

NYC PUBLIC SCHOOLS: WHAT CAN THE COLLEGE READINESS INDEX TEACH US?

Michelle Cronin

GA - DAT30

Final Presentation – March 16, 2016

THE ISSUE

- Although graduation rates in NYC public schools have risen over the years, the rate of students prepared for college has remained low (73% vs 29% on average, from my data)
- Thus, NYC public schools are graduating many students each year who are not considered college-ready
- **QUESTION:** Can we predict which schools are graduating college-ready students at a rate above NYC's average graduation rate (*within one standard deviation*)? ← High-performing schools

DAILY NEWS NEW YORK NEWS | POLITICS | SPORTS | ENTERTAINMENT

NYC CRIME | BRONX | BROOKLYN | QUEENS | MANHATTAN | EDUCATION | WEATHER | OBITUARIES | NEW YORK PIC

School grades: Department of Education data show less than a third of New York City high school students are college-ready

City Education Department releases A's to F's for city high schools; numbers are up to 29% from 25% last year

BY BEN CHAPMAN, VERA CHINESE, RACHEL MONAHAN / NEW YORK DAILY NEWS /

Updated: Tuesday, November 27, 2012, 1:19 AM

AAA

The New York Times **N.Y. / Region**

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS |

Most New York Students Are Not College-Ready

By SHARON OTTERMAN
Published: February 7, 2011

SEARCH

NEW YORK POST

METRO

f t g+ e

Most NYC high school graduates at CUNY community colleges get remedial help

By Aaron Short
July 5, 2015 | 9:54am

KEY BACKGROUND INFO ON THE DATA

- Data gathered from NYC Open Data website and the NYC Department of Education website over past two school years
 - Ever since de Blasio administration changed school evaluation methods
- Data summarizes schools – NOT info on individual students
- College Readiness Index (CRI)
 - Percentage of students who meet CUNY's standards of readiness in math and English based on test scores (Regents, SATs, APs)
 - Derived cut-off scores based on students needing to enroll in remedial classes in CUNY community colleges (found boundary)
 - 75% of enrollees in 2011¹
 - NYC public schools are some of the most racially and socioeconomically segregated schools in the nation²

¹ FODERARO, Lisa W. "CUNY Adjusts Amid Tide of Remedial Students." The New York Times, 3 Mar. 2011.

² Wong, Alia. "How to Solve the Diversity Problem at NYC's Public Schools." The Atlantic, 5 Mar. 2015.

HYPOTHESIS

- Schools with large populations of historically privileged groups will graduate students who are college-ready at a higher rate than schools with large amounts of populations that have been historically marginalized or have experienced structural inequalities.



PEAK AT THE DATA

- Mostly measurable characteristics of NYC public high schools (performance, demographics, location, etc.):
 - 2013_2014_HS_SQR_Results_2015_03_02.xlsx
 - 2014_2015_HS_SQR_Results_2016_01_07.xlsx
 - DemographicSnapshot201011to201415Public_FINAL-2.xlsx
 - 2015_Graduation_Rates_Public_School-2.xlsx- DOE_High_School_Directory_2013-2014.csv
 - DOE_High_School_Directory_2014-2015.csv- 2014 Public Data File SUPPRESSED.xlsx
 - 2015 Public Data File.xlsx- Location_Information_Report.csv
 - seven_major_felony_offenses_by_precinct_2000_2014.xls
 - violation-offenses-by-precinct_2000-2014.xls
 - Zip_MedianValuePerSqft_AllHomes.csv

```
data.shape  
(975, 51)
```

male_percent	poverty_percent	avg_home_value_sqft	district_admin_code	grade8_english	grade8_math	regents_algebra	regents_english	regent
0.623529	0.866667	1321.166667	1	2.18	2.06	64	66	53
0.621711	0.891447	1321.166667	1	2.27	2.37	64	69	67
0.530030	0.735736	1324.500000	1	2.66	2.63	65	75	64
0.432507	0.845730	1321.166667	1	2.28	2.09	61	69	63
0.484726	0.273199	1321.166667	1	3.50	3.53	80	91	87

