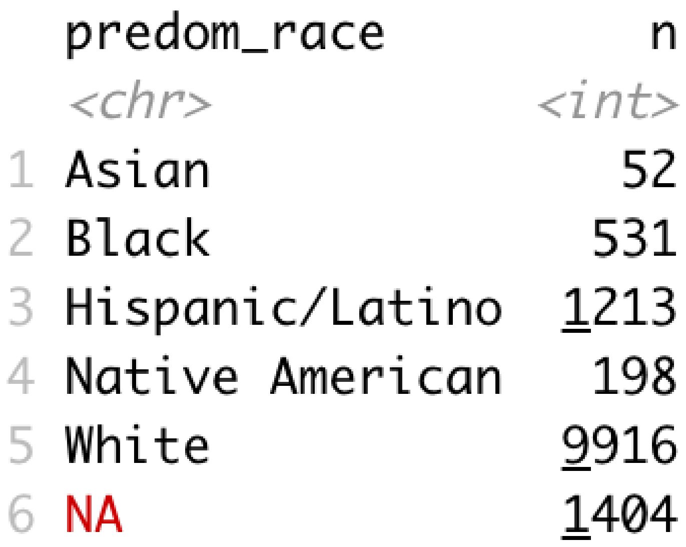


- We did not want to disregard the native population, so we turned our percent native variable into a dummy variable

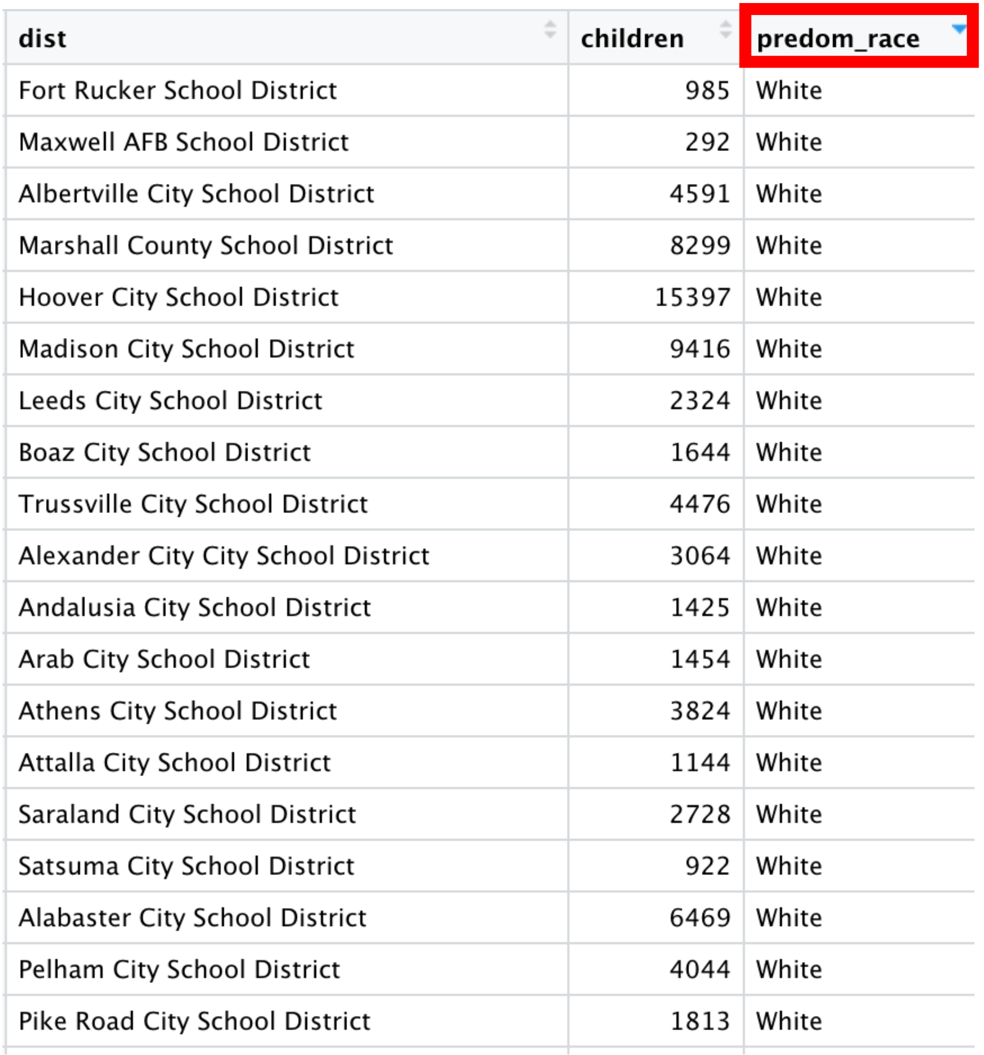
  - If native students comprise most of the district population,
    we might assume that the district is on a reservation (proxy)

Difficult to "group by" racial/ethnic groups continued...

