Final Project

CUNY MSDS DATA 609

Cesar L. Espitia
Duubar Villalobos Jimenez

December 09, 2018

Contents

TITLE

SATs and student demographic and enrollment data from the NYC Department of Education 2011 – 2016.

ABSTRACT

Education is a very important topic in any society. One of the unique aspects that makes the US education system so dynamic are the levels of diversity currently present in our public school systems. Over the past decade, research at various levels in public and private sectors have determined that diversity is important to success and it all starts with a person's education.

This final project focused on NY City Public School demographic data and SAT Test scores in order to determine the effects of a school's student demographics on SAT scores. The data-set contains 8867 records that encompass the entire school system. The variables for the data pertain to school demographic information such as number of enrollments, school names, breakout by grade and percentages, but the SAT data is only available for one of the 5 years. The purpose for this project is to analyze the data, perform any data manipulation / clean-up and use two (2) methods learned in the class which in this case was a linear model to predict SAT Scores (enhanced by employing a generalized model in R) and then difference equations to predict the demographics for the year 2016-17 and then predict the SAT scores for that given year using the improved demographic data. The final model provided an AIC = 3600.7.

The data was obtained from the NYC Open Data portal (data.cityofnewyork.us/).

- `2011 - 2016 Demographic Snapshot`
- `2012 SAT Results`

Data Source:

https://data.cityofnewyork.us/Education/2011-2016-Demographic-Snapshot/8mzw-jfss

https://data.cityofnewyork.us/Education/2012-SAT-Results/f9bf-2cp4

The following is the analysis and write-up based upon our interpretation of the data in order to predict the average SAT scores based upon demographic school data.

KEYWORDS

NY City schools, NYC SAT, NYC student demographics.

DATA EXPLORATION

The purpose of this step is to get a 'feel' for the data-set. The following information describes the data from different angles including completeness, statistical summaries, visuals to determine the shape and effect of each variable and other items deemed pertinent.

Summary Statistics

The first step is to look at the data to determine some items including completeness and the shape of each variable. The following are the results of summarizing the data in a table and the visualization of each variables density function (PDF).

Table 1: Summary Statistics for NY School Demographic Data.

	Min	1st Qu	Median	Mean	3rd Qu	Max
Grade.9	0	0	0	52.43	69	1457
Grade.10	0	0	0	50.93	67	3692
Grade.11	0	0	0	39.15	44	1393
Grade.12	0	0	0	38	39	1380
X..Female	0	143	238	291.4	357	2356
X..Female.1	0	46.3	48.8	48.46	51.4	100

Note: Data taken from NY Data portal.
Name Differences: [1] X..VARIABLE represents counts. [2] X..VARIABLE.1 represent percentages.

In looking at the above Table 1, Figure 1 and Appendix B (correlation matrix) together, we can note specific items that may skew our model building results. In this model, there was 10% of the data that was NA/null.

PDF: Figure 1 shows the PDF of some variables, this allows us to see if the data is normally distributed or not; this means that we might have to remove the effects of severe skewness. All other variables were left as is because the shape didn't warrant it.

Correlation: We looked for correlated variables that we can make decisions on and determine which variable might be closely related to others either due to col-linearity or other underlying factors that are visible at first glance in the data-set. The following variables were removed "Female", "Male", "Asian", "Black", "Hispanic", "Other", "White", "Students.with.Disabilities", "English.Language.Learners", "Poverty", "Grade4", "Grade5", "GradePK", "GradeK", "Grade1", "Grade2", "Grade3", "Grade6", "Grade7", "Grade8". The demographic ones were counts that also had percentages. For the grades they did not add value to the model as they don't affect the value of SAT scores.

DATA PREPARATION

The purpose of this step is to take the findings from the exploration and transform the data as needed. The following information describes the transformations done in order to prepare the data for model building and model selection.

For this analysis, 10% of the data had NAs and were imputed using the mean of the data-set. No variables had any transformations due to any sever skewness in the PDF graphs above. No new variables were created as there was nothing that was missing in the data-set. With this in mind, no secondary correlation check was done.

