Abhiram Kolal

Professor Singh

Data Analytics

7/17/20

Assignment 3

I allow sharing of this report with other students in future versions of this class.

**Background and Proposal:**

The SAT is a standardized test widely used for college admissions in the United States. Each year, many high school students take this exam and most students tend to spend a substantial amount of time preparing for this exam, as it is one of the most important tests one would take throughout their lifetime. There has been considerable controversy around the SAT test, with some critics saying that the “only thing the SAT predicts well now is socioeconomic status” (Sacket et. al., 2009), rather than assess a level of college-readiness which is its main purpose. In order to address these concerns, the College Board had introduced an “adversity score” for the year 2019, which ranked individual students based on their socioeconomic status. These scores were not released to the public and did not affect a student’s actual score on the exam itself but were sent to colleges for their consideration. The “adversity score” had not been implemented into any tests taken in 2019, and the idea has been withdrawn from the College Board after much backlash (Hartocolis, 2019).

I aim to research the various factors of a student’s upbringing that can contribute towards their SAT Score. It would be very helpful to predict the outcome of certain students’ SAT scores as a result of their environment, such as the overall affluence of their neighborhood as well as access towards the internet, which can be used for finding an endless amount of resources that can contribute to SAT scores. To gain a broader sense of a student’s environment, I have decided to look at geographic and socioeconomic data. These sources of information can paint an accurate picture of the environment a student is in and comparing this data to the average SAT score of the area would contribute to finding a correlation between them. The three main components I would like to analyze about a student’s environment are the quality of the schools that they attend, their level of access to the internet, as well as the education level of the people they are surrounded by. Therefore, **I propose to predict the average SAT score of students as a function of average amount of funding given to public schools, number of households with internet access, median household income, and number of individuals with bachelor's degrees, utilizing a state-by-state unit of analysis.**

**Method:**

This study will use a multi-linear regression model, with four distinct independent variables as the basis of predicting the dependent variable. These variables are all quantifiable, and the proposed variables are as follows:

Independent Variables:

Total expenditures for public and secondary education (in $ - US Dollars)

Number of households with internet access (Count)

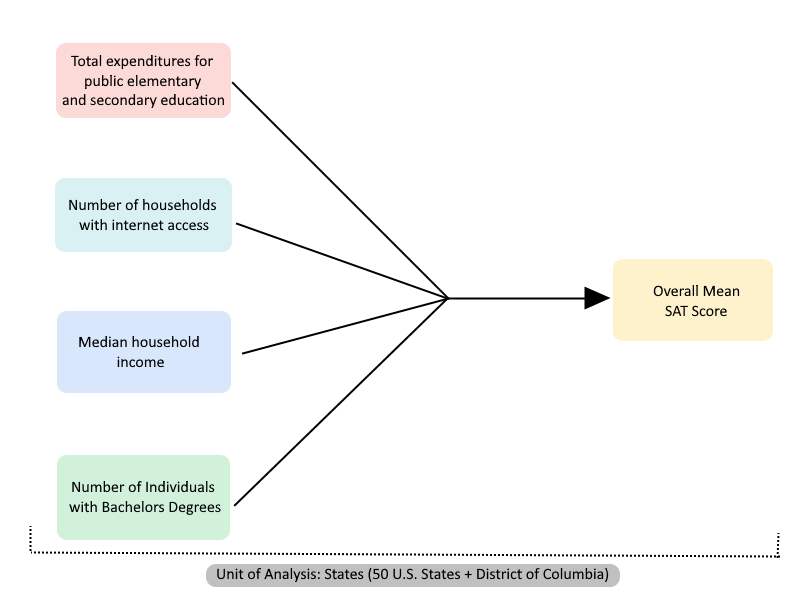
Median household income (in $ - US Dollars)

Number of Individuals with Bachelor's Degrees (Count)

Dependent Variable:

Average SAT Score by state (Count)

The unit of analysis that I will be measuring these variables by are by US state, as the main purpose of this study is to analyze the different social environments and their effect on SAT scores. A visual representation of this model is shown below:



