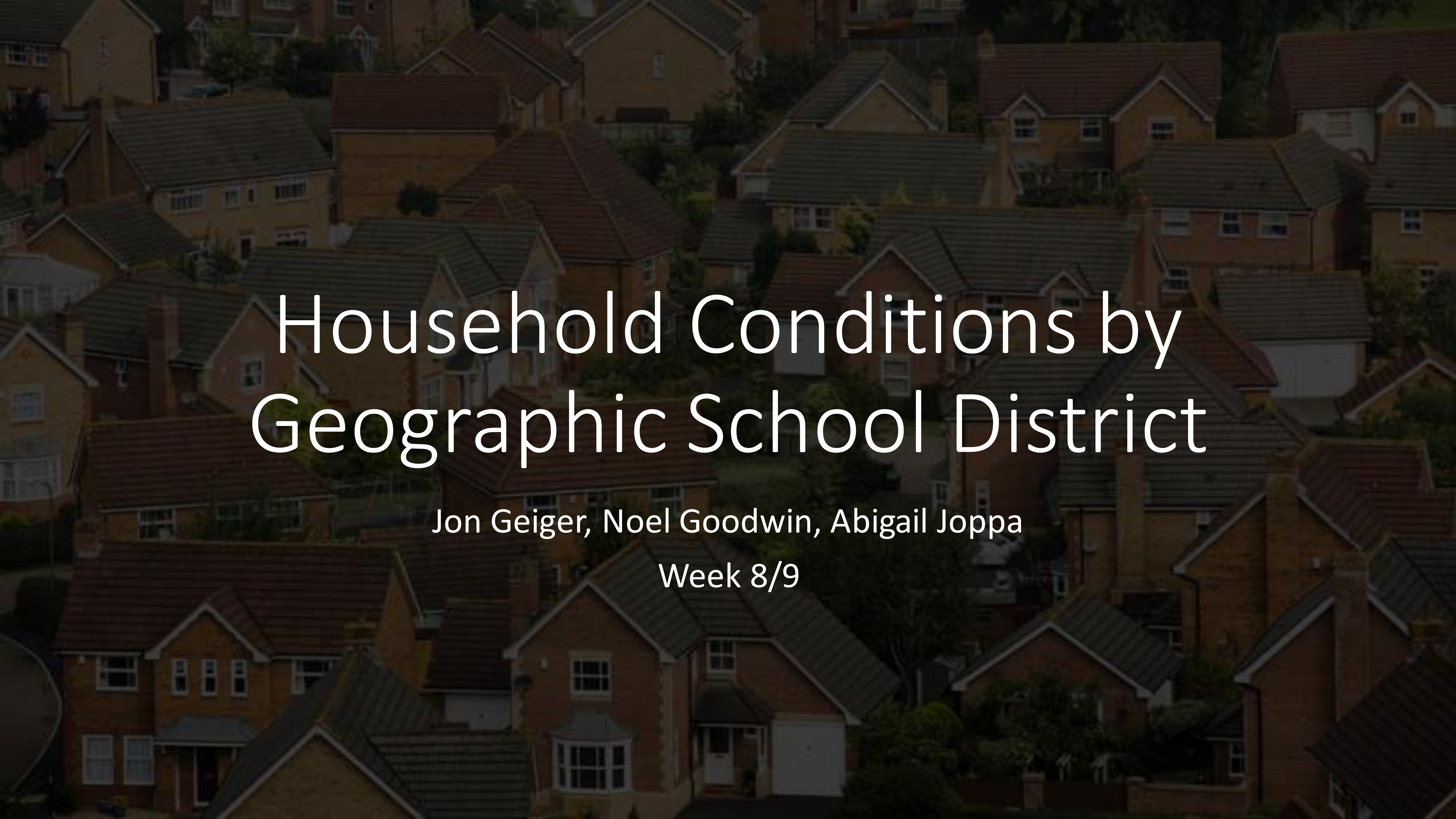


Household Conditions by Geographic School District

Jon Geiger, Noel Goodwin, Abigail Joppa

Week 8/9

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 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

