HarvardX_PH125.9x_Capstone_Predict_Test_Scores_Project

Janalin Black

7/12/2021

1. Introduction

1.1 Assignment

The purpose of this project is to train a machine learning algorithm to predict posttest scores with a given set of variables. The expectation is to use different approaches to create multiple models, each of which predicts a likely outcome. Specifically, the goal is to go beyond a simple linear regression model and explore more advanced techniques available for data prediction.

This report includes three different models, all successful at predicting reasonable posttest scores. All models include a variety of approaches including ANOVA, Chi-square, Akaika Information Criterion (AIC), Random Forest, Factor Analysis of Mixed Data (FAMD), Emsembles, regularization, and Variable Importance.

1.2 DataSet

All three models are created using the Predict Test Scores of Students dataset found at Kaggle.com. This dataset is also loaded into GitHub at Predict Test Scores Dataset. The original dataset is split three ways: "scores" dataset is used for training, "scores_testing", which includes 20% of "scores" dataset, is used to test potential models, and "validation", which includes 20% of the remaining "scores" dataset, is used only to determine the accuracy of final predictions.

2. Methods and Analysis

2.1 Data Cleaning

A review of the original dataset shows a data.frame with 11 columns and 2133 observations and no missing values. The 11 variables include 8 factors and 3 numerical classifications. Since the variable student_id is a randomly assigned value, it shouldn't be considered for analysis and is changed to a character classification.

Original data prior to cleaning

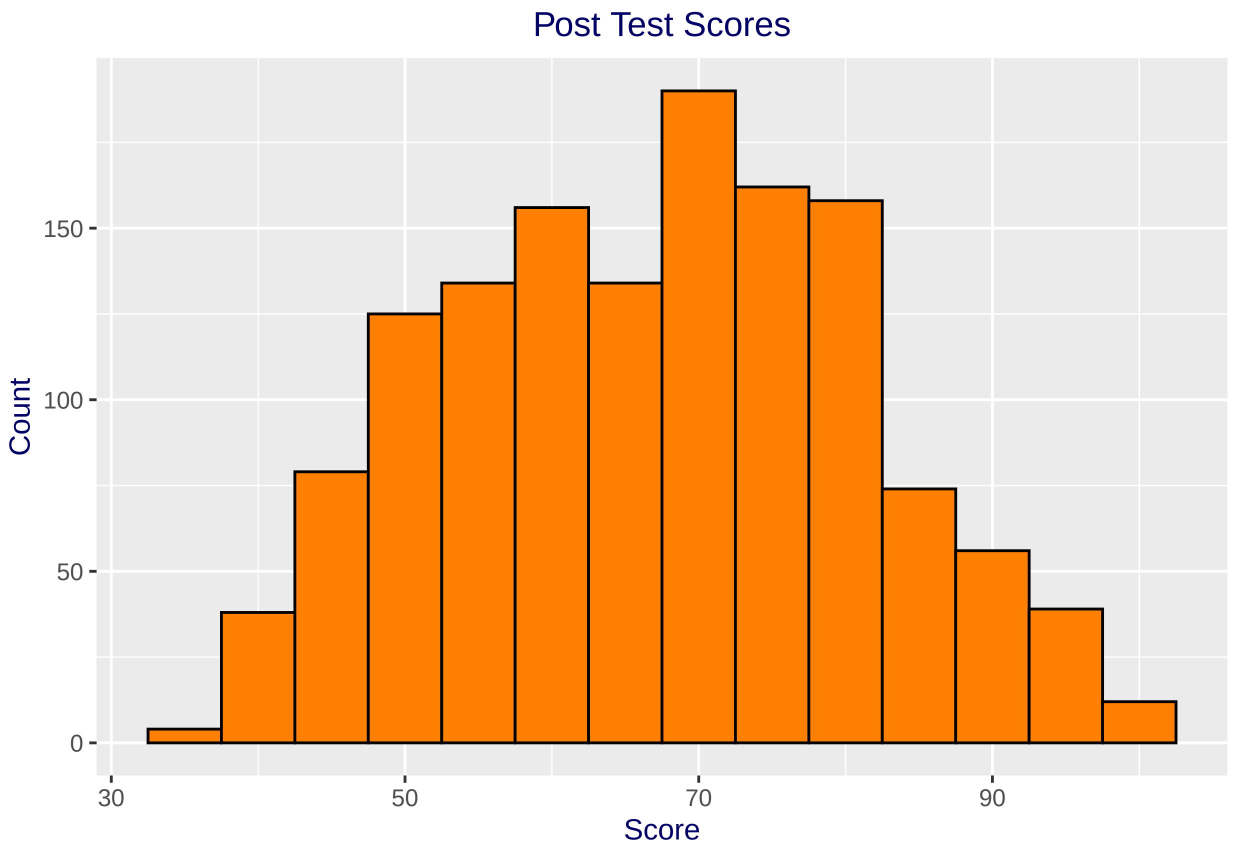
```
##   school school_setting school_type classroom teaching_method n_student
## 1  ANKYI          Urban  Non-public       6OL        Standard        20
## 2  ANKYI          Urban  Non-public       6OL        Standard        20
## 3  ANKYI          Urban  Non-public       6OL        Standard        20
## 4  ANKYI          Urban  Non-public       6OL        Standard        20
## 5  ANKYI          Urban  Non-public       6OL        Standard        20
## 6  ANKYI          Urban  Non-public       6OL        Standard        20
##   student_id gender            lunch pretest posttest
## 1      2FHT3 Female Does not qualify      62       72
## 2      3JIVH Female Does not qualify      66       79
## 3      3XOWE   Male Does not qualify      64       76
## 4      556OO Female Does not qualify      61       77
## 5      74LOE   Male Does not qualify      64       76
## 6      7YZO8 Female Does not qualify      66       74
```

Wrangled training dataset This is the updated scores dataset used for training the model.

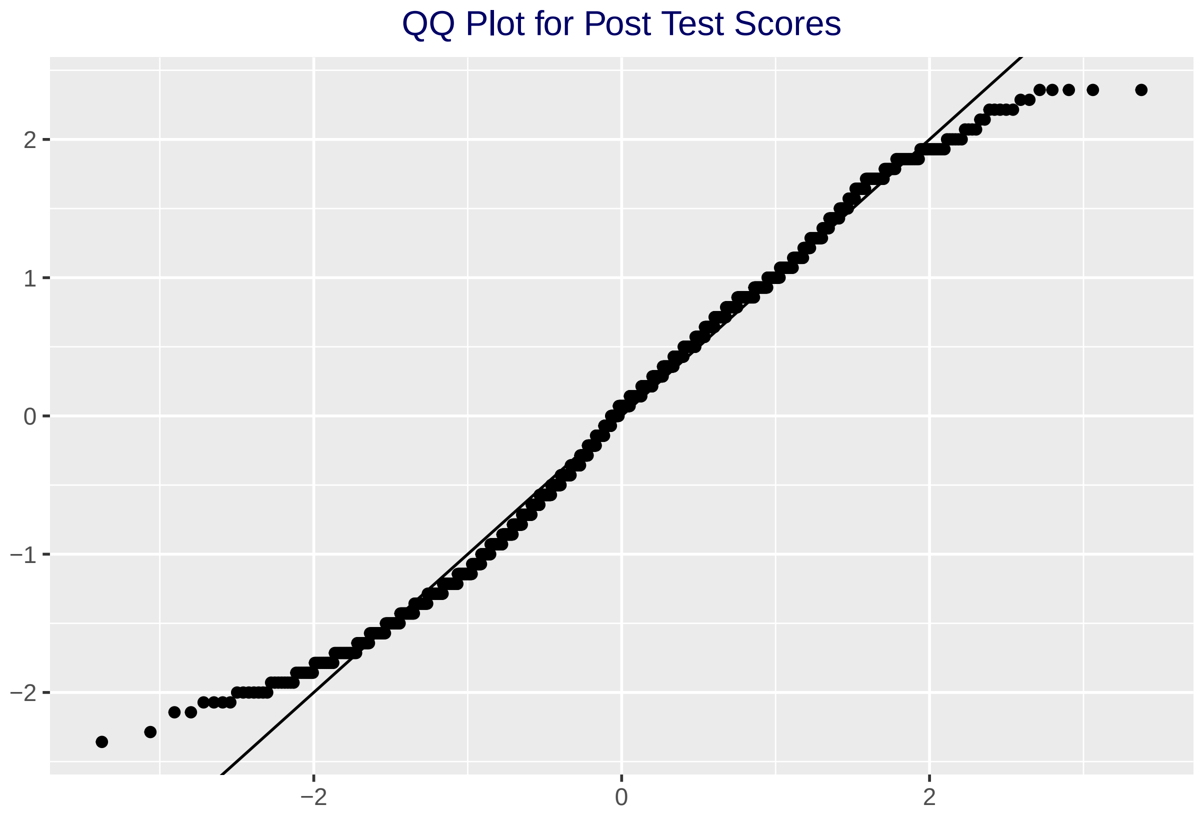
	school	school_setting	school_type	classroom	teaching_method	n_student	student_id	gender	lunch	pretest	posttest
1	ANKYI	Urban	Non-public	6OL	Standard	20	2FHT3	Female	Does not qualify	62	72
2	ANKYI	Urban	Non-public	6OL	Standard	20	3JIVH	Female	Does not qualify	66	79
5	ANKYI	Urban	Non-public	6OL	Standard	20	74LOE	Male	Does not qualify	64	76
6	ANKYI	Urban	Non-public	6OL	Standard	20	7YZO8	Female	Does not qualify	66	74
7	ANKYI	Urban	Non-public	6OL	Standard	20	9KMZD	Male	Does not qualify	63	75
11	ANKYI	Urban	Non-public	6OL	Standard	20	DZMKU	Male	Does not qualify	61	73

