Prediction of Scores for Public Schools in California

Ahrim Han, Ph.D.
June 22 2019

Contents

- Introduction
- Data Wrangling
- Data Visualization
- Exploratory Data Analysis
- Machine Learning Modeling
- Recommendations
- Conclusion

Introduction

- California Assessment of Student Performance and Progress (CAASPP)
    - Measure how well students are achieving academic standards in English language arts/literacy and mathematics

Problem Statement

- Strong need to find more informed and granular causes that impact the test achievements of schools
- We aim to predict and find the inferior groups of schools that indeed need help
  - Schools should strive to create an environment where all students feel valued and all students are learning to high standards

Expected Beneficiaries

- Administrators of the school districts/state departments of education or other organizations
  - Can allocate budgets and human resources for tutoring, mentoring, extracurricular programs, and educational consultants
- Teachers
  - Can put much more effort into the under-performing groups to reduce the achievement gaps
- Parents
  - Can select a good school that meets the high academic standards

Data Wrangling

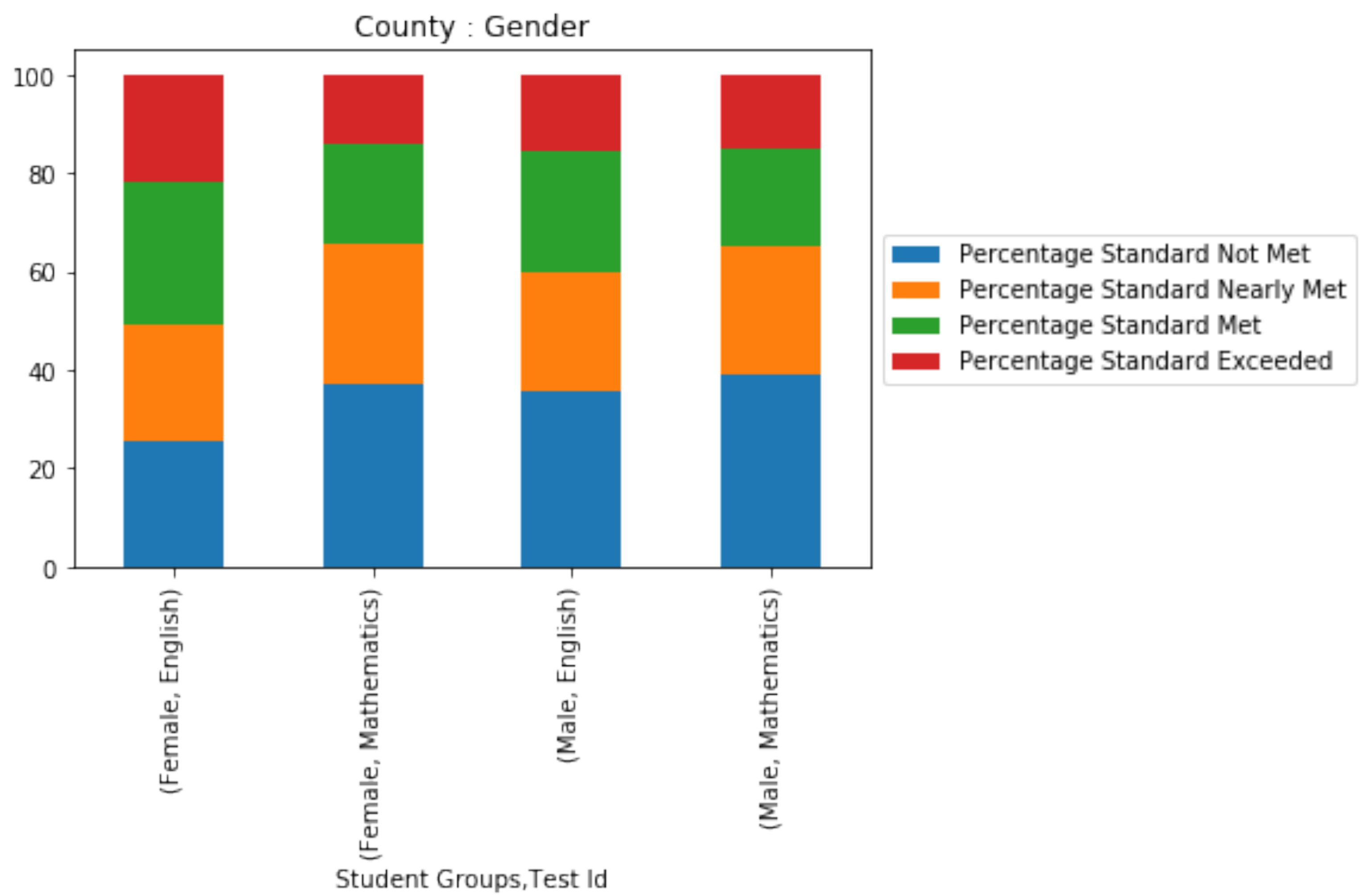
- Collecting and cleaning data
  - CAASPP test score data in 2018 (California Department of Education)
  - House prices (Zillow research data)
- Fixing missing values
  - Imputed using the statistics of the *mean* of each column in which the missing values are located
- Adding new variables
  - By manipulating or merging existing variables to tell new insights or to reduce the dimensionality

Data Visualization

- Research Questions
  - RQ1. How the students are different in achievement levels?
    - Compared for each category of gender, ethnicity, English-language fluency, economic status, disability status, and parent educations using the bar plots
  - RQ2. What features can you find in the top and bottom performance groups?
    - Compared the best and worst 10% performing counties using the bar plots
  - RQ3. Are house prices correlated to the exceeded scores or the inferior scores?
    - Analyzed the correlations using scatterplots