Adding Variables to Model

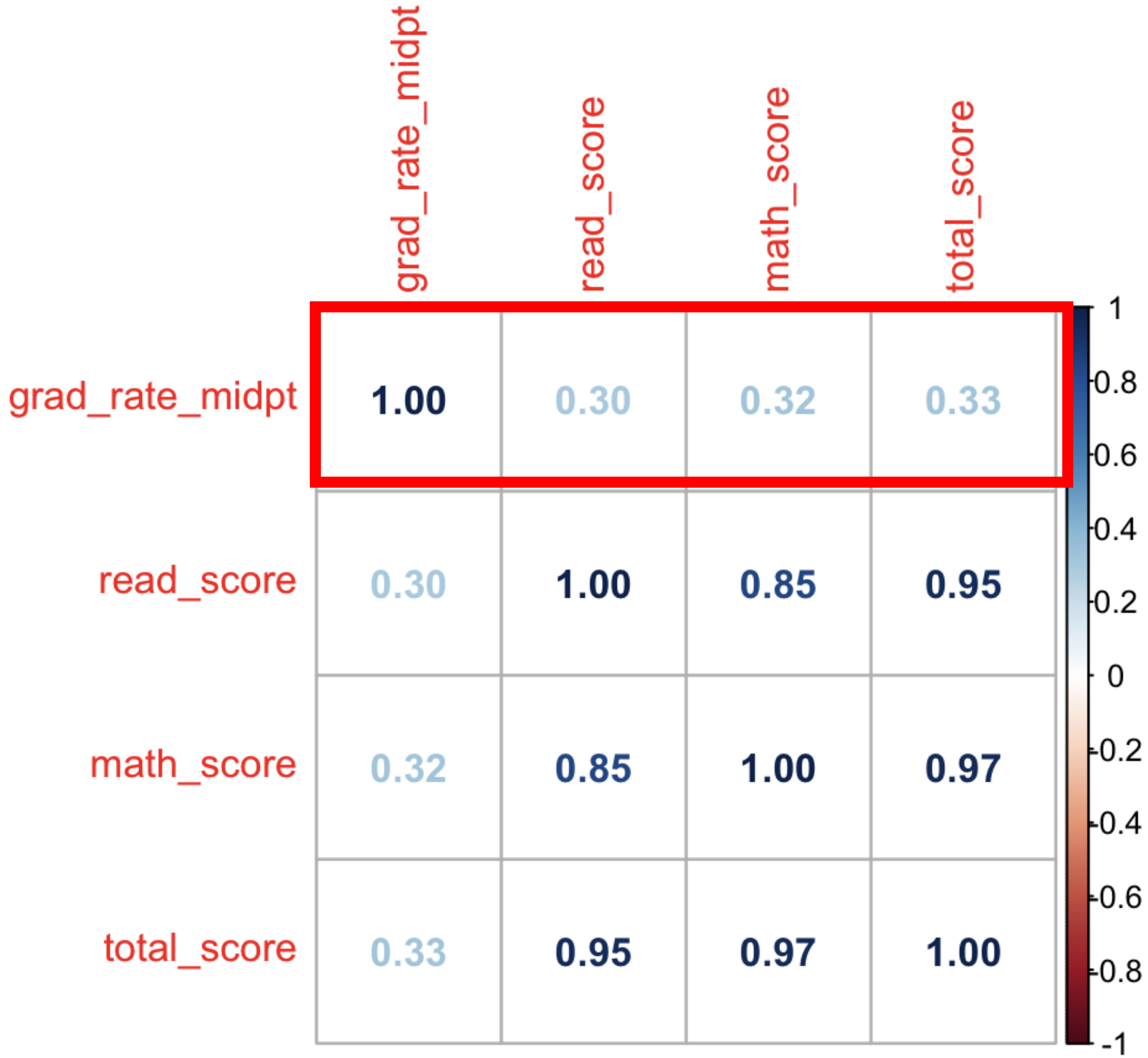
Current Variables:

- Household Conditions
- Racial/Ethnic Distribution
- Finance Data

Variables to add:

- Assessment Data
- Existing state-wide graduation rates?

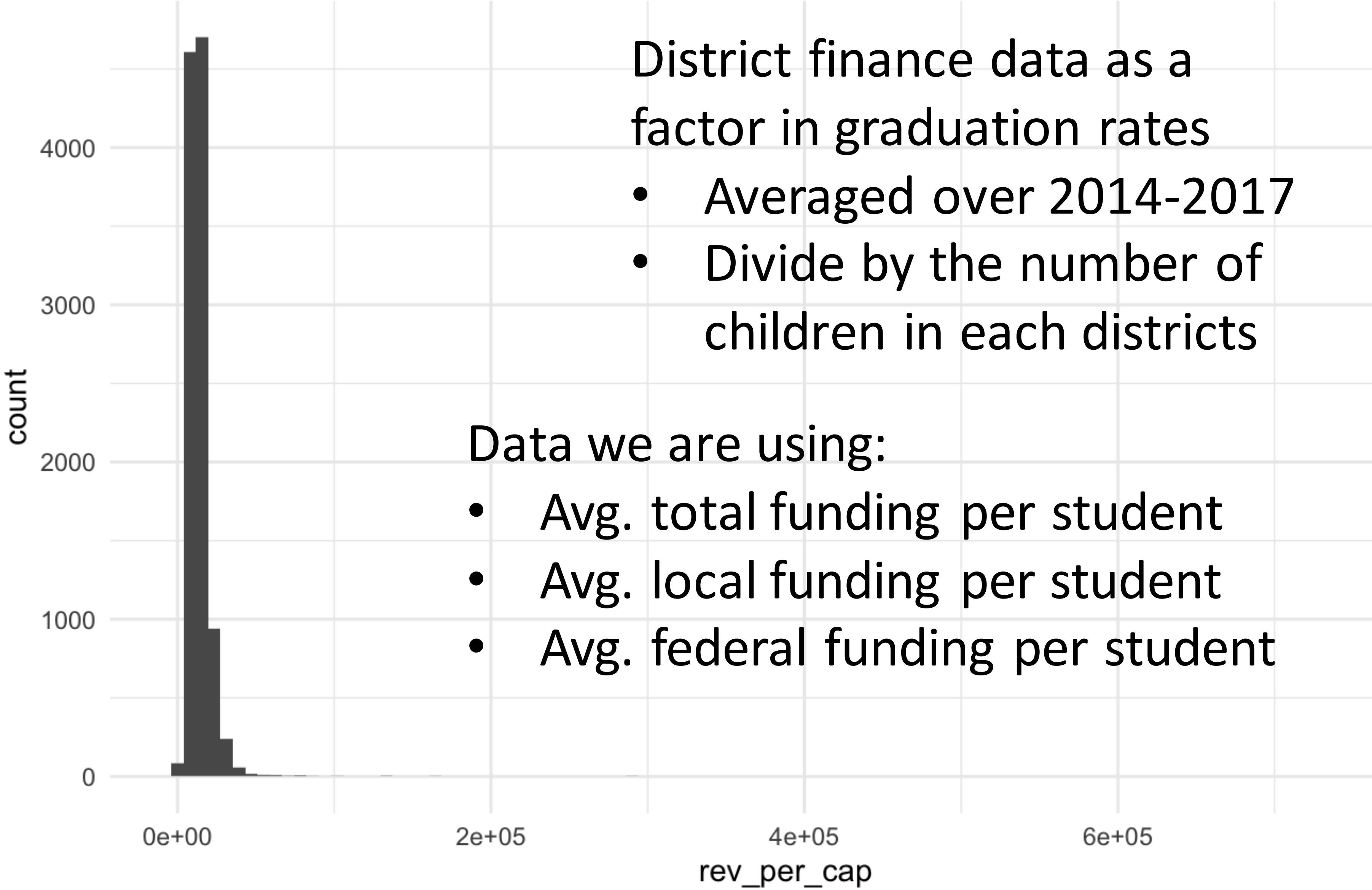
Assessments and Grad Rates



	grad_rate_midpt	read_score	math_score	total_score
grad_rate_midpt	1.00	0.30	0.32	0.33
read_score	0.30	1.00	0.85	0.95
math_score	0.32	0.85	1.00	0.97
total_score	0.33	0.95	0.97	1.00