**Analysis and Assumptions:**

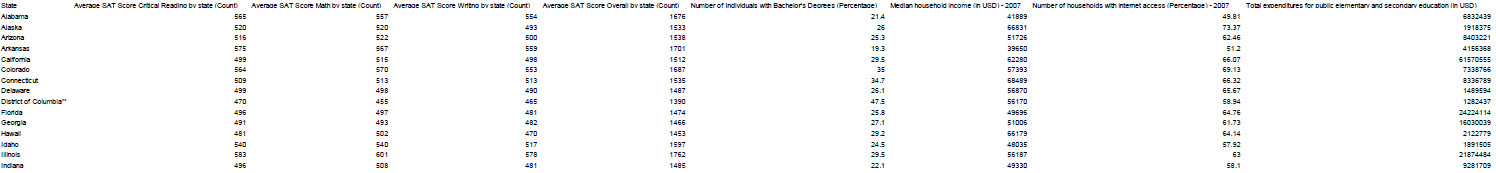
Although there is a wide range of areas that can dictate an environment for a student, I have limited my definition of an environment based on the four variables stated above. I believe that the aforementioned variables are important, as well as most significant, because they cover many aspects of a student’s upbringing, although it cannot account for everything.

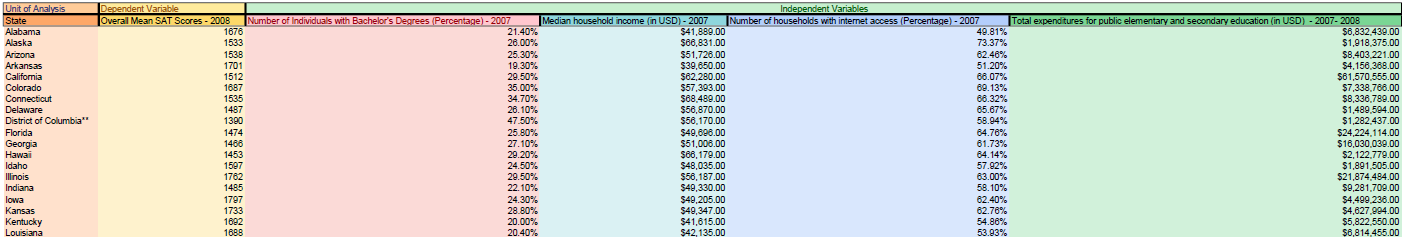
The average amount of funding in schools, as well as median household income should roughly directly correlate with one another, as it shows the overall affluence of the geographic area, which is part of the student’s environment. An assumption could be made that a higher median income can allow for the student to access more effective ways for preparation for the SAT, such as taking private classes or being tutored, but this analysis will only look at public school data simply for more control over the data being collected and analyzed. The purpose of including internet access is to further gain insight on the student’s learning environment. I am placing an assumption that because the internet is heavily used for academic purposes, if a student is able to access the internet, they will be able to access test preparation materials very easily and utilize those resources to study for the SAT. If students are unable to access the internet, they cannot view these resources as easily as ones that do have internet access. The purpose of including people that have bachelor's degrees can also affect a student’s environment, and an assumption could be made that people that hold bachelor’s degrees have a higher level of education who favor college as an option for high school students. With this assumption, another assumption could be made that if a student is interacting with people that hold bachelor’s degrees, whether it be family, friends, teachers, coaches, etc., these people would encourage students to go to college and in turn, take the SAT seriously and motivate them to do their best on the exam.

Although the test itself is based on student performance, and a variety of factors such as the student’s individual learning behavior, GPA, studying habits between them and their peers, and much more, I believe that the aforementioned variables would also be able to predict these values as well, because the external factors in a student’s life can both implicitly and explicitly place great impact on a student’s academic performance. For the purpose of this analysis, the student’s individual performance will not be included, and the focus will be placed solely on SAT scores. All the aforementioned variables can place significant impacts predicting how well students will do on the SAT exam.