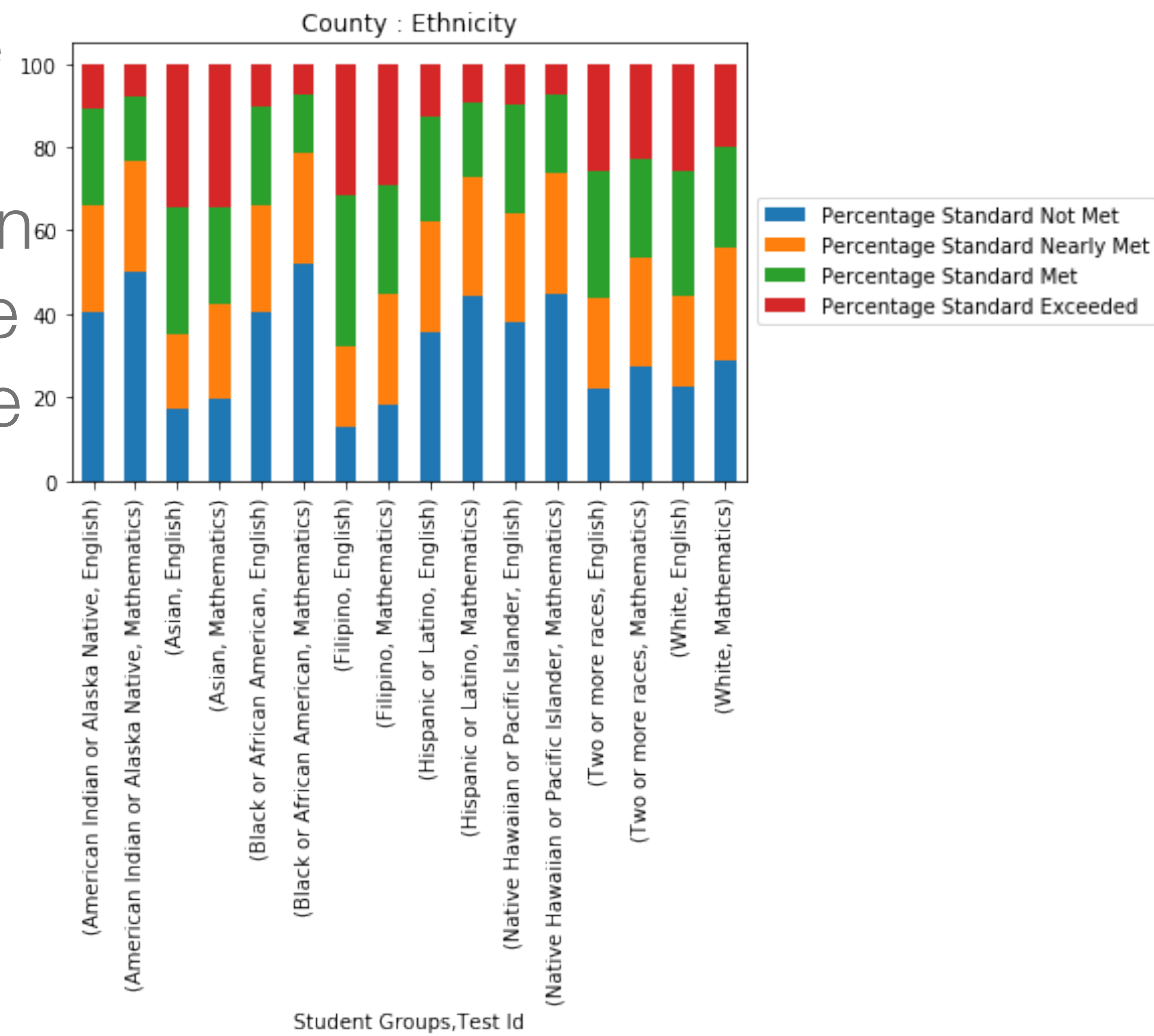
Achievement Levels by Gender

- Female students exceed male students in English, while male students exceed female students in Mathematics.



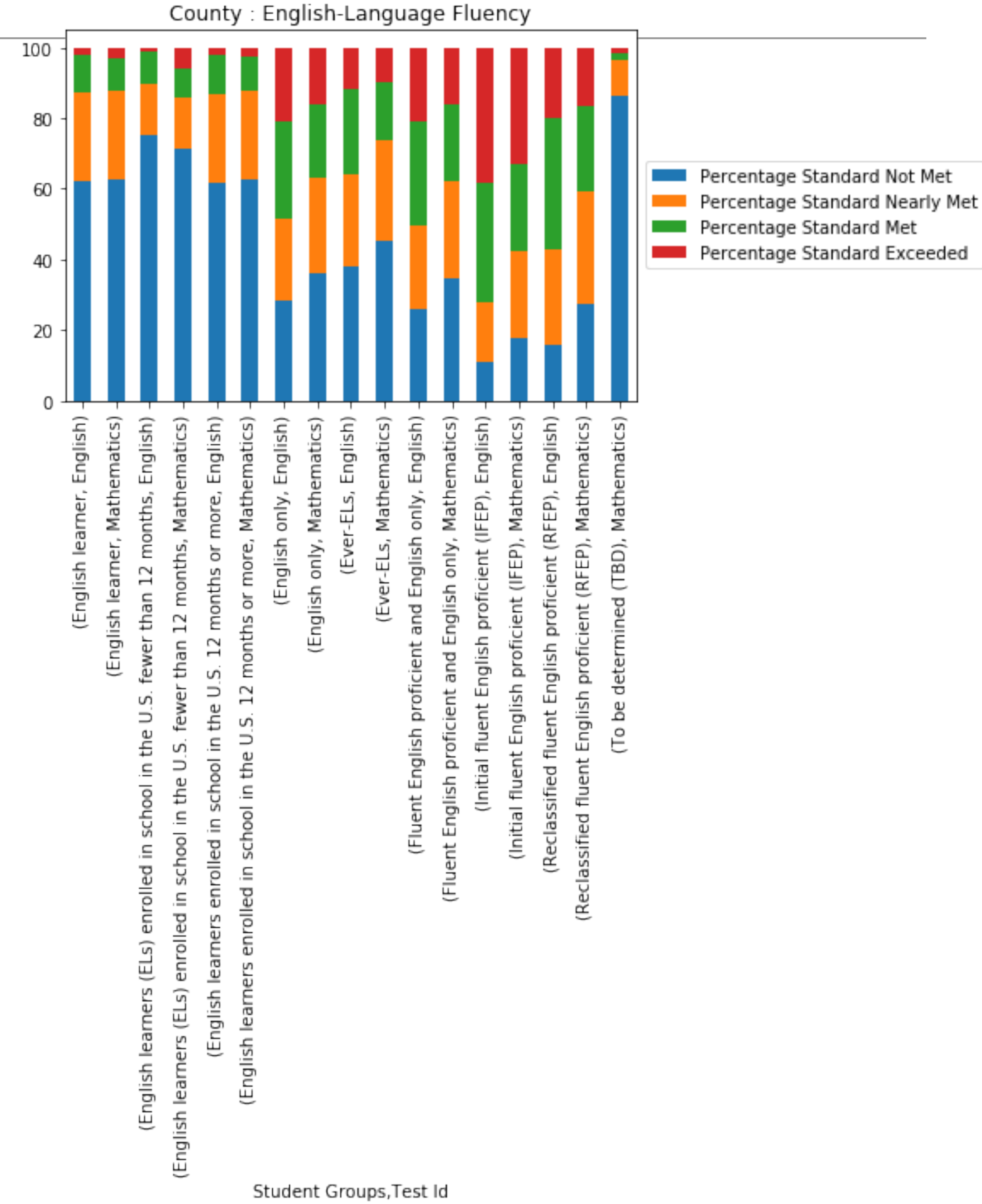
Achievement Levels by Ethnicity

- Asian students achieve the best performance, while Black or American Indian students achieve the lowest performance in both English and mathematics.