2.2 Data Exploration and Visualization

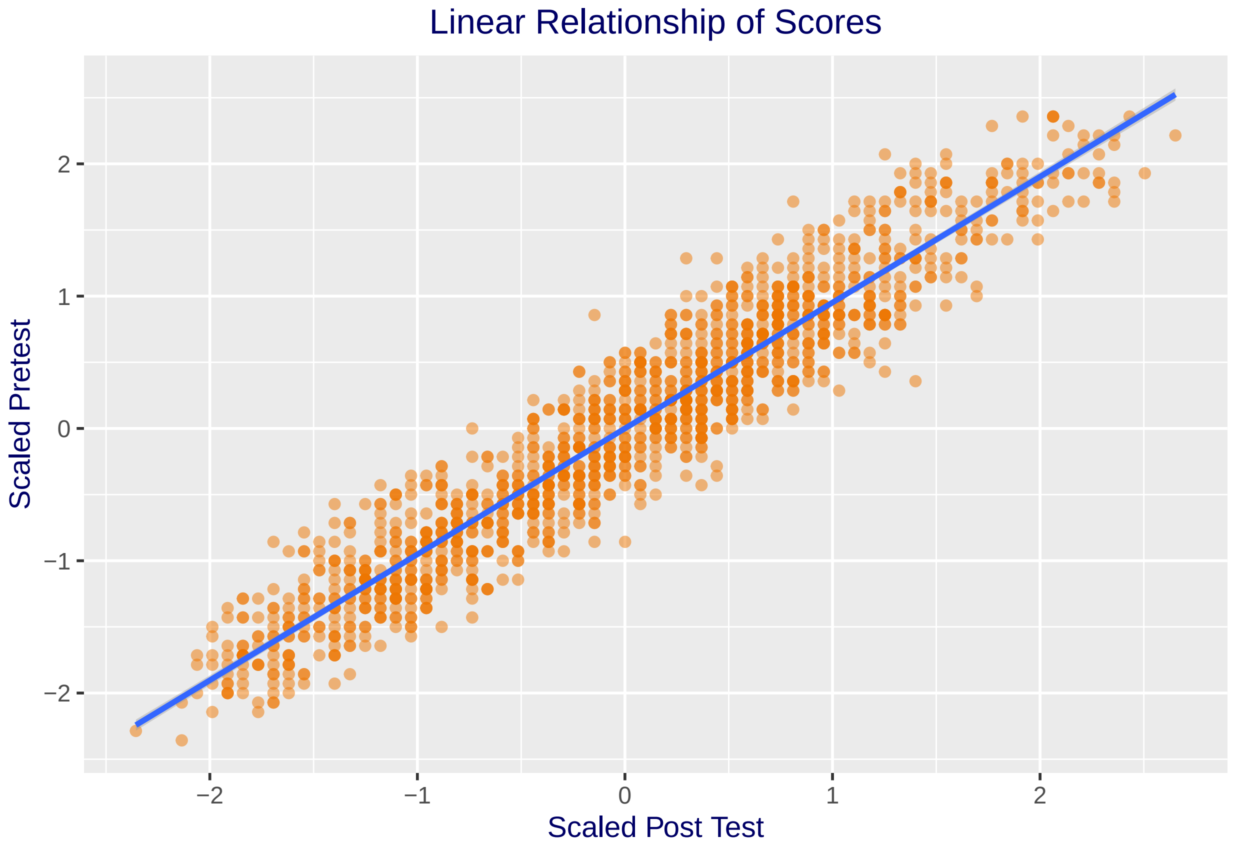
A review of the training set (scores dataset) shows posttest scores with a mean of 67, a median of 68, and a range between 34 and 100. Additionally, the following histogram seems to show a normal distribution of the dependent variable, posttest scores.



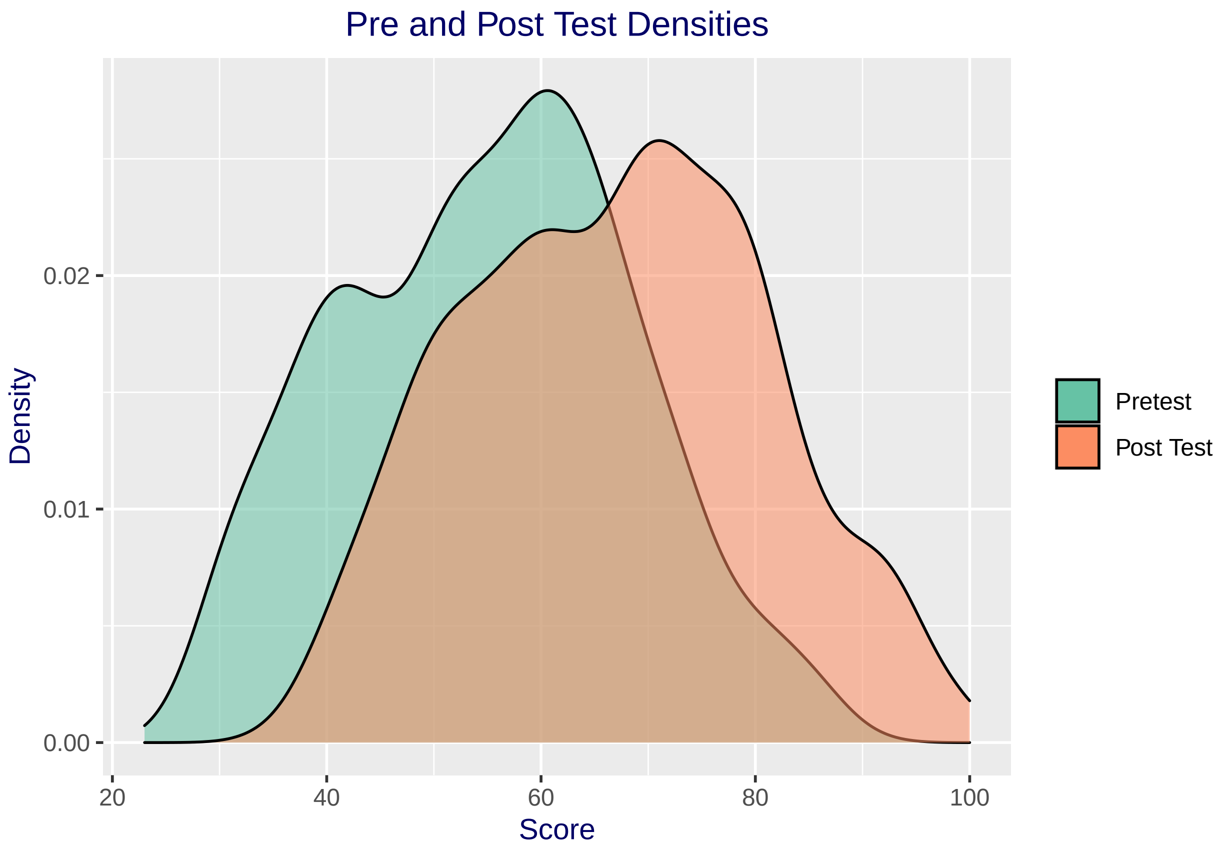
Further exploration of posttest numbers reveals a qqplot that also appears to show a normal distribution of scores.



