**Massachusetts K-12 Schools Data: A Graphical Exploration and Analysis**

Glenn Turner

Volgenau School of Engineering, George Mason University

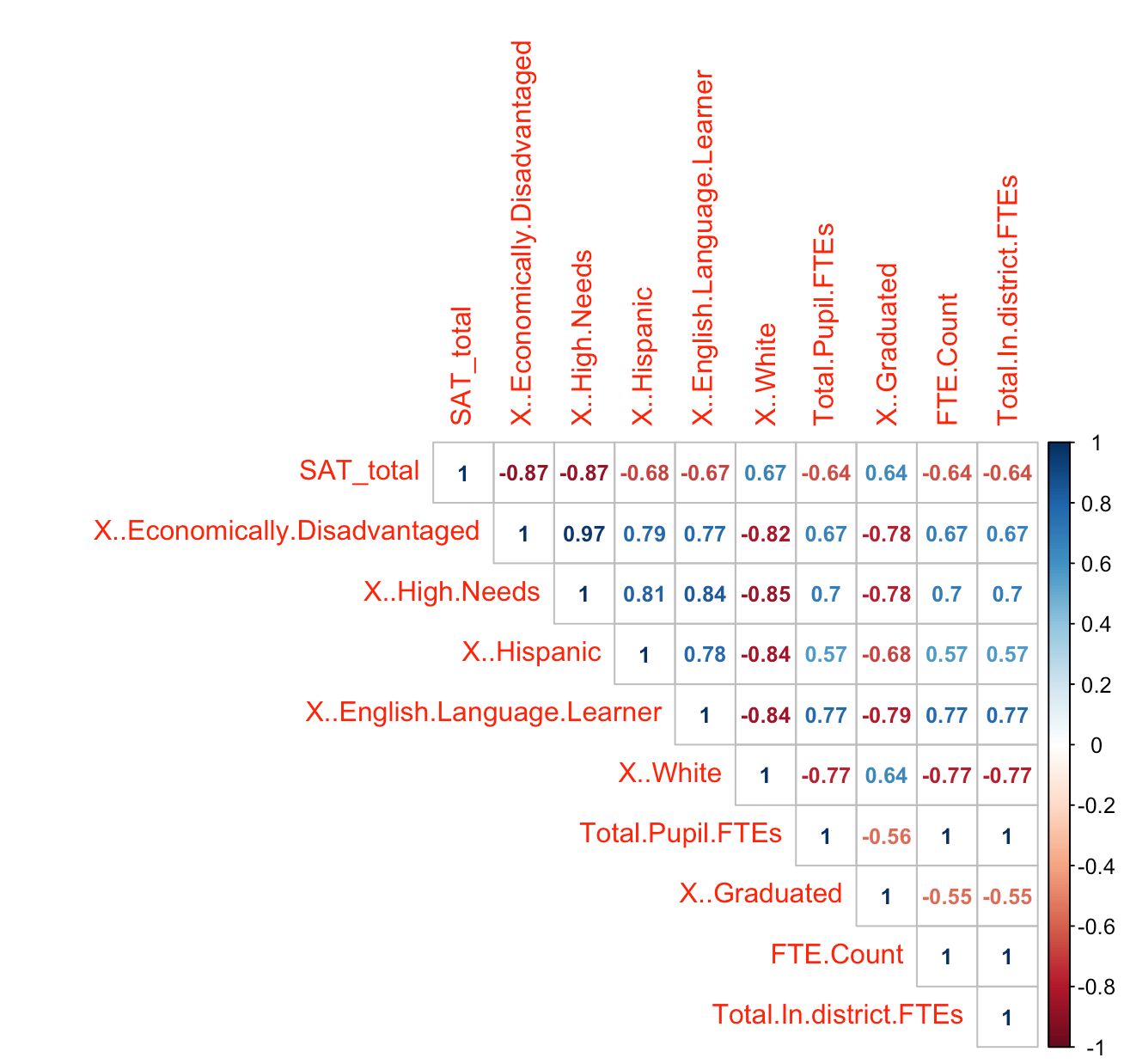
STAT 515: Applied Statistics and Visualization for Analytics

Dr. Tokunbo Fadahunsi

December 9, 2020

For my final project in this course, I have selected a dataset on K-12 schools in Massachusetts. I selected this dataset for a variety of reasons including size, the kinds of data that it contains, and its subject. The dataset has an approximate size of 1.4 MB which is comprised of data points in 302 columns and 1862 rows. It doesn’t quite fall into the category of big data, but it is certainly much too large and complex to be effectively analyzed without advanced techniques. I viewed this as a necessity in my selection process, as I want to be challenged to achieve things which I could not have before this course. The dataset also includes numerical, categorical, and geospatial data. This variety creates opportunity for a wide array of analytical techniques. Finally, the subject of K-12 education interests me greatly as my current occupation is teaching in a local public school.