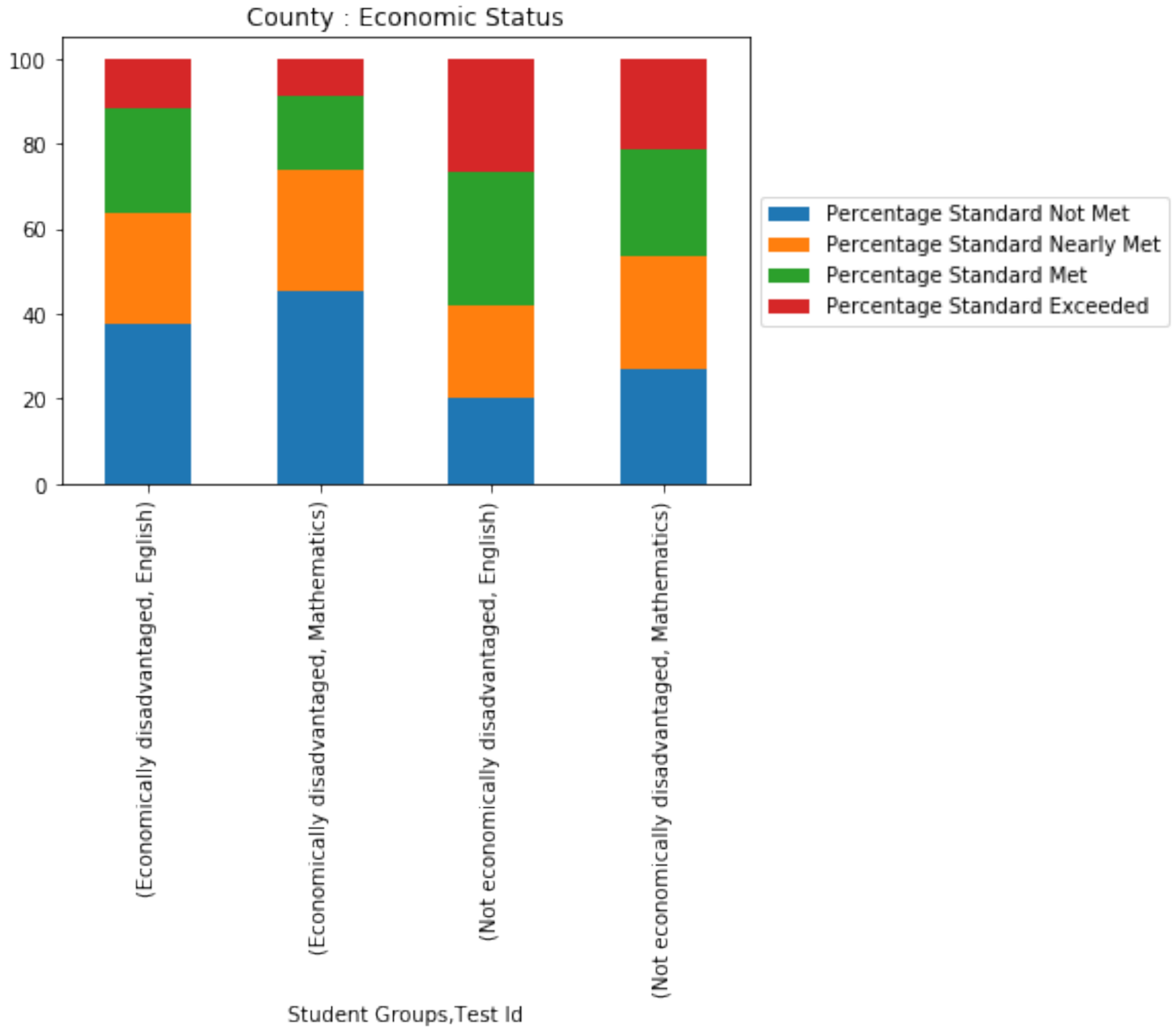
Achievement Levels by English-Language Fluency

- Initial Fluent English Proficient (IFEP) students achieve the best performance in both English and mathematics.
    - I could observe that this trend becomes more obvious in the districts where many Asian immigrants live.
    - I can insist that immigrants have high educational interests and efforts.



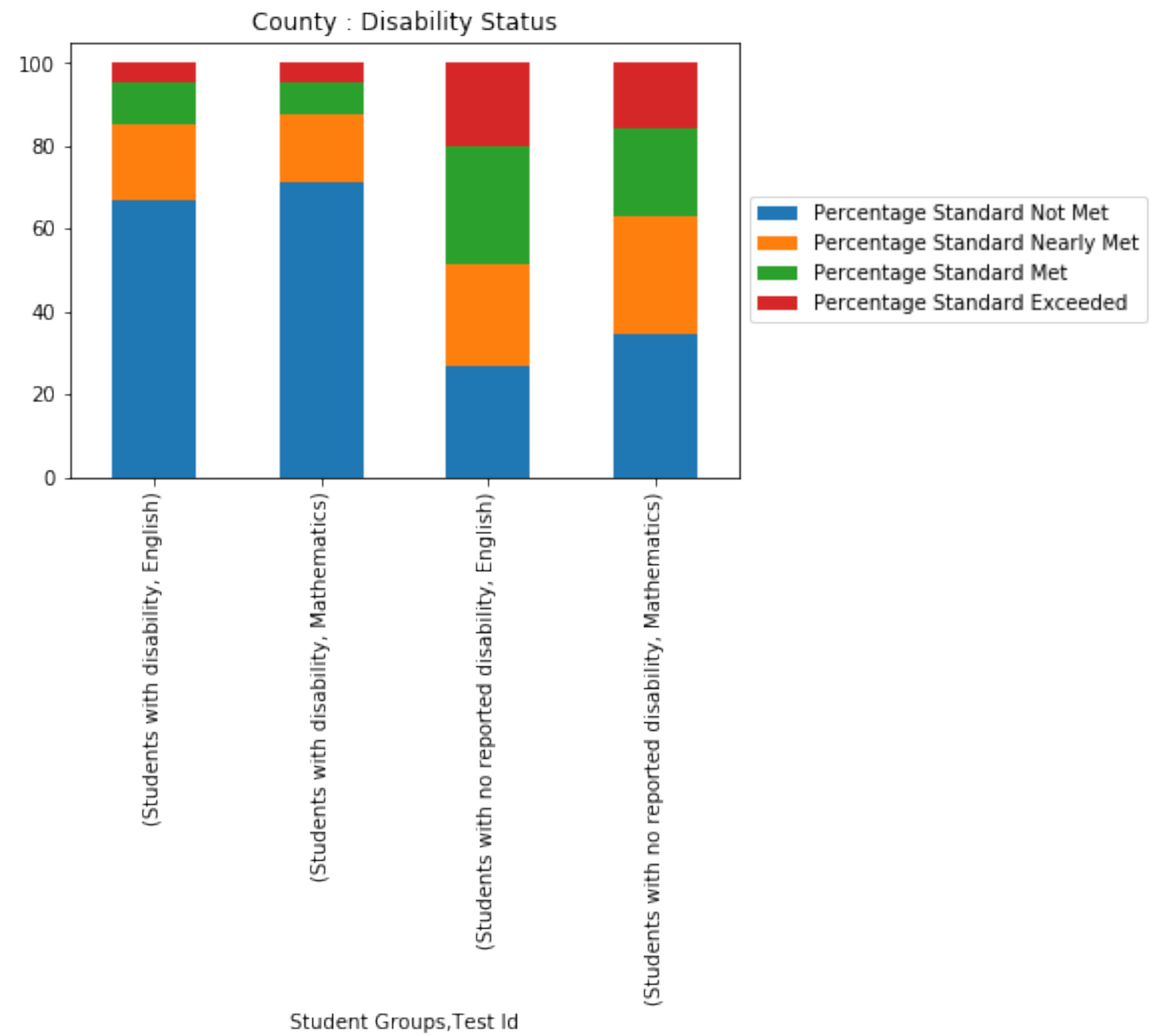
Achievement Levels by Economic Status

- Economically disadvantaged students have much more difficulties than not-economically disadvantaged students.