Source: edfacts

School District Funding



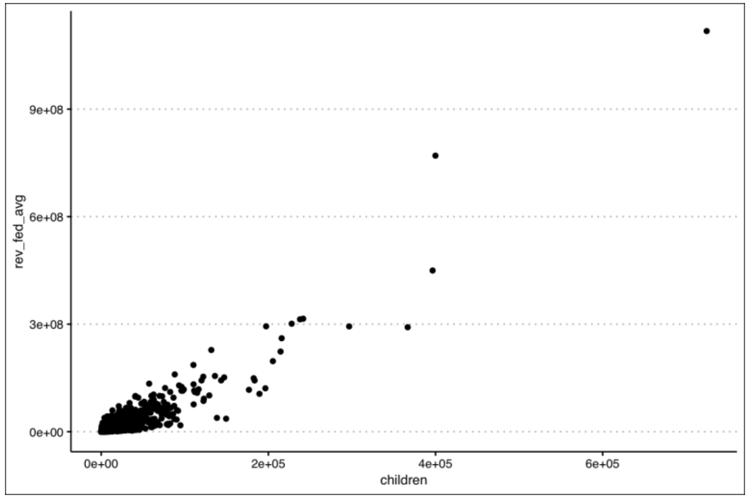
District finance data as a factor in graduation rates

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

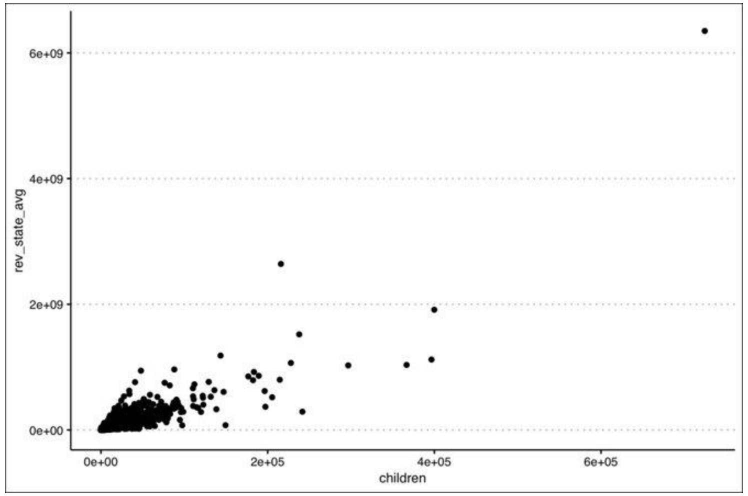
Data we are using:

- Avg. total funding per student
- Avg. local funding per student
- Avg. federal funding per student

