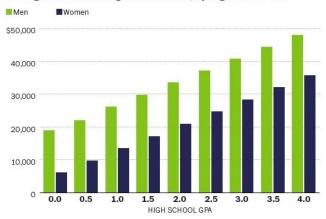
# **An Analysis of Student Grades**

## **Student Grades**

Student high school grades are important because they increase chances of college admission, number of colleges they can get into, merit based scholarships, and lifetime earnings.

#### Average annual earnings in adulthood, by high school GPA



SOURCE: University of Miami GRAPHIC: The Washington Post. Published May 20, 2014

## What affects student grades?

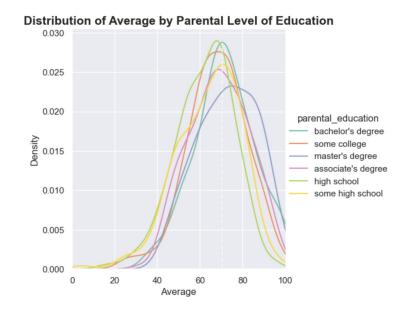


It's no doubt that grades play a role in a student's future, but what can we do to help students optimize their grades?

To help answer this question, I am analyzing a dataset from Kaggle containing information on math, reading, writing scores, as well as other factors, such as parental education, gender, standard lunch, race/ethnicity, and completion of a test prep program.

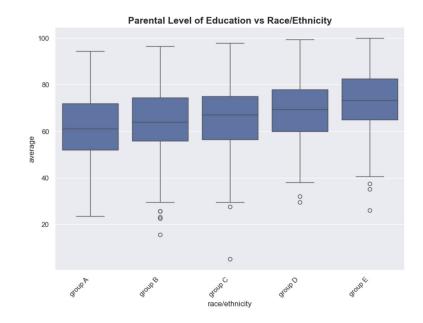
## How does parental level of education affect grades?

- The highest average comes from parents with master's degrees.
  This can be seen in the navy bell curve. It's center is more to the right, which indicates a higher average for that group.
- The lowest average comes with parents with only high school degrees.



## Do race/ethnicity affect student grades?

 By comparing averages across 5 different race/ethnicity we can see that group E has highest median average and group A has lowest median average.



#### **Correlation Matrix**

- The correlation matrix tells us how strongly each variable relates to one another.
- How does each variable relates to average?
  - Reading score has the highest correlation with average at 0.93
  - Having standard lunch, completion of test prep, parental education, and being part of race/ethnicity group E have positive, but low correlations with average.

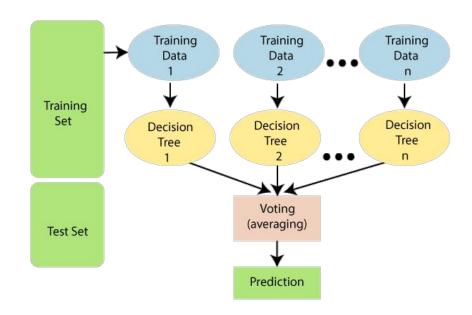
reading_score	1		0.18	0.24	-0.24	-0.096	-0.06	-0.0031	0.035	0.11	-0.23	0.23	0.24	-0.24
average	0.93		0.2	0.07	-0.07	-0.11	-0.086	-0.044	0.07	0.16	-0.31	0.31	0.26	-0.26
parental_education_encoded	0.18	0.2	1	0.044	-0.044	-0.054	-0.079	0.025	0.04	0.05	0.024	-0.024	-0.016	0.016
gender_female	0.24	0.07	0.044	1	-1	-0.071	0.028	0.063	-0.031	-0.02	0.021	-0.021	-0.006	0.006
gender_male	-0.24	-0.07	-0.044			0.071	-0.028	-0.063	0.031	0.02	-0.021	0.021	0.006	-0.006
race/ethnicity_group A	-0.096	-0.11	-0.054	-0.071	0.071	1	-0.15	-0.21	-0.19	-0.13	0.032	-0.032	-0.0063	0.0063
race/ethnicity_group B	-0.06	-0.086	-0.079	0.028	-0.028	-0.15	1	-0.33	-0.29	-0.2	0.0083	-0.0083	-0.00011	0.00011
race/ethnicity_group C	-0.0031	-0.044	0.025	0.063	-0.063	-0.21	-0.33	1	-0.41	-0.28	0.0034	-0.0034	0.013	-0.013
race/ethnicity_group D	0.035	0.07	0.04	-0.031	0.031	-0.19	-0.29	-0.41	1	-0.24	0.0095	-0.0095	-0.056	0.056
race/ethnicity_group E	0.11	0.16	0.05	-0.02	0.02	-0.13	-0.2	-0.28	-0.24	1	-0.052	0.052	0.059	-0.059
lunch_free/reduced	-0.23	-0.31	0.024	0.021	-0.021	0.032	0.0083	0.0034	0.0095	-0.052	1	-1	0.017	-0.017
lunch_standard	0.23	0.31	-0.024	-0.021	0.021	-0.032	-0.0083	-0.0034	-0.0095	0.052			-0.017	0.017
test_prep_completed	0.24	0.26	-0.016	-0.006	0.006	-0.0063	-0.00011	0.013	-0.056	0.059	0.017	-0.017	1	-1
test_prep_none	-0.24	-0.26	0.016	0.006	-0.006	0.0063	0.00011	-0.013	0.056	-0.059	-0.017	0.017		
	reading_score	average	_education_encoded	gender_female	gender_male	ace/ethnicity_group A	ace/ethnicity_group B	ace/ethnicity_group C	ace/ethnicity_group D	ace/ethnicity_group E	lunch_free/reduced	lunch_standard	test_prep_completed	test_prep_none