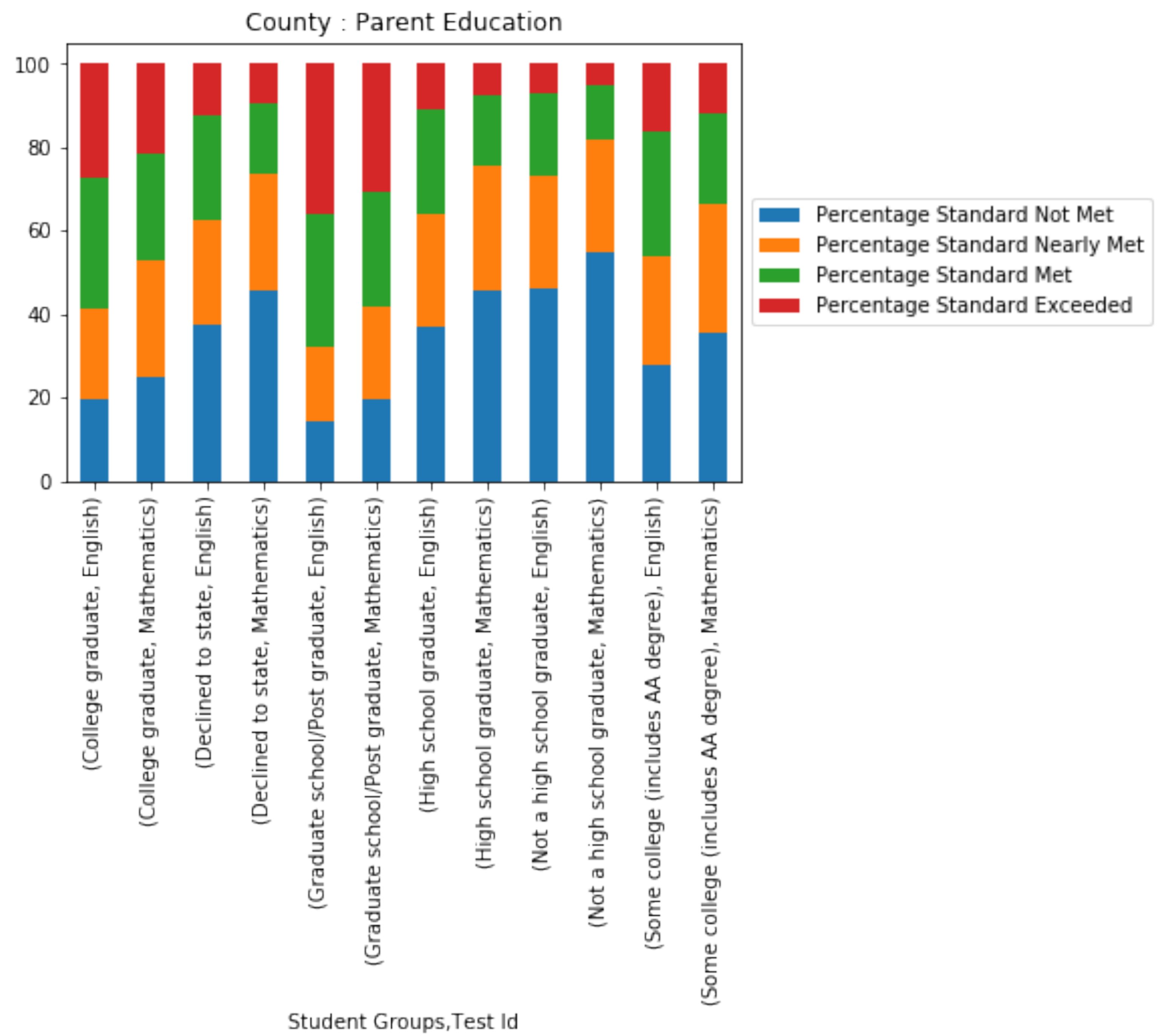
Achievement Levels by Disability Status

- Only the small number of students with disabilities (English: 4.6\%, mathematics: 4.5\%) could achieve the best performance.



