Middle School Alcohol Abuse & Demographics in the United States

by Taylor Corbalis, Courtney Sutton, and Johnny Whitaker

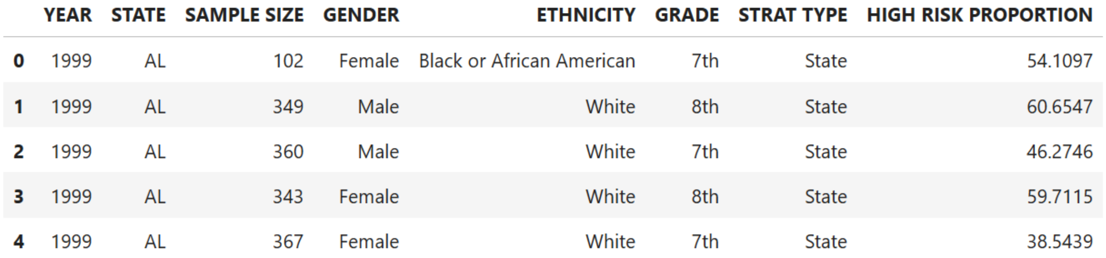
Introduction

Middle school student alcohol usage is a problem in the United States, and not enough attention has been given to the issue. This makes it difficult to measure and improve. Our group did our final project on measuring student alcohol use data with demographic information to discover more about what correlations are significant and to brainstorm ideas about how to fix this problem. Substances are unfortunately somewhat accessible to underage kids and students ages 11-13 are easily influenced by their peers and their own curiosity when the opportunity to use alcohol becomes available. Alcohol use at a young age can cause addiction, create serious health problems such as impaired learning, brain development and brain functioning, and can cause a lack of motivation in school. This brings about risk of absenteeism, lowering grades, and a greater risk of students dropping out early. We obtained our data from a publicly shared file from the Centers for Diseases Control and Prevention, which was collected through online surveys polling middle school students on their alcohol use between 1999-2017. Our ultimate goal was to measure the correlation between demographics of middle schoolers to the risk factor of their alcohol usage. Our study is important because it affects the economy. Lack of student motivation caused by alcohol use can lead to more school dropouts, bringing about a decrease in the qualified pool of job applicants in the workforce. This can also lead to lower rates of college enrollment, further affecting local economies. Student alcohol use also affects the ability of local, state and federal governments to efficiently allocate resources. The amount of financial and infrastructural capital invested in students' educations is an important part of government budgets. Students not graduating means there is an under-utilization of those resources.

Data cleaning/pre-processing

The initial data set from the CDC had over 33,000 entries and featured responses to subjective surveys of middle school students on their alcohol use. Based on their responses to the survey questions, the students were divided proportionally into Greater Risk and Lesser Risk proportions, then further divided into confidence limits within each respective risk proportion. We initially wished to compare these proportions, sans demographic information, to the high school graduation rate of each state. However, upon conferring with Professor Piri, we discovered the structure of the data based solely on the proportions was not sufficient for predictive analytics, and was instead closer to hypothesis testing.

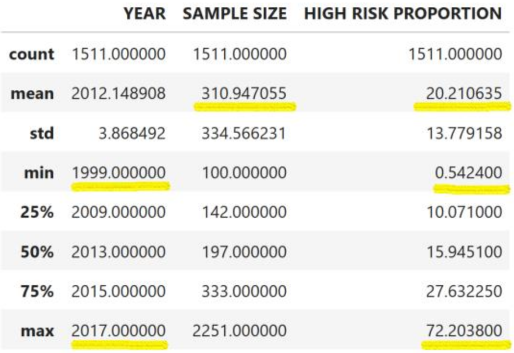
To adjust our data to be used effectively for predictive analytics, we removed the secondary confidence limit divisions as well as the Lesser Risk Proportion and adjusted our target variable to the Greater Risk Proportion. We then reincorporated demographic data and proceeded with cleaning the remaining. We used only statewide data, choosing to omit data for cities as the cities chosen did not lie in the states chosen and we felt it would serve as a poor comparison. We then removed all surveys with less than 100 respondents, and removed any surveys with non-homogenous respondents. Remaining were approximately 1500 entries, each with specific Gender, Ethnicity, and Grade information. Finally, we converted to string and encoded with dummies the following variables: State, Gender, Ethnicity, Grade, and Stratification Type.