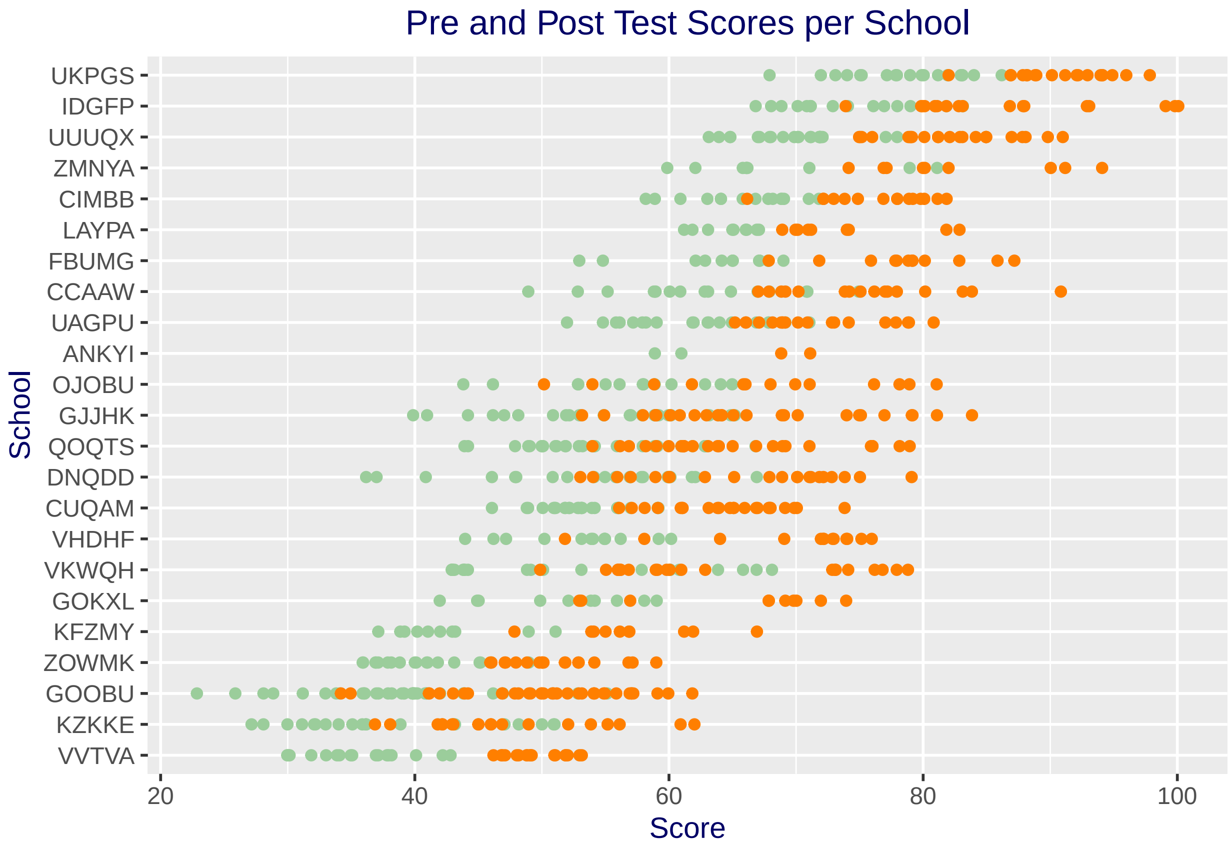
2.2a Pretest Scores Evaluation of pretest scores reveal a high linear relationship with posttest scores. Although the mean score of 55 for the pretest is significantly lower than the mean score of the posttest at 67, a linear regression model indicates a large portion of the variance in posttest scores are explained by pretest scores with an R-squared value of 0.91; 91% of the variability can be explained with pretest scores. Additionally, Pearson's correlation indicates a positive relationship of 95% and a significant P-value with an alpha level below .05%. Further discovery indicates individual students score similarly on the pretest and the posttest. 39 pupils who scored in the top 50 for the posttest also scored in the top 50 for the pretest and 30 students who scored in the bottom 50 for the posttest also scored in the bottom 50 for the pretest.



A density plot comparing pretest and posttest scores indicate similar distributions of variables, overlap of pre and posttest scores, and higher scores for the posttest.



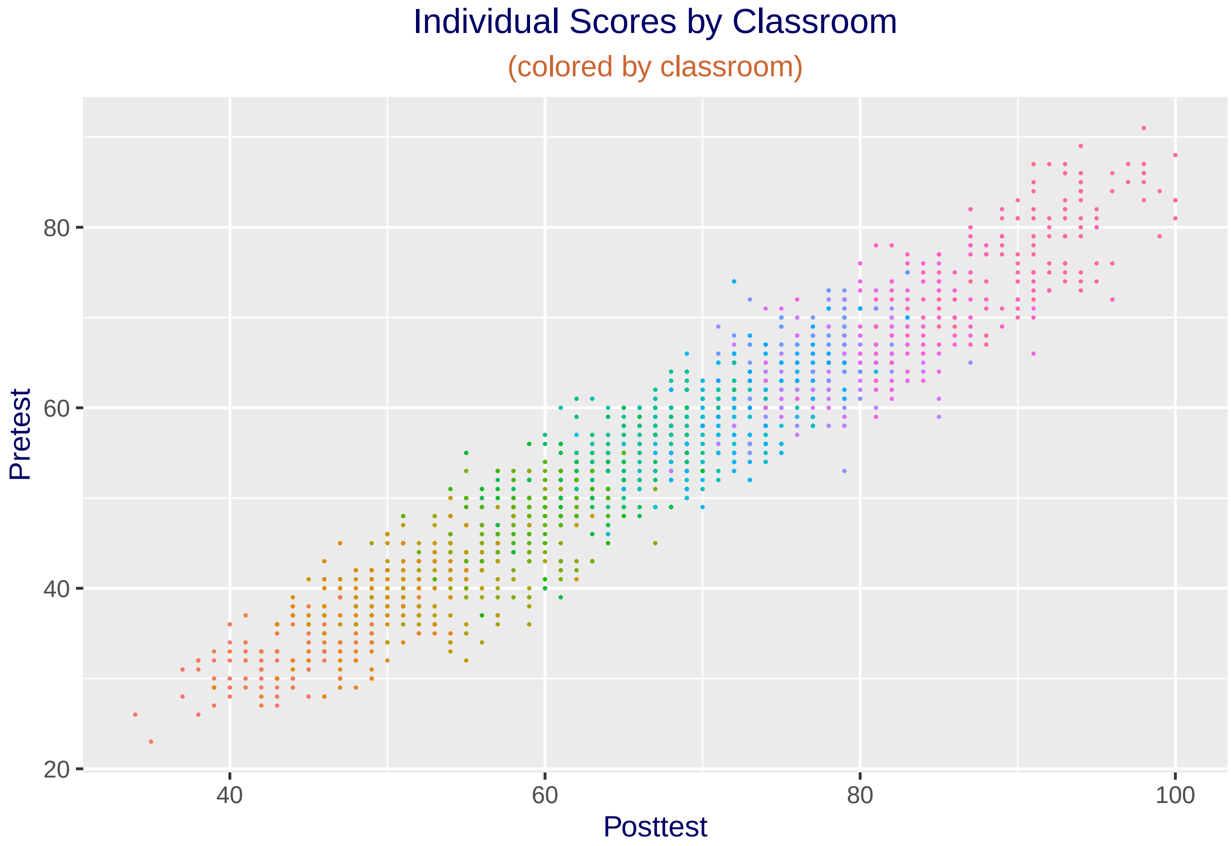
2.2b Schools Observations in the original dataset include 23 different schools with posttest school mean scores ranging from 47 to 91. Schools appear to score similarly on the pretest and the posttest as 4 schools who scored in the top 5 for the posttest also scored in the top 5 for the pretest. In addition, the following plot created from a sample of the scores dataset shows some schools do better than others on both pretest and posttest scores.