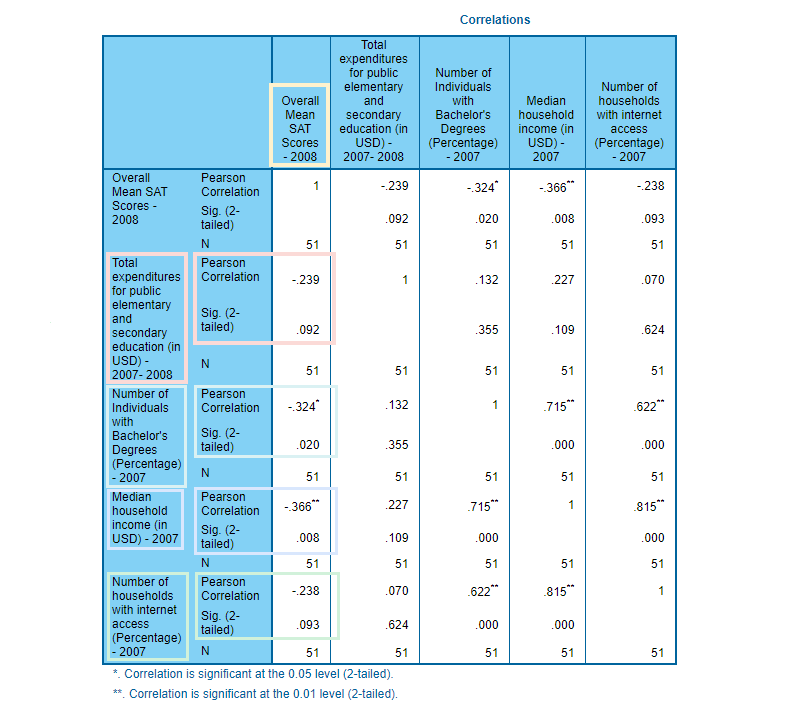
**Curation Process**

The data that is relevant to this analysis was found through government sources as well as the College Board archives. One particular challenge I found was that the latest data reported for a detailed analysis of public school expenditures on the National Center for Education Statistics website was for the 2007-08 fiscal year, so due to that I had to limit my findings of other input variables to that timeframe. As the primary unit of analysis was by state, I was able to find the relevant values for each state and distributed this data accordingly. I was able to find the data for the number of households with internet access through the National Telecommunications and Information Administration website, where I separated the values by state and found it easier to convert the values into percentages of the population rather than a total count. This action was taken to provide a better understanding of the overall socioeconomic status of the state, and it works together with factors such as the median household income, number of individuals with bachelor's degrees, and total expenditures in public and secondary education. The median household income for the US by state in 2008 was found through an official Census survey published in 2009. The number of individuals with bachelor’s degrees for 2007 was found from the Education Resources Information Center, where I changed the values from a count to a percentage. The dependent variable, SAT scores, were calculated as an average of all test-takers per state in 2008. This information was provided from the College Board, and their reports only included the scores taken in the three assessments for the test - Critical Reading, Math, and Writing. Using the averages, I calculated the student’s final scores by taking the sum of these scores. All of these values had been reported in a pdf format rather than in spreadsheet form, so I had to manually input all of these data points into an Excel spreadsheet in order for better accessibility and usage for my final analysis.  I had also noticed that many of the datasets I encountered had more than the initial 50 states, including U.S. territories as well. I had filtered out the territorial data, but I had included the District of Columbia as I found there were significant differences between D.C. and the surrounding states, which is still considered to be part of the U.S. mainland. All in all, this brought my unit of analysis to 51 states (including Washington, DC). After curating the main data and listing them all out in an Excel spreadsheet, matching values state-by-state, I had to “clean up” the data by labeling units of analysis, independent variables, and dependent variables and organizing the datasets in that manner. The screenshots provided below show my initial data curation, as well as the simplified and calculated changes I have made which I hope to use for the next step of my analysis.

**Before:**

**After:**

**SPSS Analysis:**



**Correlation Scores**

* Total expenditures for public elementary and secondary education & Overall SAT Score 🡪 r = -0.239
* Number of Individuals with bachelor’s degrees & Overall SAT Score 🡪 r = -0.324
* Median household income & Overall SAT Score 🡪 r = -0.366
* Number of households with internet access & Overall SAT Score 🡪 r = -0.238

**Significant Values**

* Total expenditures for public elementary and secondary education & Overall SAT Score 🡪 p-value = 0.092
  + Not statistically significant (p-value > 0.05)
* Number of Individuals with bachelor’s degrees & Overall SAT Score 🡪 p-value = 0.020
  + Statistically significant (p-value < 0.05)
* Median household income & Overall SAT Score 🡪 p-value = 0.008
  + Statistically significant (p-value < 0.05)
* Number of households with internet access & Overall SAT Score 🡪 p-value = 0.093
  + Not statistically significant (p-value > 0.05)

**Correlation Directions**

* Total expenditures for public elementary and secondary education & Overall SAT Score 🡪 negative direction
* Number of Individuals with bachelor’s degrees & Overall SAT Score 🡪 negative direction
* Median household income & Overall SAT Score 🡪 negative direction
* Number of households with internet access & Overall SAT Score 🡪 negative direction

**Correlations: Interpretation and Analysis**

For each of the independent variables’ correlations with the dependent *Overall SAT Score* variable, the R-values were both negative and their absolute values were closer to zero than one. The negative direction indicates that there is a negative relationship with the independent and dependent variables, and the absolute values being closer to zero indicate that there is a less linear, and more scattered relationship between them.