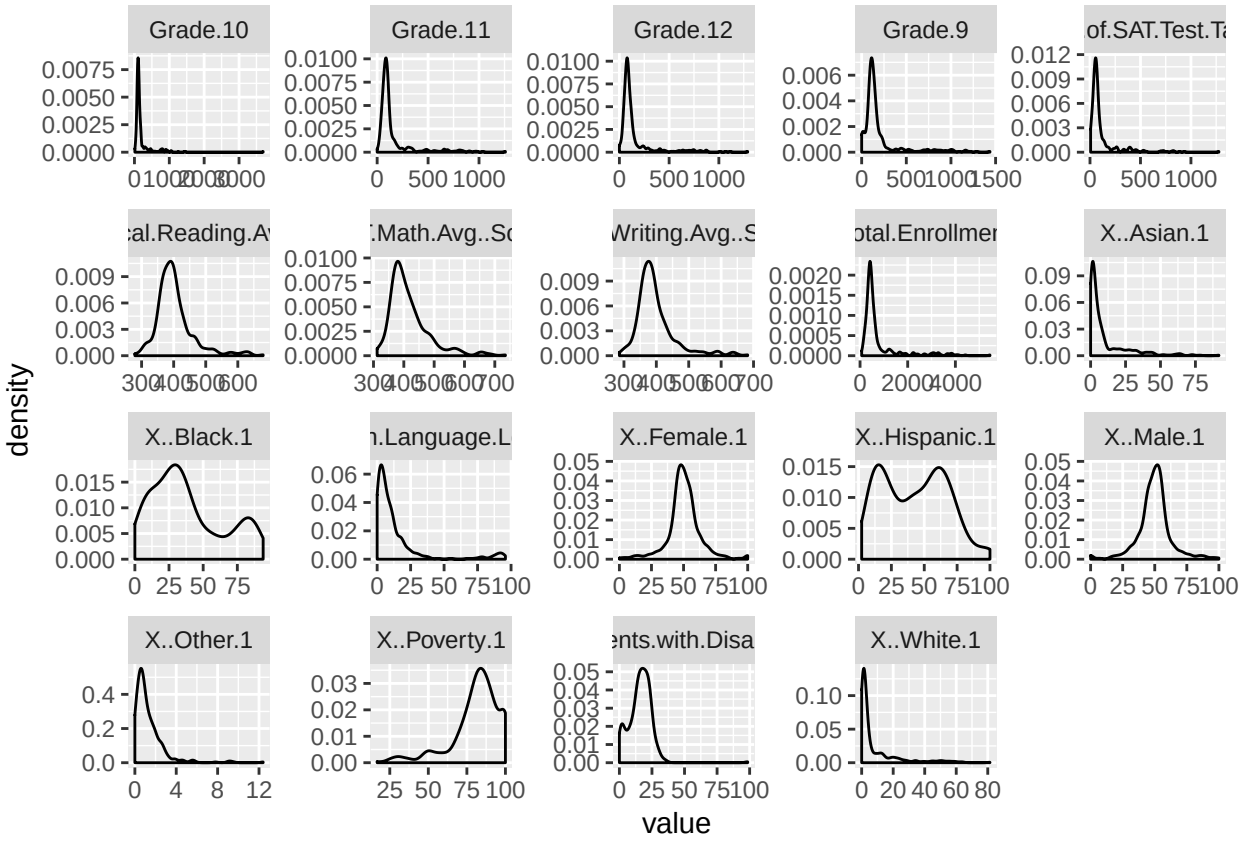

Figure 1: PDF for each variable.

MODEL BUILDING

The purpose of this step is to take the modified data-set and begin exploring potential models that will be used on the final data-set provided. The following information describes the two (2) models built for this step and the relevant analysis to provide reasons for model selection in the next step.

MODEL 1: Linear Regession.