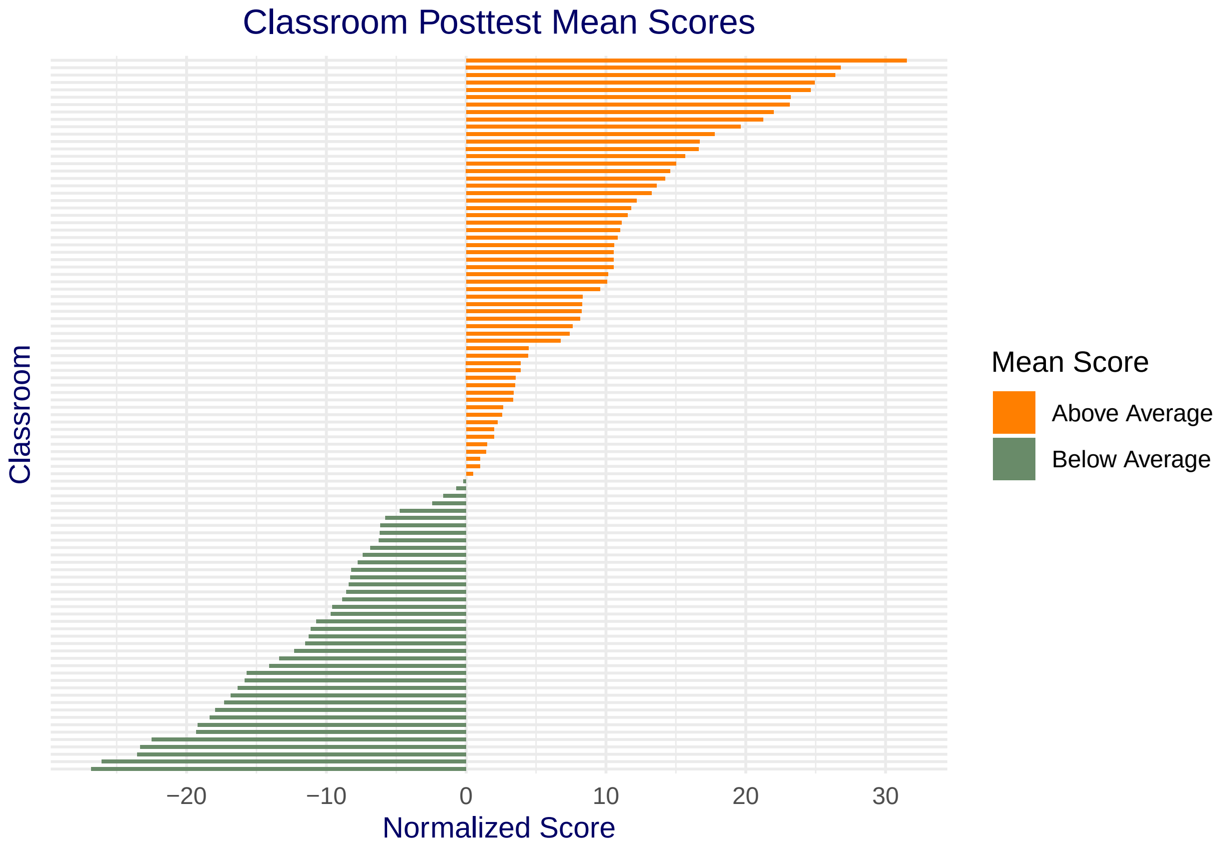
Further exploration of schools shows there are significant differences between the means of some posttest scores. Bartlett's test of homogeneity of variances indicates there are significant differences between schools with a p-value of 0; Welch's one-way analysis of means also shows a p-value of 0. Bonferroni's correction of pairwise comparisons using t tests indicate some school mean scores are significant and others are not.

2.2c Classroom Another potential predictor is different classrooms. The original dataset has observations from 97 classroom classifications with mean posttest scores ranging between 40.2 and 98.5.

The following plot shows differences between scores based on the classroom, with some classrooms scoring significantly higher than others.

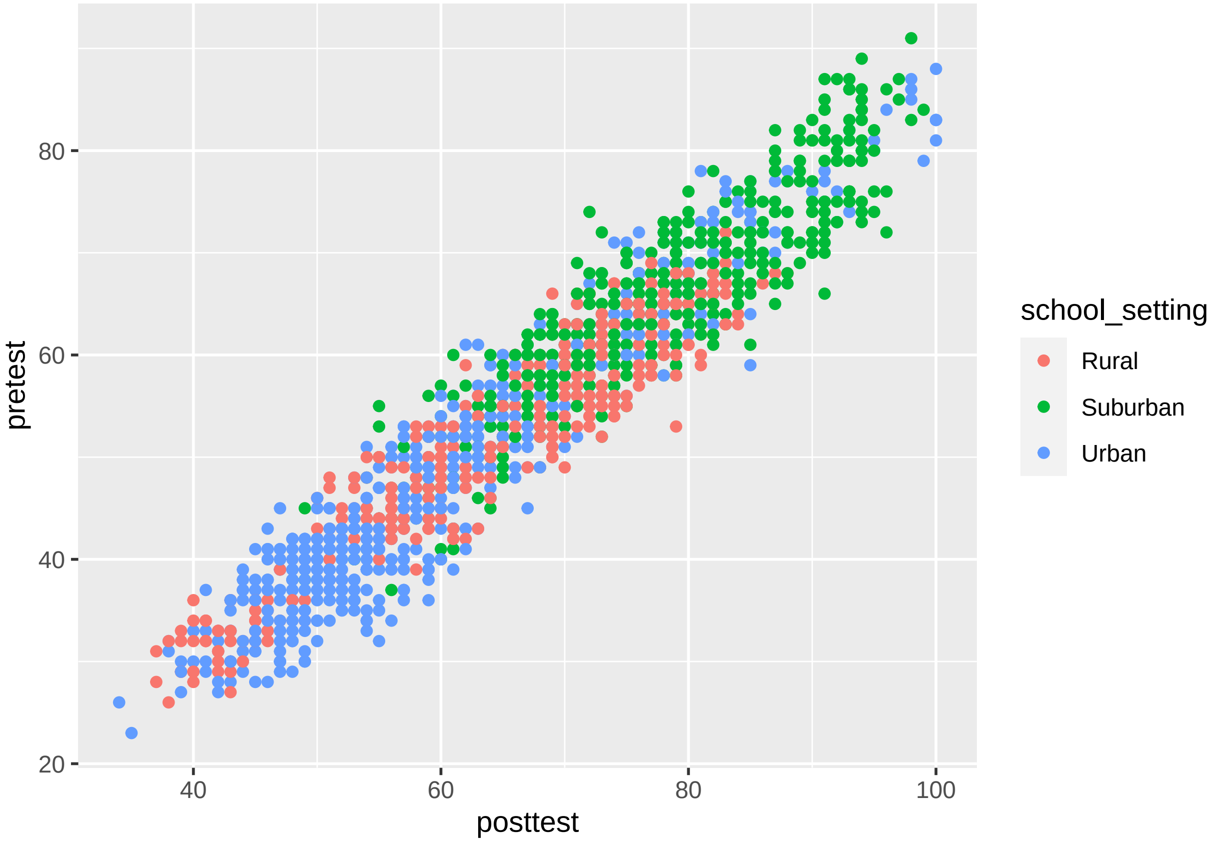


A diverging plot with normalized classroom posttest scores also shows there are differences in scores based on the classroom.

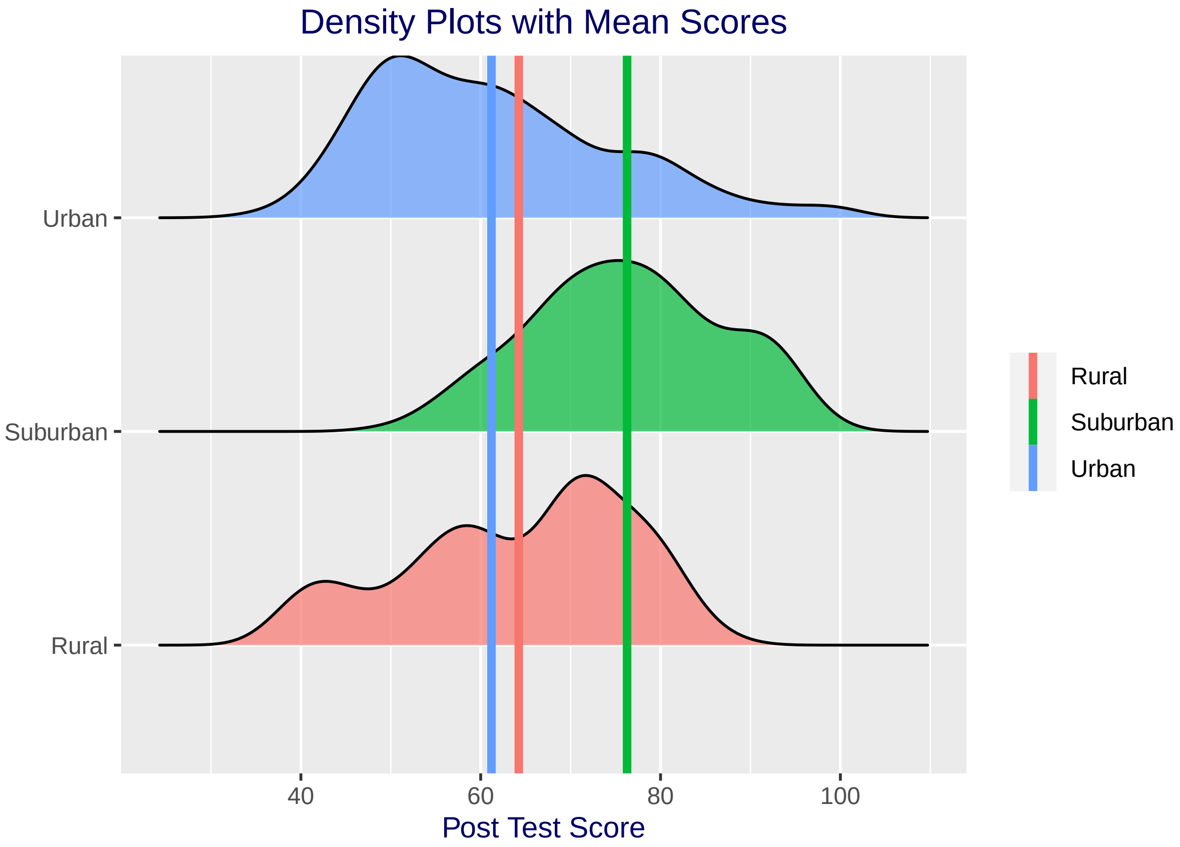


2.2d School Setting The variable school setting has three levels, with rural schools representing 24%, suburban schools at 34%, and urban schools with 42%. Further exploration shows there are differences between the means of posttest scores for all three groups. Bartlett's test of homogeneity of variances indicates there are significant difference between school settings with a p-value of 0 and Welch's one-way analysis of means also shows a p-value of 0. Bonferroni's correction of pairwise comparisons using t tests indicate significant differences between all three groups with the highest p-value at 0.001 when comparing rural to urban schools.