## How well can we predict a student's grade average?

- Based on some EDA from previous slides, it looked like there was a relationship between certain features and student's average, but how well do these predict grades?
- To answer this we are using three models:
  - Random Forest
    - Averages the results of a group of decision trees.
  - Gradient Boosting
    - Reduces bias of weak learner (underfit).
  - Linear Regression

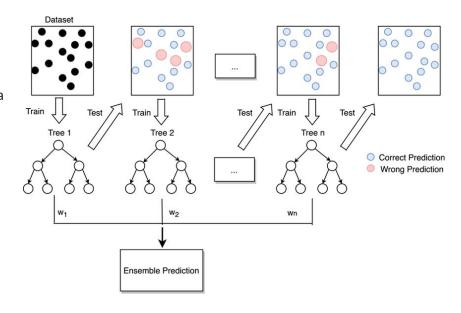
#### What is Random Forest?

- A group of decision trees that are forced to be as unique as possible through bootstrap sampling and random prediction selection.
  - Bootstrap Sampling: sampling with replacement.
  - Random Prediction Selection: RF selects random sample of predictors instead of using all predictor variables.
- The averaging of decision trees makes Random Forest successful because it reduces the high variance of decision trees.



## What is Gradient Boosting?

- A type of machine learning boosting which is used to solve regression and classification problems.
- Target outcomes for each case are set based on the gradient of the error with respect to the prediction. Each new model takes a step in the direction that minimizes prediction error.
- The target outcome for each case in the data depends on the how much a change in prediction affects overall error:
  - If a small change in the prediction for a case causes a large drop in error, then next target outcome of the case is a high value.
  - If a small change in the prediction for a case causes no change in error, then next target outcome of the case is zero. Changing this prediction does not decrease the error.



### Results

Random Forest (all predictors)	Random Forest (top predictors	Gradient Boosting (all predictors)	Gradient Boosting (top predictors	Linear Regression (only Reading Score)
RMSE:	RMSE:	RMSE:	RMSE:	RMSE:
4.95	5.55	4.79	5.31	5.125
MSE:	MSE:	MSE:	MSE:	MSE:
24.54	30.81	22.97	28.18	26.27
R-Squared: 0.889	R-Squared: 0.860	R-Squared : 0.896	R-Square d: 0.872	R-Squared: 0.881

- Gradient Boosting Consistently performed better.
- The best scoring model included all predictors and gradient boosting, however the simple linear regression is favored due to its simplicity and high r-squared.
- 88.1% of variation in student averages can be explained by the relationship between reading score and student average.