# Predicting SAT Scores using Regression Modeling

By: Yejoong (Paul) Kim

Advised by: Xiaoyan Li

Department of Computer Science

Princeton University

22 April 2022

**I pledge my honor that this represents my own work in accordance with University regulations.**

*Yejoong Kim*

# Abstract

*What kinds of correlations exist between socioeconomic factors and SAT performance? Can a machine learning model predict SAT performance based on city socioeconomic data? In order to solve this problem, machine learning regression models were used to predict average scores for the Scholastic Assessment Test (SAT) administered by College Board. Such models utilized city and SAT data collected from government census data and state education department SAT data. Although the models did perform better than the baseline metric, they were not as accurate as expected. Despite these limited results, they do indicate a slight correlation, which may provide further insight for future researchers with more resources at their disposal.*


## Introduction

The SAT is a nationwide, standardized test administered by the College Board organization testing a student in evidence-based reading, English writing conventions, and high-school level mathematics. Until 2016, the tests were scored out of 2400; afterwards, the SAT was redesigned to be scored out of 1600.[1] The test plays a fairly important role in college admissions, providing a standardized metric by which college admissions may judge students.[2] Intuitively, the higher the SAT score, the higher the chance a student has in receiving an acceptance letter. Given the stakes, it is no wonder that many high school students pour many hours into preparing for these tests.

However, recent events have called the test's efficacy and fairness into question. The coronavirus pandemic, emerging in 2020, forced many SAT test centers to shut down.[3] Given these drastic changes, many students were unable to take the test, making evaluation of students' performance much more difficult. At this time, many colleges made headlines in the education

world with decisions to make standardized tests such as the SAT optional in the admissions process.[4] This, in turn, seemed to have encouraged discussion regarding the SAT.

## Motivation and Goal

Personal experiences have helped in developing the goal of this project. Years spent tutoring students in various high school subjects, including standardized test preparation, uncovered many insights, notably the main hurdles to receiving proper mentoring and tutoring. These hurdles, often financial or environmental, prevented students from preparing for standardized tests to their fullest potential. With the discussion surrounding the SAT heightening more and more, the motifs of financial hurdles during these experiences became much more noticeable. Ultimately, this personal motivation crystallized into the topic of this research paper, aiming to utilize machine learning regression techniques to investigate the correlations between various socioeconomic factors and SAT performance by city.

## Related Work

Previous work regarding the SAT has appeared in the academic and public horizon over the years. Skeptics of the SAT have often raised doubts about the SAT's fairness, citing statistics detailing the disadvantages low-income students face in preparing for the SAT. One such skeptic is Mark Kantrowitz. In his statistical analysis, Mark posits that families with over $100,000 in income are much more likely to get competitive SAT scores (i.e. satisfactory for selective colleges) than do families of income less than $50,000.[5] This makes sense, as lower-income families would often have fewer resources as their disposal than richer students in preparing for the SAT.

Although not relevant to the SAT, previous work dealing with predicting life outcomes does exist. In particular, one paper, "Measuring the Predictability of Life Outcomes with a Scientific Mass Collaboration," aims to predict the life outcomes of families as part of a mass-collaboration effort. Although many of the models initially failed to give satisfactory performance (i.e. only marginally beat a baseline linear regression model), the mass collaboration effort did give rise to some models that slowly improved over time.[6] However, previous work has not specifically touched upon predicting SAT scores by socioeconomic metrics. Therefore, this paper aims to utilize specific machine learning regression techniques in an attempt to reasonably predict a likely SAT score, based on a collection of data linking socioeconomic data to SAT performance.