The correlation value (R-value) of *Total expenditures for public elementary and secondary education* and *Overall SAT Score* was negative and close to zero. This would imply that as the dependent variable (*Overall SAT Score*) increases, the independent variable (*Total expenditures for public elementary and secondary education*) decreases. However, this assumption can vary because the R-value is closer to zero than one, which implies a less linear, more scattered relationship.

The R-value of *Number of Individuals with bachelor’s degrees* and *Overall SAT Score* was negative and the absolute value was close to zero. This would imply that as the dependent variable (*Overall SAT Score*) increases, the independent variable (*Number of Individuals with bachelor’s degrees*) decreases. However, this assumption can also vary because the R-value is closer to zero than one.

The R-value between *Median household income* and *Overall SAT Score* was also negative and close to zero. This would imply that as the dependent variable (*Overall SAT Score*) increases, the independent variable (*Median household income*) decreases. However, this assumption can also vary because the R-value is closer to zero than one.

Finally, the R-value between *Number of households with internet access* and *Overall SAT Score* was also negative and close to zero. This would imply that as the dependent variable (*Overall SAT Score*) increases, the independent variable (*Number of households with internet access*) decreases. However, this assumption can also vary because the R-value is closer to zero than one.

For the P-values of each association, it appears that two out of the four relationships created a P-value lower than the 0.05 critical value. This indicates that two out of the four independent variables’ correlations with the dependent variable are statistically significant and provide an accountable relationship between the two variables.

The P-value derived from the relationship between *Total expenditures for public elementary and secondary education* and*Overall SAT Score* was above the 0.05 critical value. This was not statistically significant, which shows that there is not a genuine relationship between the variables, and more a more varied range of values.

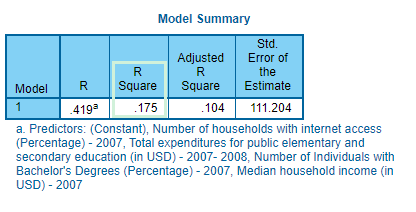
The P-value derived from the relationship between *Number of Individuals with bachelor’s degrees* and *Overall SAT Score* was below the 0.05 critical value. This was one of two independent variables to be statistically significant. This implies that there is a greater likelihood of this value being a genuine correlation, rather than a coincidence.

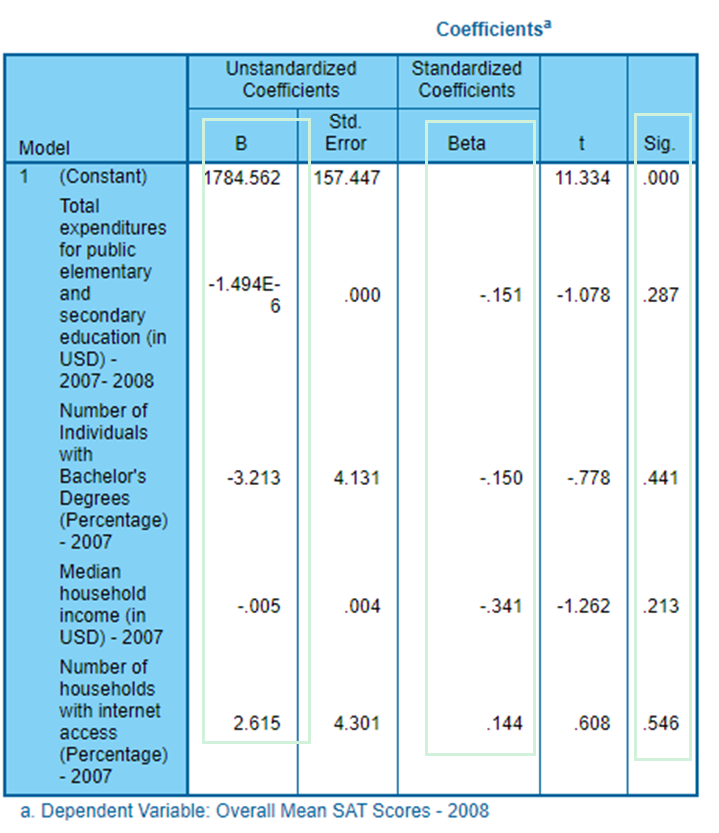
The P-value derived from the relationship between *Median household income* and *Overall SAT Score* was below the 0.05 critical value. This was the second of two independent variables to be statistically significant. This implies that there is a greater likelihood of this value being a genuine correlation, rather than a coincidence.