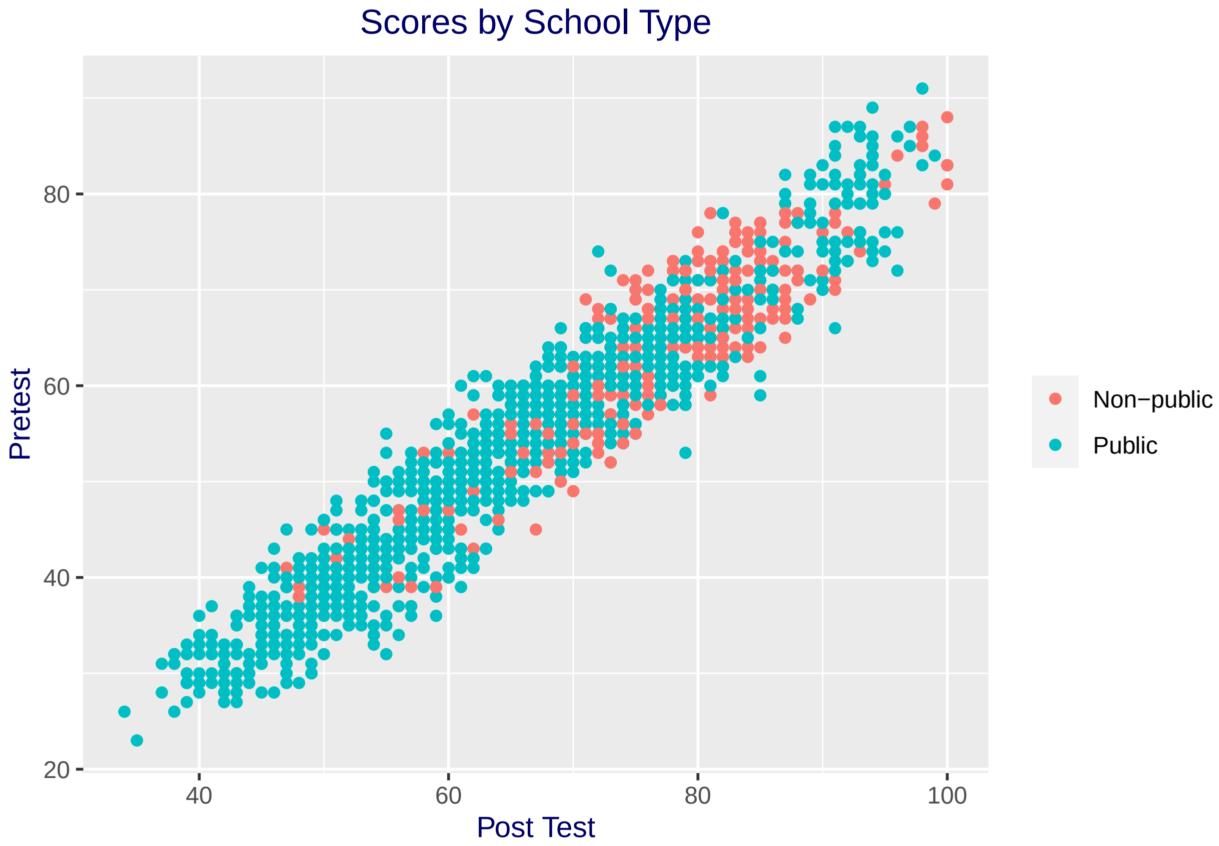
This plot shows differences in posttest and pretest scores versus school setting. Many suburban schools score higher than urban schools with only a few suburban schools schools with low scores.



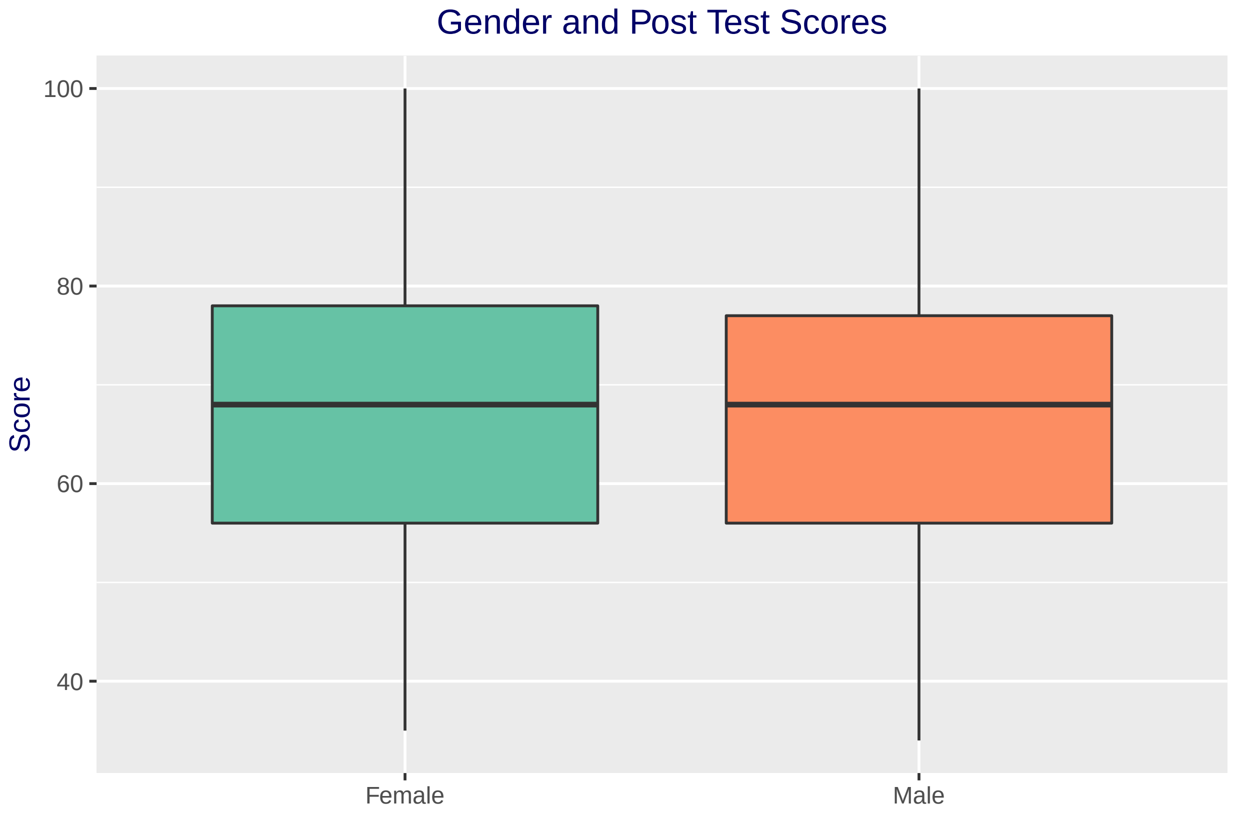
This second image shows that suburban schools score higher on posttest scores than both rural and urban schools. Additionally, mean scores for rural and urban schools are similar, which agrees with Bonferroni's results.



2.2e School Type Categories in school type include non-public schools with 26% and public schools with 74% of observations. There appears to be differences between school type with non-public schools outperforming public schools.