## Implementation

The general approach taken to accomplish this paper's goal was as follows:

1. Collect school data
2. Collect city data
3. Link school data with city data
4. Perform pre-processing on data set
5. Model training
6. Model testing and evaluation

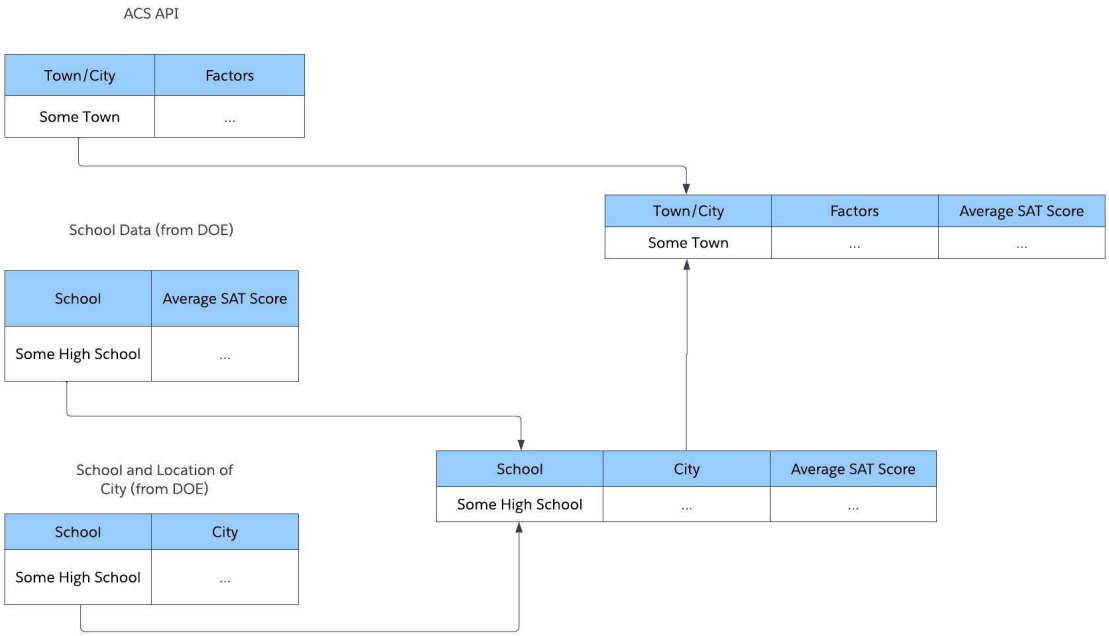


*Figure 1: Dataset construction visualized*

Given the nature of the work, the data collection process would be far more intensive than the average machine learning data set. Several steps would need to be taken for the final data set to be constructed, as visually shown in Figure 1.

## Collecting School Data

To collect SAT scores (from 2015-2016) by school, it became necessary to manually search within various U.S. states' departments of education. Much of this data was provided under the ".csv" format, allowing libraries such as pandas to process and manipulate the data quite easily. These files usually contained various types of data surrounding the school, but four in particular are relevant for this discussion: school name, average SAT math score, average SAT

reading score, and average SAT writing score. The three section scores (math, reading, writing) were simply added, like in the SAT, to provide an average composite score.

However, in order for a school to be linked to its city of residence, another vital part of the school data collection process was collecting files listing the city each school was located in. For this, various states' school directory query pages often accompanied each SAT report. These pages were helpful in finding such locations; without them, it would become time-inefficient to automate the web search of thousands of schools along with their cities of residence.

Once every school and its accompanying city was found, the data was later combined into a school data set such that it contained the school, the average SAT score, and the accompanying city utilizing the pandas library.

## Collecting City Data

City socioeconomic data was collected from the United States Census Bureau's American Community Service (ACS) API, specifically from the year 2015.[7] The API allows a developer to save the data into JSON format, making it easily accessible by pandas.

The following factors were taken into consideration for the city:

- Median Household Income
- Median Value (of Home)
- Total Number of Residents with Less than HS Diploma – Total Number of Residents with HS Diploma
- Total Number of Residents with Some College
- Total Number of Residents with Bachelor's

- Total Number of Residents with Graduate or Professional – Total Number of Labor Force Unemployed
- Total Number of Residents with Income Below Poverty Level Total

## Linking School Data with City Data

Once sufficient school and city data were collected, the data was linked via the school's city, creating a data set with city, socioeconomic data, and SAT scores. During this process, over 1,500 final data points were collected.