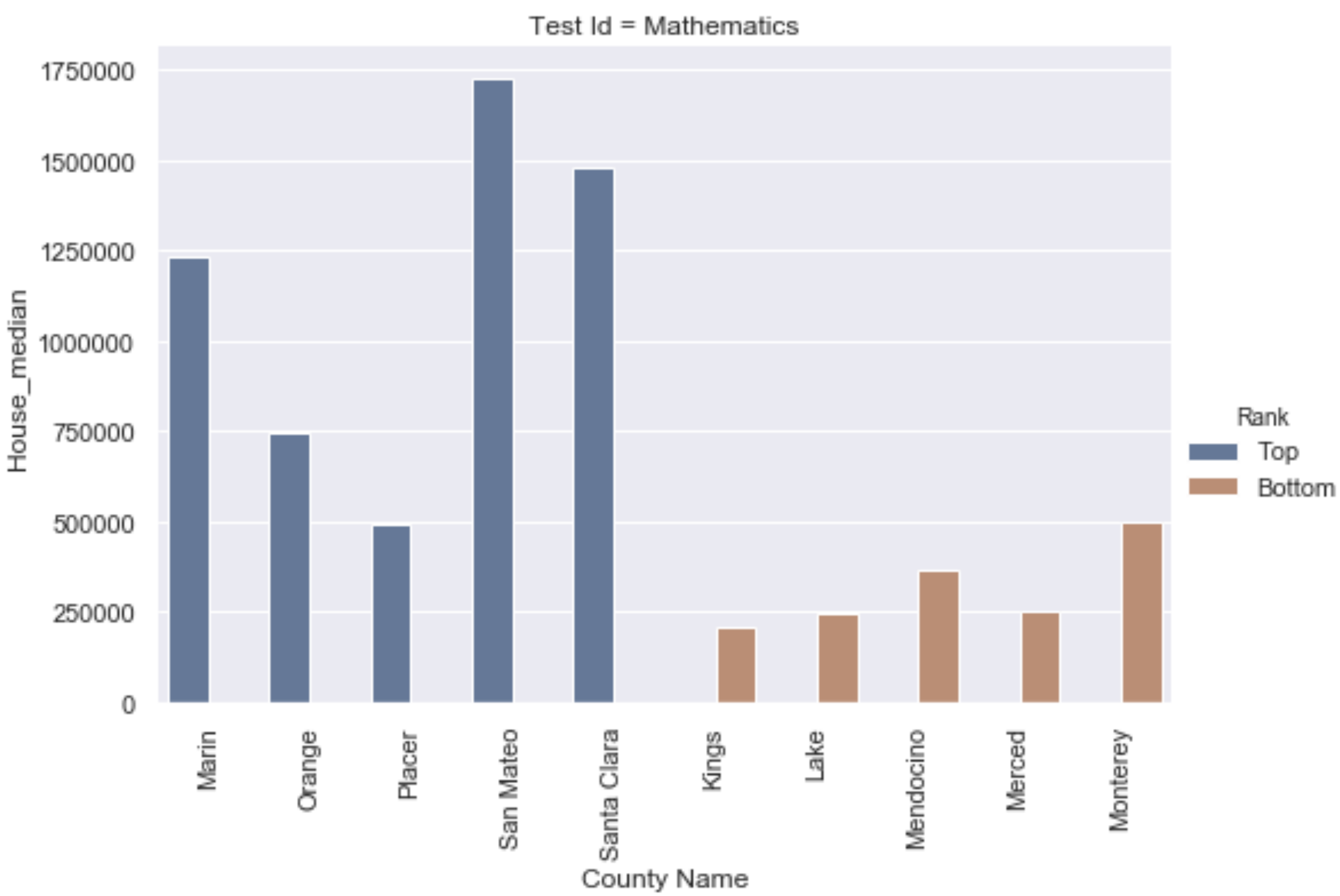
Achievement Levels by Parent Education

- The higher the level of parental education, the higher the achievement of students.
    - Students' achievement is the highest in the parents' education of "graduate school/post graduate”.



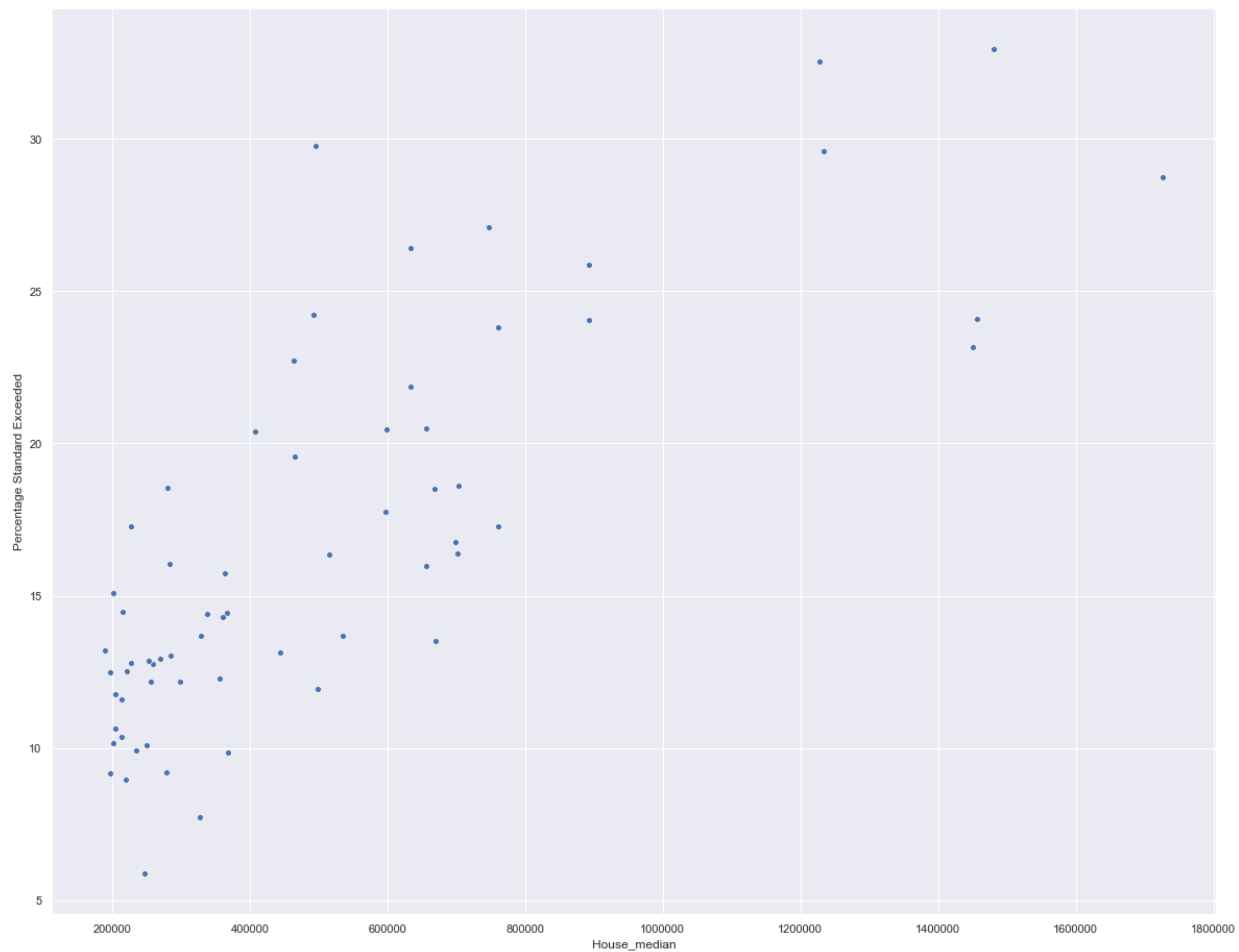
House Prices in Best and Worst 10% Performance Counties

- Test performance is closely related to the economic capabilities of the family to which the student belongs.