- Created a column for each district that indicates the majority racial group

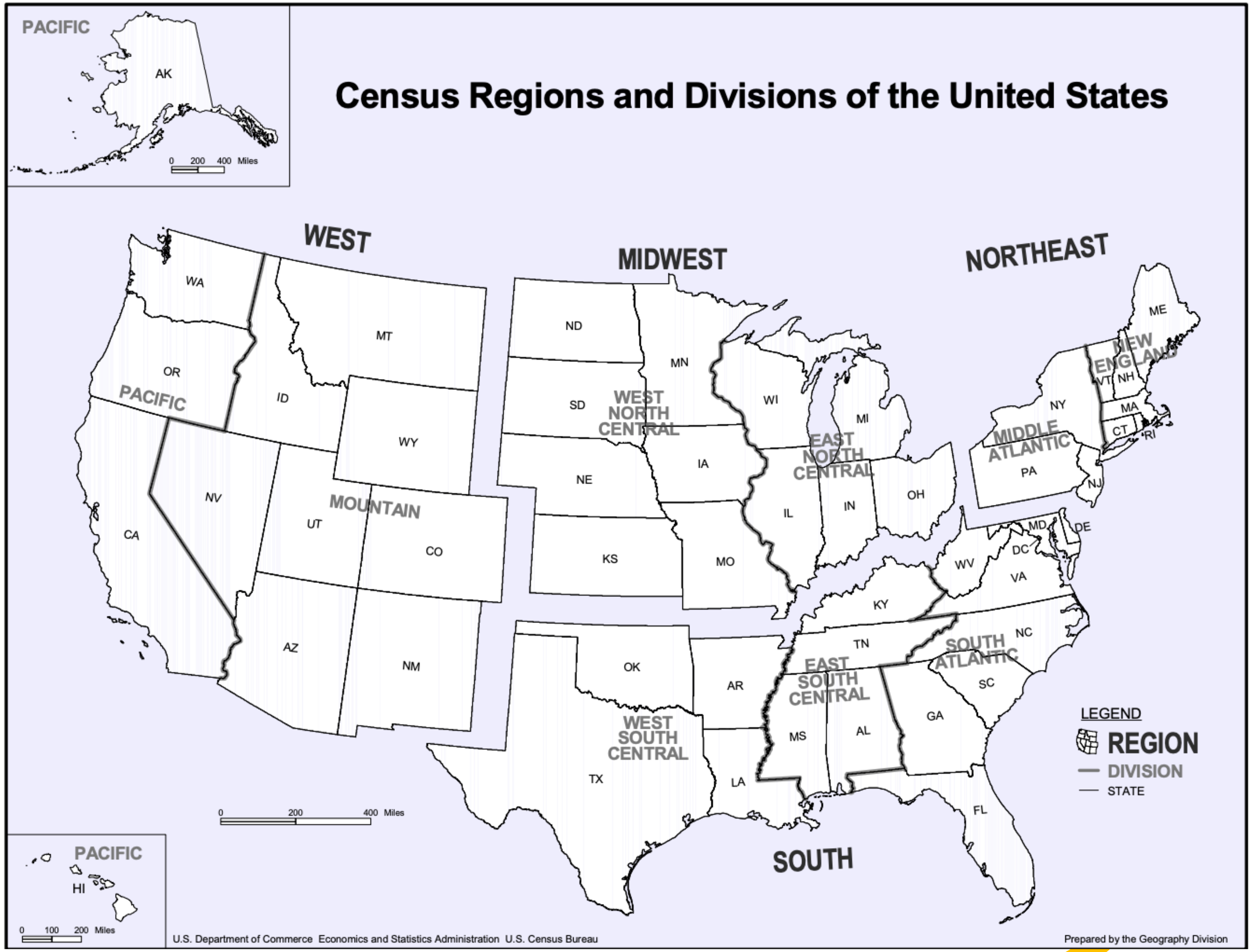


```
  predom_race          n
  <chr>            <int>
1 Asian               52
2 Black              531
3 Hispanic/Latino   1213
4 Native American    198
5 White             9916
6 NA                1404
```