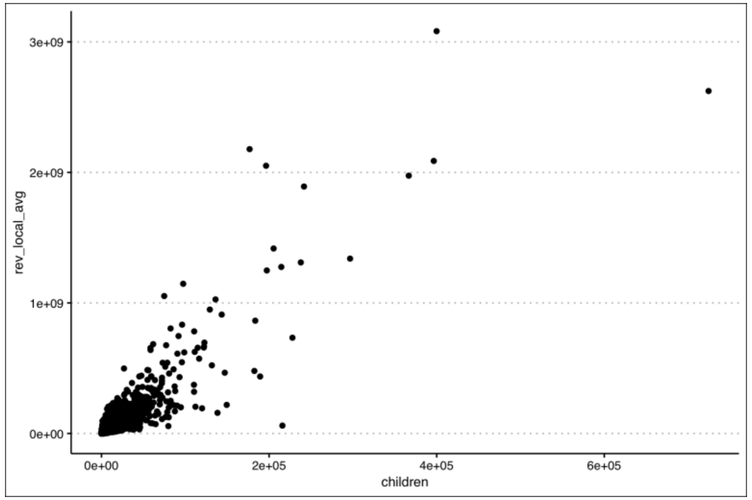
Federal Funding



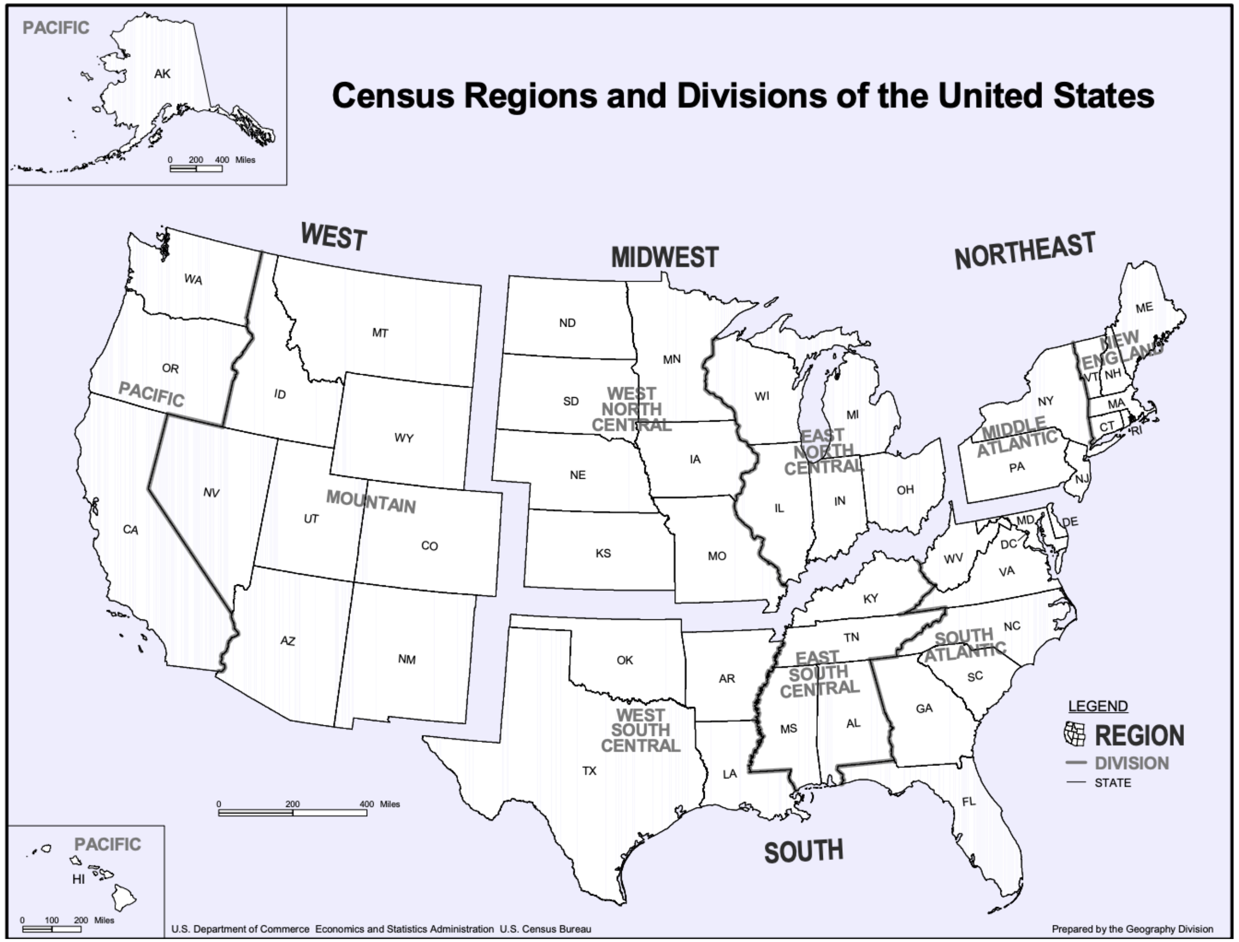
State Funding



Local Funding



Defining Regions

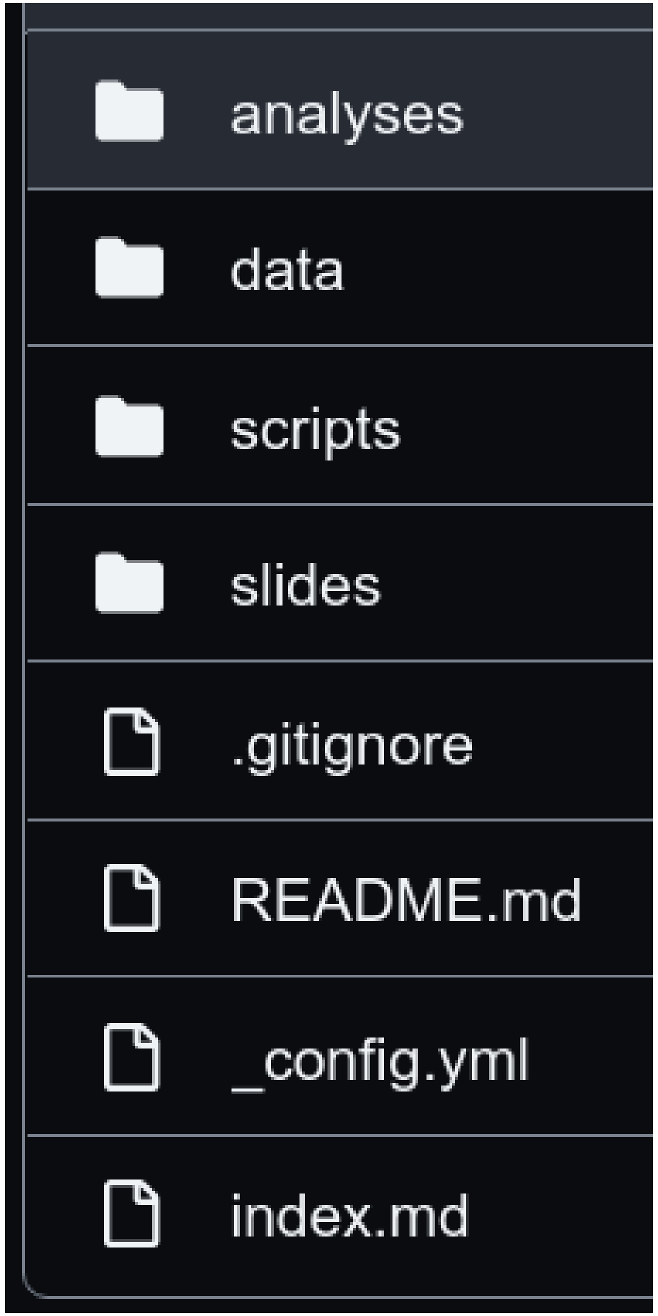


R "State" dataset

- Categorized each district to a region (West, South, North Central, and Northeast)