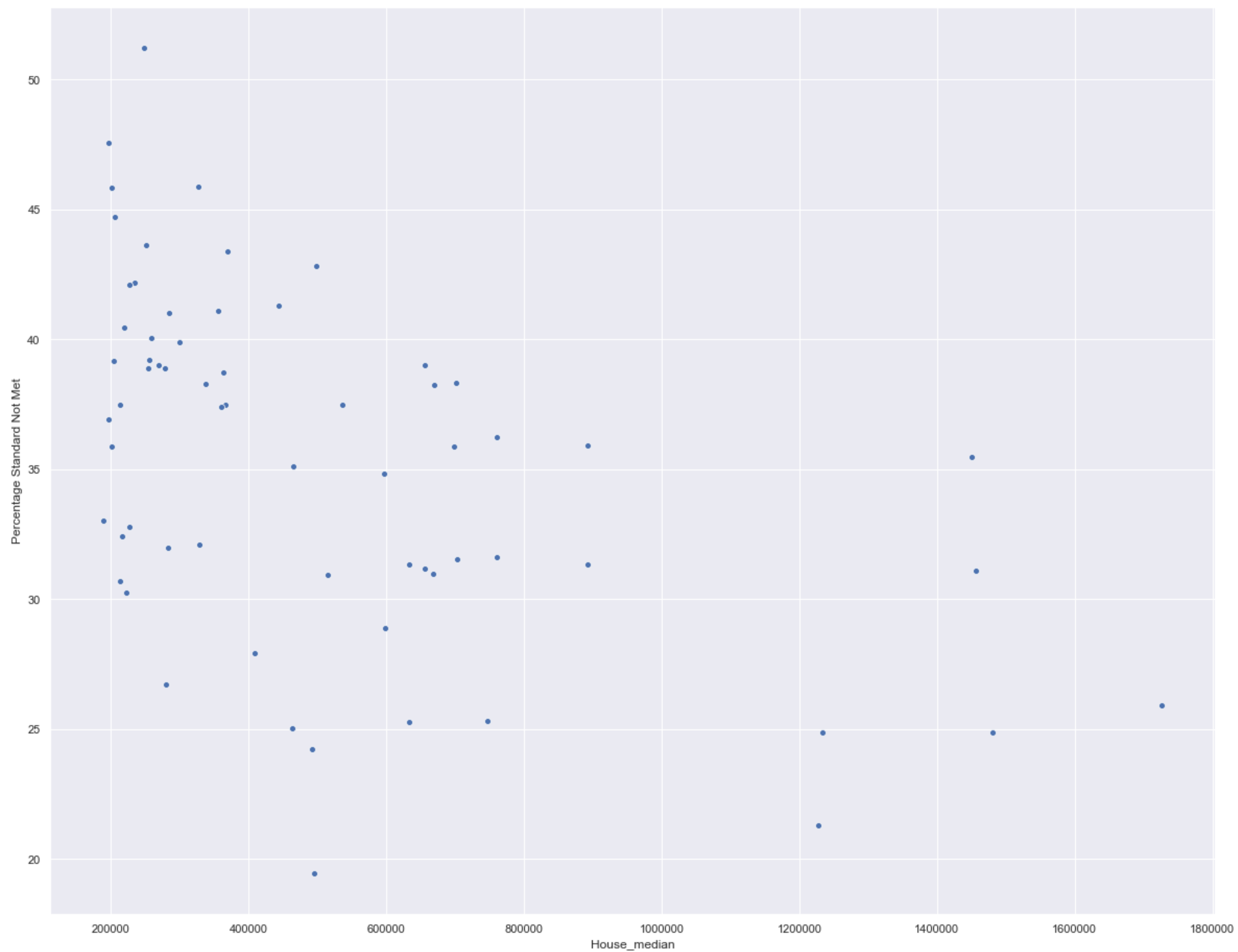
Correlations Between
Test achievements and House Prices

- Strong **positive** correlations between "percentage of **standard exceeded**" and house prices



Correlations Between Test achievements and House Prices

- Strong **negative** correlations between "percentage of **standard not met**" and house prices



Exploratory Data Analysis

- Significant number of features can be redundant and irrelevant, therefore it is important to apply feature selection/dimension reduction

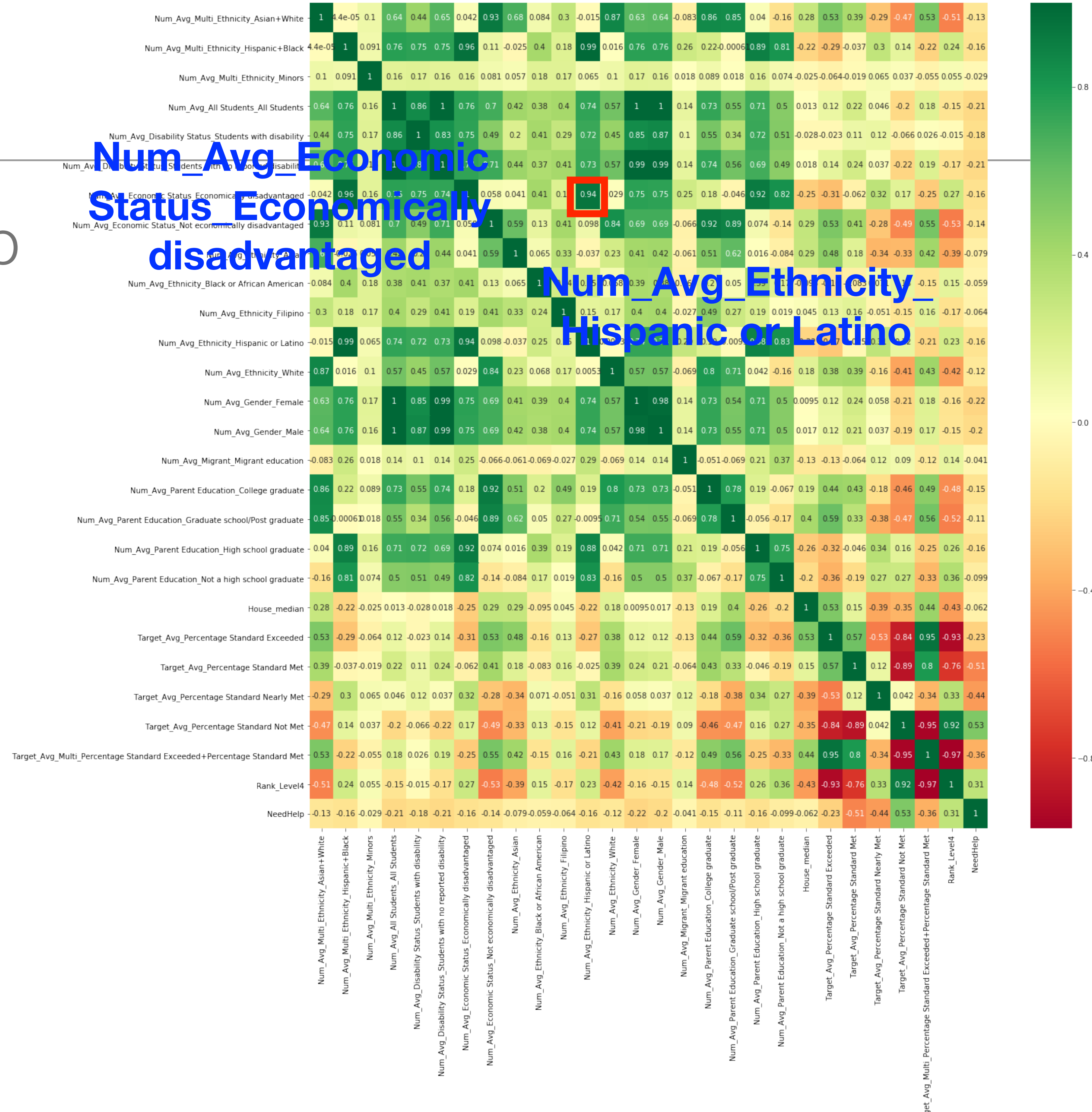
- Methods
  - Statistical hypothesis testing
  - Correlation test
  - Feature selection