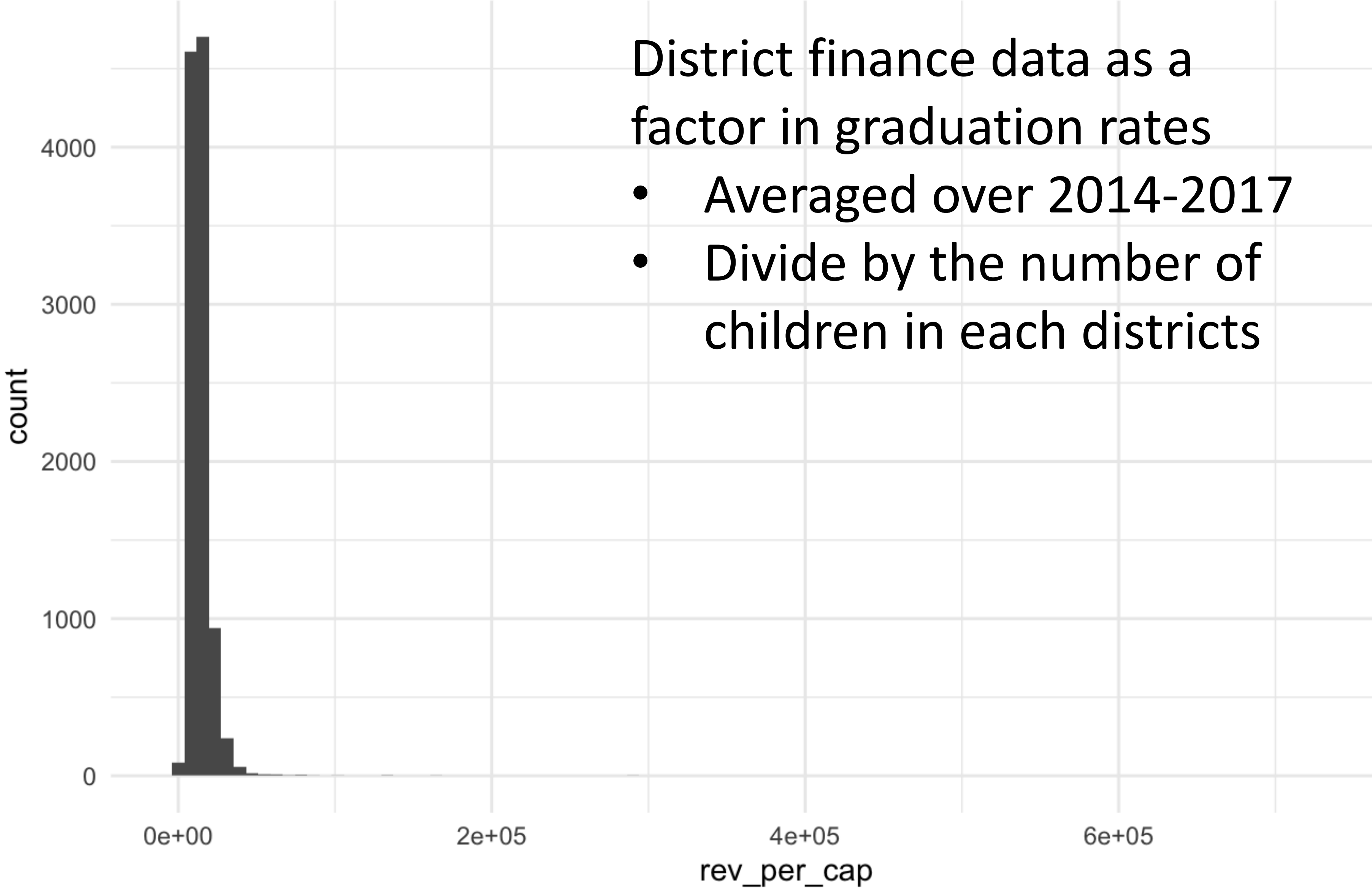


dist	children	predom_race
Fort Rucker School District	985	White
Maxwell AFB School District	292	White
Albertville City School District	4591	White
Marshall County School District	8299	White
Hoover City School District	15397	White
Madison City School District	9416	White
Leeds City School District	2324	White
Boaz City School District	1644	White
Trussville City School District	4476	White
Alexander City City School District	3064	White
Andalusia City School District	1425	White
Arab City School District	1454	White
Athens City School District	3824	White
Attalla City School District	1144	White
Saraland City School District	2728	White
Satsuma City School District	922	White
Alabaster City School District	6469	White
Pelham City School District	4044	White
Pike Road City School District	1813	White