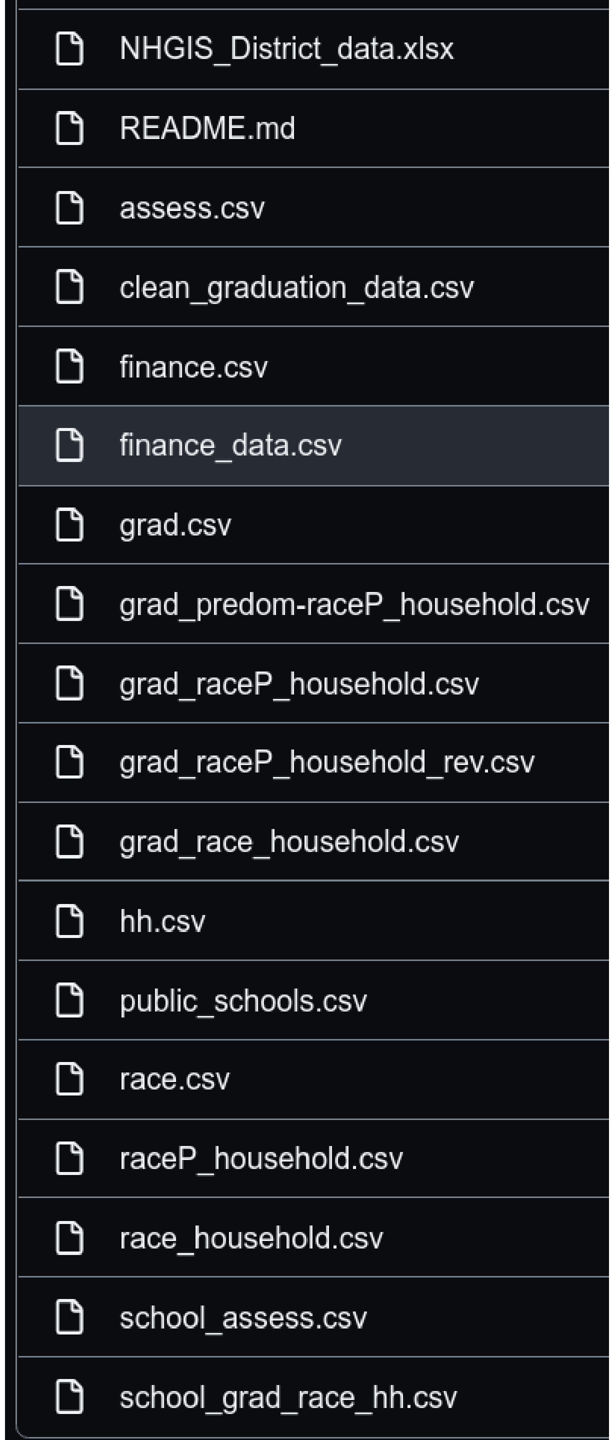
Data Files!

Our Repo



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

Our Data Directory



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

The data sets we actually have:

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

Goal:
Create a sort of "CSV Database" where we can join data sets by LEAID before analysis

Modeling!

Current Preprocessing Steps

1. Made interaction terms between numeric predictors

2. Made dummy variables for predominant race and region

3. Centered and Scaled predictors

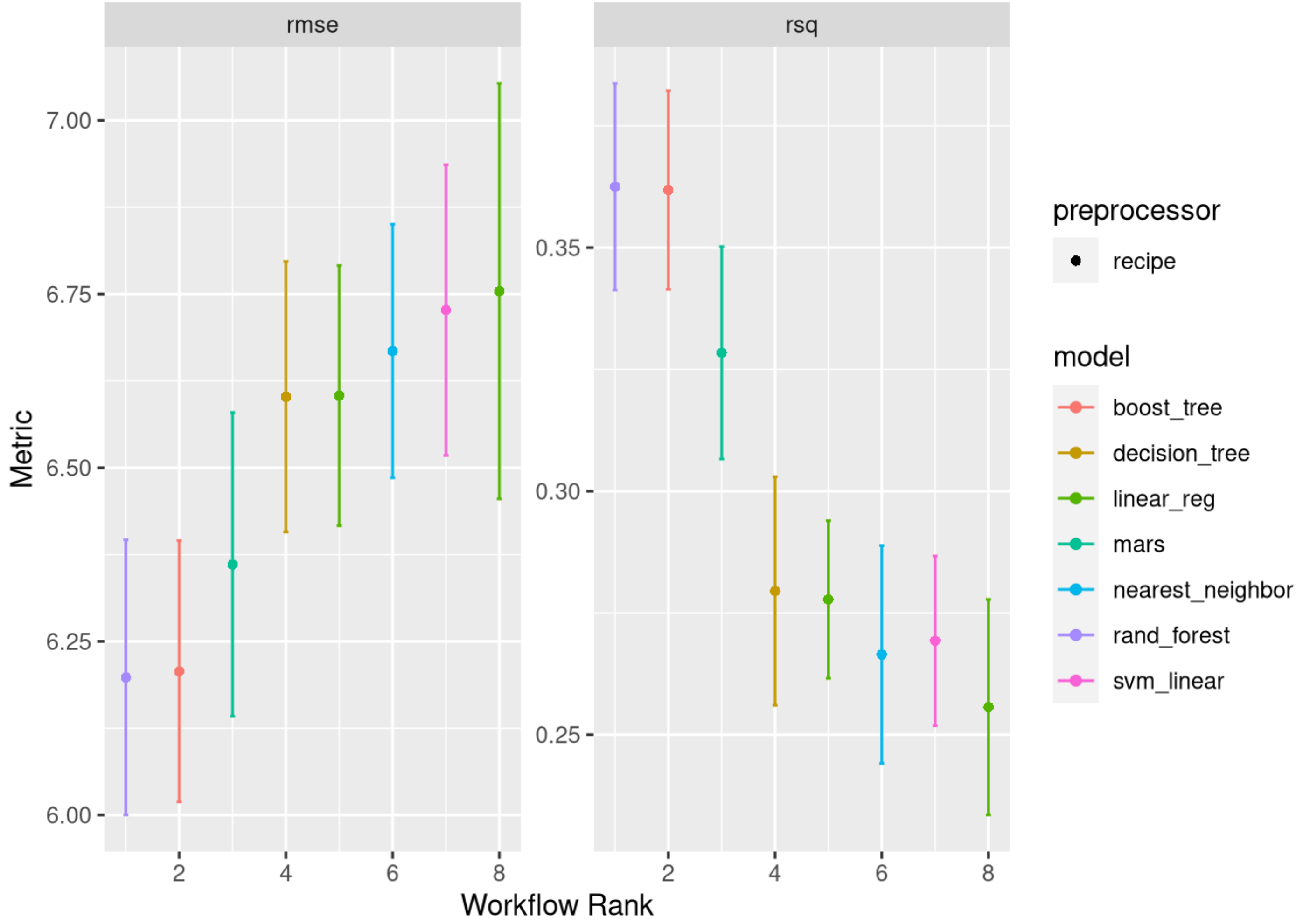
4. Removed predictors with near-zero variance