Finally, the P-value derived from the relationship between *Number of households with internet access* and*Overall SAT Score* was above the 0.05 critical value. This was not statistically significant, which shows that there is not a genuine relationship between the variables, and more a more varied range of values.

Overall, based on bivariate correlation there is not much linearity within the given variables. As discussed above, there is a negative correlation between the variables, and two out of the four variables proved to be statistically significant based on P-value.

**Multiple Linear Regression**





**Percentage of Variance Explained by the Created Model**

* R-Square = 0.175 – 17.5%
  + 17.5% of the variance was explained by the created model

**Significant Variables and Coefficients**

* **Significance**
  + Total expenditures for public elementary and secondary education & Overall SAT Score 🡪 p-value = 0.287
    - Not statistically significant (p-value > 0.05)
  + Number of Individuals with bachelor’s degrees & Overall SAT Score 🡪 p-value = 0.441
    - Not statistically significant (p-value > 0.05)
  + Median household income & Overall SAT Score 🡪 p-value = 0.213
    - Not statistically significant (p-value > 0.05)
  + Number of households with internet access & Overall SAT Score 🡪 p-value = 0.546
    - Not statistically significant (p-value > 0.05)
* **Unstandardized Coefficients**

An increase of the independent variables by an index of one, will reflect a change in index for the dependent (outcome) variable in the amount of the specified unstandardized coefficients.

* + ***Total expenditures for public elementary and secondary education:*** -1.494E-6 (Negative)
    - For every 1 unit of change in *Total expenditures for public elementary and secondary education*, the dependent variable (*Overall SAT Scores*), decreases by -1.494E-6.
  + ***Number of Individuals with bachelor’s degrees:*** -3.213 (Negative)
    - For every 1 unit of change in *Number of Individuals with bachelor’s degrees*, the dependent variable (*Overall SAT Scores*), decreases by -3.213
  + ***Median household income:*** -0.005 (Negative)
    - For every 1 unit of change in *Median household income*, the dependent variable (*Overall SAT Scores*), decreases by -0.005
  + ***Number of households with internet access:*** 2.615 (Positive)
    - For every 1 unit of change in *Number of households with internet access*, the dependent variable (*Overall SAT Scores*), increases by 2.615

**Multiple Linear Regression: Interpretations, Analysis and Policy Making**

The model summary proved to only show that 17.5% of the variance of the dependent variable (Overall SAT Score) can be explained by the independent variables, which is represented by R-square value. This is somewhat low, and it means that over 80% of the data cannot fit using this model, but the presence of an R-square value shows that there still is a correlation between the dependent and independent variables. There were many assumptions that were to be made in order to control the data curation of this model, and given that there is more factors to be seen with SAT scores that were ignored due to lack of quantifiable data such as a student’s intellectual level and test-taking ability, it is expected that there can be varied data using these specific variables.

All the independent variables were not statistically significant when viewing the multiple regression P-values.

* The P-value for *Total expenditures for public elementary and secondary education* and *Overall SAT Score* was 0.287, which is far greater than the critical value of 0.05.
* The P-value for *Number of Individuals with bachelor’s degrees* and *Overall SAT Score* was 0.441, which is also far greater than the critical value of 0.05.
* The P-value for *Median household income* and *Overall SAT Score* was p-value = 0.213, which is far greater than the critical value of 0.05.
* The P-value for *Number of households with internet access* and *Overall SAT Score* was p-value = 0.546, which is far greater than the critical value of 0.05.

These values being much greater than the critical value all imply that there is a greater chance of coincidence rather than genuine correlation.

While analyzing the coefficients of each variable, for both the standardized and unstandardized coefficients, the only positive relationship found was between *Number of households with internet access* and *Overall SAT Score.* This indicates that the dependent *Overall SAT Score* variable increases in index (for the unstandardized coefficient) and standard deviation (for the standardized coefficient) for *Number of households with internet access* increase in these relationships. The other three independent variables, *Total expenditures for public elementary and secondary education, Number of Individuals with bachelor’s degrees,* and *Median household income,* all demonstrate a negative relationship with respect to the dependent *Overall SAT Score* variable. This indicates that the dependent *Overall SAT Score* variable decreases in index (for the unstandardized coefficient) and standard deviation (for the standardized coefficient).

After creating the regression model, the results were very surprising to see. The model implies that much of the data is varied and not statistically significant, which leads to strong variance in the data itself when building a multi-linear regression. This makes it hard to prove that there is much of a correlation between the data itself, but the analysis does show an overall negative relationship with the input variables and output variable.