## Data Preparation and Pre-Processing

Throughout each step of the data collection stage, the individual data sets were cleaned of invalid or duplicate values. Any duplicate values found in the school data sets were dropped, keeping the first occurrences of such values intact. Invalid SAT scores (i.e. empty or negative numbers) were also ignored, as such values would not make sense. This significantly whittled down the number of data points, as many of the data sets contained duplicates and invalid values.

Once the data set was cleaned, pre-processing began taking place on the data. Initially, min-max normalization was used to normalize the data set. However, since min-max normalization is often volatile to outliers,[8] standardization was ultimately used on the data. With the data cleaned and processed, the project moved on to model training and evaluation.

## Model Training

Over the semester, four different regression models - two linear, two non-linear - were trained and evaluated:

- Linear models
    - Linear regression
    - Lasso regression
- Non-linear models
    - Decision tree regression
    - Random forest regression

For each model, the data was split into a standard 80:20 partition, where 80 percent of the data was used to train the model and 20 percent of the data was used to test the model.[9] The model would then train on the training set and be evaluated with the testing set.

There were considerations to use other types of machine learning algorithms, notably neural networks. However, such algorithms would need much more time, perhaps infeasible given the time frame of the semester, and the scope of this project made the complexity of such algorithms unnecessary.[10] Ultimately, the choice was made to use specific regression models.

## Tools Used

Certain tools proved to be extremely valuable in data collection and model training and evaluation. Python's numerous and popular data science packages simplified and executed such processes efficiently, including:

- pandas
    - Useful for combining and merging school and city data sets.
- sklearn
    - Provided the code to train various regression models.
- numpy

    - Provided the code to manipulate numerical data.
- matplotlib
    - Provided the code to make visual plots.

## Evaluation and Results

Several metrics were used in evaluating the regression models[11]:

- Mean absolute error (MAE)
- Mean squared error (MSE)
- Root mean squared error (RMSE)
- $R^2$ score
- Adjusted $R^2$ score

The first three measures are common metrics used to evaluate the error for a regression model. However, these metrics alone cannot provide a scale of correlation, useful only in comparing various models' errors with each other. Thus, the last two metrics listed above, both the $R^2$ and the adjusted $R^2$ scores, will be especially important to the evaluation of models, as these two metrics can judge the strength of the correlation between the models and SAT performance irregardless of other machine learning models.[12]

## Baseline Model

Ultimately, upon consultation with Professor Xiaoyan Li, the baseline used was the mean of the SAT scores in the data set. Intuitively, when asked to guess the SAT score of a city, the average person would provide the mean SAT score, as it is one of the simplest descriptive statistics one could use for a set of data. It is also somewhat accurate, as by definition the mean

would represent the central tendency of the data set. This mean was then evaluated with the same five metrics as the machine learning models, generating a baseline set of scores for the machine learning models to outperform.

## Linear Model Results

With the data set constructed, the next step was to train and evaluate the linear regression models. As shown in Table 1, the linear models did perform better than the baseline, achieving scores that surpassed those of the baseline model. Nevertheless, the results indicated a weak correlation, with low $R^2$ and Adjusted $R^2$ scores.

| **Linear Model Performance** | | | |
|---|---|---|---|
| Model | Baseline | Linear Regression | Lasso Regression |
| MAE | 141.31375106958998 | 116.23500296485798 | 116.75613643994146 |
| MSE | 32249.512884175685 | 23273.072853936814 | 23288.160352954073 |
| RMSE | 179.58149371295386 | 152.55514692705984 | 152.60458824345378 |
| $R^2$ Score | 0.0 | 0.27834342994506023 | 0.27787559345179447 |
| Adjusted $R^2$ Score | -0.04313725490196085 | 0.24721314653092552 | 0.24672512885559728 |

*Table 1: Linear and Lasso Regression Performance*

Upon analysis of the predicted vs. actual plots for the two linear models (Figures 2 and 3), it seems clear that there are many data points where the model severely underestimates or

overestimates, some of which differ by over 400 points. No particular visual pattern is apparent in either of the two linear plots.