PRE-PROCESSING

- Dropped and imputed missing values
- Feature Selection – Necessary due to overlapping features in sources as well as presence of irrelevant features
- Feature Engineering:
 - Length of School Day (End of day – Start of day)
 - Diversity Score (Sum of the squared reported race/ethnicity percents [black, white, Hispanic, Asian])
 - Scaled since percent sums not always equal to 100%
 - .25 = Most Diversity | 1 = Least Diversity/Most Homogeneity
 - Overall SAT score (Reading SAT + Writing SAT + Math SAT average scores)
 - Overall Regents score (Algebra + English + Living Environment + Global History Regents average scores)
 - Grade 8 Proficiency score (Grade 8 Math + Grade 8 English scores)
 - All Crime (Major crimes + violation offenses)
- Tidied the names
- Dummy categorization (separately)
 - Zip codes
 - Administrative Districts

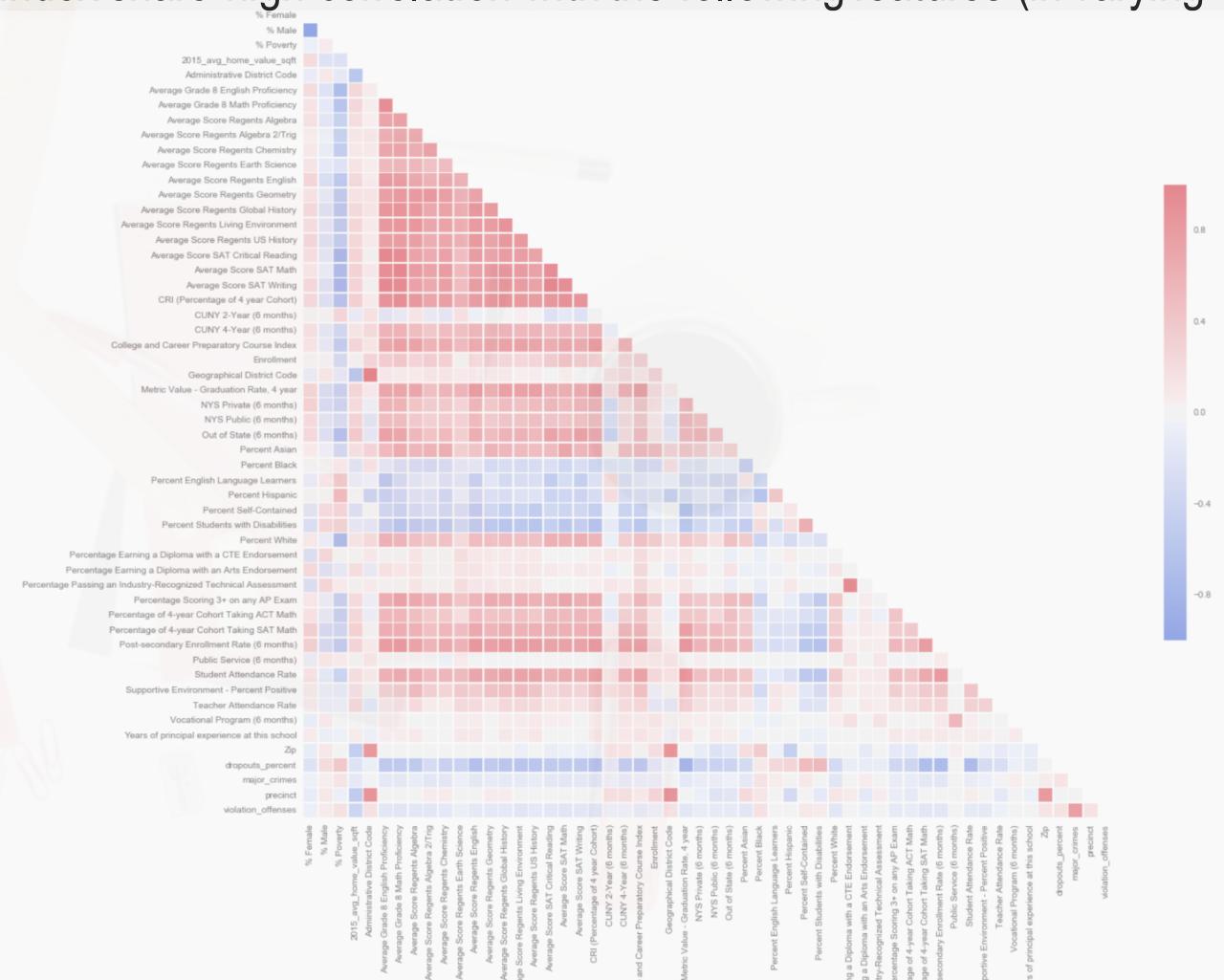
FEATURE SELECTION

- Created visuals to look at correlation (heatmap) → Removed highly correlated features (ex: many Regents exams)
- Created visuals to view missing values → Removed those missing 50%+ (ex: pupil-teacher-ratio)
- Viewed feature variance → removed those with lowest variance (ex: Public Service percentage attendance)