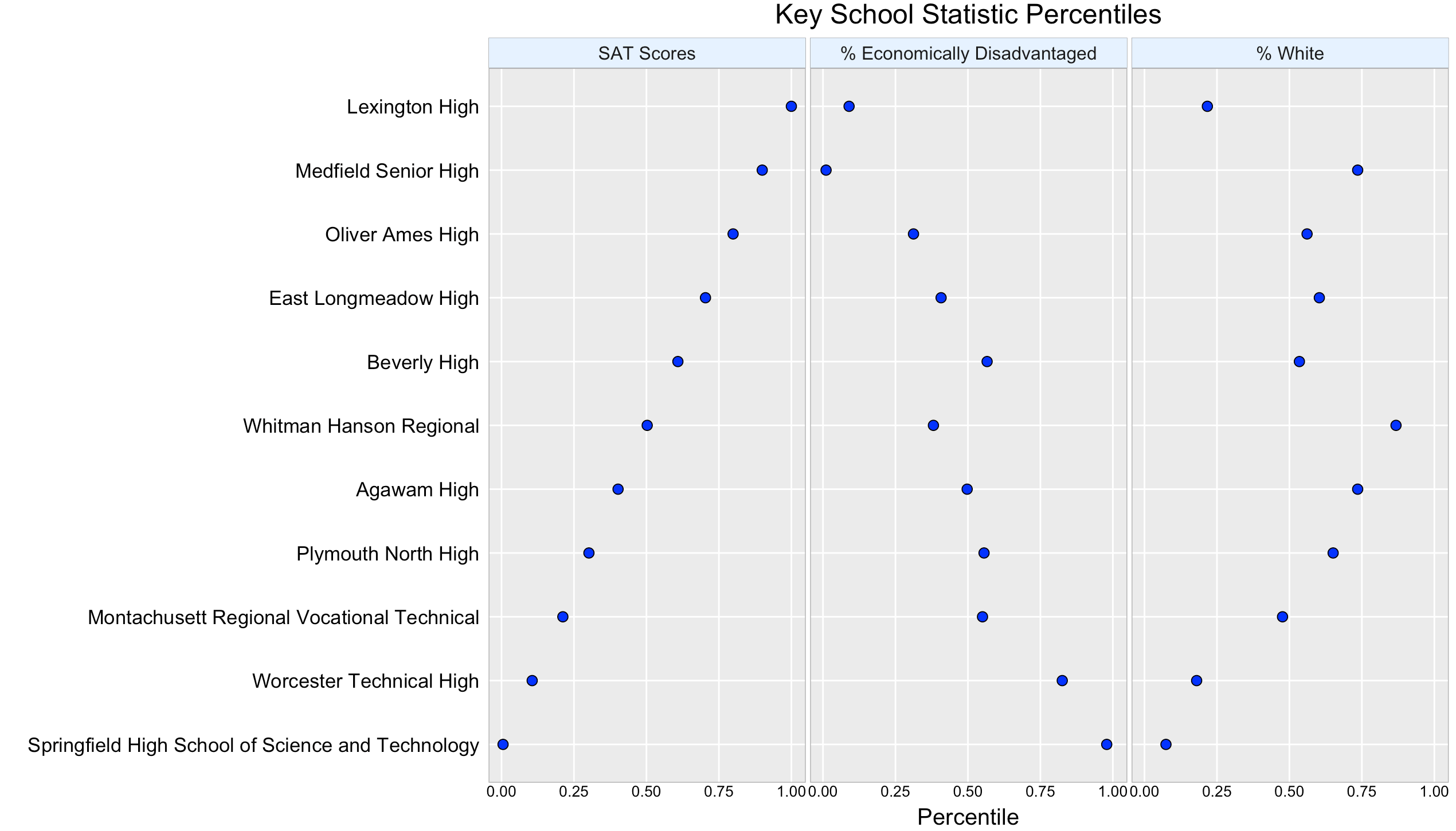
The data was obtained through the Massachusetts Department of Education website, where it is publicly available (“Statewide Reports,” 2017). The data is well-formatted, but still requires a substantial amount of processing to work with. Among the 302 columns, most rows have NA values for at least 50. This is largely due to the fact that some school metrics are totally inapplicable to specific types of schools. For example, elementary schools have NA values in SAT and AP scores. As a result, a significant amount of filtering was required to reduce unnecessary or problematic rows and columns for parts of the analysis. This process of filtering the data for analysis had to be repeated several times due to the need for several different subsets of data. Additionally, SAT scores were only provided in the 3 sub-categories of the test, so in order to calculate the total SAT score, an additional computed column had to be generated.

The two key variables addressed in this project are high school mean SAT scores, and school district mean teacher salaries. For mean SAT scores, the goal of this project was to model and predict SAT scores at each high school using a variety of methods, compare modeling results, and ultimately create the most effective model possible. The goal for teacher salaries was to visualize and describe the distribution across the state, and then create a choropleth map of teacher salaries within Massachusetts.