The first model takes in the data as manipulated in step two. In this first model, we have an AIC of 3169.7.

```
##
## Call:
## glm(formula = SATTotal ~ . - SAT.Critical.Reading.Avg..Score -
##     SAT.Writing.Avg..Score - SAT.Math.Avg..Score - Num.of.SAT.Test.Takers,
##     family = gaussian(), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -223.11   -55.93    -2.72    45.22   327.72
##
## Coefficients: (1 not defined because of singularities)
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -3.720e+03  8.416e+03  -0.442 0.658871
## Total.Enrollment                -8.556e-03  4.627e-02  -0.185 0.853458
## Grade.9                         -5.325e-02  8.810e-02  -0.604 0.546093
## Grade.10                        -1.398e-01  1.019e-01  -1.372 0.171211
## Grade.11                         2.918e-01  1.375e-01   2.123 0.034747 *
## Grade.12                        -8.910e-03  1.162e-01  -0.077 0.938943
## X..Female.1                      8.805e-03  4.252e-01   0.021 0.983494
## X..Male.1                               NA         NA      NA       NA
## X..Asian.1                       5.804e+01  8.412e+01   0.690 0.490858
## X..Black.1                       5.204e+01  8.414e+01   0.618 0.536816
## X..Hispanic.1                    5.344e+01  8.415e+01   0.635 0.525935
## X..Other.1                       5.432e+01  8.374e+01   0.649 0.517160
## X..White.1                       5.483e+01  8.413e+01   0.652 0.515172
## X..Students.with.Disabilities.1 -2.472e+00  7.109e-01  -3.477 0.000595 ***
## X..English.Language.Learners.1  -3.894e+00  3.467e-01 -11.230  < 2e-16 ***
## X..Poverty.1                    -3.928e+00  6.475e-01  -6.066 4.72e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 6908.505)
##
##     Null deviance: 9488471  on 269  degrees of freedom
## Residual deviance: 1761669  on 255  degrees of freedom
## AIC: 3169.7
##
## Number of Fisher Scoring iterations: 2
```

No variables seem in terms of predictability and therefore no values will be removed for the second method.