- In the end, removed features directly related to the response variable due to leakage
 - Graduation Rate
 - College-Readiness Index (CRI) (and the highly-correlated college and career readiness index)
 - Test scores related to the CRI: all SAT scores, all Regents values, percentage of students scoring 3+ on AP exam
 - Also removed forward-looking features (post-secondary enrollment rate and related indicators)
 - Created four different datasets for modeling:

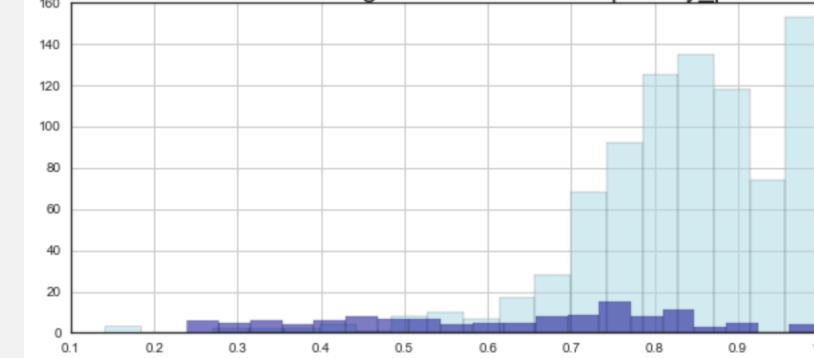
DataFrame Name	Description	Number of Features
data_noexam	Base	30
data_admin	Contains administrative district codes + base	63
data_zip	Contains zip codes + base	151
data_rses	Contains only features directly related to hypothesis	12

OBSERVATIONS

- High school's graduation rate and college readiness index share high correlation with the following features (in varying strengths):
 - Post-Secondary enrollment rate
 - Student Attendance Rate
 - Percentage of students who take the SAT
 - Grade 8 Proficiency Scores
 - Poverty and Disability (Negatively correlated)

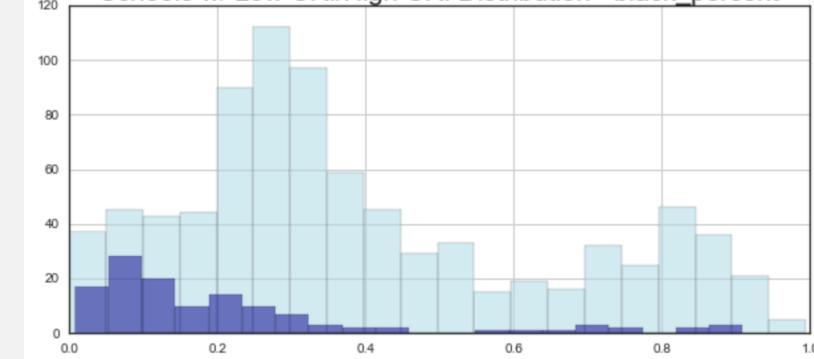


Schools w/ Low CRI/High CRI Distribution - poverty_percent



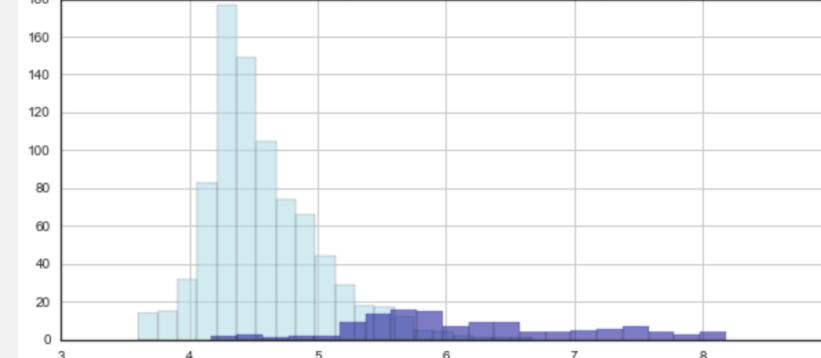
Median=Dataset: 83% | Low CRI: 84% | High CRI: 66%