Defining Regions



School District Funding

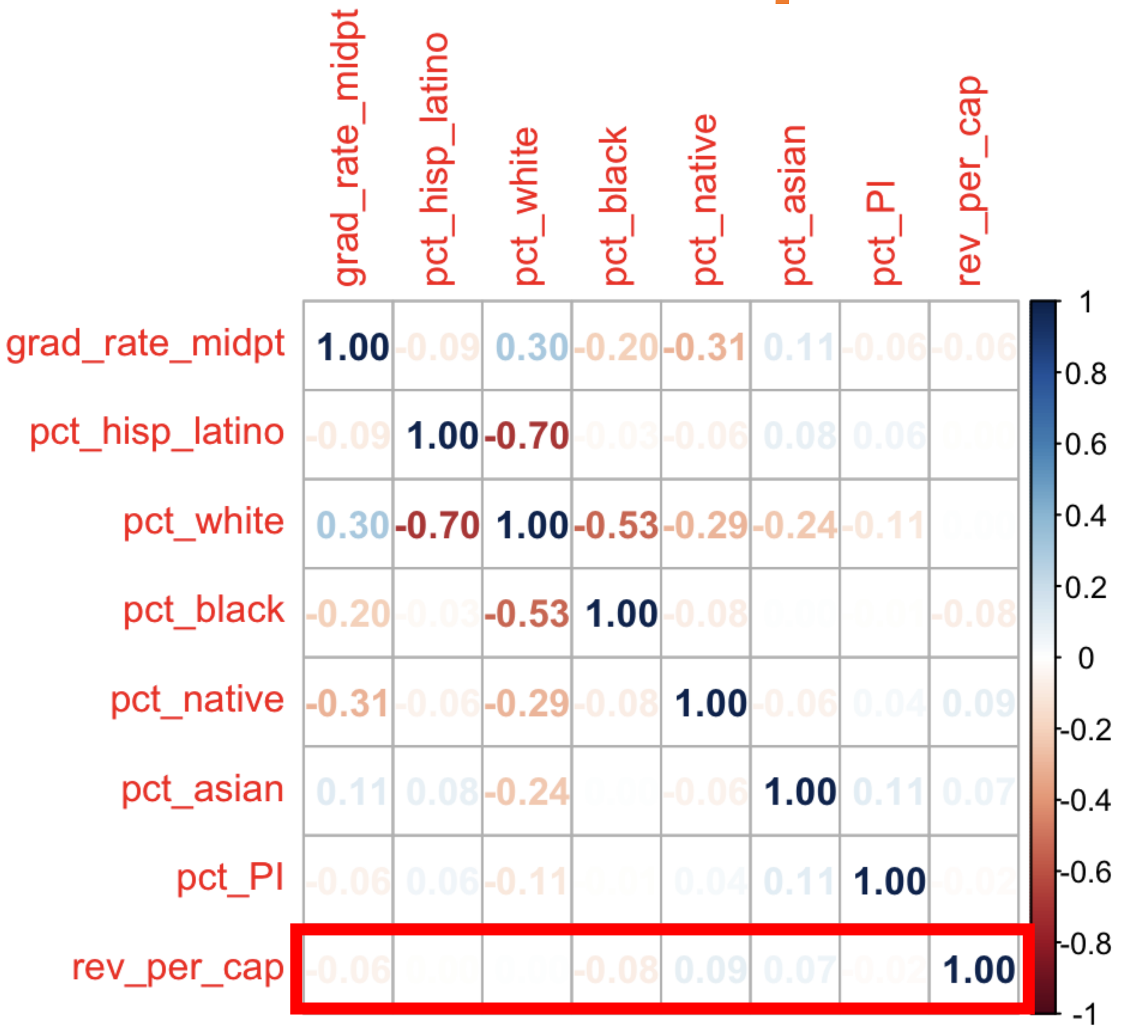


District finance data as a factor in graduation rates

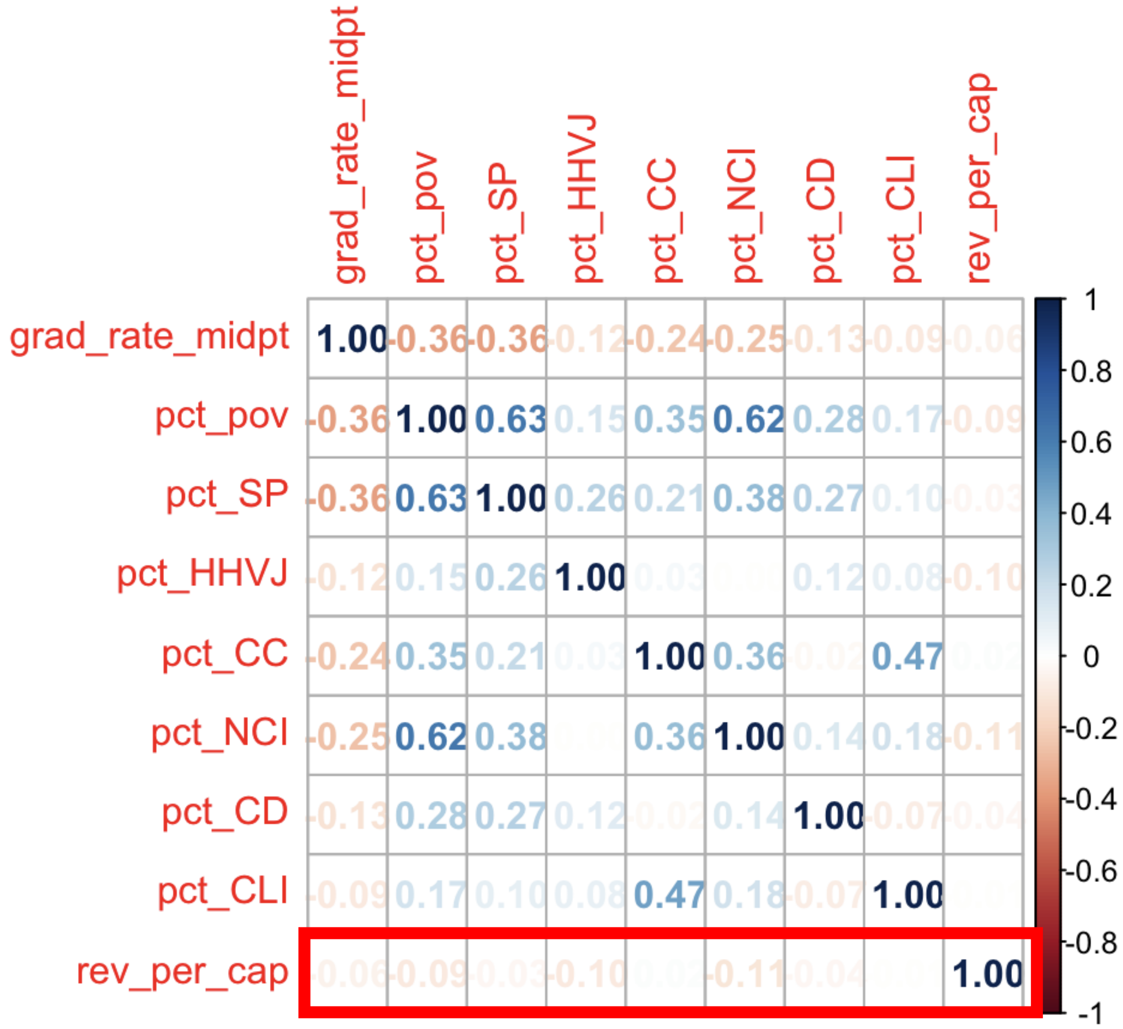
- Averaged over 2014-2017
- Divide by the number of children in each districts

Correlated?

no.