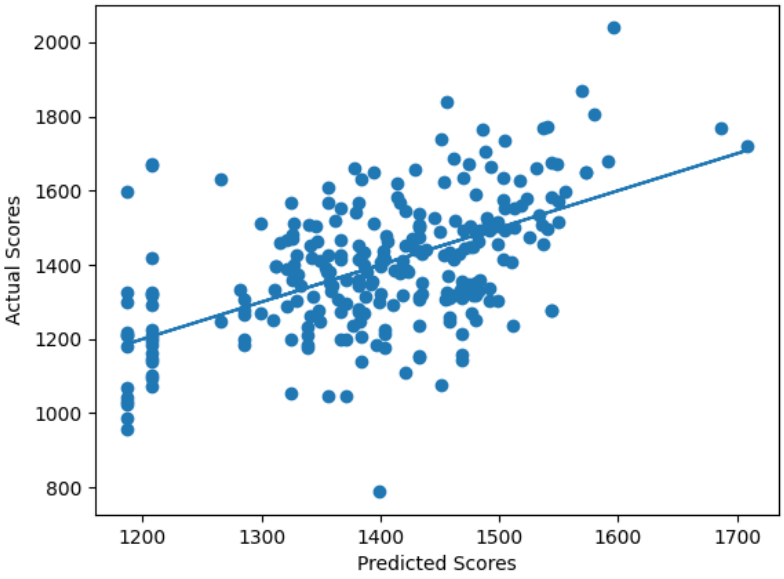


*Figure 2: Predicted vs. Actual Scores for Linear Regression*

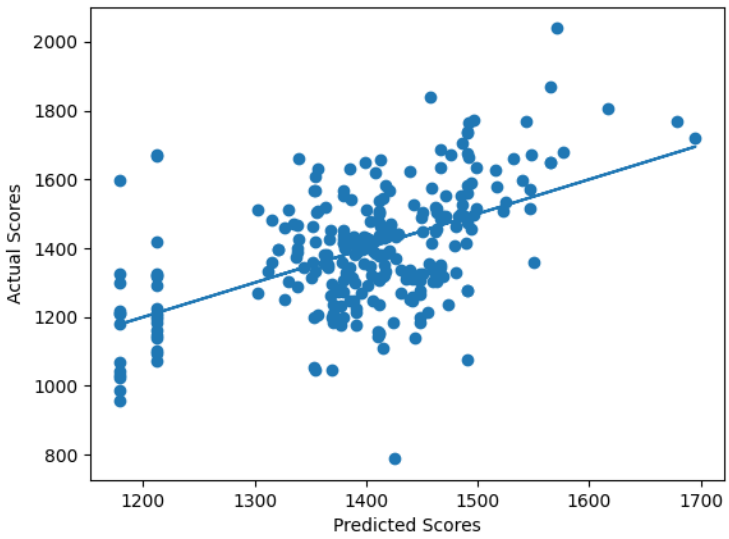


*Figure 3: Predicted vs. Actual Scores for Lasso Regression*

## Non-Linear Model Results

Following investigation of the linear models' performance, the two non-linear models were trained to see if they would perform better on the data set. Following the aforementioned process of training and evaluating the models, the non-linear models' results were slightly better than those produced by the linear models. Despite this, the non-linear models still provided relatively weak correlation, as the Decision Tree and Random Forest Regression models still only provided $R^2$ and Adjusted $R^2$ Scores of around 0.3.