Schools w/ Low CRI/High CRI Distribution - black_percent



Median=Dataset: 30% | Low CRI: 32% | High CRI: 14%

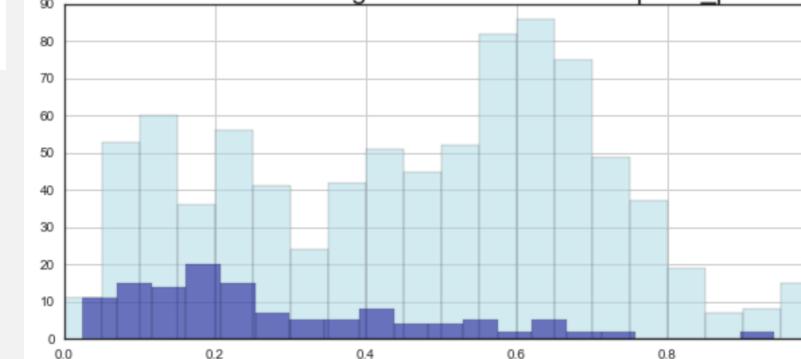
Schools w/ Low CRI/High CRI Distribution - grade8_proficiency



Median=Dataset: 4.5 | Low CRI: 4.5 | High CRI: 6

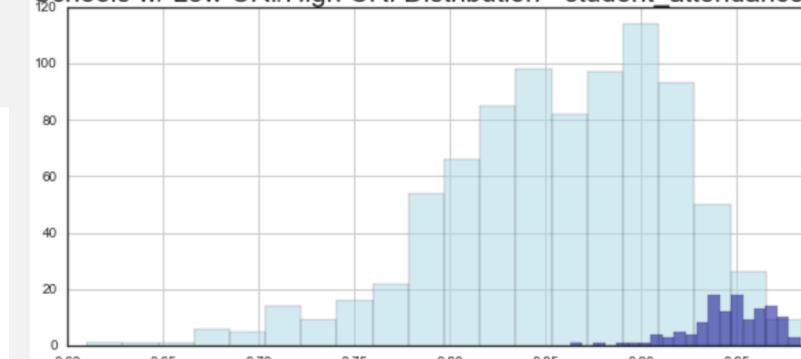
OBSERVATIONS (CONT'D)

Schools w/ Low CRI/High CRI Distribution - hispanic_percent



Median=Dataset: 46% | Low CRI: 51% | High CRI: 22%

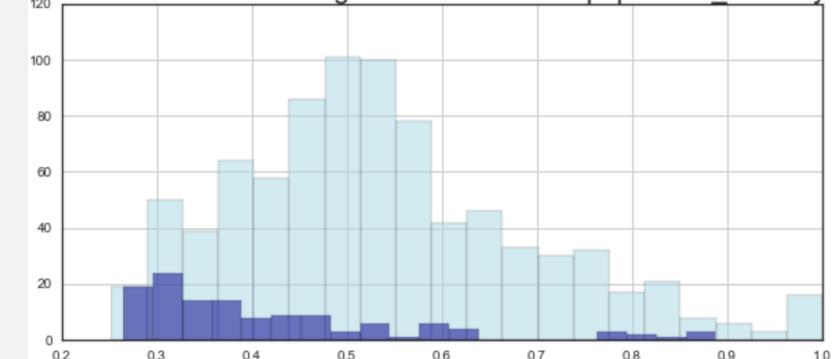
Schools w/ Low CRI/High CRI Distribution - student_attendance_rate