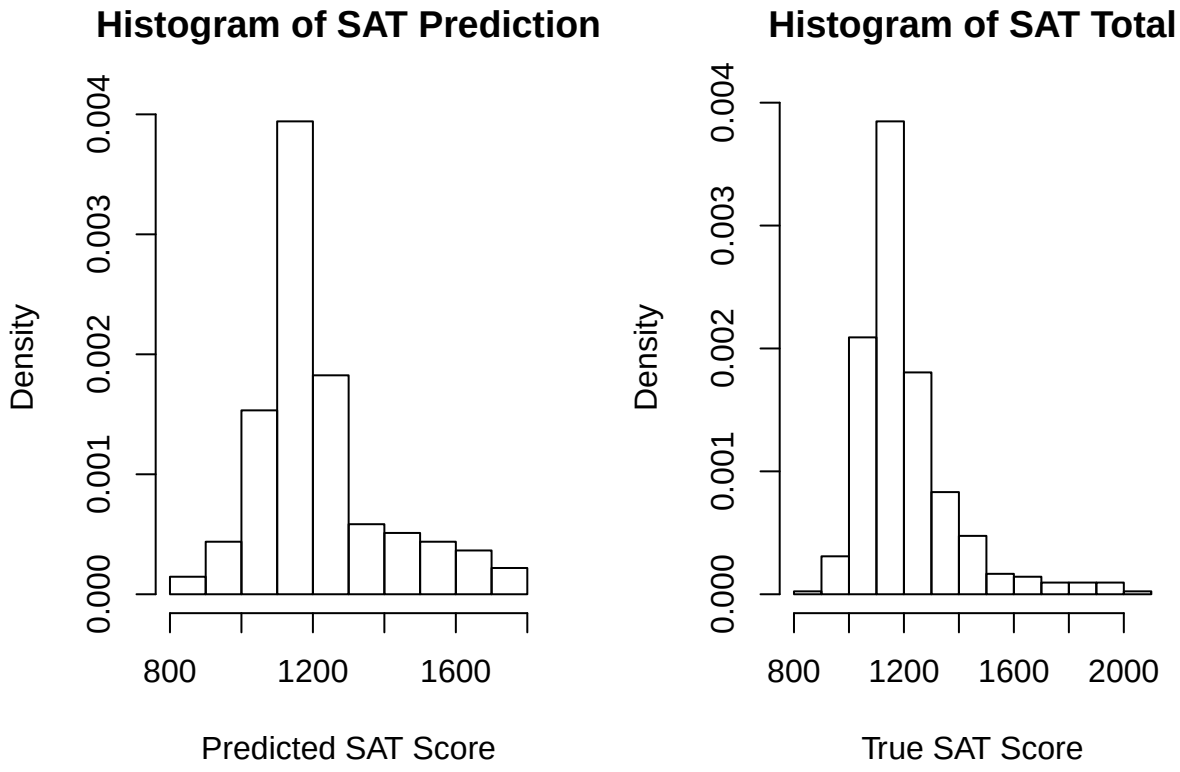

Figure 2: Histogram of Model 1 Prediction of SAT Total Scores

MODEL 2: Difference equations and restrictions.

This is the second model of two.

Due to the limited values given in our data set, we are going to create a model able to predict a possible SAT score for the given year 2016-17. For this, we will employ the SAT scores provided for the year 2012-13 in order to predict the 2016-7 values. Please note that in order to do so, we will employ difference equations approximations with a few restrictions.

- The number of students can not be negative.
- The percentages can not be below zero.
- The percentages can not be over 100.

Table 2: Years reported and number of schools in that group.

nYears	count
1	41
2	41
3	76
4	58
5	1647

Note: Data taken from NY Data portal.
Meaning: [1] nYears: Number of years. [2] Count: Number of schools

In figure 3, we can observe the predictive SAT scores for 2016-17 school year. The scores were obtained by employing a linear model as seen in the Appendix. It is noted that a second linear model was chosen over the above for the second model analysis. The linear model was employed after automated values for 2016-17 were generated by employing difference equations and restrictions as noted above.