	YEAR	STATE	SAMPLE SIZE	GENDER	ETHNICITY	GRADE	STRAT TYPE	HIGH RISK PROPORTION
0	1999	AL	102	Female	Black or African American	7th	State	54.1097
1	1999	AL	349	Male	White	8th	State	60.6547
2	1999	AL	360	Male	White	7th	State	46.2746
3	1999	AL	343	Female	White	8th	State	59.7115
4	1999	AL	367	Female	White	7th	State	38.5439

Pictured above is the table of our final variables; High Risk Proportion as the Target Variable.

Descriptive analytics



	YEAR	SAMPLE SIZE	HIGH RISK PROPORTION
count	1511.000000	1511.000000	1511.000000
mean	2012.148908	310.947055	20.210635
std	3.868492	334.566231	13.779158
min	1999.000000	100.000000	0.542400
25%	2009.000000	142.000000	10.071000
50%	2013.000000	197.000000	15.945100
75%	2015.000000	333.000000	27.632250
max	2017.000000	2251.000000	72.203800

As mentioned above, our data set takes surveys from 1999 to 2017, reflected in the highlighted statistics in the left column of the above table. The middle column reflects the average sample size of approximately 310 respondents and a maximum of 2,251 which we believe provides our data with nominal strength. Lastly, pictured in the right column is a mean High Risk Proportion of approximately 20%, a minimum of 0.5% (a positive but non-startling statistic), and a maximum of 72% (a negative and troubling statistic). For clarification, an average High Risk Proportion indicates that on average, 20% of respondents were determined to be at high risk of abusing alcohol based on the answers to the survey questions.

Predictive analytics
Finally, you need to predict the target value for a couple of test data points

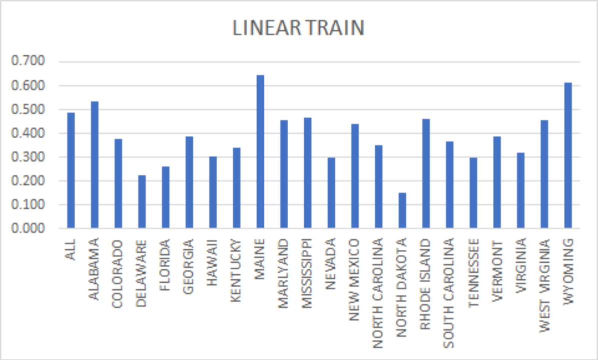
As our target variable was a proportion, we used regression analyses with our data set. For all states combined, and each of the twenty-one states individually, we created Linear Regression, K-Nearest Neighbor Regressor, Lasso Regression, Ridge Regression, Decision Tree Regressor, and Random Forest Regressor models, for a total of 132 models. To find the optimal R^2 values for each model we experimented heavily with various parameter values for each model; the optimal specific parameter values for each model are detailed in the table below.

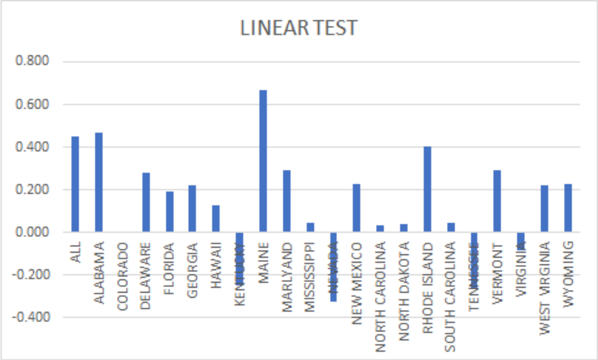
All States		
Model Type	**Parameter Value**	
Linear Regression	Alpha = Default	
KNN Regressor	N Neighbors = 300	
Lasso Regression	Alpha = 0.01	
Ridge Regression	Alpha = 0.01	
Decision Tree Regressor	Max Leaf Nodes = 10	Random State = 0
Random Forest Regressor	N Estimators = 250	Random State = 0