| **Non-Linear Model Performance** | | | |
|---|---|---|---|
| Model | Baseline (Mean of Entire Dataset) | Decision Tree Regression | Random Forest Regression |
| MAE | 141.31375106958998 | 110.88939358339954 | 114.77699367973595 |
| MSE | 32249.512884175685 | 21589.999159860687 | 22021.667369928644 |
| RMSE | 179.58149371295386 | 146.93535707875313 | 148.39699245580633 |
| $R^2$ Score | 0.0 | 0.3305325498279216 | 0.3171472868747013 |
| Adjusted $R^2$ Score | -0.04313725490196085 | 0.3070881644133352 | 0.29323415684307597 |

*Table 2: Decision Tree and Random Forest Regression*

Upon analysis of the predicted vs. actual plots for the two non-linear models (Figures 4 and 5), it seems clear that there are many data points where the model severely underestimates or overestimates, some of which differ by over 400 points, as with the linear models. No particular

visual pattern is apparent in either of the two linear plots. Only some of the points seem to fall within 100 points of the models' predictions, explaining the weak (but still present) correlation the models have with SAT performance. The decision tree regression model seems to have a notable vertical pattern in its plot of predicted vs. actual scores, often dividing into several "lines" of data points. The random forest regression model's plot looks similar to those of the linear models, where no clear visual pattern seems to exist.

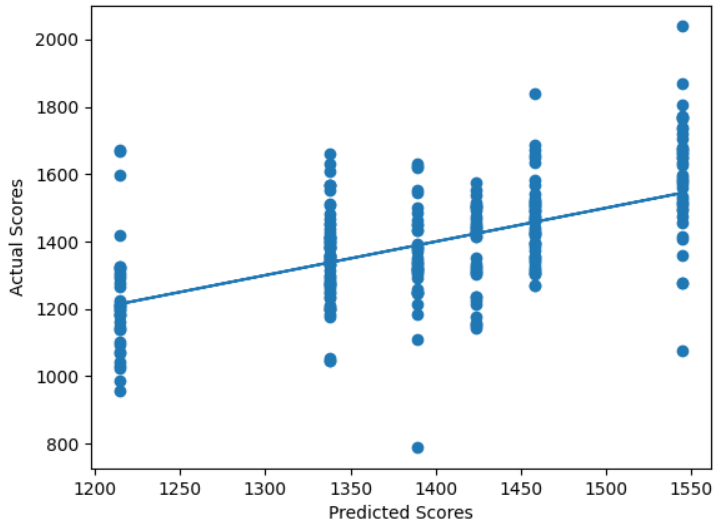


*Figure 4: Predicted vs. Actual Scores for Decision Tree Regression*

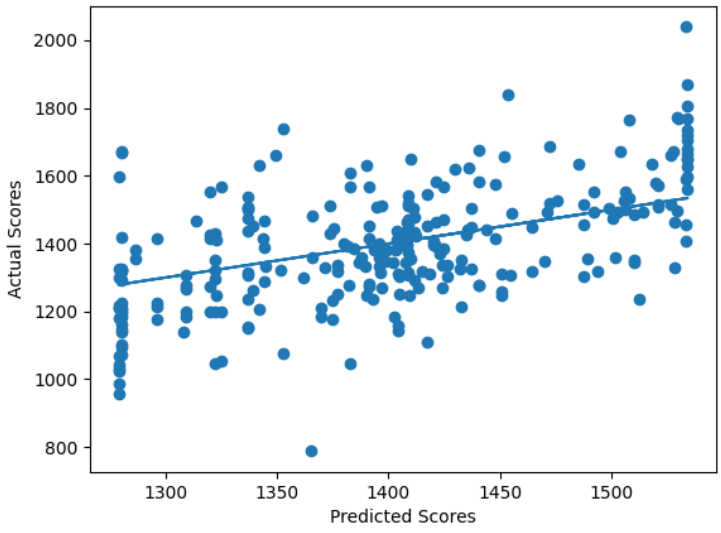

*Figure 4: Predicted vs. Actual Scores for Random Forest Regression*

## Post Model Analysis

Upon observing the relatively weak correlations between the models and SAT performance data, the factors' relationship with SAT performance were re-analyzed. Specifically, each factor used in the data set was plotted against SAT performance to see if any particular visual trends may have appeared. Upon further observation, seven factors were found to have visually very little correlation with SAT performance, such as total number of bachelor's degree holders, as shown in Figure 6. These factors would often take patterns in the shape of vertical lines, indicating little to no correlation.