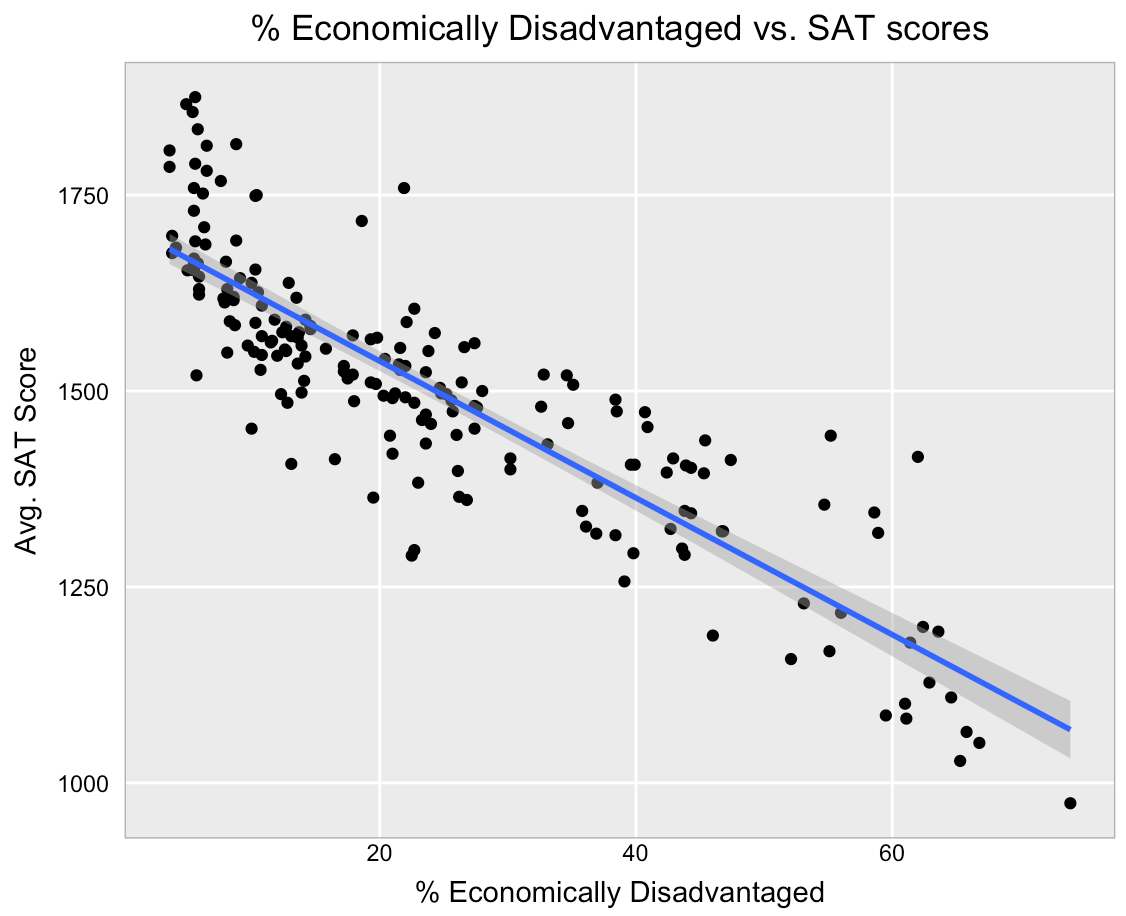
**Figure 1:** SAT scores corrplot – A numerical correlation matrix

To begin exploration and modeling of SAT scores, I first created a correlation matrix of the 10 strongest correlations to the target variable. Then I visualized this matrix by using the package “corrplot” (Fig. 1). The top row of this matrix displays correlations to SAT scores in order of absolute value. Using this plot, I quickly isolated two key predictors: percent of economically disadvantaged students (tied with percent high needs), and percent of white students. These predictors represent the highest negative and positive correlations, respectively.