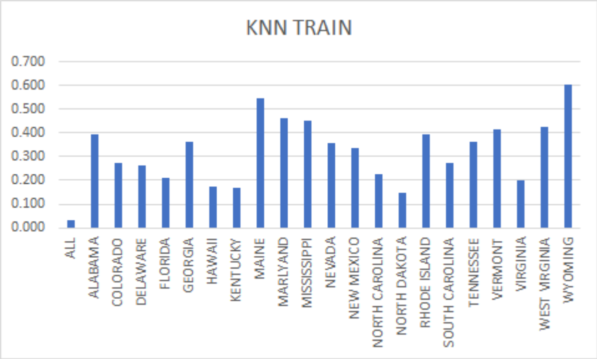
Individual States		
Model Type	**Parameter Value**	
Linear Regression	Alpha = Default	
KNN Regressor	N Neighbors = 3	
Lasso Regression	Alpha = 0.01	
Ridge Regression	Alpha = 0.01	
Decision Tree Regressor	Max Leaf Nodes = 2	Random State = 0
Random Forest Regressor	N Estimators = 100	Random State = 0

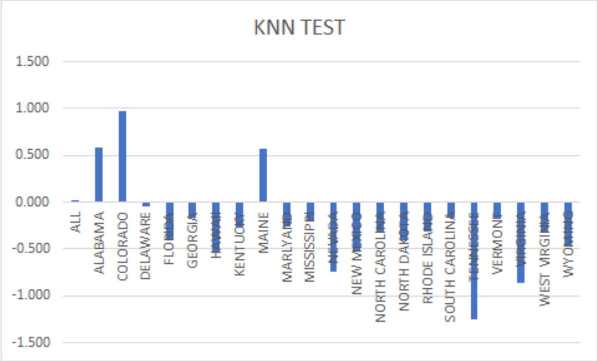
The best performing model was the Linear Regression model, which for most of the states was extremely close to the results of both Lasso and Ridge Regression models as well; approximately in the 40% range for the all states regression R^2 values. The worst performing was the Random Forest Regressor models, which overfitted significantly and produced almost ubiquitously negative test results. Featured in the tables below are graphs of all of the R^2 values for each model type, on train and test data each, across all states, and each state individually.

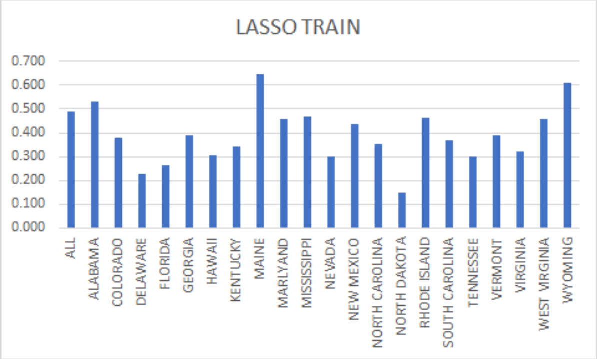
Of important note, is that every model performed relatively poorly and the data set had little strength evidenced in the R^2 values.