Models

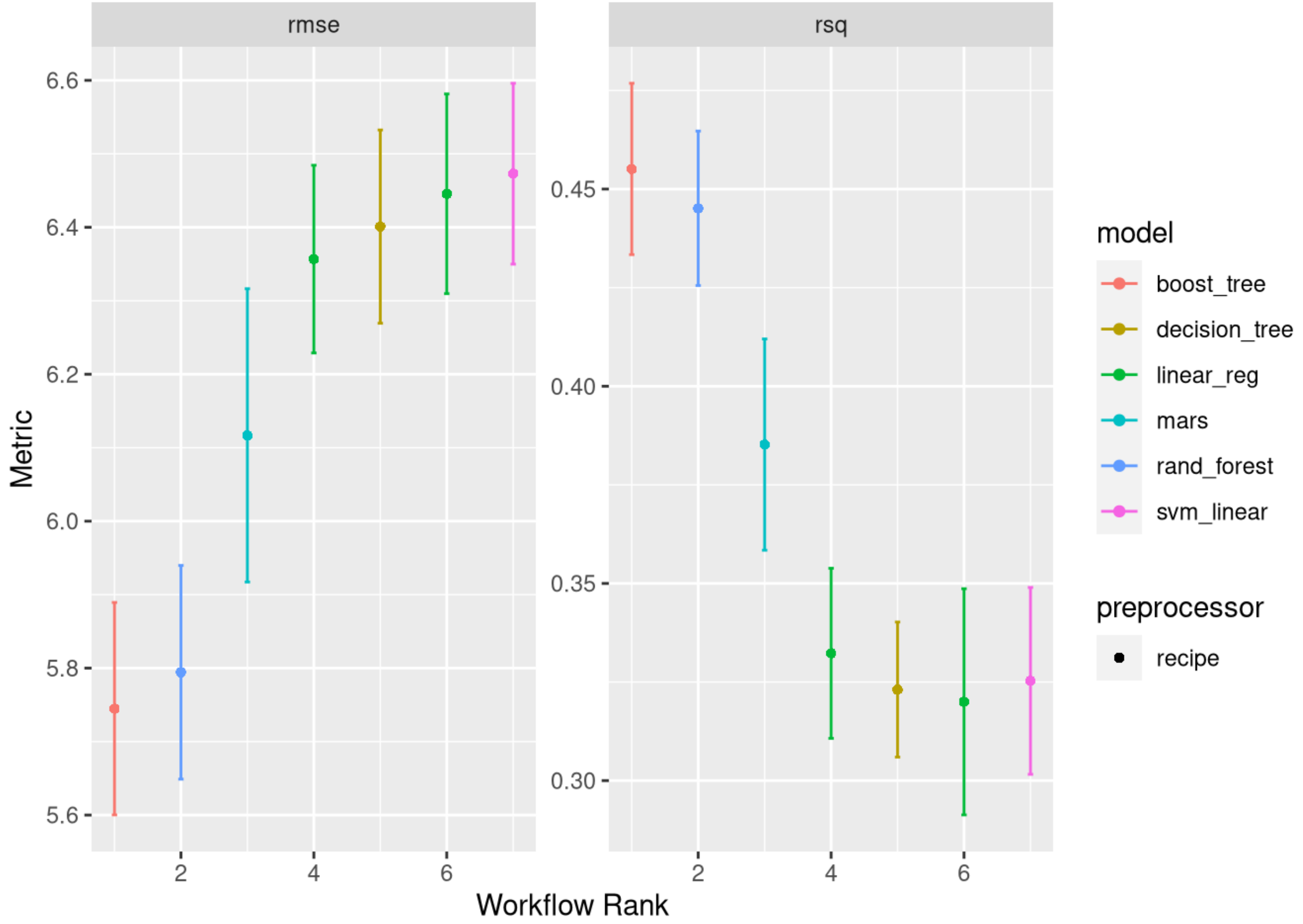
Tuned parameters with 10-fold cross-validation and random grid search