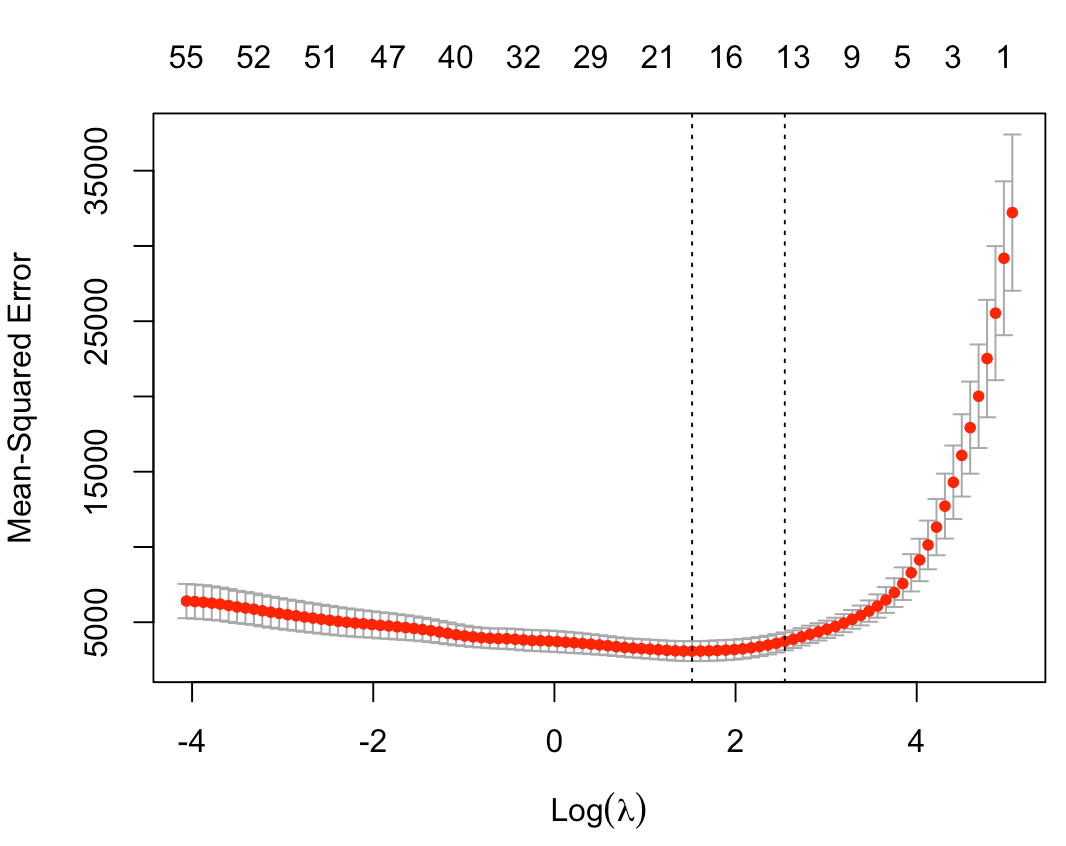
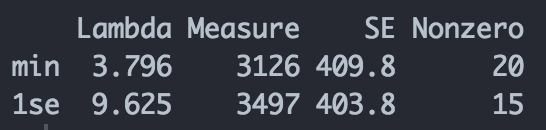
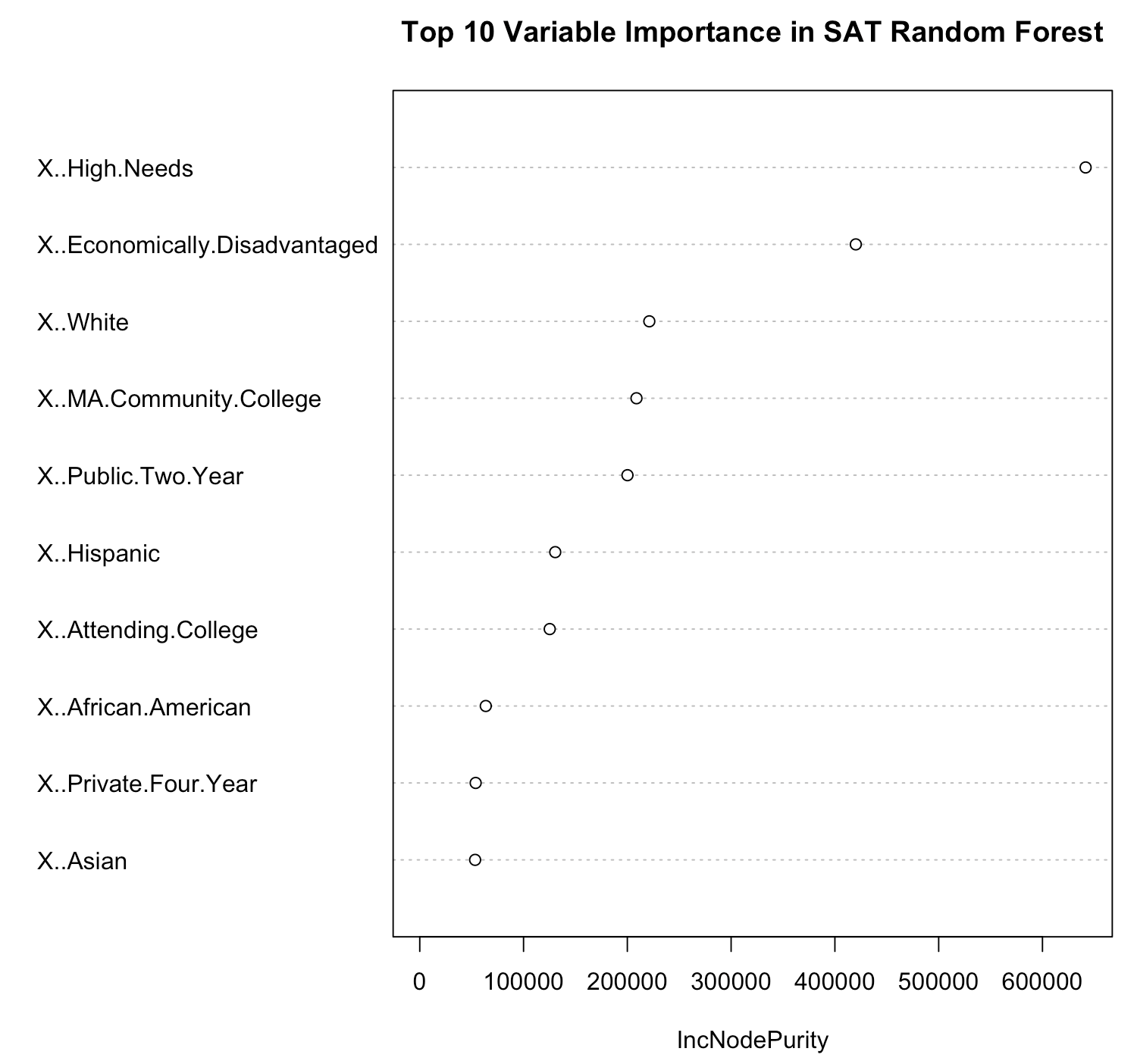
It is also important to note that the two leading correlations (% economically disadvantaged and % high needs) are extremely well correlated (.97). In the state of Massachusetts, students who are classified as having an “economic disadvantage” are also automatically classified as “high needs,” and make up by far the largest sub-group of “high needs” students in the state (“Student Group Data Definitions”). Due to their high correlation and data overlap, these variables can be thought of as similar predictors.

To better display these specific correlations, I then created a dot plot showing the school-specific values for these two variables plus SAT scores using ggplot2’s “facet\_grid” (Fig. 2). The schools for the dot plot were chosen by their proximity to decile breaks in SAT scores, therefore the best and worst SAT-scoring schools are plotted along with 9 evenly spaced decile ranks in between. In order to standardize the units of measure for all three variables, I also computed and plotted “percent economically disadvantaged” and “percent white” in term of their percentiles. Using this dot plot, the correlations become a lot more visually apparent, and seem very strong with the notable anomaly of Lexington High, which is in the 100th percentile for SAT scores, but only in the 22nd percentile for percent white.

**Figure 2:** Key statistics dot plot – all values are percentiles

 Taking the strongest correlation in the set (% Economically Disadvantaged), I then created a simple linear regression to begin modeling SAT scores. The result of that linear regression is visualized by a scatterplot with the regression line added (Fig. 3). Using simple linear regression with the best-case variable, this first model achieved a root mean square error (RMSE) of 84.68. This result serves as the baseline against which more sophisticated techniques can be measured.