	grad_rate_midpt	pct_hisp_latino	pct_white	pct_black	pct_native	pct_asian	pct_PI	rev_per_cap
grad_rate_midpt	1.00	-0.09	0.30	-0.20	-0.31	0.11	-0.06	-0.06
pct_hisp_latino	-0.09	1.00	-0.70	-0.03	-0.06	0.08	0.06	0.00
pct_white	0.30	-0.70	1.00	-0.53	-0.29	-0.24	-0.11	[illegible]
pct_black	-0.20	-0.03	-0.53	1.00	-0.08	0.00	-0.01	-0.08
pct_native	-0.31	-0.06	-0.29	-0.08	1.00	-0.06	0.04	0.09
pct_asian	0.11	0.08	-0.24	0.00	-0.06	1.00	0.11	0.07
pct_PI	-0.06	0.06	-0.11	-0.01	0.04	0.11	1.00	-0.02
rev_per_cap	-0.06	0.00	[illegible]	-0.08	0.09	0.07	-0.02	1.00



	grad_rate_midpt	pct_pov	pct_SP	pct_HHVJ	pct_CC	pct_NCI	pct_CD	pct_CLI	rev_per_cap
grad_rate_midpt	1.00	-0.36	-0.36	-0.12	-0.24	-0.25	-0.13	-0.09	-0.06
pct_pov	-0.36	1.00	0.63	0.15	0.35	0.62	0.28	0.17	-0.09
pct_SP	-0.36	0.63	1.00	0.26	0.21	0.38	0.27	0.10	-0.03
pct_HHVJ	-0.12	0.15	0.26	1.00	0.03	0.00	0.12	0.08	-0.10
pct_CC	-0.24	0.35	0.21	0.03	1.00	0.36	-0.02	0.47	0.02
pct_NCI	-0.25	0.62	0.38	0.00	0.36	1.00	0.14	0.18	-0.11
pct_CD	-0.13	0.28	0.27	0.12	-0.02	0.14	1.00	-0.07	-0.04
pct_CLI	-0.09	0.17	0.10	0.08	0.47	0.18	-0.07	1.00	[illegible]
rev_per_cap	-0.06	-0.09	-0.03	-0.10	0.02	-0.11	-0.04	[illegible]	1.00

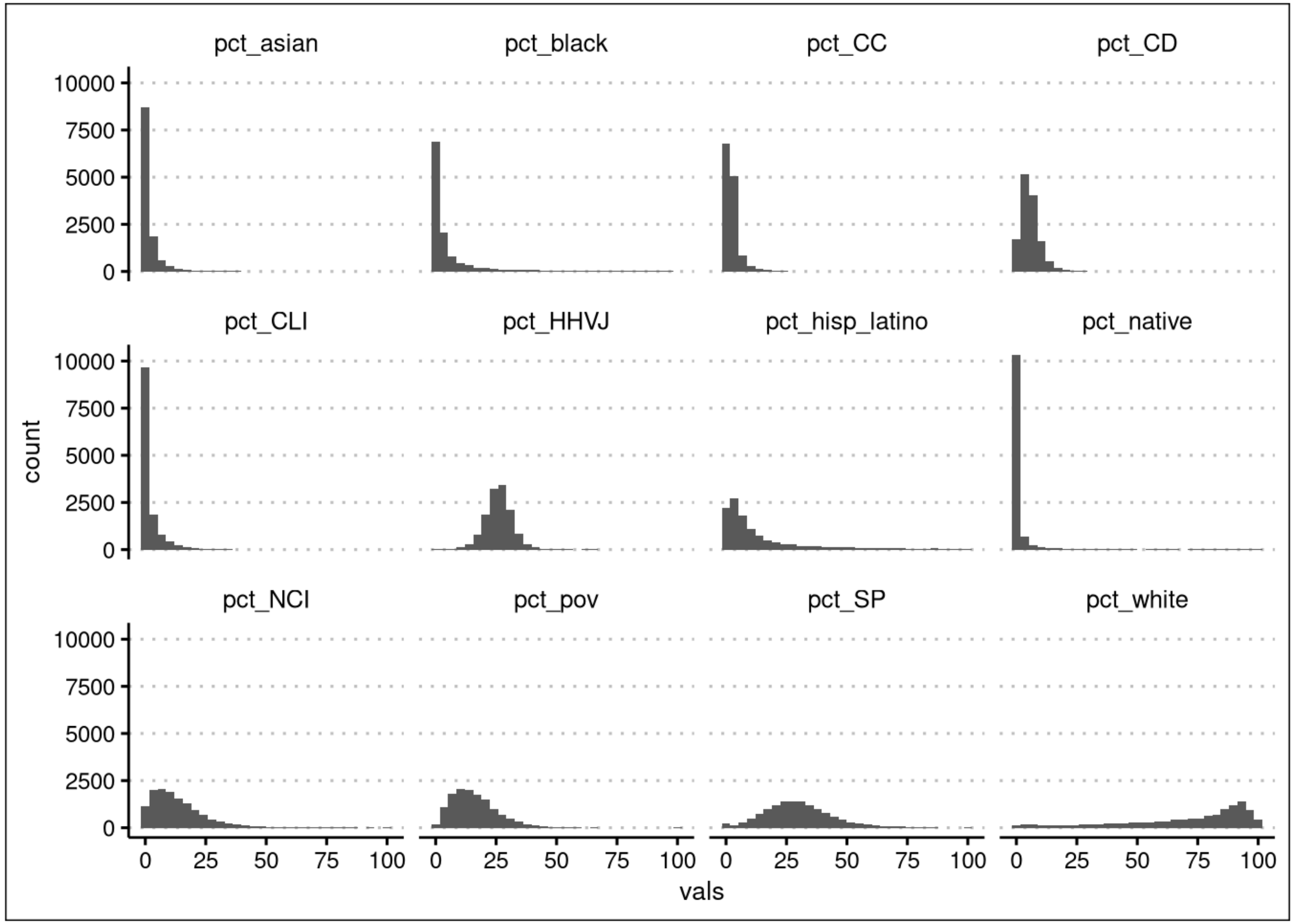
Regression!