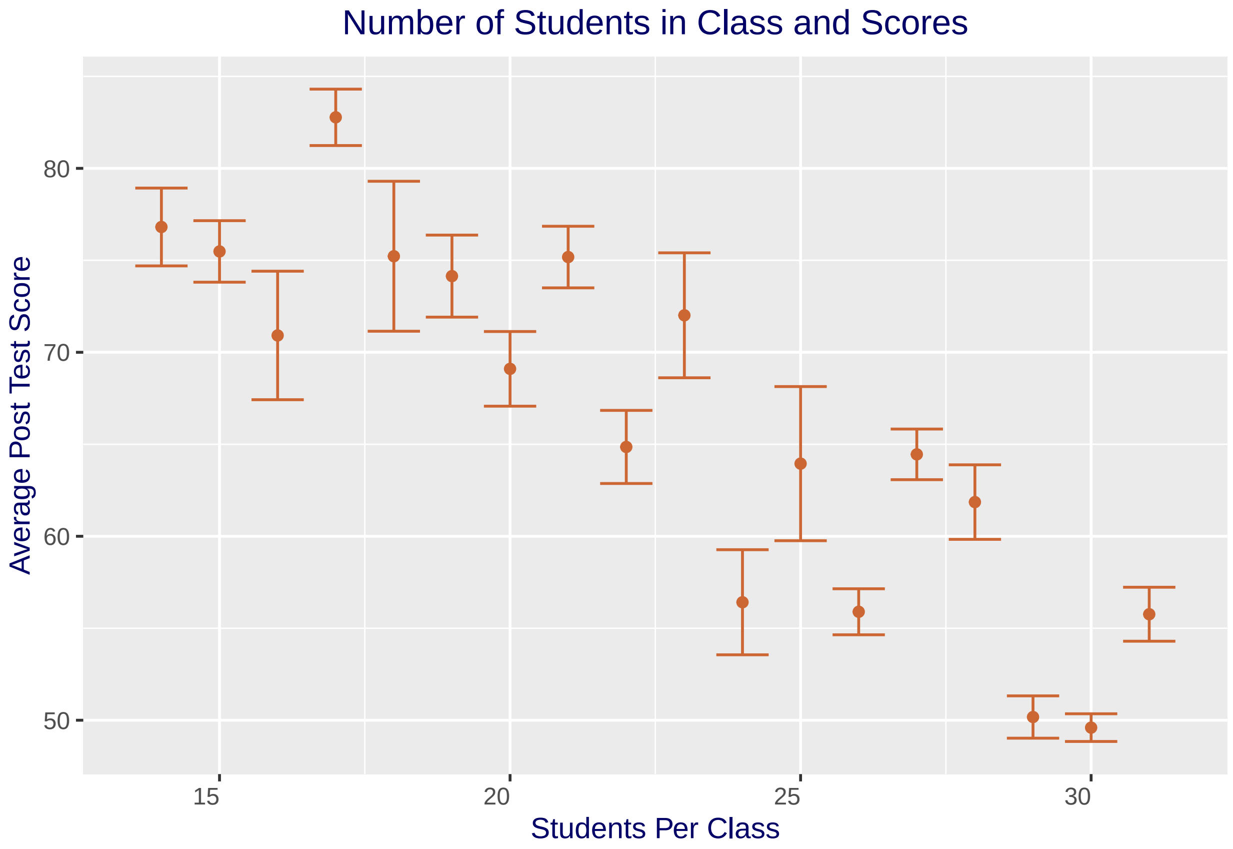


2.2f Gender The gender variable indicates that sexes are nearly equally split with 49.8% female and 50.2% male. The initial evaluation of gender as a predictor is not promising as the following plot shows nearly identical boxplots on posttest scores for both genders.



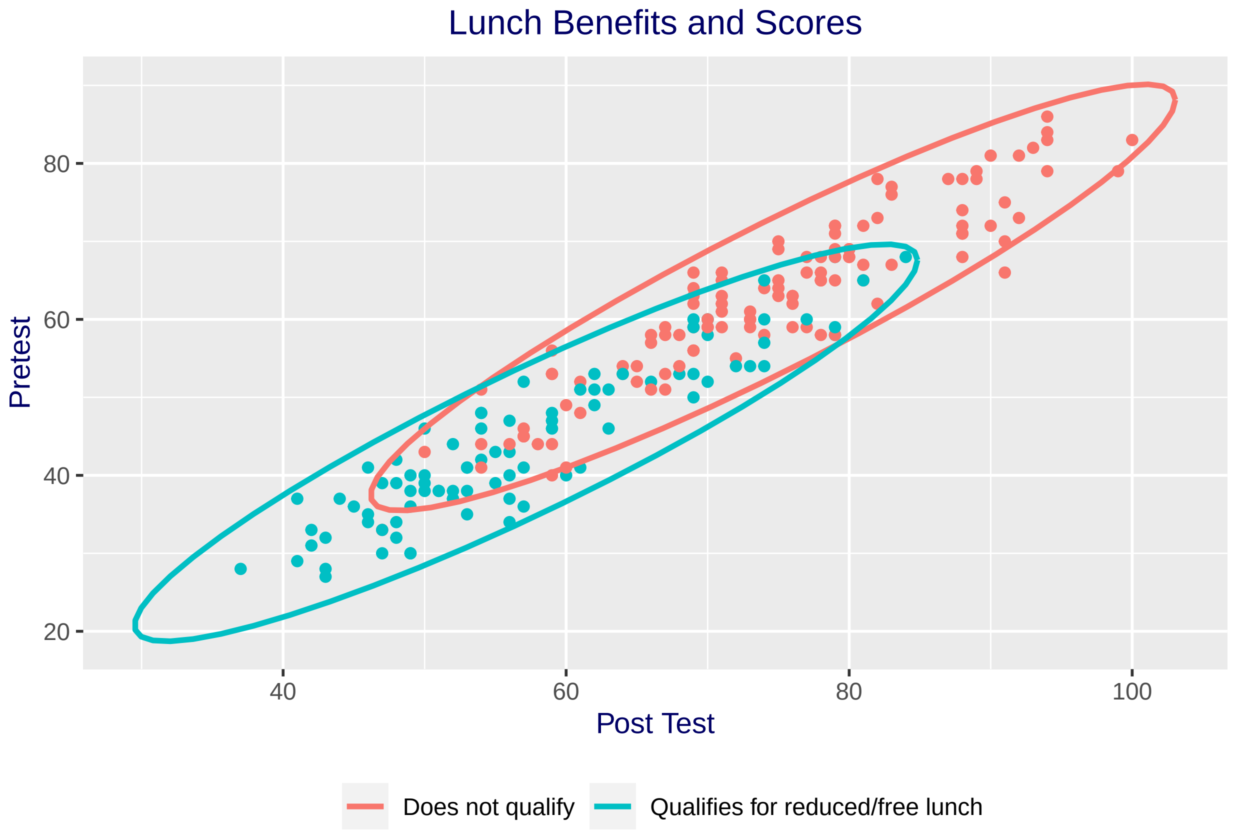
2.2g Number of Students in the Classroom The number of students per classroom ranges between 14 and 31 with the average number at 23. Correlation between the number of students in a classroom and posttest scores is negative at -0.51, which indicates classrooms with higher numbers of students have lower scores.

The following plot visualizes the negative relationship between classroom numbers and test scores. As the number of students in a classroom increases, posttest scores decrease.



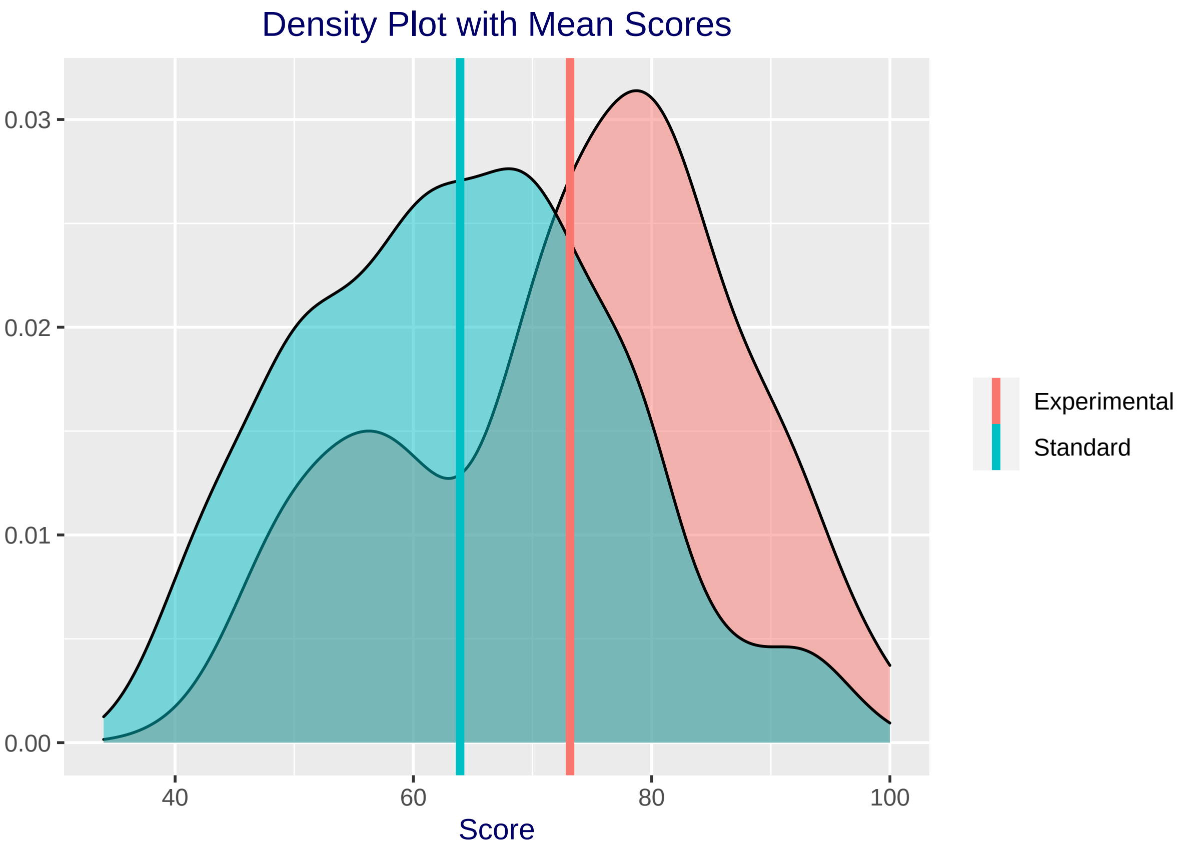
2.2h Free and Reduced Lunch Another potential variable is whether or not a student qualifies for free or reduced lunch. From the dataset, 43% of all students qualify for free or reduced lunch and the other 57% do not.