**Figure 3:** Linear regression scatterplot – strong negative correlation

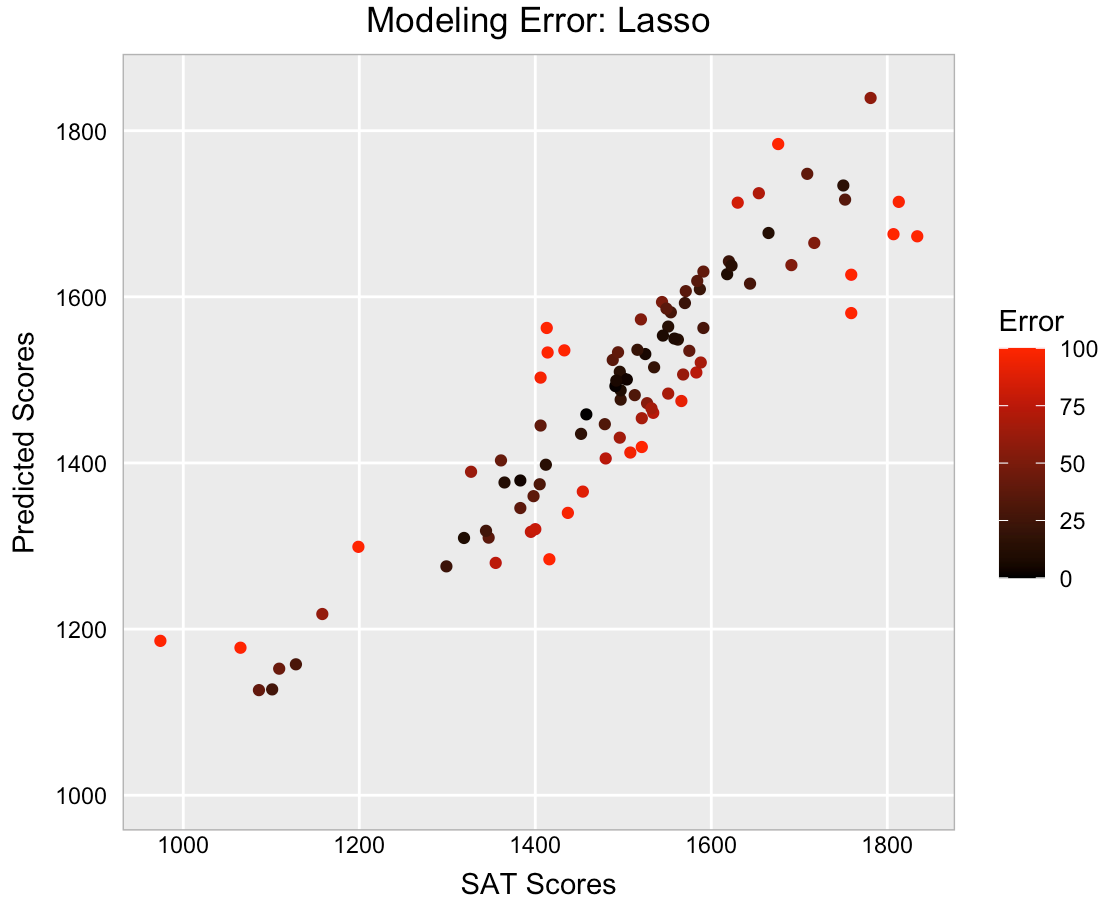
 To improve upon simple linear regression, I utilized a lasso regression capable of considering multiple explanatory variables. After splitting the data 50/50 into training and test data, I ran the cross-validated cv.glmnet() command to find the best lambda (λ) value for the final model (Fig. 4, Fig. 5). The cross-validation shows that using a lambda value of 3.796 provides the lowest MSE, and leaves 20 non-zero coefficient predictors. In order to compare this model with the simple linear regression, the final step was to predict SAT scores for the test set using the ideal lambda value. This yielded an RMSE of 68, which is a notable improvement over the simple linear model!

**Figure 5:** Summary of suggested λ values – Best λ value for MSE is 3.796

**Figure 4:** Cross-validated glmnet plot

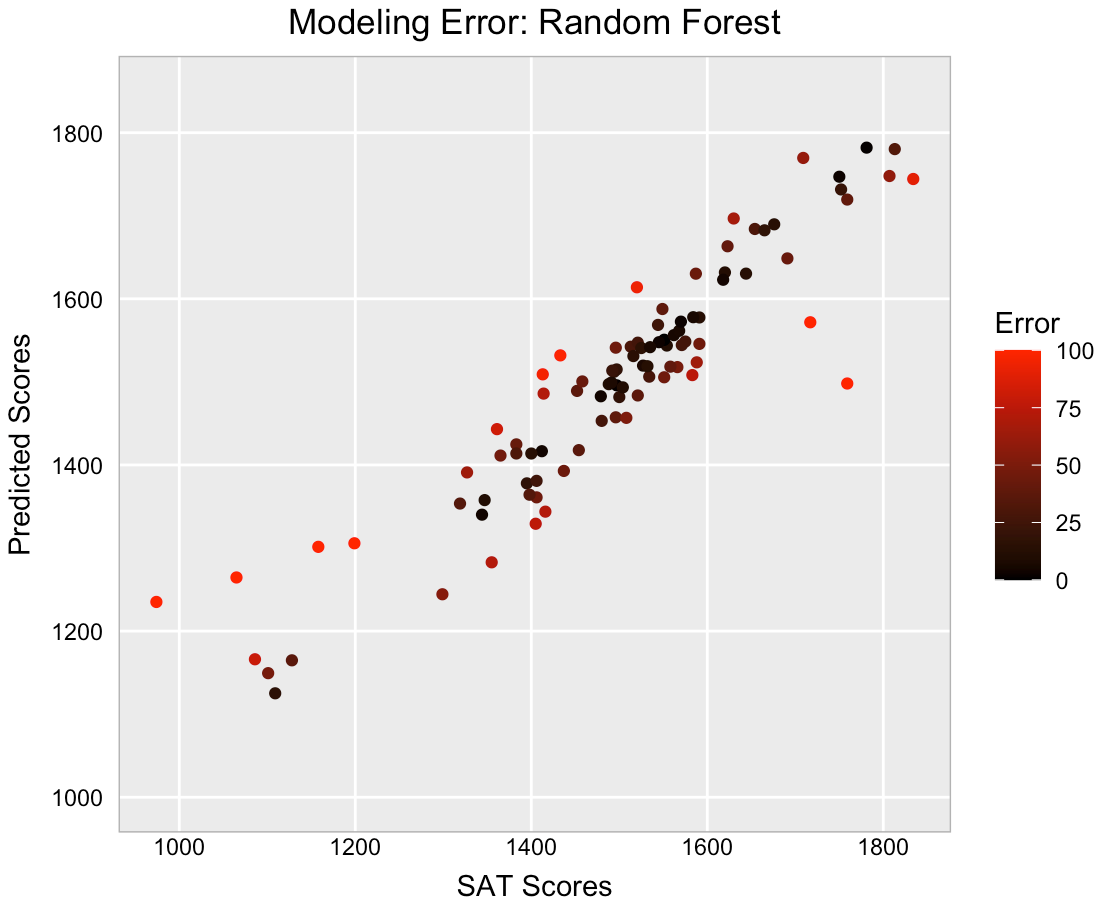
The final model that I attempted to construct was a random forest model. In order to do this, the same 50/50 test/training data split from the lasso regression was reused. Figure 6 shows the variable importance plot, which is available in the package “randomForest.” Here again, percent economically disadvantaged, percent high needs, and percent white are at the top of the list. This affirms the validity of the initial correlation matrix. After creating the model with training data, I used it to predict SAT scores for the test data. The RMSE of the randomForest model was 63.91, which represents a modest improvement over the lasso regression.