Median=Dataset: 88% | Low CRI: 87% | High CRI: 95%

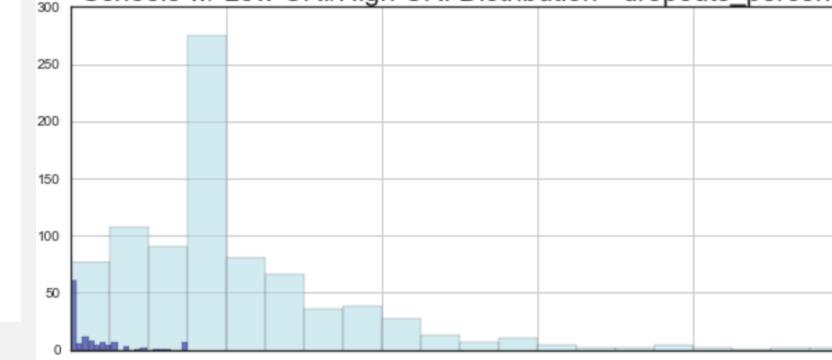
Key: = Schools with Low CRI
 = Schools with High CRI

Schools w/ Low CRI/High CRI Distribution - population_diversity



Median=Dataset: 50% | Low CRI: 52% | High CRI: 37%
(25% = Most Diversity | 100% = Least Diversity)

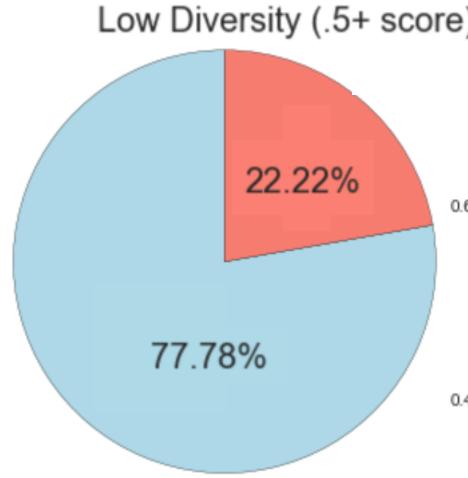
Schools w/ Low CRI/High CRI Distribution - dropouts_percent



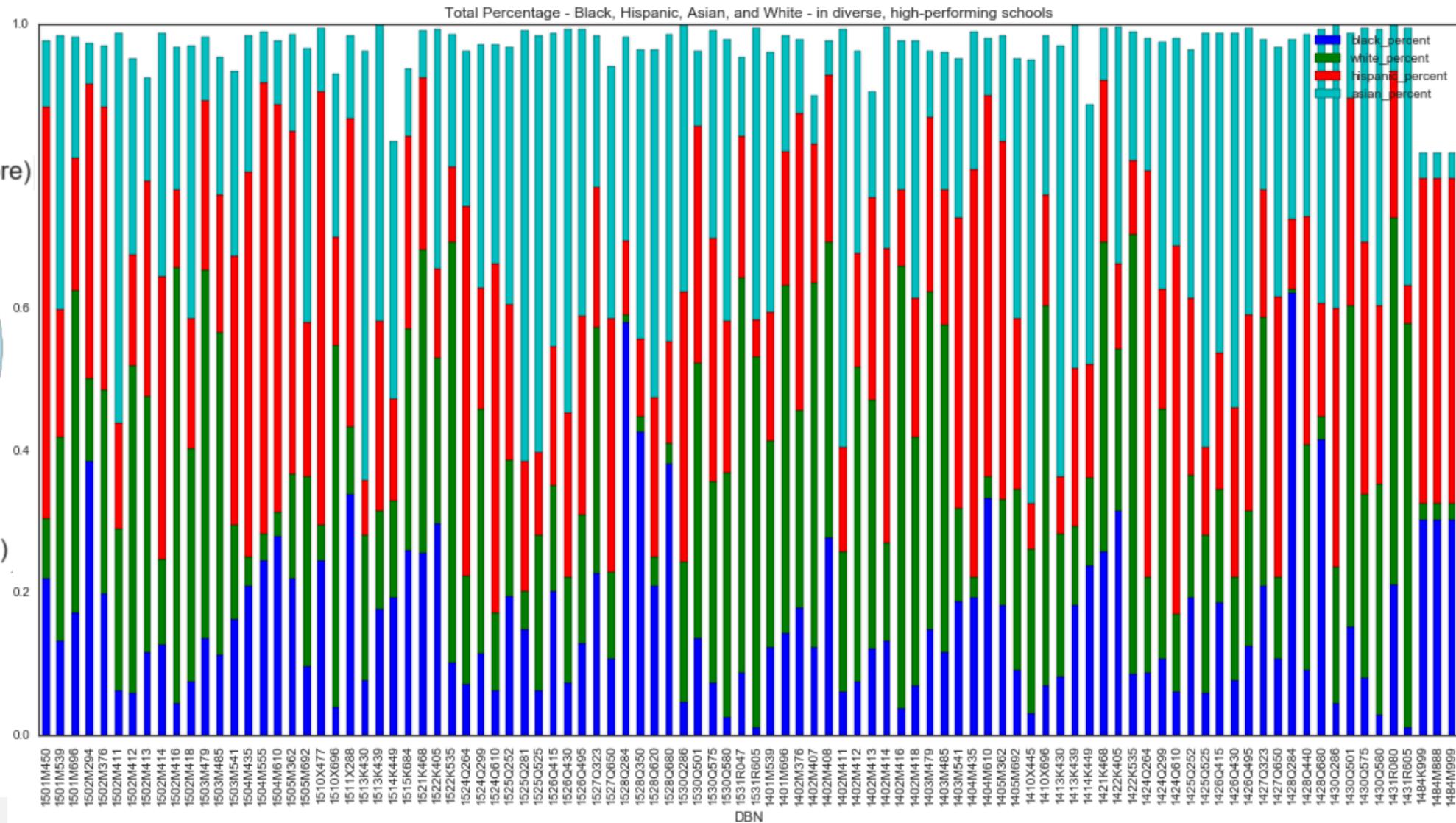
Median=Dataset: 8% | Low CRI: 8% | High CRI: 0.4%