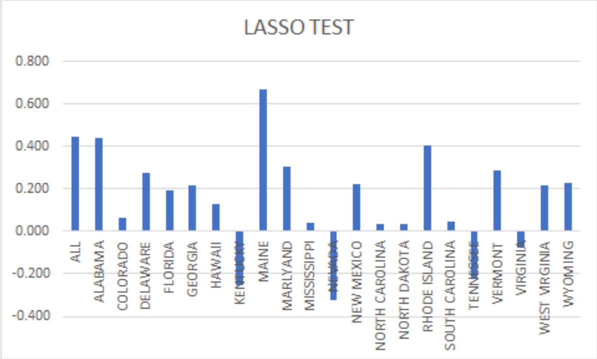


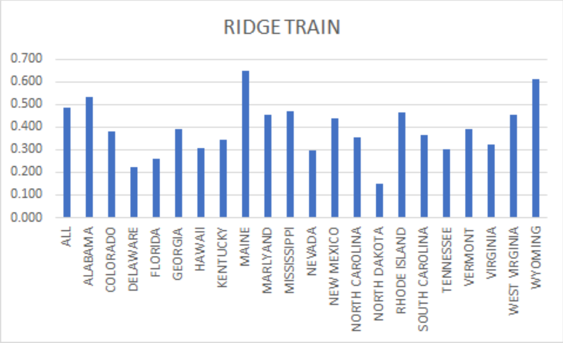


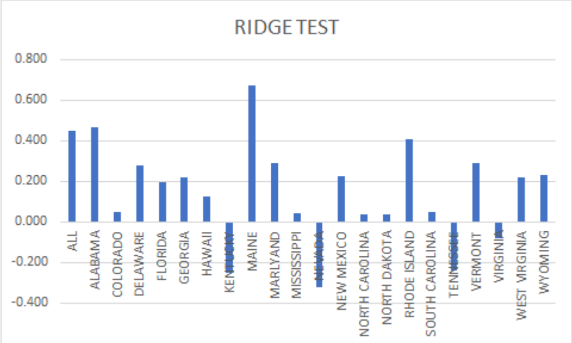

KNN TRAIN
0.700
0.600
0.500
0.400
0.300
0.200
0.100
0.000
ALL
ALABAMA
COLORADO
DELAWARE
FLORIDA
GEORGIA
HAWAII
KENTUCKY
MAINE
MARLYAND
MISSISSIPPI
NEVADA
NEW MEXICO
NORTH CAROLINA
NORTH DAKOTA
RHODE ISLAND
SOUTH CAROLINA
TENNESSEE
VERMONT
VIRGINIA
WEST VIRGINIA
WYOMING

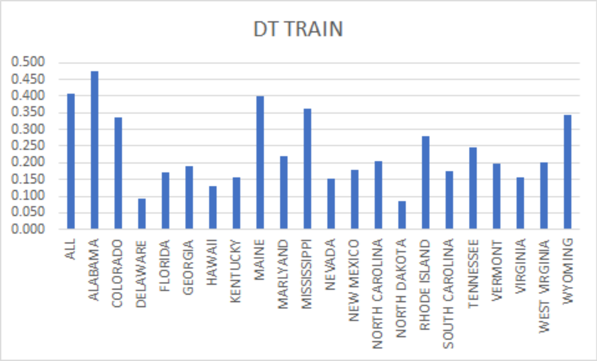

KNN TEST
1.500
1.000
0.500
0.000
-0.500
-1.000
-1.500
ALL
ALABAMA
COLORADO
DELAWARE
FLORIDA
GEORGIA
HAWAII
KENTUCKY
MAINE
MARLYAND
MISSISSIPPI
NEVADA
NEW MEXICO
NORTH CAROLINA
NORTH DAKOTA
RHODE ISLAND
SOUTH CAROLINA
TENNESSEE
VERMONT
VIRGINIA
WEST VIRGINIA
WYOMING

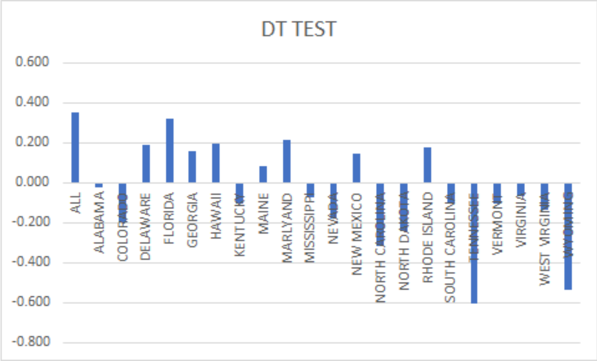

LASSO TRAIN
0.700
0.600
0.500
0.400
0.300
0.200
0.100
0.000
ALL
ALABAMA
COLORADO
DELAWARE
FLORIDA
GEORGIA
HAWAII
KENTUCKY
MAINE
MARLYAND
MISSISSIPPI
NEVADA
NEW MEXICO
NORTH CAROLINA
NORTH DAKOTA
RHODE ISLAND
SOUTH CAROLINA
TENNESSEE
VERMONT
VIRGINIA
WEST VIRGINIA
WYOMING