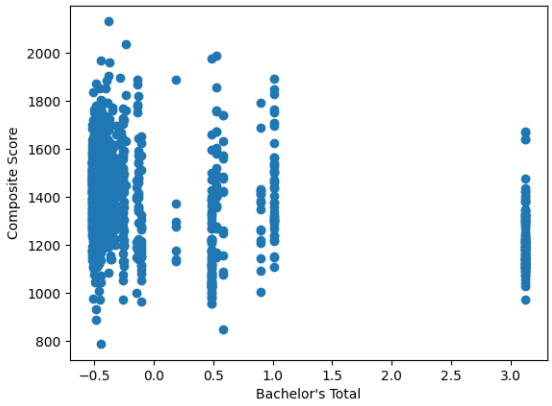


*Figure 6: Total Number of Bachelor's Degree Holders Against SAT Composite Score. Note the vertical line patterns forming in the plot.*

Two factors - Median Household Income and Median Value of Homes - were found to have a clearer positive correlation with SAT performance, however, as seen in Figures 7 and 8. Unlike the other factors, the two factors had a clear visual direction. However, it seems that even in these plots, there are many points which would stray far from a

regression model.

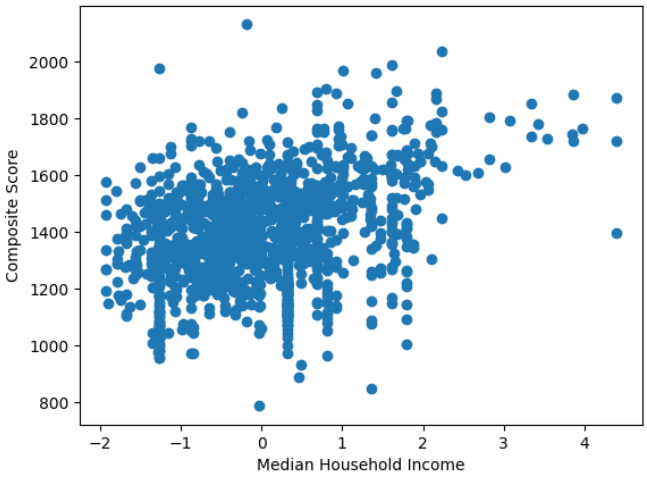


*Figure 7: Median Household Income Against SAT Composite Score*

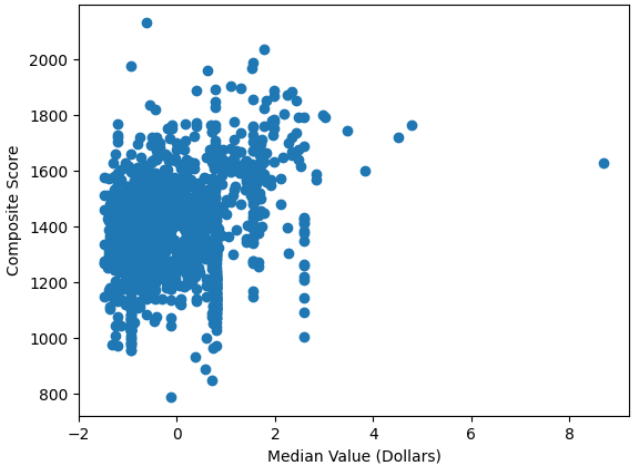


*Figure 8: Median Household Income Against SAT Composite Score*

## Limitations

There were many hurdles that came about over the course of the research project. Most notably, many states' departments of education did not release SAT score data, limiting the

number of states by about half. Even among states that did provide SAT data, many of them provided statistics unsuitable for machine learning (i.e. the SAT scores were only displayed as statewide averages, not as SAT scores by school). Furthermore, among the states that clearly provided SAT scores by school, some states, such as Florida, did not provide a time-efficient directory of schools by city. This made any linking between school and city relatively difficult and time-inefficient. Only four states were found within the time span of the semester that provided the necessary statistics to construct a detailed data set: Washington, Massachusetts, California, and Pennsylvania.