**Figure 6:** Random Forest variable importance plot – key predictors appear at top

 To conclude the modeling and visualization of SAT scores, I created plots comparing the predictions of my 3 models against the actual SAT scores from the test set (Fig. 7-9).

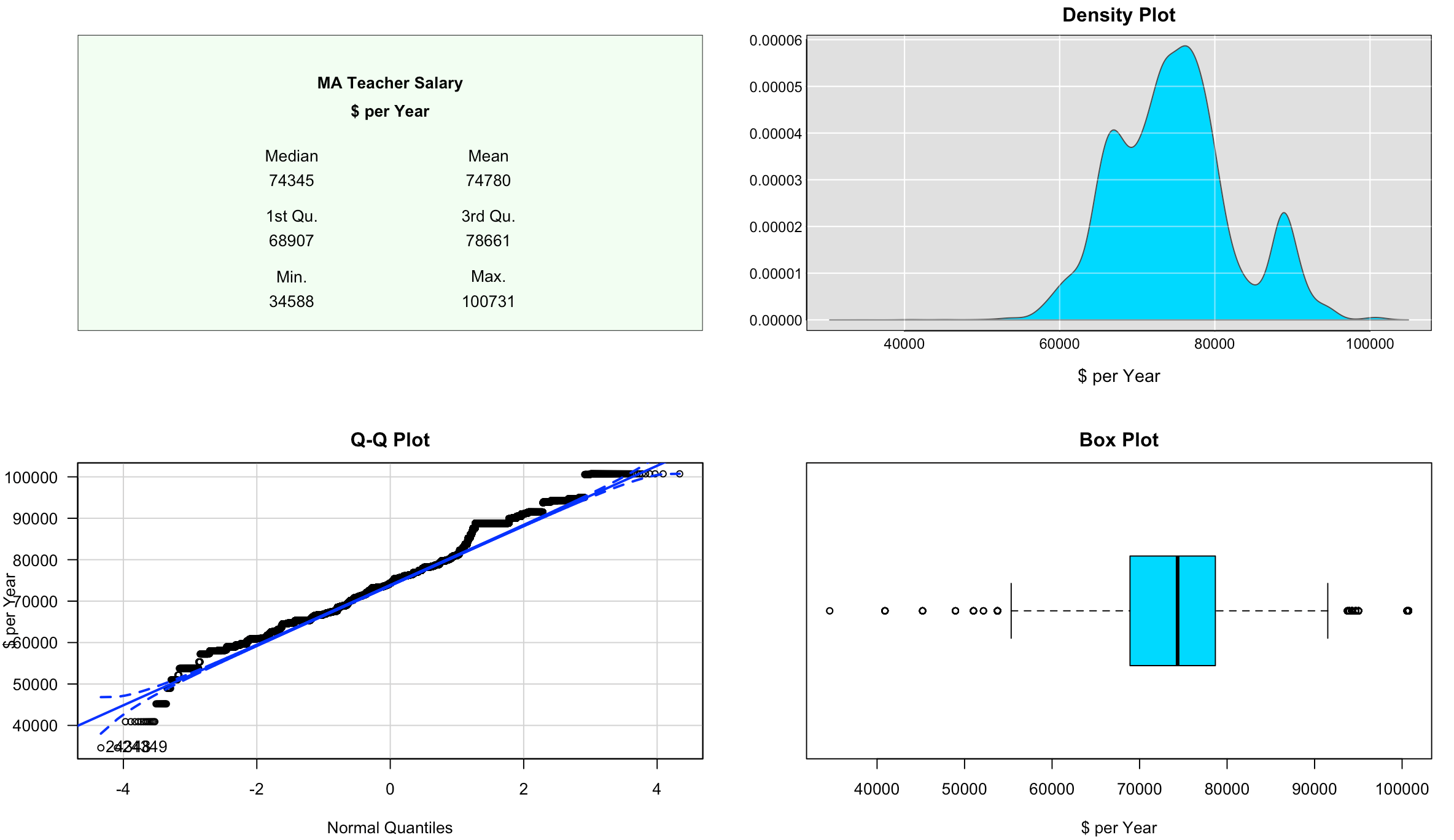
**Figure 7:** Linear modeling error of SAT scores

**Figure 8:** Lasso modeling error of SAT scores



**Figure 9:** Random Forest modeling error of SAT scores

The modeling error plots show points with the greatest prediction error in red, which gets increasingly bright as the size of the error increases. Scatter plots like these are a more user-friendly and accessible alternative to modeling diagnostics like the aforementioned RMSE. With these plots, the consumer/viewer can quickly assess the accuracy of each model by the amount of red they see, as well as scan for systematic biases in a model. The plot for the linear model (Fig. 7) is clearly the least accurate of the group, and randomForest (Fig. 9) also stands out as having the least red data points in the set. These are the final plots which show randomForest’s superiority with modeling SAT scores in this dataset.