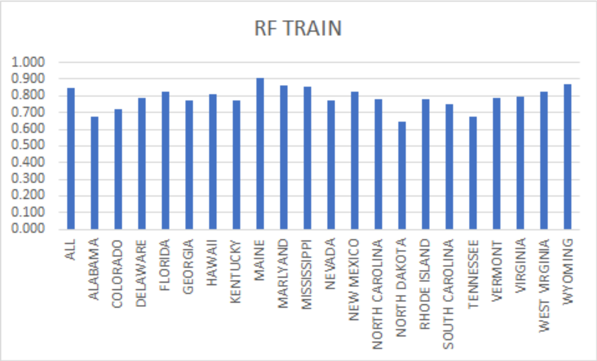

LASSO TEST
0.800
0.600
0.400
0.200
0.000
-0.200
-0.400
ALL
ALABAMA
COLORADO
DELAWARE
FLORIDA
GEORGIA
HAWAII
KENTUCKY
MAINE
MARLYAND
MISSISSIPPI
NEVADA
NEW MEXICO
NORTH CAROLINA
NORTH DAKOTA
RHODE ISLAND
SOUTH CAROLINA
TENNESSEE
VERMONT
VIRGINIA
WEST VIRGINIA
WYOMING

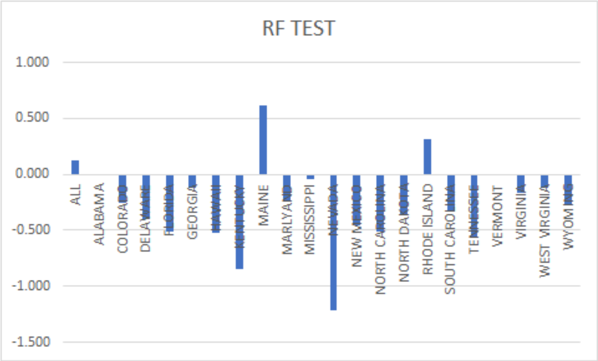

RIDGE TRAIN
0.700
0.600
0.500
0.400
0.300
0.200
0.100
0.000
ALL
ALABAMA
COLORADO
DELAWARE
FLORIDA
GEORGIA
HAWAII
KENTUCKY
MAINE
MARLYAND
MISSISSIPPI
NEVADA
NEW MEXICO
NORTH CAROLINA
NORTH DAKOTA
RHODE ISLAND
SOUTH CAROLINA
TENNESSEE
VERMONT
VIRGINIA
WEST VIRGINIA
WYOMING


RIDGE TEST
0.800
0.600
0.400
0.200
0.000
-0.200
-0.400
ALL
ALABAMA
COLORADO
DELAWARE
FLORIDA
GEORGIA
HAWAII
KENTUCKY
MAINE
MARLYAND
MISSISSIPPI
NEVADA
NEW MEXICO
NORTH CAROLINA
NORTH DAKOTA
RHODE ISLAND
SOUTH CAROLINA
TENNESSEE
VERMONT
VIRGINIA
WEST VIRGINIA
WYOMING