Compounding this issue was an issue of modernity. Specifically, the scores used in the SAT data sets were from the 2015-2016 academic year. The SAT had since then undergone a major change in 2016, changing the types of questions, the number of multiple-choice answers, etc. Though an analysis of the 2015-2016 academic year could provide interesting results, an analysis of more recent years (i.e. 2018-2019) would provide a much more modern and robust analysis of the current landscape of the SAT.

## Future Work

Although the results proved to be relatively weak, they do provide clear and interesting potential steps for any future research. The most important direction that can be taken is to find more data points through the College Board organization (the one that administers the SAT), which already provides state-wide (but not school wide) statistics through their own website. Should such SAT data be obtained, the only step necessary would be to link each school with its city of residence, which is likely recorded by the College Board along with the school. At that point, SAT performance can be easily linked to

the city data generated by the Census Bureau's API. Such a step, though dependent on College Board's approval, could provide better insight and perhaps train the models to become more accurate.

Even without College Board's intervention, future steps can be taken with cites' socioeconomic data. Although the socioeconomic factor of unemployment seems to have very little correlation with SAT performance, it may be possible that certain occupations (i.e. total number of jobs in technology, education, etc) may have stronger correlation with SAT performance. Given the intuitive speculation that a student's environment may often influence their academic performance, this may perhaps lead to more interesting results. Therefore, investigation of these factors may prove to be worthwhile for future researchers.

## Conclusions

Despite many limitations for the research project, the results seem to provide some insight regarding socioeconomic factors and SAT performance. Linear and lasso regression were more accurate than the baseline metric, but proved to have a weak correlation with SAT performance. Non-linear models - such as decision tree and random forest regression - performed slightly better than the linear models, but such models still provided only a loose correlation with SAT performance. During this project, interesting patterns were found between Median Household Income and Median Value of Homes, confirming what much of the public sphere has said regarding the SAT. Future research may be able to bypass both the data and time limitations of this paper and perhaps find more definitive patterns.

This, in turn, may have many implications for policy. For instance, such research may result in an increased demand for more accessible tutoring for low-income families.

Such initiatives are already happening even in the Princeton community through groups such as HatchTutors[13], a group dedicated to providing tutoring to low-income students. In addition, this kind of research, though not directly aiming to call for such action, may contribute to the call for the complete elimination of the SAT requirement in college admissions. Many post-secondary institutions have already made the SAT optional in their admissions processes[14], after all, and such research may provide sufficient impetus for them to continue to do so.

Regardless, this paper serves to illuminate to readers the socioeconomic effects surrounding a student in preparation for the SAT. While the models may have had only loose correlations between the socioeconomic factors and SAT performance, further research may perhaps show a much better picture of the climate of SAT test preparation.

**Footnotes:**

1. DeGeurin
2. DeGeurin
3. Saul
4. Saul
5. Kantrowitz, 8
6. Salganik
7. Bureau, US Census, 2015 ACS Survey
8. Li, Xiaoyan, “Data Preparation II”
9. Li, Xiaoyan, “Model Training and Evaluation I”
10. Li, Xiaoyan, “Model Training and Evaluation I”

11. Li, Xiaoyan, “Model Training and Evaluation II”
12. Li, Xiaoyan, “Model Training and Evaluation II”
13. Saul
14. HatchTutors is an organization dedicated to providing cost-effective tutoring solutions for low-income students.

# Bibliography

"2020-21 SAT Performance Report - All Students." Massachusetts Department Of Elementary And Secondary Education - 2020-21 SAT Performance Report - All Students Statewide Report, https://profiles.doe.mass.edu/statereport/sat.aspx.