Because of circumstances such as the R-squared score being 0.175, where only 17.5% of the data can be expressed using this linear regression model, and the fact that most of the P-values were not statistically significant, I would argue that it is insufficient to accurately predict SAT scores solely based on external variables, which was the purpose of this study. Policy makers such as the College Board and local officials that fund our public schools cannot use this data to argue that an increase in total expenditures for schools will lead to a decrease in SAT scores, but it is interesting to see this result. However, one thing I saw that was also interesting was that there was a positive correlation between Number of households with internet access and Overall SAT Scores. This piece of information can be useful as it is practical that students that can access the internet will be able to study for the SAT more effectively, and it could encourage policy makers that govern accessibility and affordability to the internet to increase their influence over high school students, and get as many high school students as they can to be able to access the internet.

**Ethical Implications and Study Limitations**

All the data curated during this study was found using federal databases as well as the College Board’s archive of SAT scores for the years of 2007-2008. Although the data collected were analyzing human characteristics and demographics, everything used in this study was aggregated to counts or percentages and was also public information with consent by each individual participant. The data used in this study had no way of compromising an individual’s identity or having a breach of privacy because of these factors.

There were many limitations that were placed on this study, due to both time constraints and curation constraints, which have affected the overall result of this analysis greatly. One limit to the data was that I was using a given schoolyear’s data rather than an aggregation of multiple years. The reason this limit exists is due to both time constraints as well as lack of sufficient data. Due to the nature of this assignment, especially given the accelerated summer schedule, the time it would take to curate the data for multiple years of these variables would not allow me to deliver this study on time. Seeing as I had to curate the data manually, I would have been unable to aggregate all of the data in the given timeframe. Data curation was one of the longest parts of this assignment, and if I had been given more time I would be able to collect all the information needed, and with multiple years of data I could possibly find a better regression model due to being able to analyze these trends over time rather than a single year. One of the main reasons that I was unable to find a more statistically significant data model was due to the fact that I chose to only include external factors in this research, which were defined as the Total expenditures for public elementary and secondary education, Number of Individuals with bachelor’s degrees, Median household income, and Number of households with internet access. The reasons I had chosen these variables were stated above, but looking at the result it is clear that these variables are not enough to predict a correlation between a student’s SAT score. There are much more factors that can contribute to a student’s performance on the SAT, which are not necessarily defined from the external factors of a student. Factors such as a student’s intellectual level and test-taking ability are some examples of what can determine a student’s performance on the SAT, but these values were not included in this study mainly due to the lack of quantifiable data.

Although the results of this study were not what I had initially expected, they still can serve as a baseline for the time that this study had analyzed. Seeing a negative overall trend for the year 2008 regarding the external factors of a student vs. their SAT score can be used as a snapshot of data and future researchers can do the same analysis with different years to see what has changed since then.

Works Cited

Literature

Hartocollis, Anemona. “SAT 'Adversity Score' Is Abandoned in Wake of Criticism.” The New York Times, The New York Times, 27 Aug. 2019, www.nytimes.com/2019/08/27/us/sat-adversity-score-college-board.html.

Sackett, Paul R., et al. “Socioeconomic Status and the Relationship Between the SAT® and Freshman GPA: An Analysis of Data from 41 Colleges and Universities.” College Board, 2009, files.eric.ed.gov/fulltext/ED562860.pdf.

Data Sources

*College-Bound Seniors 2008*, College Board, 16 Feb. 2018, research.collegeboard.org/programs/sat/data/archived/cb-seniors-2008.

Crissey, Sarah R. “Educational Attainment in the United States: 2007.” *Education Resrouces Information Center*, 2009, files.eric.ed.gov/fulltext/ED505040.pdf.

*Households Using the Internet in and Outside the Home, by Selected Characteristics: Total, Urban, Rural, Principal City, 2007*. National Telecommunications and Information Administration, 2007, www.ntia.doc.gov/files/ntia/publications/table\_householdinternet2007.pdf.

“Revenues and Expenditures for Public Elementary and Secondary Education: School Year 200708 (Fiscal Year 2008).” *National Center for Education Statistics (NCES) Home Page, a Part of the U.S. Department of Education*, 2008, nces.ed.gov/pubs2010/expenditures/tables.asp.

Semega, Jessica. “Median Household Income for States: 2007 and 2008 American Community Surveys.” *United States Census Bureau*, 2009, www2.census.gov/library/publications/2009/acs/acsbr08-02.pdf.