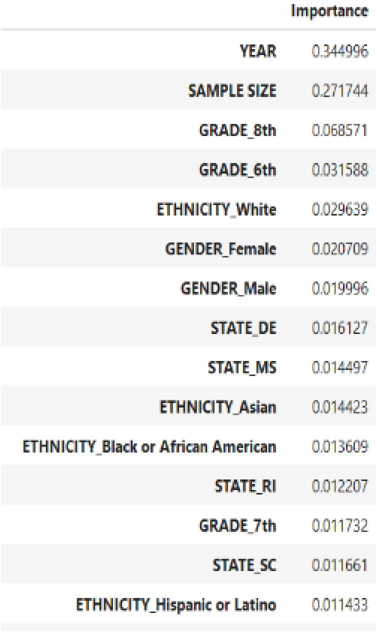

DT TRAIN
0.500
0.450
0.400
0.350
0.300
0.250
0.200
0.150
0.100
0.050
0.000
ALL
ALABAMA
COLORADO
DELAWARE
FLORIDA
GEORGIA
HAWAII
KENTUCKY
MAINE
MARLYAND
MISSISSIPPI
NEVADA
NEW MEXICO
NORTH CAROLINA
NORTH DAKOTA
RHODE ISLAND
SOUTH CAROLINA
TENNESSEE
VERMONT
VIRGINIA
WEST VIRGINIA
WYOMING


DT TEST
0.600
0.400
0.200
0.000
-0.200
-0.400
-0.600
-0.800
ALL
ALABAMA
COLORADO
DELAWARE
FLORIDA
GEORGIA
HAWAII
KENTUCKY
MAINE
MARLYAND
MISSISSIPPI
NEVADA
NEW MEXICO
NORTH CAROLINA
NORTH DAKOTA
RHODE ISLAND
SOUTH CAROLINA
TENNESSEE
VERMONT
VIRGINIA
WEST VIRGINIA
WYOMING





We believe the notable outliers of Maine and Rhode Island are due to the racial homogeneity of the states; each was principally composed of white respondents which was the ethnicity variable of highest importance, likely due to its status as a majority of respondents. Featured below is the full list of feature importance derived from the Random Forest Regressor model for all states.



	Importance
YEAR	0.344996
SAMPLE SIZE	0.271744
GRADE_8th	0.068571
GRADE_6th	0.031588
ETHNICITY_White	0.029639
GENDER_Female	0.020709
GENDER_Male	0.019996
STATE_DE	0.016127
STATE_MS	0.014497
ETHNICITY_Asian	0.014423
ETHNICITY_Black or African American	0.013609
STATE_RI	0.012207
GRADE_7th	0.011732
STATE_SC	0.011661
ETHNICITY_Hispanic or Latino	0.011433

To help predict the risk factor proportion we chose three sub-groups within our middle schoolers in the United States demographic. Among the groups, we changed four defining variables, State, gender, age, and ethnicity to measure if one group was more at risk for alcohol consumption than others. Additionally, each group was tested in two different years (1999 and 2007), to observe the trend over time.

Group	**Characteristics**				**Results**	
	State	**Gender**	**Ethnicity**	**Grade**	**1999**	**2007**
1	Tennessee	Female	Hispanic or Latino	7th	***41.92%***	***34.03%***
2	Maine	Male	White	6th	28.71%	18.28%
3	Florida	Male	African American or Black	8th	***59.52%***	***31.50%***

As you can observe from the table above, we found that since 1999 there was a decrease in risk proportion in all groups. One thing we found in the tests above is that the state variable has very little significance when manipulated compared to other variables. Because of this we can apply our findings to the country as a whole and say that overall since 1999 there has been a decrease in risk of alcohol consumption for middle schoolers living in the United States. To contradict this, Maine displayed no change in proportion. We attribute this to lack of diversity in the data for the state of Maine. Most of the data provided for Maine was largely white males.

Overall we found some promising results. Since 1999 there has been a decrease in risk of alcohol consumption for middle schoolers living in the United States. We recommend that the community be aware that there is still an alcohol abuse problem for middle schoolers that needs to be addressed. Alcohol use is not only bad for development of children at a young age but it can also have an impact on the economy. When a large percentage of students are not graduating high school it decrease the amount of qualified applicants in the workforce. It also has an impact on how public school funds are getting distributed. Another important finding to be aware of is the correlations that persist among different demographics group. Some communities can be more at risk than others but can also display the biggest change over time. What we do not want people to do solely with this data is take action. This data should only be used for general perspective to lead into more research, and should not be used as a reason to target certain ethnic or gender groups to fix the problem. The CDC has provided a very subjective data set that gave poor data results in the context of our predictive analytics. In order for this data to be of any value the surveyors need to incorporate more objective variables to provide a more complete data set, capable of producing actionable results.