*About Us*, HatchTutors, https://www.hatchtutors.org/about_us.

Brownlee, Jason. "How to Use StandardScaler and MinMaxScaler Transforms in Python." *Machine Learning Mastery*, 27 Aug. 2020, https://machinelearningmastery.com/standardscaler-and-minmaxscaler-transforms-in-python/.

Bureau, US Census. "American Community Survey 5-Year Data (2009-2020)." *Census.gov*, 8 Mar. 2022, https://www.census.gov/data/developers/data-sets/acs-5year.2015.html.

"California SAT Report 2015-2016 - Dataset by Education." *Data.world*, 16 Aug. 2017, https://data.world/education/california-sat-report-2015-2016.

DeGeurin, Mack. "Here's How the SAT Has Changed over the Past 90 Years and Where It Might Be Heading." *Insider*, Insider, 9 Aug. 2019, https://www.insider.com/how-the-sat-has-changed-over-the-past-90-years-2019-8.

Kang, Jay Caspian. "Why Sat Test Prep Doesn't Help Who You Might Think It Helps." *The New York Times*, The New York Times, 9 Sept. 2021,

https://www.nytimes.com/2021/09/09/opinion/sat-standardized-tests-ucs.html.

Kantrowitz, Mark. “Admissions Tests Discriminate Against College Admission of Minority and Low-Income Students at Selective Colleges.” Student Aid Policy Analysis Papers, 21 May 2021.

Li, Lorraine. “Introduction to Linear Regression in Python.” *Medium*, Towards Data Science, 5 Feb. 2019, https://towardsdatascience.com/introduction-to-linear-regression-in-python-c12a072bedf0.

Li, Xiaoyan. “Analysis & Presentation.” IW03: Data Science.

Li, Xiaoyan. “Common Mistakes in Machine Learning.” IW03: Data Science.

Li, Xiaoyan. “Data Preparation I.” IW03: Data Science.

Li, Xiaoyan. “Data Preparation II.” IW03: Data Science.

Li, Xiaoyan. “Model Training and Evaluation I.” IW03: Data Science.

Li, Xiaoyan. “Model Training and Evaluation II.” IW03: Data Science.

“Merge, Join, Concatenate and Compare.” *Merge, Join, Concatenate and Compare - Pandas 1.4.2 Documentation*, https://pandas.pydata.org/docs/user_guide/merging.html.

“Numpy Reference.” *NumPy Reference - NumPy v1.22 Manual*, https://numpy.org/doc/stable/reference/index.html#reference.

"SAT and ACT Scores." *Department of Education*, https://www.education.pa.gov/K-12/Assessment%20and%20Accountability/SAT-ACT/Pages/default.aspx.

Salganik, Matthew J., et al. "Measuring the Predictability of Life Outcomes with a Scientific Mass Collaboration." Proceedings of the National Academy of Sciences, vol. 117, no. 15, Apr. 2020, p. 8398, https://doi.org/10.1073/pnas.1915006117.

Saul, Stephanie. "Put down Your No. 2 Pencils. Forever." *The New York Times*, The New York Times, 25 Jan. 2022, https://www.nytimes.com/2022/01/25/us/sat-test-digital.html.

"Transformer-based Machine Learning for Fast SAT Solvers and Logic Synthesis" by Feng Shi, Chonghan Lee, Mohammad Khairul Bashar, Nikhil Shukla, Song-Chun Zhu, Vijaykrishnan Narayanan

"Tutorials." *Tutorials - Matplotlib 3.5.1 Documentation*, https://matplotlib.org/stable/tutorials/index.html.

"User Guide." User Guide - Pandas 1.4.2 Documentation, https://pandas.pydata.org/docs/user_guide/index.html.

"User Guide: Contents." *Scikit*, https://scikit-learn.org/stable/user_guide.html.

*Washington - Public Schools - K12.Wa.us*. https://www.k12.wa.us/sites/default/files/public/advancedplacement/pubdocs/districtresul

tscollegeboardexamswa2016.pdf.