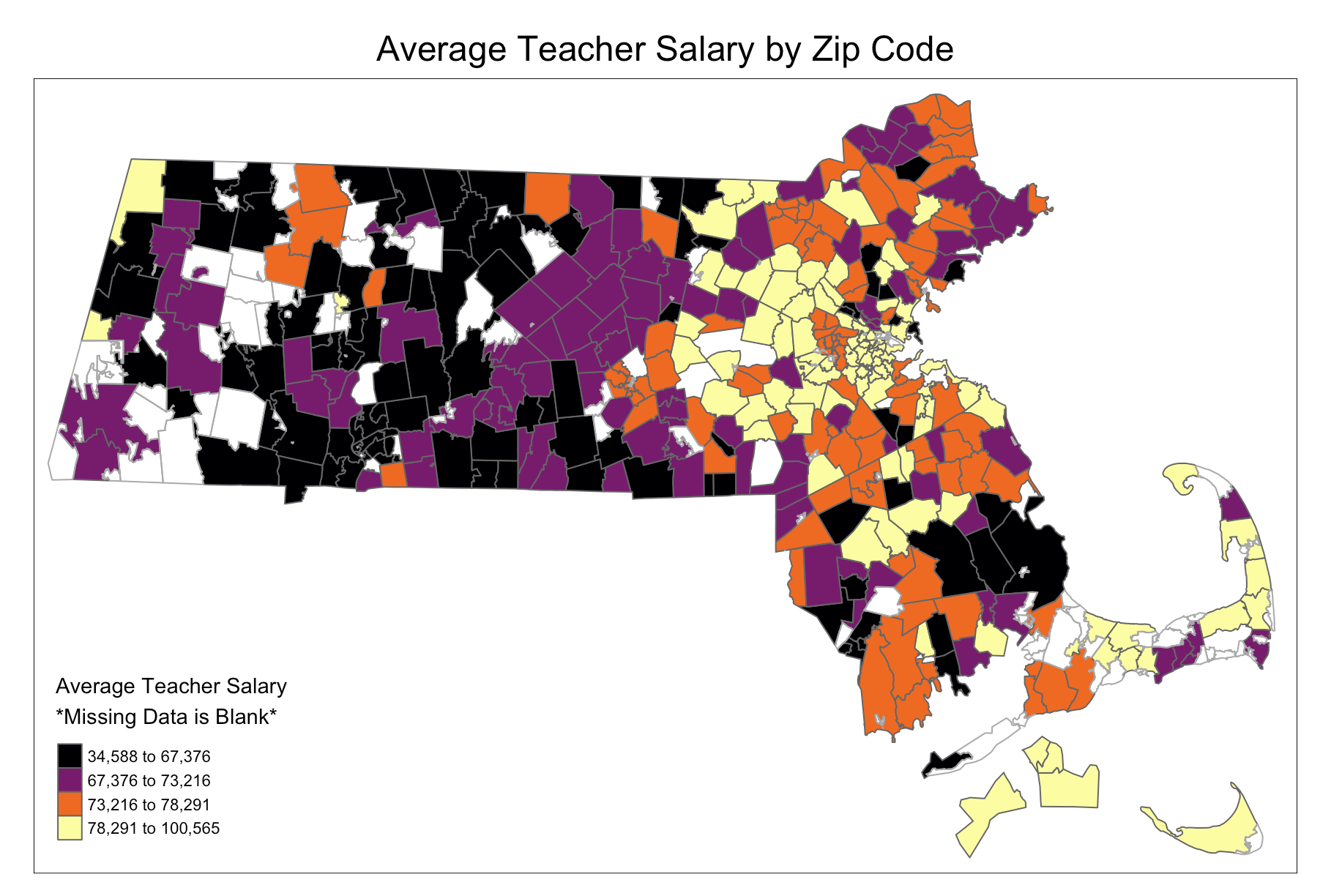
 The second key variable I selected in this dataset was teacher’s salary. For this variable, the goal was to visualize the distribution of salaries across the state. The first step in this process was to view the distribution using common statistical plotting techniques. I chose to do this by utilizing a modified version of the class EDA plot for distributions (Fig. 10). It is important to know that this data is a distribution of school district means, not individual salaries. There are likely individual salaries outside of the range of this data. The school districts also vary greatly in size, so to approximate the actual teacher salary distribution, each district mean salary was multiplied by the number of teachers in the district. This creates an effective weighted average that allows us to analyze teacher salaries in their true proportions.

**Figure 10:** Exploration Data Analysis plot of mean teacher salary by school district – districts are weighted by number of employees

In the upper-left of the EDA plot, a standard 5-number summary is displayed alongside the mean value for the distribution. It is immediately clear from this summary that there is great variability among teacher salaries per district in Massachusetts, with a range from 34,588 to 100,731. Shifting attention to the density plot in the upper-right, some finer details can be seen. The distribution appears approximately bell-shaped. The most distinctive feature of the density plot is the “bump” at ~$90,000. This is a result of highly populous and well-paying districts in the Boston area.