- 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}

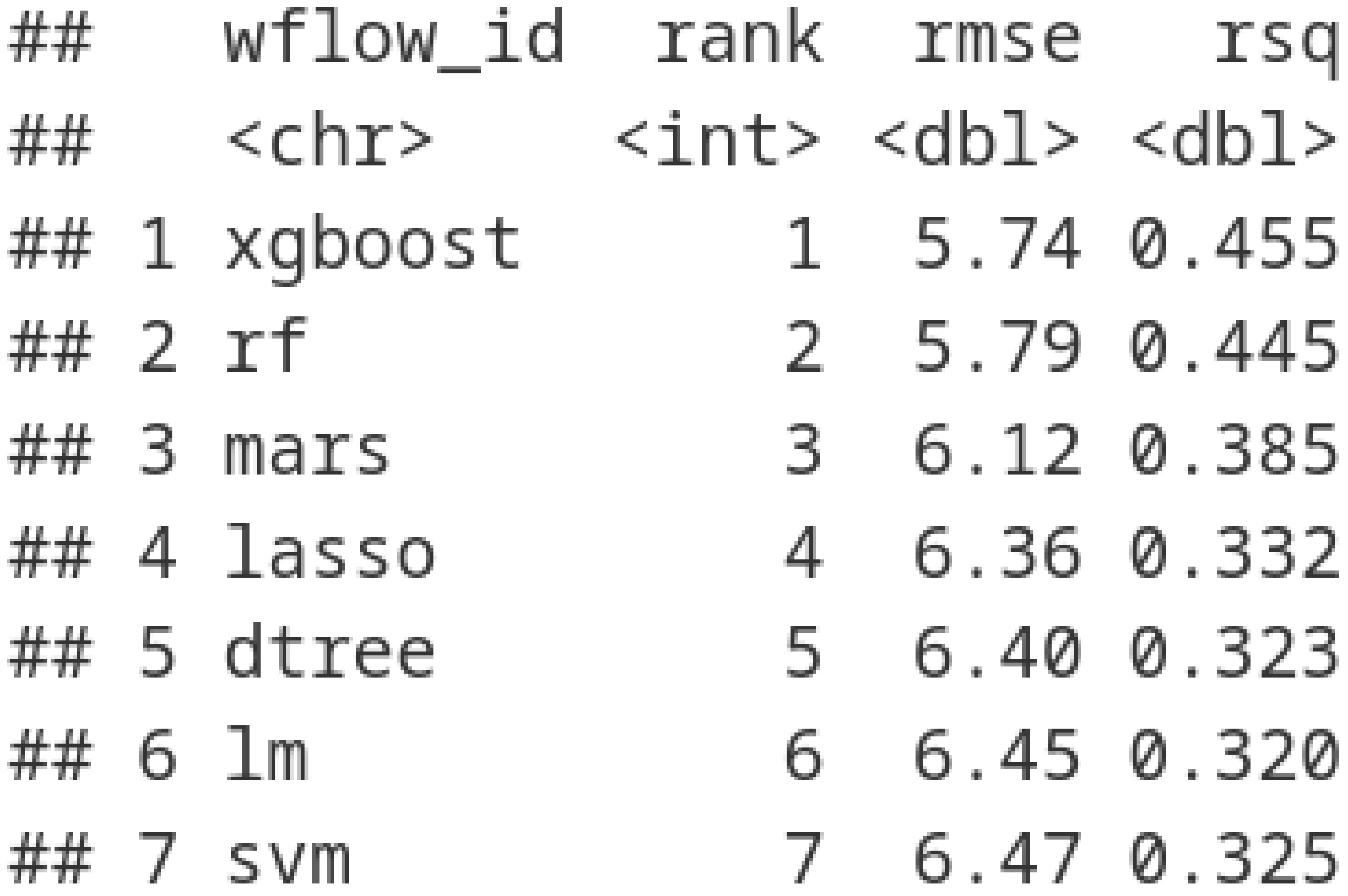
Last week's results



This week's results



Result Specifics



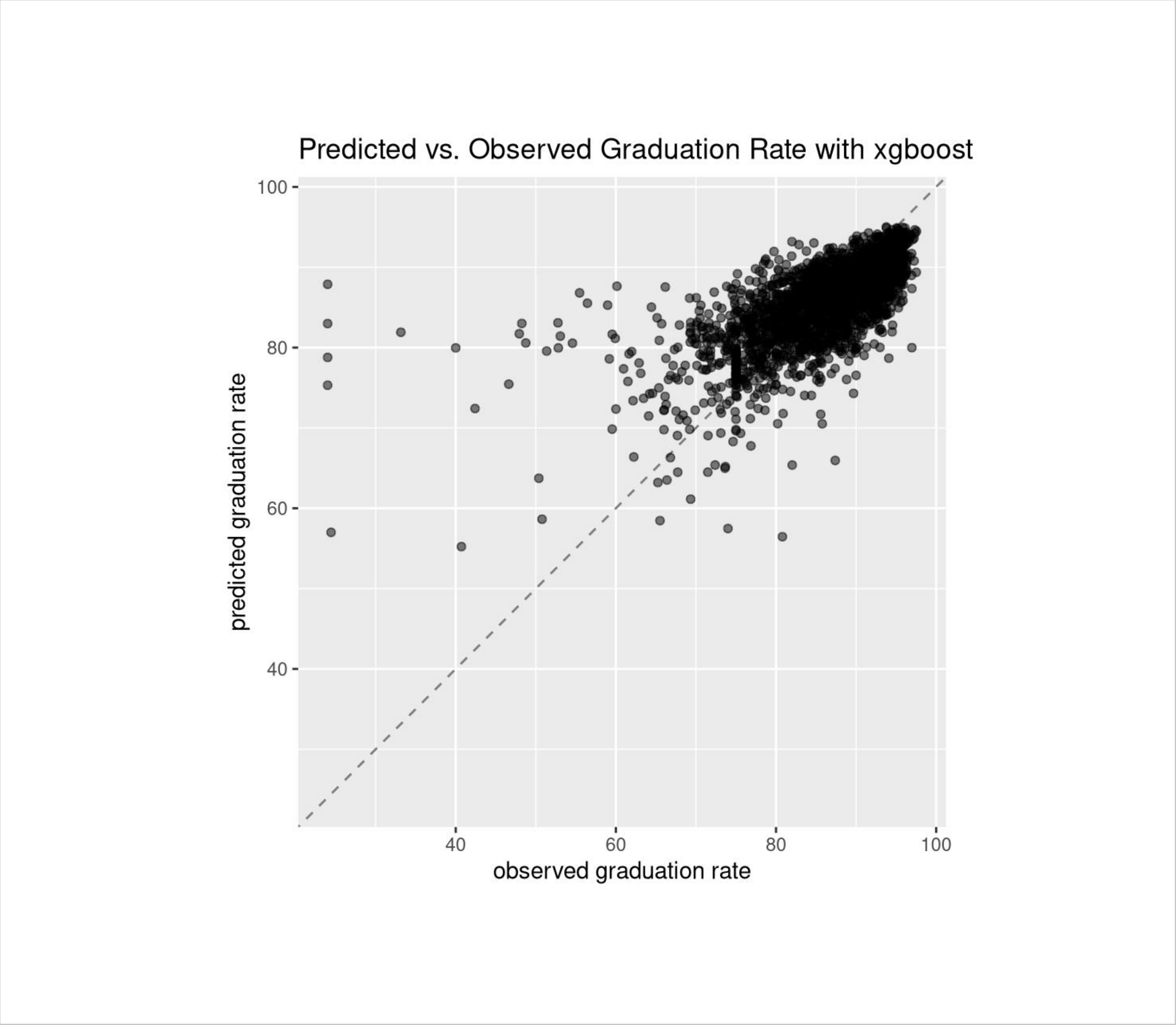
```
##   wflow_id  rank  rmse   rsq
##   <chr>    <int> <dbl> <dbl>
## 1 xgboost      1  5.74 0.455
## 2 rf           2  5.79 0.445
## 3 mars         3  6.12 0.385
## 4 lasso        4  6.36 0.332
## 5 dtree        5  6.40 0.323
## 6 lm           6  6.45 0.320
## 7 svm          7  6.47 0.325
```