Statistical Hypothesis Testing

- T-Test for means of two independent samples
  - Process
    - Tests whether the means of two independent samples are significantly different
    - If there is no difference (p-value is greater or equal than $\alpha = 0.05$), then we eliminate or merge the weak affecting student group indicators
  - Decisions for variables
    - Delete the meaningless indicators such as, 'To be determined (TBD)' and 'Declined to state'
    - Delete the 'Disability Status', 'Economic Status' that seem rather trivial that do not produce any new results

Correlation Test

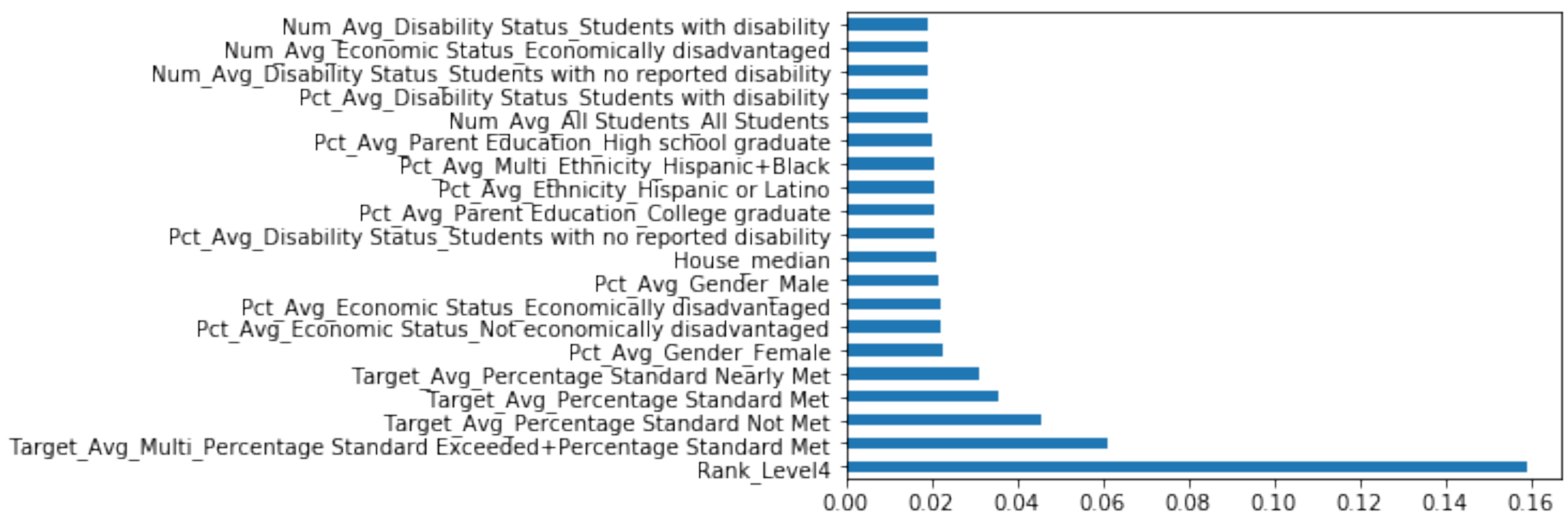
- Matrix with Heatmap
- Pearson's correlation coefficient
- Spearman's rank correlation methods



→ **Number of Hispanics is highly correlated (0.94) with the number of economically disadvantaged student**

Feature Selection

- Univariate selection
  - *SelectKBest* class using the chi-squared as a scoring function to select 20 best features
- Feature importance
  - *Extra Tree Classifier* for extracting the top 20 features for the dataset



Machine Learning Modeling

- The goal is to predict the inferior scores of schools
    - Various machine learning techniques to pick the one which performs best
- Methods
    - Regression
        - Predicts the percentage of students who do not meet the standard
    - Classification
        - Predicts if the schools “need help” or “do not need help”

Regression

- Cross Validation
  - Train/Test Split, Leave One Out (LOO), K-Fold CV
- Evaluation Metrics
  - Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R^2
- Algorithms
  - Linear Regression
  - Random Forest Regressor
  - Gradient Boosting Regressor

Results of Accuracy for Regression Models

- The Random Forest Regressor worked best

Model Name	RMSE	MAE	R^2
Linear Regression with 1 folds Train and test split	11.2853	8.2113	0.6614
Linear Regression with 8,768 folds Leave One Out (LOO)	11.3417	8.2913	0.0000
Linear Regression with 10 folds CV	11.7262	8.5554	0.6233
Random Forest Regressor with 10 folds CV	10.7661	7.6911	0.6763
Gradient Boosting for Regression with 10 folds CV	11.4108	8.3881	0.6368

Classification: Preprocessing data

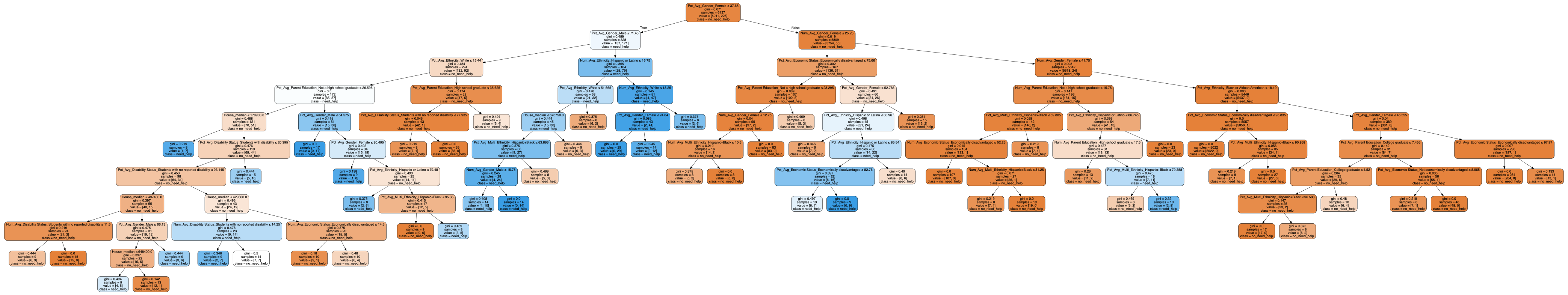
- New binary target variable, **“NeedHelp”,** indicating a school needs help or not
    - ***80% of the standard not met students as*** **1**, otherwise 0
- Data splitting into train data and test data of 70% and 30%
    - For parameter tuning, we use the cross validation in the train data and build the machine learning model, then validate the model with the remained test data
- Scaling
    - For the K-Nearest Neighbor algorithm, we scale the independent variables into the range such that the range is now between 0 and 1