Looking at the bottom-left of the EDA plot, a Q-Q plot is displayed. This plot has slightly curved ends which deviate significantly from the diagonal blue line. This phenomenon is commonly referred to as “heavy tails,” and it indicates that the extreme values in the distribution (tails) are more frequent than would be expected from a normal distribution. Finally, looking to the bottom-right of the EDA plot shows a box plot. This is the only plot in the set which is capable of displaying individual values, and it gives insight on the “heavy tails” that are present in the Q-Q plot. The box plot clearly shows several outliers both on the high and low side of the dataset, which explain the “heavy tails” and the extreme range seen in the 5-number summary. Taken together, the graphs shown in this EDA plot offer a detailed, multifaceted view of the distribution.

With a much better understanding of the shape and features of the distribution, the next step is to map the salary data using geospatial information. In order to do this, I used the package “tigris” to download shape files for Massachusetts. I then used the package “tmap” to create the underlying choropleth map of Massachusetts. The most precise geospatial data in this dataset is zip codes, so it is what I used to map the distribution of salaries across the state. This affords a much more detailed view than using counties.

The map created by “tmap” is filled with the average teacher salary of the school district centered in that zip code (Fig. 11). In order to counteract the aforementioned “heavy tails,” this map uses color bins based on quartiles. Making color bins of equal width would result in the vast majority of the zip codes being colored one or two colors. This method provides a more even distribution of color for quick visual analysis.

A few things are immediately apparent looking at this map. First, some zip codes do not have a school district centered within them. Those codes are left blank in this map. Second, there is a large zone of dense, high-paying school districts in the central-eastern portion of the state. This is Boston, where that concentration of wealthy districts also caused the “bump” in the density plot of the EDA (Fig. 10). More generally, nearly all of the first quartile districts are in the West, and nearly all of the fourth quartile districts are in the east. This is just one of the many significant results of the analysis of this dataset.

There were significant results found in both portions of the analysis. In the first section centered around SAT scores, there were two key results. The first was that all models pointed to the same predictors as the most important. Namely, these are: Percent economically disadvantaged students, percent high needs students, and the percent of white students at a high school. Sadly, two out of three have a negative correlation. This indicates that the most important predictors of SAT scores are negative situations that a student can find themselves in, rather than actions they can personally take to increase their scores. The second key result of the SAT analysis was that the randomForest model was the most accurate in predicting scores from the same set of predictors. Based on this analysis, a randomForest model would be the ideal tool for school systems to use to predict their SAT scores.

In the second section of the analysis, this project explored the distribution of teacher salaries in Massachusetts. Again, there were two key results from this analysis. First, the distribution of salaries is very broad, and has a significant number of high and low outliers. Teachers in Massachusetts could probably significantly change their salary by working for a different district within the state. Second, the Boston region of Massachusetts accounts for the vast majority of the high-paying districts. In the future, it could be very beneficial to compare average district salary to cost of living data, or the local consumer price index. Boston is a very dense and expensive area. It is quite possible that the local cost of living in Boston is so high that the increased salary is necessary to sustain the same quality of living. Considering this additional data could allow a future analysis to answer this question.

**References**

A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

Andri Signorell et mult. al. (2020). DescTools: Tools for descriptive statistics. R package version 0.99.38.

Dean Attali (2020). shinyjs: Easily Improve the User Experience of Your Shiny Apps in Seconds. R package version 2.0.0. https://CRAN.R-project.org/package=shinyjs

Fadahunsi, T (2020) hw.R Plotting Theme Code [Source code]. https://mymasonportal.gmu.edu/webapps/blackboard/execute/displayLearningUnit?course\_id=\_406427\_1&content\_id=\_11803944\_1

Fadahunsi, T (2020) classeda.R Plotting Code [Source code]. https://mymasonportal.gmu.edu/webapps/blackboard/execute/displayLearningUnit?course\_id=\_406427\_1&content\_id=\_11803944\_1

Fadahunsi, T (2020) classdensity.R Plotting Code [Source code]. https://mymasonportal.gmu.edu/webapps/blackboard/execute/displayLearningUnit?course\_id=\_406427\_1&content\_id=\_11803944\_1

H. Wickham. ggplot2: Elegant Graphics for Data Analysis.

Springer-Verlag New York, 2016.

Hadley Wickham and Lionel Henry (2020). tidyr: Tidy Messy Data. R package version 1.1.0. https://CRAN.R-project.org/package=tidyr

Jerome Friedman, Trevor Hastie, Robert Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22. URL

http://www.jstatsoft.org/v33/i01/.

Kyle Walker (2020). tigris: Load Census TIGER/Line Shapefiles. R package version 1.0. https://CRAN.R-project.org/package=tigris

R Core Team (2020). R: A language and environment for statistical

computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

School Listings / Boston International High School. (n.d.). Retrieved December 06, 2020, from https://www.bostonpublicschools.org/school/boston-international-high-school

Statewide Reports. (2017). Retrieved December 9, 2020, from https://profiles.doe.mass.edu/state\_report/

Student Group Data Definitions. (n.d.). Retrieved December 07, 2020, from https://www.renniecenter.org/subgroupdefinitions

Taiyun Wei and Viliam Simko (2017). R package "corrplot": Visualization of a Correlation Matrix (Version 0.84). Available from https://github.com/taiyun/corrplot

Tennekes M (2018). “tmap: Thematic Maps in R.” \_Journal of Statistical Software\_, \*84\*(6), 1-39. doi: 10.18637/jss.v084.i06 (URL:

https://doi.org/10.18637/jss.v084.i06).