OBSERVATIONS (CONT'D)

Population Diversity in High-Performing Schools



High Diversity (under .5 score)



MODELING

- Minimum Benchmark using Dummy Classifier:

Precision	Recall	F1-Score
.79	.81	.80

- Random Forest
 - Protect against overfitting (great because I have a considerable amount of features)
 - Will tell me most important features (which I can feed into other models)
- Logistic Regression
 - Lasso/ridge will drop/let me know unimportant features
 - Will allow me to view feature coefficients
- k-means Clustering:
 - I may be able to discover distinguishable groups of schools to find some counterintuitive combinations of characteristics as well as obvious ones

```
list(data_rses)
```

```
['female_percent',
'male_percent',
'poverty_percent',
'avg_home_value_sqft',
'asian_percent',
'black_percent',
'ell_percent',
'hispanic_percent',
'self-contained_percent',
'disability_percent',
'white_percent',
'population_diversity',
'label']
```

MODELING – RANDOM FOREST

- RESULTS:

	Precision	Recall	Dummy Classifier F1-Score	F1-Score	% Change	Time
data_noexam	.86	.86	.80	.86	+7.5%	3.99 ms per loop
data_admin	.86	.86	.80	.86	+7.5%	4.42 ms per loop
data_zip	.82	.86	.80	.84	+5%	17.3 ms per loop
data_rses	.64	.82	.80	.72	-10%	2.87 ms per loop

- Three out of the four datasets beat the dummy classifier
- data_rses (with student demographics/characteristics) is worst-performing
- data_noexam (which has no geographic info) is marginally best-performing



Features	Importance Score	Features	Importance Score
grade8_proficiency	0.149885	ell_percent	0.193844
grade8_math	0.125332	asian_percent	0.150225
grade8_english	0.124609	disability_percent	0.142821
sat_percent_took_exam	0.071781	poverty_percent	0.115429
student_attendance_rate	0.068905	white_percent	0.071317
dropouts_percent	0.061824	black_percent	0.063662

MODELING – LOGISTIC REGRESSION

- RESULTS:

	Precision	Recall	Dummy Classifier F1-Score	F1-Score	% Change	Time
data_noexam	.82	.86	.80	.84	+5%	47.7 Microsecs per loop
data_admin	-	-	-	-	-	-
data_zip	-	-	-	-	-	-
data_rses	.60	.76	.80	.67	-16%	44.8 Microsecs per loop

- Logistic Regression was not as successful as Random Forest
- Large amount of features in data_admin and data_zip made LR prohibitive

coefficients	features		
1.537	student_attendance_rate		
1.322	grade8_math	-1.351	population_diversity
0.997	grade8_proficiency	-1.356	self-contained_percent
0.927	violation_offenses	-1.779	white_percent
0.917	ell_percent	-1.996	asian_percent
0.914	sat_percent_took_exam	-3.197	black_percent
0.685	school_day_length	-3.255	hispanic_percent
0.590	poverty_percent	-3.774	dropouts_percent