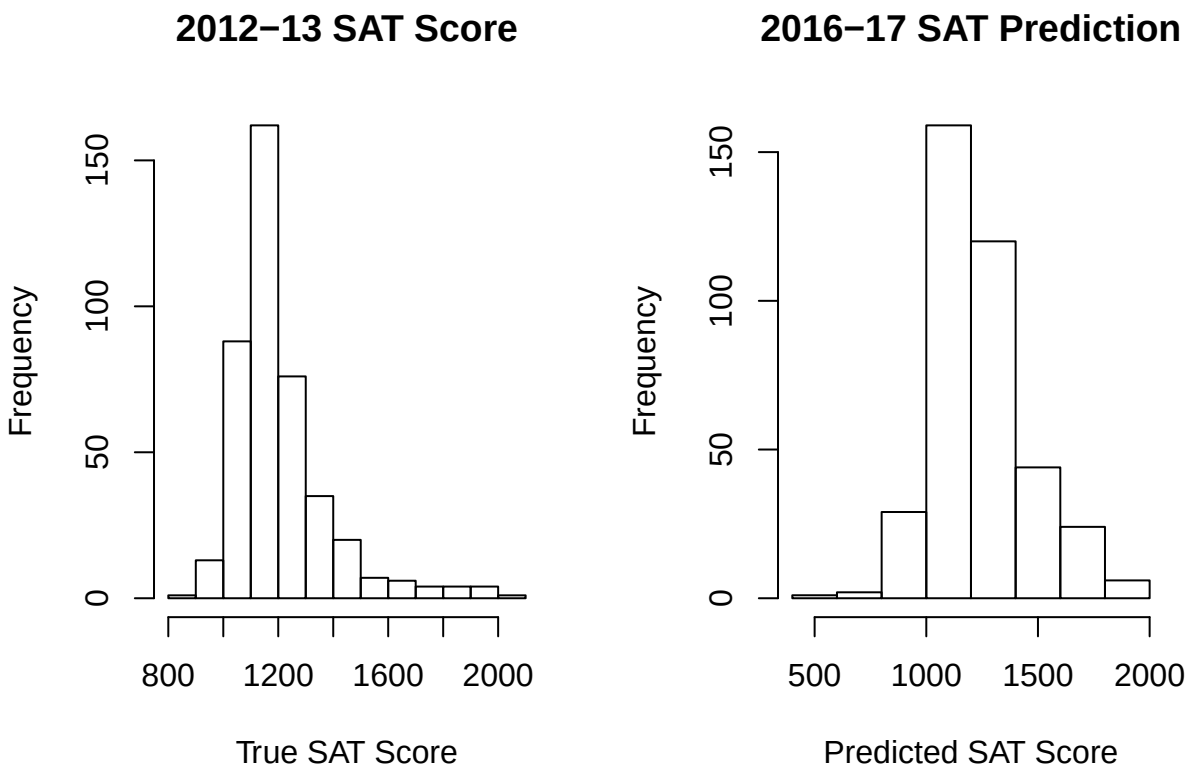

Figure 3: Histogram of Model 2 SAT Total Predictions for 2016-17 vs 2012-13 SAT Total Scores

Also, by looking at the Appendix related to SAT scores, we can observe some sort of increase in the scores, that is the tails are moving from right to left; or in other words, the scores seemed to get higher as the years passed,in this case we wer just working with population demographics percentages provided by the NYC Department of Education.

CONCLUSIONS

These two (2) models were presented after exploring and manipulating the data as necessary. With using a multi-criteria approach for this exercise, it became clear that the Model 2 was selected and provided an AIC of 3600.7 which was adequate for the data but doesn't necessarily indicate the best model if it were solely based upon AIC (Model 1 would have been chosen) which is the equivalent of R-squared for binary regression models. If more time were available, the creation of other new variables that were not correlated could have been generated with better insight into the data set.

This is a great exercise in order to test some hypothesis in which demographics is said to play an important part in SAT scores.

REFERENCES

A First Course in Mathematical Modeling, 5th Edition Frank R. Giordano, William P. Fox, Steven B. Horton

APPENDIX

Appendix A

Table 3: Full Summary Statistics for NY School Demographic Data.

	Min	1st Qu	Median	Mean	3rd Qu	Max
Total.Enrollment	1	319	483	599.6	710	5534
Grade.PK	0	0	0	14.03	30	967
Grade.K	0	0	33	47.73	85	406
Grade.1	0	0	33	49.31	88	383
Grade.2	0	0	27	47.75	86	349
Grade.3	0	0	3.5	46.1	83	373
Grade.4	0	0	0	44.19	80	380
Grade.5	0	0	0	43.34	79	398
Grade.6	0	0	0	42.75	62	812
Grade.7	0	0	0	42.06	58	819
Grade.8	0	0	0	41.79	56	839
Grade.9	0	0	0	52.43	69	1457
Grade.10	0	0	0	50.93	67	3692
Grade.11	0	0	0	39.15	44	1393
Grade.12	0	0	0	38	39	1380
X..Female	0	143	238	291.4	357	2356
X..Female.1	0	46.3	48.8	48.46	51.4	100
X..Male	0	159	249	308.2	370	3250
X..Male.1	0	48.6	51.2	51.54	53.7	100
X..Asian	0	4	14	92.78	70	3340
X..Asian.1	0	1.1	3.5	10.72	12	94.7
X..Black	0	47	115	169.5	218	1552
X..Black.1	0	9.8	27.8	34.48	55.8	98.8
X..Hispanic	1	79	169	241.1	311.8	2478
X..Hispanic.1	0.4	18.9	38.2	41.45	62.4	100
X..Other	0	2	5	9.514	11	310
X..Other.1	0	0.5	1	1.661	2	38.3
X..White	0	4	12	86.69	72	3230
X..White.1	0	1	2.6	11.69	13.5	93.6
X..Students.with.Disabilities	0	56	89	110.5	135	842
X..Students.with.Disabilities.1	0	13.7	18	20.79	23.1	100
X..English.Language.Learners	0	16	40	82.27	99	1233
X..English.Language.Learners.1	0	4.1	9	13.36	17.7	100
X..Poverty	1	242	383	479.1	584	3842
X..Poverty.1	3.3	75.3	87.9	81.93	97.8	100

Note: Data taken from NY Data portal.
Name Differences: [1] X..VARIABLE represents counts. [2] X..VARIABLE.1 represent percentages.