This plot created from a sample of the scores dataset indicates there may be a relationship between student lunch benefits and test scores. Ellipses show differences between test scores for students who do and don't qualify for free and reduced lunch.



2.2i Teaching Method This last variable is the teaching method utilized, experimental or standard. In the dataset, there are 33% experimental and 67% standard teaching method groups. Looking further, the posttest mean score for the experimental group is 73.15 and 63.92 for the standard method group.

Showing these numbers visually reveals differences between experimental and standard posttest densities and mean scores, with the experimental group outperforming the standard group.



2.3 Models

Three models are trained using the scores training set to predict posttest scores. Final models for each of the three groups is then tested on the scores test dataset. Because testing on these models happens throughout the process of model creation, a final test for the winning model is done on the validation set, the final hold-out test set, to determine the best overall fit.

2.3a MODEL 1 The first model is a linear approach that utilizes ANOVA and Chi-squared tests to discover variables to consider for potential models. The Akaike information criterion (AIC) is used on five iterations for variable choices to determine the best options for the final model.

Because there are multiple promising independent variables to consider, variables that don't seem to have an impact on the regression model are omitted in different versions of potential models.

The first approach includes all variables. Since every variable, excluding gender, had promising predictive power, tossing them all in seemed a good way to start. Adjusted R-Squared indicates 96% of the variation in posttest scores are explained with this first approach. An ANOVA analysis indicates >.01 significant p-values for all variables analyzed except gender, which has an insignificant p-value of 1.

The variable school wasn't utilized in the calculation. A Chi-squared test measuring the relationship between school and classroom indicates a p-value of 0, or highly significant. It may be the case that the variables school and classroom have collinearity as classrooms are within schools.

The second and third approaches to this model utilize what was learned from the first analysis: gender, school, and classroom variables either have interactions with each other or aren't significant to posttest prediction. The second approach includes all significant variables from the first approach; all variables are considered except gender and school. The third approach excludes gender but includes the variable school by excluding classroom.

Variables excluded in the fourth approach are school, gender, and lunch. It was found with ANOVA-between variables that there is a significant interaction between lunch and school setting with a p-score of 0. It's possible that there may be a difference between socio-economic status of urban, suburban, and rural schools which could create correlation with qualifications for reduced/free lunch.

The fifth approach excludes variables school, lunch, and school setting. Since there is a significant interaction between school type and school setting with a p-score of 0, school setting was removed from this final model.

MODEL 1 RESULTS The Akaike information criterion (AIC) is used to test models for best fit. AIC determines the overall value of each model by comparing the explained variation with the number of parameters. The lowest AIC value indicates the most information explained of all models.

Applying AIC to the five approaches shows AIC scores ranging from 6810.7 to 6996.6, with the lowest score for the first model. This suggests the first model is the best fit and will achieve the lowest RMSE of the five options. From AIC, the first model yields an AIC weight of 0.843, which indicates that 84.3% of total variation in the posttest score can be explained by this model.

Testing the first approach, Model 1, show the following RMSE score for the scores test set.

Model 1	RMSE Score
Test Set	2.74

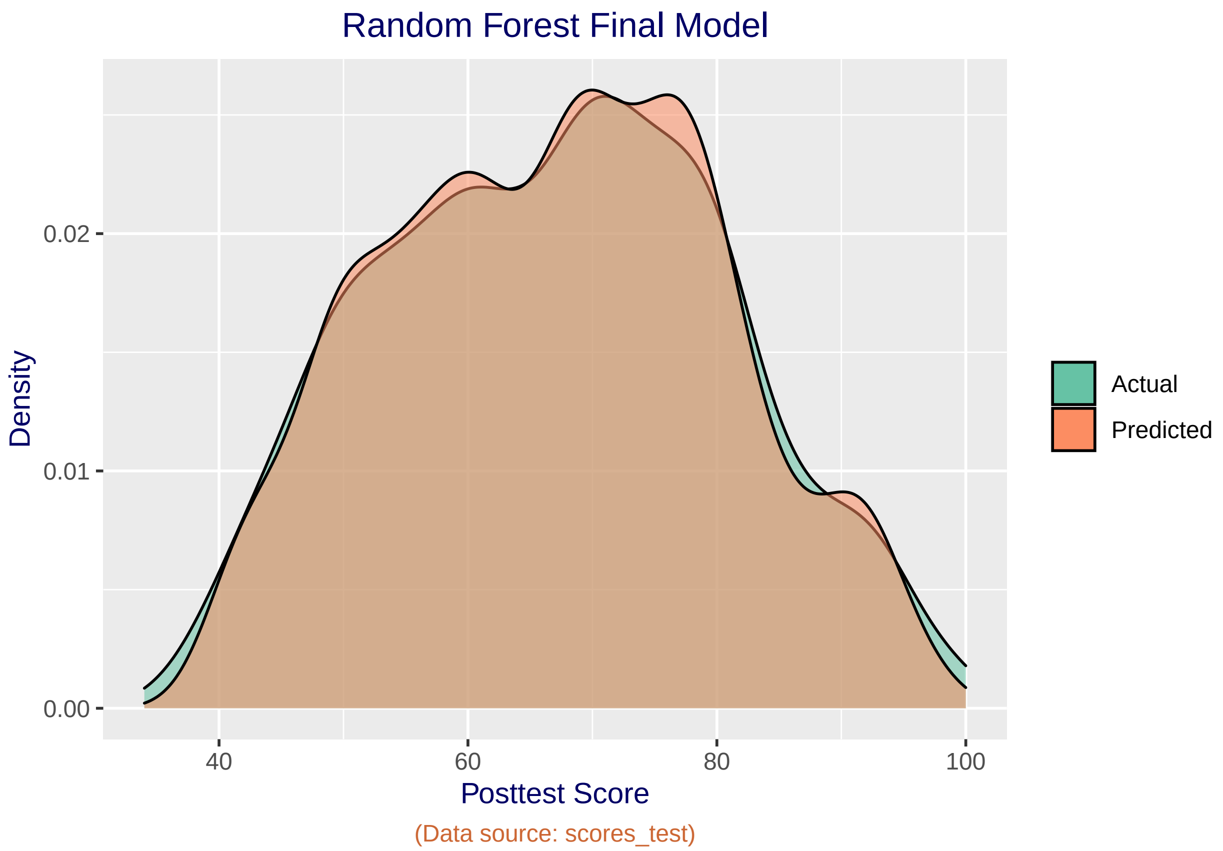
2.3b MODEL 2 The second approach to creating a best-fit model is using Random Forests. The Random Forest tool is selected because it is robust when there are correlated predictors, which seems to be a problem with several of the variables in this dataset. Random Forest is also capable of finding the importance of variables and will be helpful in designing Model 3. One drawback of using Random Forest is it's limitation on groups within variables. The selected Random Forest calculator limits the number of groups to 53, while the predictor classroom has 97. Since the classroom variable was significant for Model 1 results, this will be a limitation for Model 2.

Four models are created using Random Forest, each with a different set of predictors. After training and testing each of these models, the last model, Fit 4, produces the best RMSE score. From here, parameters for Fit 4 are trained for number of variables sampled (mtry) and node size. Results indicate that Fit 4 with added parameter adjustments produced the best fit model.