	0	1	2	3
female_percent	0.472	0.435	0.531	0.532
male_percent	0.528	0.565	0.469	0.468
poverty_percent	0.946	0.867	0.822	0.531
avg_home_value_sqft	420.117	385.978	515.952	583.631
grade8_english	1.845	2.219	2.469	3.128
grade8_math	2.112	2.100	2.375	3.054
enrollment	385.818	481.699	599.564	1473.037
asian_percent	0.171	0.033	0.079	0.282
black_percent	0.083	0.477	0.396	0.127
ell_percent	0.819	0.117	0.070	0.031
hispanic_percent	0.687	0.451	0.450	0.234
self-contained_percent	0.009	0.094	0.034	0.020
disability_percent	0.039	0.252	0.174	0.109
white_percent	0.053	0.027	0.053	0.329
cte_percent	0.032	0.021	0.034	0.026
arts_percent	0.021	0.008	0.031	0.050
act_percent_took_exam	0.064	0.038	0.088	0.227
sat_percent_took_exam	0.569	0.481	0.704	0.828
student_attendance_rate	0.885	0.818	0.895	0.938
school_environment_survey	0.885	0.812	0.852	0.876
teacher_attendance	0.972	0.961	0.968	0.968
principal_year_exp	4.727	4.461	5.062	5.355
dropouts_percent	0.102	0.142	0.062	0.022
major_crimes	1772.795	1713.769	1482.761	1333.402
violation_offenses	947.909	1008.803	851.812	743.729
school_day_length	428.727	424.368	421.139	412.243
population_diversity	0.639	0.565	0.514	0.369
grade8_proficiency	3.957	4.320	4.844	6.132
all_crime	2720.705	2722.572	2334.573	2077.131

MODELING – K-MEANS

- **Cluster 0 – A Challenging Start**
 - Low Grade 8 Proficiency Scores
 - High Hispanic Percentage and ELL rates
 - High population diversity index, suggesting low diversity
 - High Poverty percentage
 - High crime areas
 - Small schools
- **Cluster 1 – Difficulty Staying On Track**
 - Large disability percentage rates
 - Low attendance and test-taking rates
 - High drop-out rates and percentage of black students
- **Cluster 2 – On the Right Track**
 - Despite poverty figure, schools appear to be in wealthy neighborhoods
 - Slightly above average testing scores
 - Slightly above avg diversity index, suggesting slightly low diversity
 - Presents similarly like cluster 1 for race/ethnicity but with slightly smaller numbers, but has better testing, attendance, and dropouts
- **Cluster 3 – Large and High-Achieving**
 - Large Asian population (Mean of entire dataset: 9%), so overrepresentation in this cluster → same with white percent
 - Low population diversity index, suggesting high diversity
 - High Grade 8 Proficiency scores

CHALLENGES/SUCCESSES

- CHALLENGES:
 - Putting together so many different datasets was time-consuming
 - Realized late in the process that I had forward-looking features that previously were artificially boosting my score
 - Removing them significantly decreased model performance
 - Wasted time running models I couldn't use (including SVM, which would not run on new models)
 - Datasets I tested with geographic dummy categoricals had so many features that Logistic Regression wouldn't run
- SUCCESSES:
 - Interesting exploration and important features for insights
 - Some models beat benchmark, which means we can successfully answer project question

CONCLUSIONS

- INSIGHTS:

- While students who belong to historically marginalized groups also appear less frequently in schools with high college readiness rates, their presence alone is not an accurate predictor of whether a school has a high or low rate of college readiness.
- Although schools with high college-readiness rates appear to have higher amounts of diversity (or dispersion) than schools with low rates of college readiness, this feature did not score highly on RF's Feature Importance nor as a coefficient for LR.
- Grade 8 Proficiency seems to be one of strongest predictive factors for college readiness

- FUTURE STEPS:

- Complete similar work but on middle/elementary schools
 - What contributes to those Grade 8 Proficiency scores?
- Clustering with geographic data
 - More exploration
- Compare data over several years as more is released