Model Results:

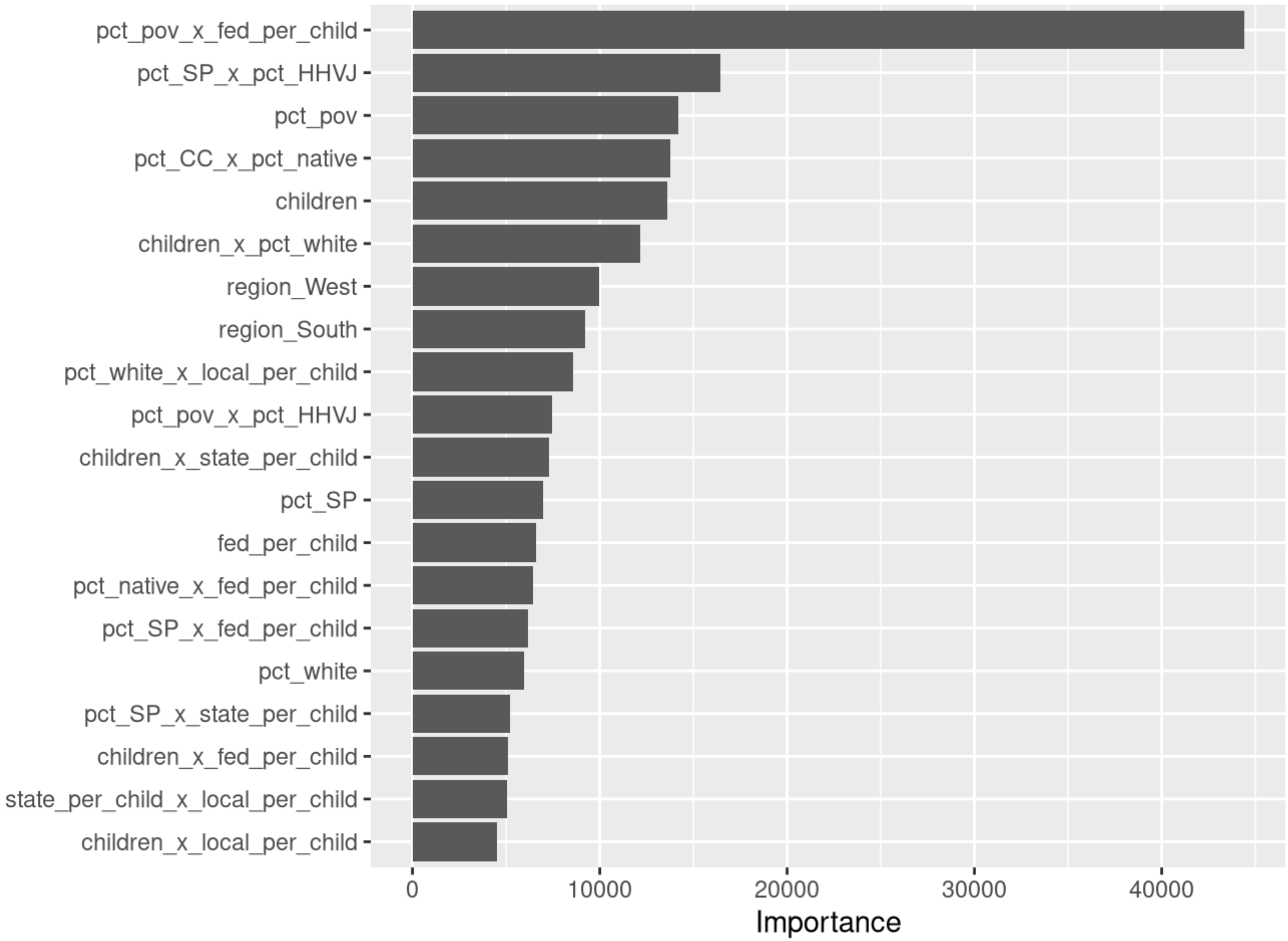
- Test RMSE: 6.77 --> 5.98
- Test R^2: 0.319 --> 0.451

Model Parameters:

- Mtry = 19
- Min_n = 13
- Tree_depth = 11
- Learn_rate = 0.00565
- Sample_size = 0.960



Variable Importance for Random Forest



Next Week's Steps

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