A summary of each random forest model and RMSE conclusions are listed in the table below. Notice the lowest RMSE from Model 2 is 2.93 which is good, but Model 1 still outperforms with an RMSE of 2.74.

Random Forest Model(classroom excluded)	Test Set RMSE
Fit 1 - pretest as only predictor	4.46
Fit 2 - best-fit predictors from Model 1	5.45
Fit 3 - doesn't include pretest	4.36
Fit 4 - all predictors	3.03
Fit 4 Adj. - adding best mtry/node size	**2.93**

Visualizing the predictive power of Model 2 shows densities of predictions compared to actual posttest scores are quite similar.



A final analysis done on Model 2, the Random Forest model, is to evaluate the variable importance for predictors included in the winning model. These values will be helpful in constructing the final model, Model 3.

The table below shows the variable importance for the best model using Random Forest. Teaching method is listed as having the most importance and gender the least.

	Overall
teaching_method	84.25
pretest	62.62
school	24.27
n_student	21.49
lunch	13.13
school_type	11.97
school_setting	9.81
gender	4.20

2.3c MODEL 3 This final model utilizes information learned about the variables from Model 1 and Model 2 to construct a learning algorithm.

The first step used for this model is including the generalized linear model (glm) for the variable pretest. From Model 1, it was assessed that pretest had a strong linear relationship with the dependent variable, posttest. Additionally, constructing an ensemble of potential methods for predicting scores from a numeric variable indicated that the glm approach was best.

Results in the table below indicate that "glm" is the best model for predicting posttest scores from the pretest.

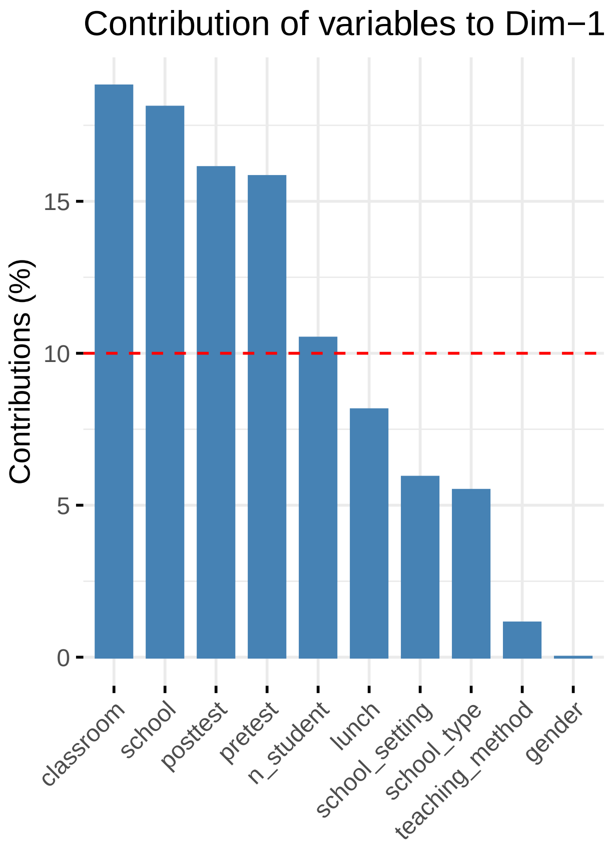
Model	RMSE
lm	4.31
glm	**4.29**
rlm	5.30
knn	4.47
rf	4.46
svmLinear	4.33

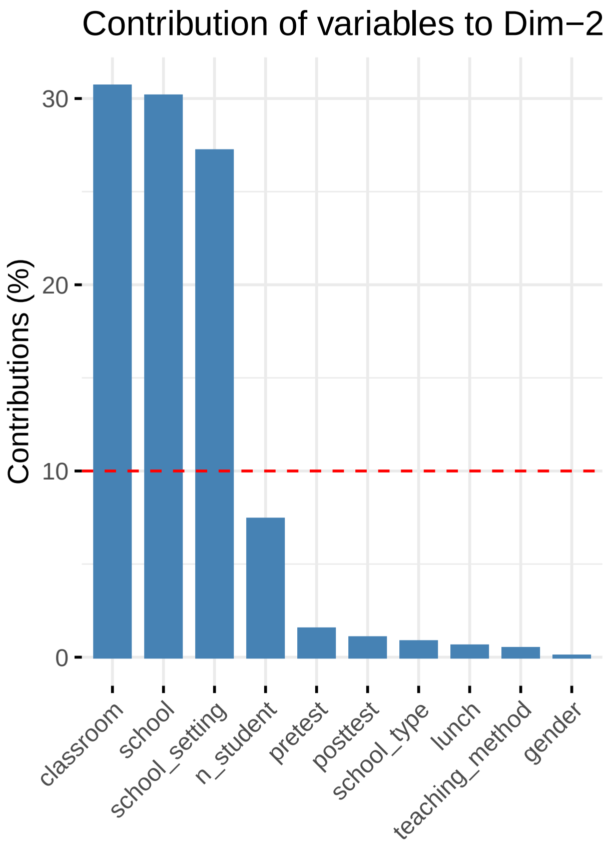
Adding the pretest variable to the average of posttest scores yields improvement over just the average of all scores when applying it to the test dataset. However, this RMSE result is still higher than what was found in both previous models.

Model 3	RMSE
Posttest Average	13.83
Pretest glm Model	4.46

Because the Random Forest model, Model 2, doesn't include the predictor "classroom" in the analysis, the variable importance of results may be missing valuable information from this variable. Because of this, Factor Analysis of Mixed Data (FAMD) is used to further analyze potential contributions to the scores dataset, which has both quantitative and qualitative predictors.

From FAMD results, the two tables below indicate classroom is the highest contributor for both dimensions 1 and 2. From these results, the classroom predictor will be added next to our model using typical error loss. Residual Mean Squared Error (RMSE) will be used on the predictions using the scores test set.





The table below shows a lower RMSE score with classroom added as a predictor. Additionally, from FAMD, some classrooms are used as predictors and others are not. Regularization is applied to classroom to account for some classrooms with limited posttest scores. This results in a slightly lower RMSE number.

Model 3	RMSE
Posttest Average	13.8304
Pretest glm Model	4.4597
Classroom Effect	3.3041
Regularized Classroom Effect	3.3029

From here, the remaining predictors are added to our error loss model ordered according to Variable Importance from the Random Forest model. Results using this approach indicate improved RMSE scores overall. However, some variables increase RMSE scores.

Model 3	RMSE
Posttest Average	13.8304
Pretest glm Model	4.4597
Classroom Effect	3.3041
Regularized Classroom Effect	3.3029
Teaching Method Effect	3.3024
Scool Effect	3.3020
n_Students in Class Effect	3.3033
Lunch Effect	3.2817
School Type Effect	3.2833
School Setting Effect	3.2842
Gender Effect	3.2720

Since the variables school type and school setting raise RMSE scores, they are removed from Model 3. This results in a slightly lower score as shown by the following table.

Model 3	RMSE
Posttest Average	13.8304
Pretest glm Model	4.4597
Classroom Effect	3.3041
Regularized Classroom Effect	3.3029
Teaching Method Effect	3.3024
Scool Effect	3.3020
n_Students in Class Effect	3.3033
Lunch Effect	3.2817
Revised Gender Effect	**3.2700**

3. Results

Comparing the three models, the table below shows that Model 1 has the losest RMSE score of 2.74 when evaluated using the scores test dataset.

Model	Test Set RMSE
Model 1	**2.74**
Model 2	3.03
Model 3	3.27

From here, Model 1 is applied to the final hold-out, the validation dataset. This is the only time the validation set has been used in this project. All prior RMSE scores were obtained using the train and test datasets. Prediction differences between the test set and the validation set are small with a difference of 0.14. From this model, test scores should be predicted within 2.88 points.

Model 1	RMSE Score
Test Set	2.74
Final Validation Set	**2.88**

3.1 Limitations

One challenge with Model 2 was limitations on groups within variables using random forests. The Random Forest program used in calculations limited the number of groups within one variable to 53, while classroom has 97. Since the classroom variable was significant for Model 1, including it in Model 2 was a logical next step.

A second concern was with collinearity between variables. The degree of correlation between several variables was high, which made it difficult to interpret results. Although, adding all variables to the model resulted in the best overall fit.

3.2 Recommendations for Future Study

Future studies may consider including additional predictors to the dataset. Helpful data may be age of student, race/cultural background, family income, number of siblings, location of school, and number of parents in the home. These variables may provide better predictive power than some currently included.

Additionally, this dataset could be expanded and used in a longitudinal study to predict future events such as GPA's, college entrance exam scores, college application/acceptance numbers, chosen majors, and graduation rates.