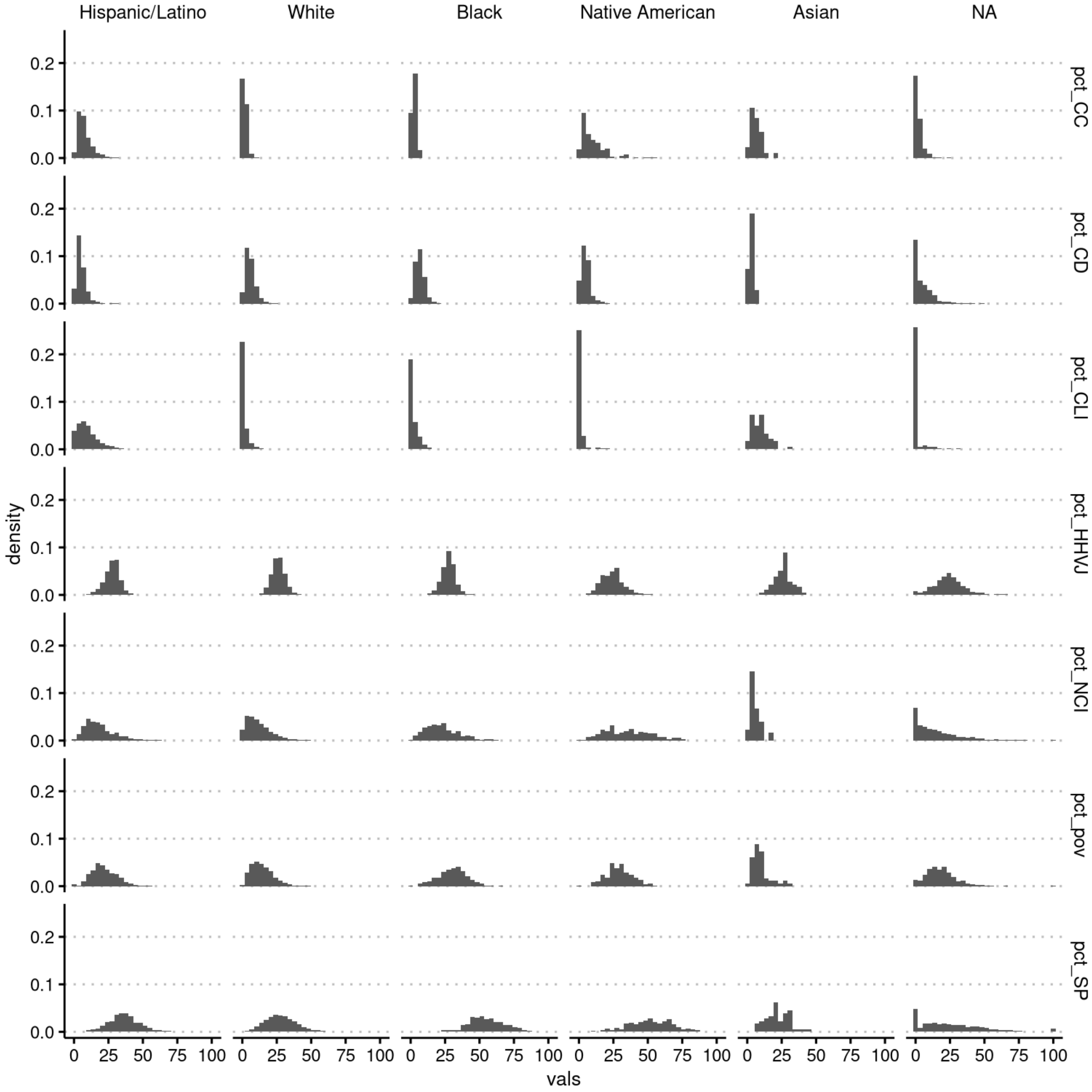
Linear Regression:

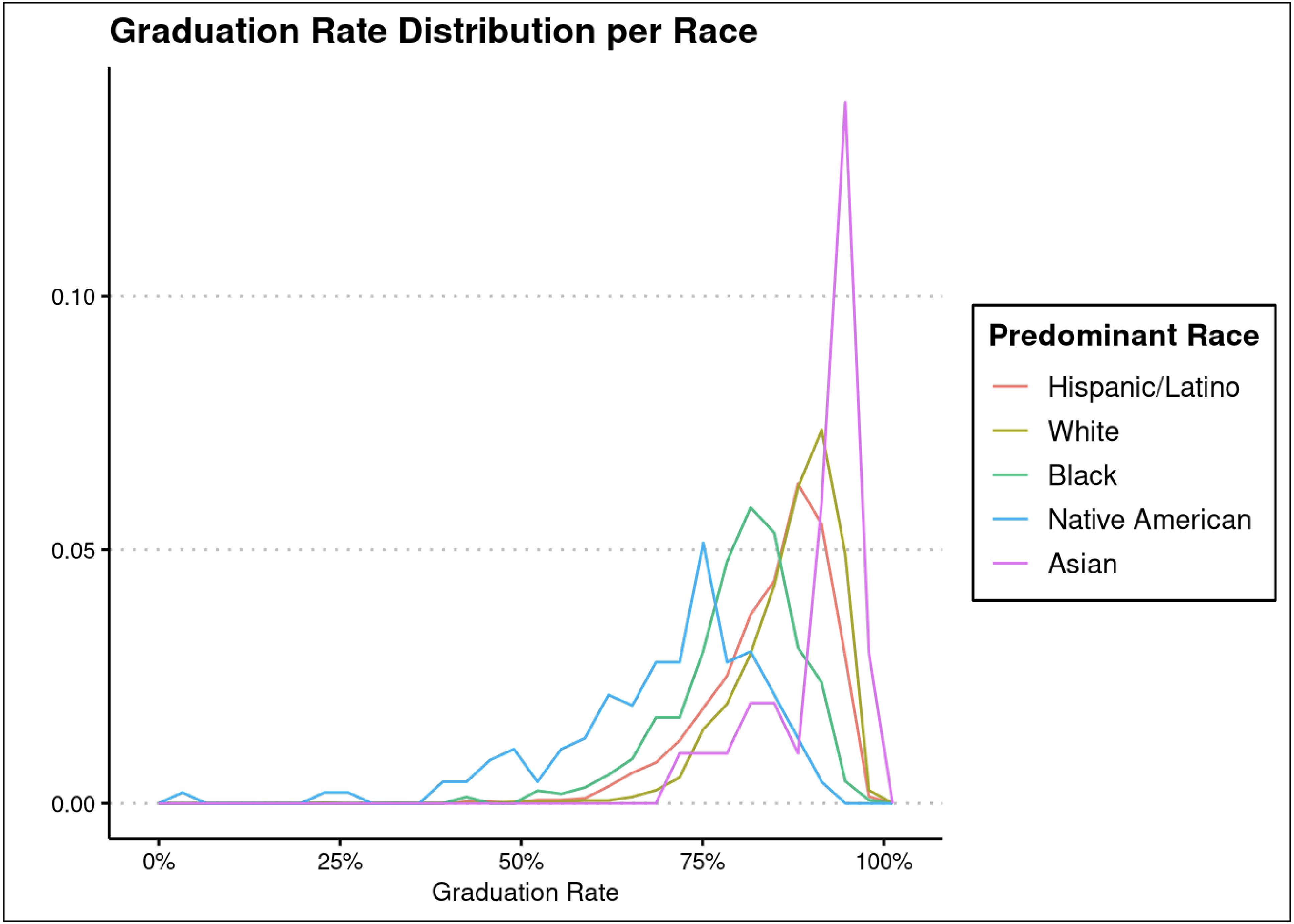
1. Household Condition Data
2. Race Data
    1. %'s of Hispanic/Latino, White, Black, Native American, Asian, and Pacific Islander
3. Combined

ANOVA to test whether adding race data significantly improves the RSS

Indicator distributions




Hispanic/Latino
White
Black
Native American
Asian
NA
pct_CC
pct_CD
pct_CLI
pct_HHVJ
pct_NCI
pct_pov
pct_SP
density
0.0
0.1
0.2
vals
0
25
50
75
100


Graduation Rate Distribution per Race
0.10
0.05
0.00
0%
25%
50%
75%
100%
Graduation Rate
Predominant Race
Hispanic/Latino
White
Black
Native American
Asian

Regression Moving Forward

- New tool: **Tidy modeling! (tidymodels)**
  - We are trying different regression models beyond just linear models to determine which one makes the most sense.
  - Jon started running models last night which took a while to run.

Regression Moving Forward

- Preprocessing Steps:

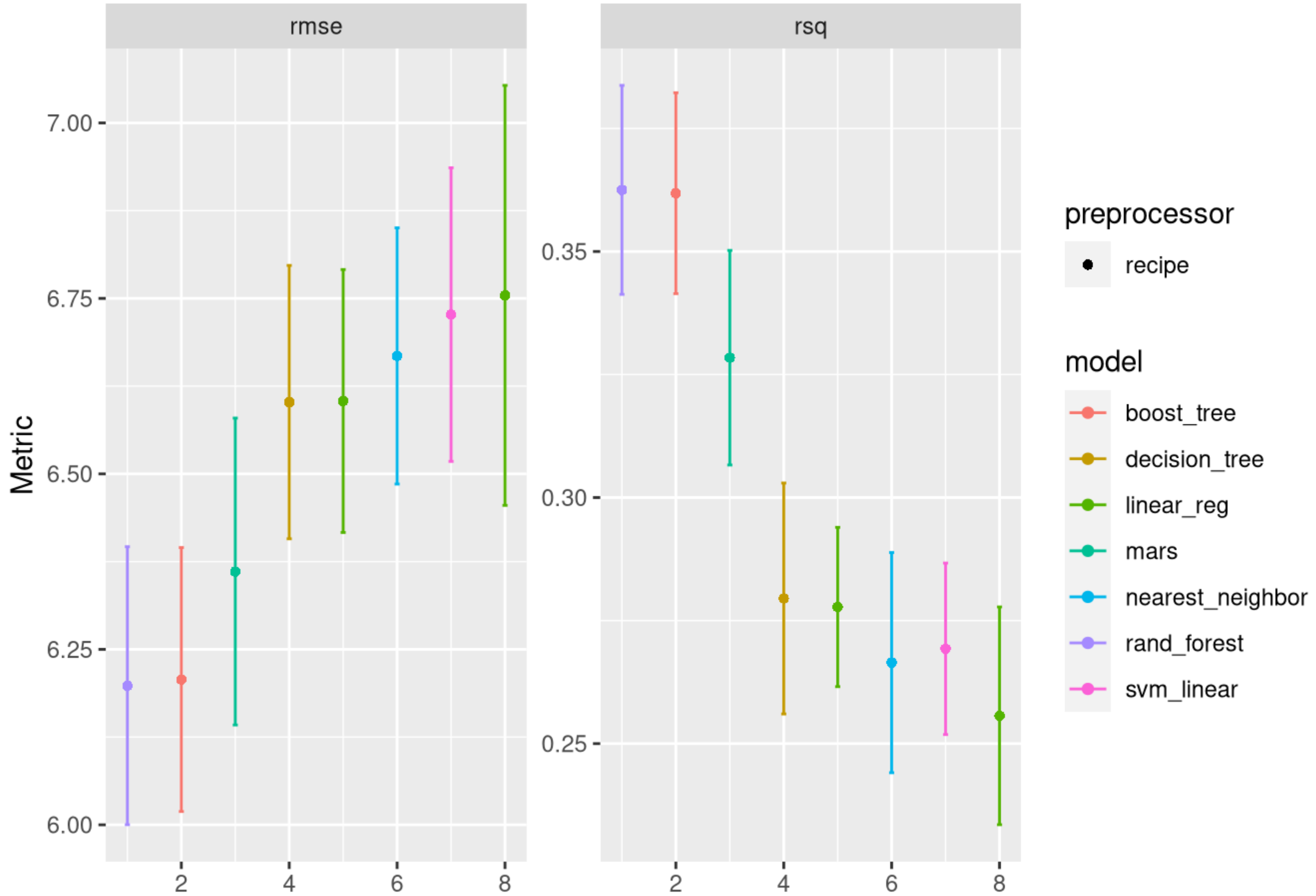
1. Made interaction terms between HH conditions and Racial data
2. Made dummy variables for predominant race
3. Centered and Scaled predictors
4. Removed predictors with near-zero variance

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}
- K-Nearest Neighbors {kknn}

root mean squared error



Model Rank by Performance

Model Results:

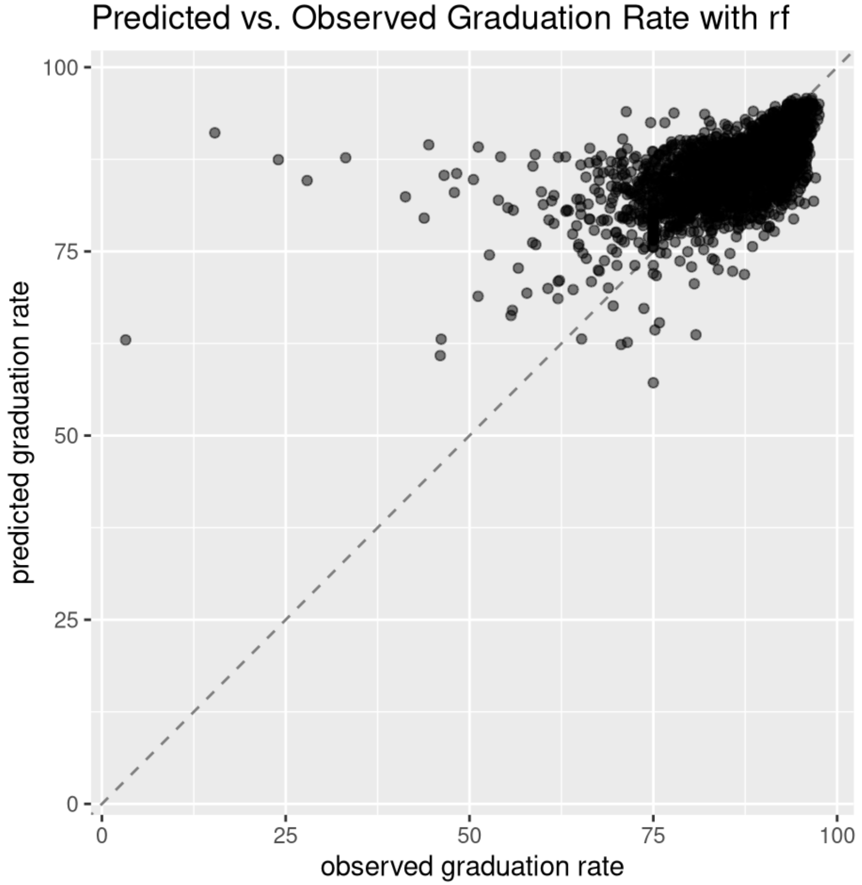
- Test RMSE: 6.77
- Test R^2: 0.319

Model Parameters:

- mtry*: 16
- min_n**: 39

*mtry: The number of predictors that will be randomly sampled at each split when creating the tree models.

**min_n: The minimum number of data points in a node that are required for the node to be split further.



Next Steps…

- Add variables to regression model (include financial information)
- Determine which type of model is most appropriate – change by region, race
- Measure most impactful variables for regression (Societal Implications)
- Assessment data