Appendix B

Linear model employed to find predicted SAT scores for 2016-17.

```
##
## Call:
## glm(formula = SATTotal ~ X..Poverty.1 + X..Asian.1 + X..English.Language.Learners.1 +
##     X..Black.1 + X..Students.with.Disabilities.1 + X..Hispanic.1 +
##     Grade.11 + Grade.10 + Grade.9, family = "poisson", data = lm.data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.8161  -1.5317  -0.0327   1.3025   9.2909
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       7.497e+00  1.331e-02 563.382  < 2e-16 ***
## X..Poverty.1                     -2.772e-03  2.143e-04 -12.937  < 2e-16 ***
## X..Asian.1                        2.403e-03  2.660e-04   9.034  < 2e-16 ***
## X..English.Language.Learners.1   -3.231e-03  1.206e-04 -26.796  < 2e-16 ***
## X..Black.1                       -2.082e-03  2.067e-04 -10.070  < 2e-16 ***
## X..Students.with.Disabilities.1 -2.029e-03  2.356e-04  -8.611  < 2e-16 ***
## X..Hispanic.1                    -1.044e-03  2.303e-04  -4.531 5.88e-06 ***
## Grade.11                          1.762e-04  3.029e-05   5.818 5.97e-09 ***
## Grade.10                         -9.621e-05  3.122e-05  -3.082  0.00206 **
## Grade.9                          -3.963e-05  2.508e-05  -1.580  0.11401
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 7153.7  on 269  degrees of freedom
## Residual deviance: 1169.1  on 260  degrees of freedom
## AIC: 3600.7
##
## Number of Fisher Scoring iterations: 3
```

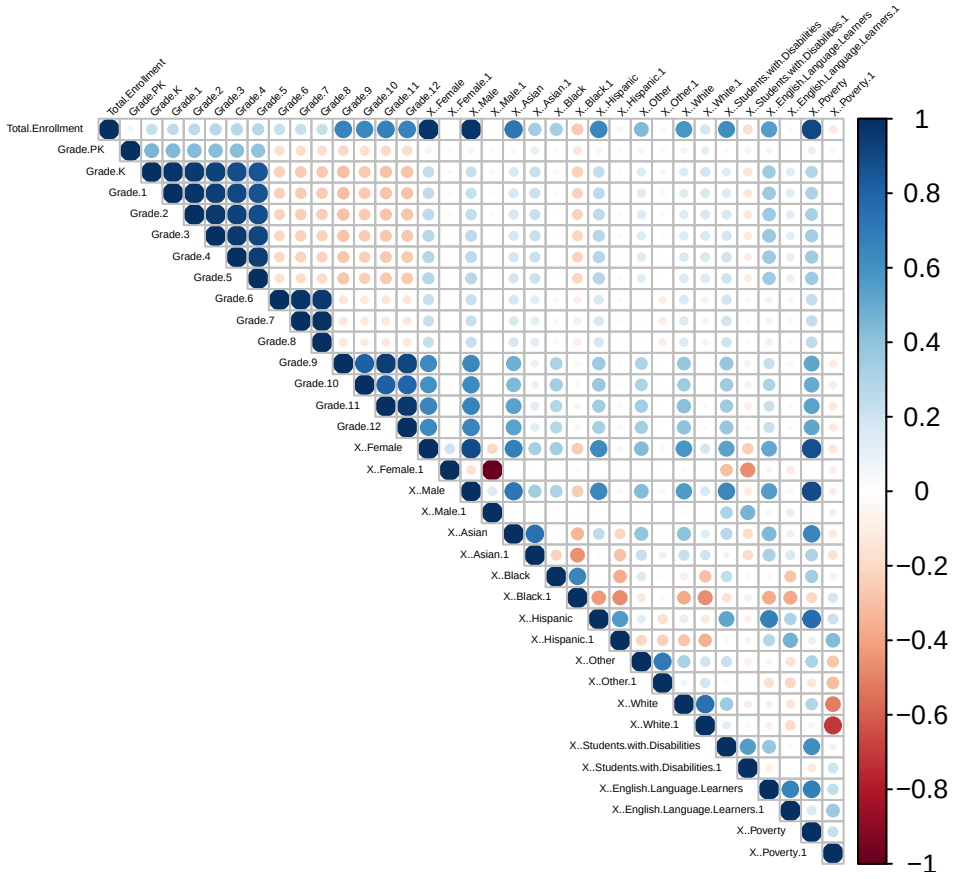

Figure 4: Demographic correlation data.