Classification

- Resolving imbalanced classes
  - Stratified K-folds cross validation
    - Ensures that the percentages of each class in your entire data will be the same within each individual fold
  - Weighted evaluation metrics to reflect the mass of the classes
- Evaluation Metrics
  - Accuracy, AUC, Precision, Recall, score F1
- Algorithms
  - Logistic Regression, Decision Tree, GridSearchCV for Parameter Tuning for Decision Tree, Random Forest Classifier, and k-Nearest Neighbors Classifier

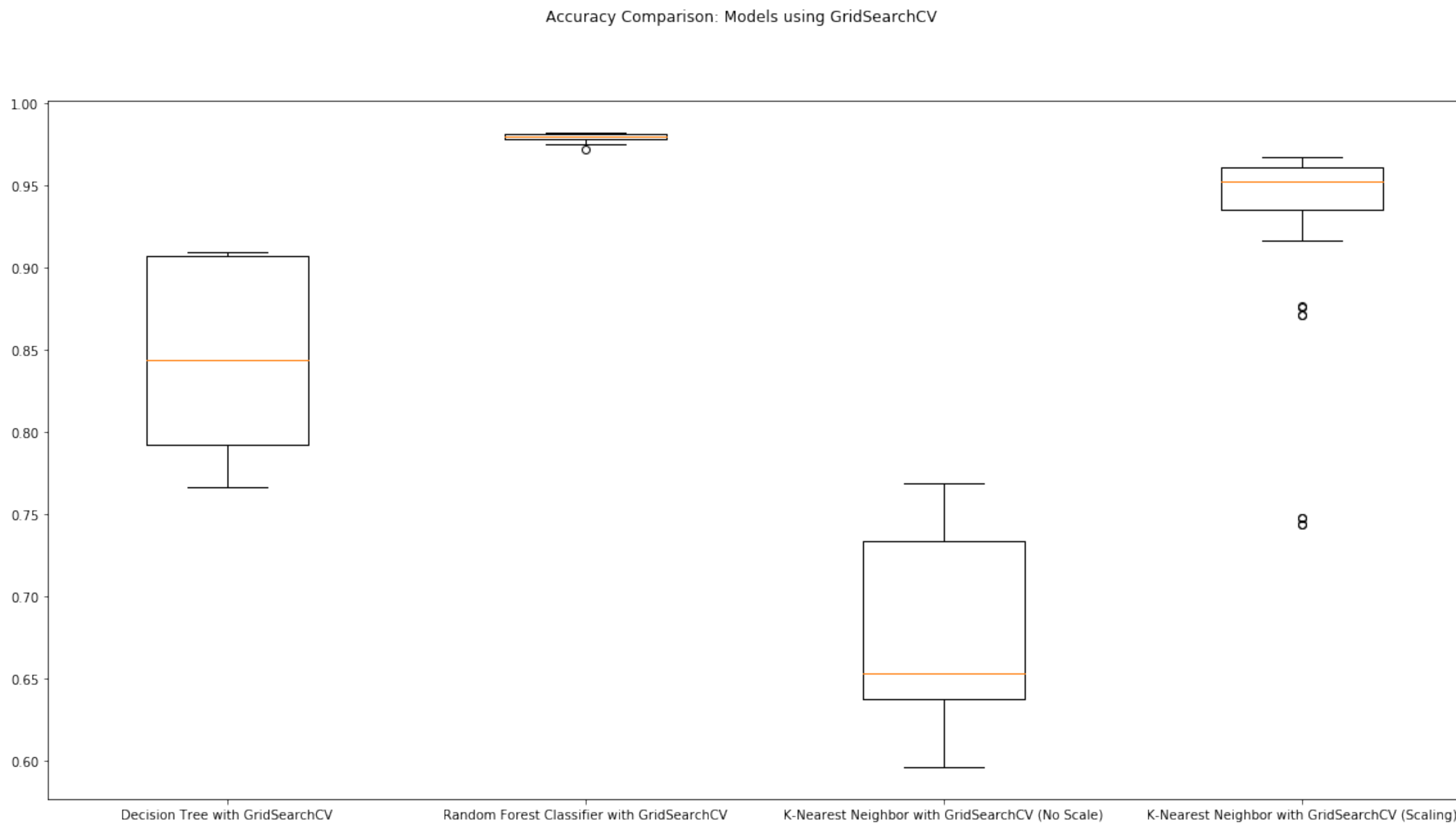
Classification

- Decision Tree with GridSearchCV (Stratified 5-Folds CV)
  - Parameters
    - {'max_depth': [50, 75, 100], 'min_samples_leaf': [1, 2, 4, 8, 10]}
  - Best parameters for the best Decision Tree model
    - {'max_depth': 50, 'min_samples_leaf': 8}.
  - Results model evaluation
    - Best accuracy: 0.9684, best roc_auc_score: 0.9070, weighted avg precision: 0.9666, weighted avg recall: 0.9684, and weighted avg f1-score: 0.9674.



Classification: Boxplots of Accuracy Comparison for GridSearch CV Models

- Random Forest Classifier model has the highest accuracy



Results for the Performance of Classification Models

- Random Forest Classifier with GridSearchCV worked best
  - Parameters: {'max_depth': 100, 'min_samples_leaf': 1, 'n_estimators': 200}
- After applying the scaler to the K-Nearest Neighbor model, the accuracy has been significantly improved

Model Name	accuracy	auc	precision	recall	f1
Logistic Regression with Stratified 5-Folds CV	0.9656	0.9656	0.9646	0.9656	0.9597
Decision Tree with Stratified 5-Folds CV	0.9596	0.7320	0.9660	0.9596	0.9614
Decision Tree with GridSearchCV	0.9684	0.9070	0.9666	0.9684	0.9674
Random Forest Classifier with GridSearchCV	0.9733	0.9774	0.9711	0.9733	0.9718
K-Nearest Neighbor with GridSearchCV (No Scale)	0.9650	0.7309	0.9556	0.9650	0.9526
K-Nearest Neighbor with GridSearchCV (Scaling)	0.9728	0.9618	0.9692	0.9728	0.9695

*** Best accuracy in Random Forest Classifier: 97.33%**

*** K-Nearest Neighbor: 96.5% (no scaling) and 97.28% (0.78% improvement)**

Recommendations

- It is obvious that that the high scores of schools are strongly correlated with the students raised in high-income families.
- In my opinion, the schools need the help
  - Schools have more than 73.14% of students of low-income families,
  - House median prices are less than $335,500 (more urgent help is needed when the house prices are when less than $194,350)
  - Parents who do not graduate high schools is more than 89.1%,
  - Parents who do not graduate colleges is more than 84.9%, or
  - Hispanic or Black students is more than 67.2%

Conclusion

- Analyzed the CAASPP score data to help predict and find the inferior groups of schools that indeed need help and provide suggestions
    - Data wrangling
    - Data visualization
    - Exploratory Data Analysis
    - Machine Learning Modeling
- Future Work
    - To identify the factors that could effectively improve the scores, we will investigate the scores of the 5 consecutive years (2014 to 2018)
        - We expect to find the important features on the schools in which the scores have been dramatically improved