Appendix C

Demographic correlation data.

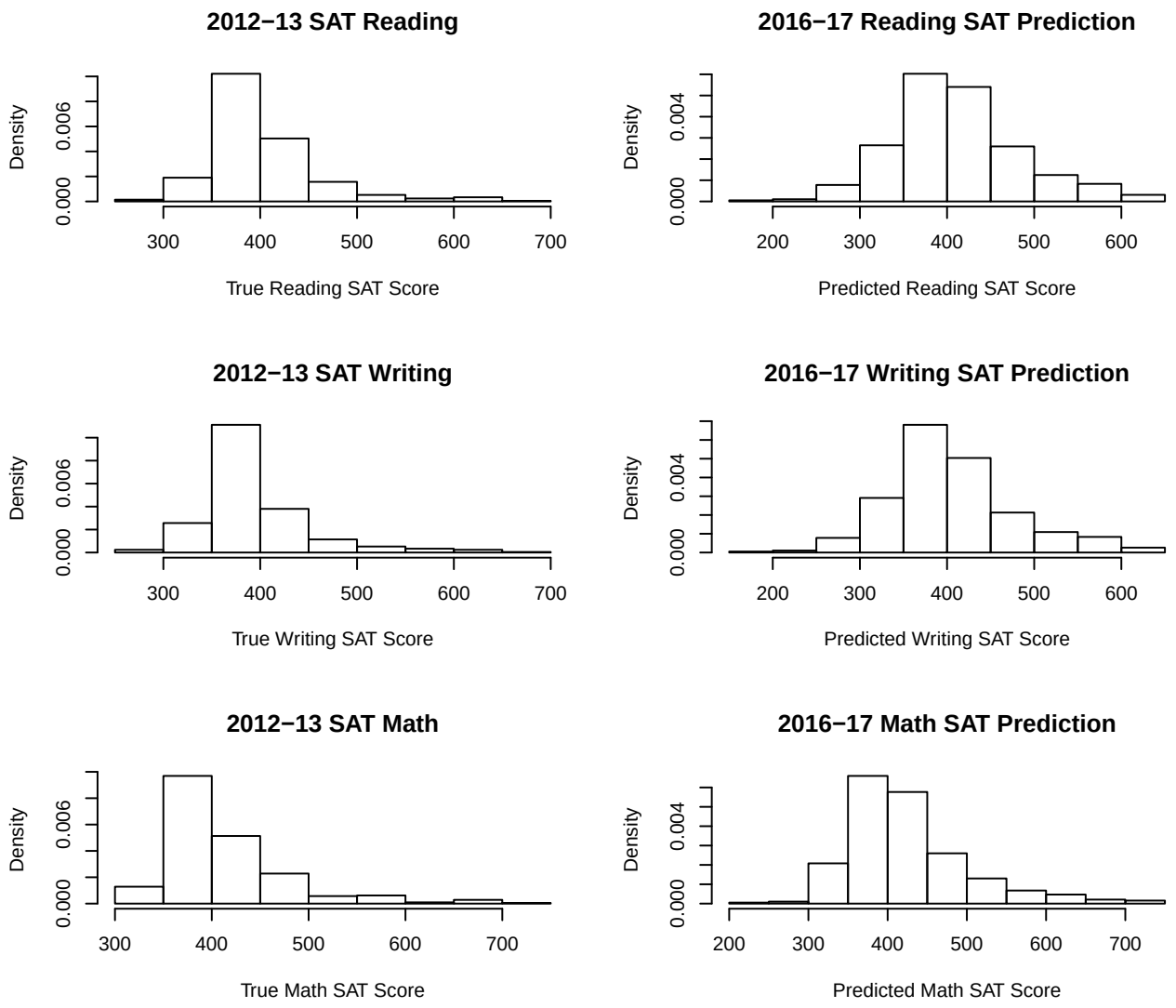

Figure 5: Histogram of Model 2 Predictions of SAT Scores by category.

Appendix D

SAT scores by year.