Impact of Public Library Resources on K-12 Education in Kansas

Group Members: Bakhbyergyen, Edina, Ina, Paul

December 10, 2023

1 Introduction

The goal of this project was to investigate whether it is possible to detect an impact of public library resources on educational outcomes in Kansas public schools. The project was requested by the *Kansas State Department of Education* (KSDE) and is one of the projects under consideration as a project for students enrolled in a Community Data Labs course for Spring 2024. The projects for the Data Lab course are recruited by the *Kansas Data Science Consortium* (KDSC) at the University of Kansas and other Kansas colleges and universities.

The KSDE would like for us to make use of the annual Public Libraries Survey, which is conducted by the Institute of Museum and Library Services, which is a U.S. Federal Government institute. Other than pointing to this survey, the KSDE gave us no other instructions or suggestions, and we were not given any other data sets. Therefore, there were no constrains on our analysis, other than to use the libraries survey data. We chose two educational outcome in our analysis: 1) the ACT test, which is taken by high school juniors and seniors who are considering attending college; 2) a Reading/Language Arts score (RLA) that is collected annually from states by the U.S. Federal Government, and is taken annually by most students in grades three through eight, as well as at least once during high school. In addition, we included data from the U.S. Census Bureau on poverty rates in each school district, because poverty rates strongly impact educational outcomes. Initially, we looked at data from 2019 (and thereby avoided any impact from the COVID 19 pandemic). We then repeated key parts of the analysis for 2015, so that we had a comparison year to check the consistency of our results.

Our goal was not to generate a model that would help predict test scores, as we do not expect this to be possible from public library resource data. Many factors impact test scores, the most important being family income (Ref. 1). Rather, we focused on on library resources, with the hope that if we are able to detect any impact, perhaps our work might play a role in guiding the allocation of library resources in the future. Our initial plan was to employ primarily correlation and regression analyses. Because our results were weak, we also explored more sophisticated classification methods, after dividing up test outcomes into groups ("classes"). This gave us experience with a wide range of methods.

2 Data

Library Resources Data Set

The key data for our analyses was obtained from the Public Libraries Survey (PLS). The data for this survey has been collected annually by the American Institutes for Research since 1988, and it is processed, analyzed, documented, and published by the Institute of Museum and Library Services. The data sets are available for download from 1998 - 2021 (Ref. 2). Our initial analysis focused on the data from 2019 (Ref. 3). We added the data set from 2015 for secondary analyses (Ref. 4). Although the PLS survey is voluntary, the response rates are very high (for Kansas, the rate for 2019 was 100%, and for 2015 it was 99.1%). There are 329 Kansas libraries in the data sets and it is the whole population of libraries (not including outlets and branches, which were handled in a separate data set that had much less detail).

The 2019 PLS data set contained 159 columns, and the majority of these somehow related to library resources. This included the numbers of different types of staff, expenditures for staff and for different types of library materials, funding amounts from various sources, circulation of different types of library materials, different types of programming, hours of operation, and so forth. We narrowed down the variables to those that logically seemed to have the most promise for detectable impact, and we eliminated variables that had too many missing values, and/or little or no variation among districts. This left 27 variables for the 2019 dataset; two of these were absent in the 2015 data set. The 27 variables and their descriptions are in a Colab notebook file (Appendix A). Most of the variables were normalized by population within the library service area (exceptions were made when this did not make much sense; for example, library service hours).

All data wrangling was done using R in RStudio; a rendering of the extensively annotated R notebook files for 2019 (Appendix B) and for 2015 (Appendix C) are attached as a single compressed folder (.zip file).

Census Data

We obtained poverty rates for school districts, as well as geographic codes to facilitate data set joining, from the U.S. Census Bureau's website (Ref. 5). The data from the Census web site must be generated according to filters that we had to choose. We used the American Community Survey data, which has estimates for 5-year blocks for all U.S. regions (and 1-year estimates from large metro areas). We generated a data set for the 2017-2021 period because this included the years we initially considered analyzing. This data set was used for our 2019 analysis. We generated a second Census data set for the 2011-2015 period, and this was used for our 2015 analysis.

Educational Outcomes Data

We initially used ACT tests scores for our educational outcomes data because this was readily available from the KSDE website (Ref. 6). The KSDE data set has composite scores (that is, a single score for all test areas) for each Kansas school district for the years 2015-2022.

The ACT scores were simple to obtain and analyze, and they have the significant advantage of standardization among states. However, the ACT scores are not a broad indicator of educational outcomes because the test is taken only by a subset of college-bound high school students. Furthermore, our initial analyses with ACT scores generated poor results. Therefore, we searched for better data sets that were readily downloadable and are at the school district level. The U.S. Department of Education makes available data sets called EdFacts, which met our requirements (Ref. 7). EdFacts data are collected annually from states and include tests for Reading/Language Arts (RLA scores). The scores indicate the percentage of students who are at, or above, the level of skill expected at their grade level, according to standards set by each state.

Geographic Reference Files

Combining our various data sets required a fifth source of data, this one also from the Census Bureau, which makes available yearly Geographic Reference Files (Ref. 8), in which various geographic territories are matched within the same file. We needed these file because our educational outcomes data contained solely school districts, whereas our libraries data contained only Zip codes at the library address (the census tracts for service area might be available upon request, because the number of tracts, as well as population within the service area, are included in the libraries data sets). Therefore, we downloaded a data set that contained school districts (names, number, and geo-id) as well as a list of zip codes within each district (between 1 and 27 zip codes per school district).

Summary of Data Wrangling

The American Community Survey data could be readily joined to educational outcomes data because both were at the school district level. This joined file was then joined to the appropriate Geographic Reference File, generating a row for each Zip code (multiple rows per school district). This file was then joined to the appropriate Libraries survey data set (2019 or 2015) by matching Zip codes. All districts matched to at least one library Zip code, and all libraries had at least one district match. The data was then folded so that each district was matched to the average from the librarie(s) to which it had matched, so that we ended up with one row per school district in our final data sets (one for 2019 and one for 2015).

3 Methodology

In the following section, we give an overview of the analyses used. Our results are presented in the final section, 4.

3.1 Correlation analysis

Correlation analysis is a statistical method used to assess the strength and direction of the linear relationship between two continuous variables. Its primary aim is to quantify how changes in one variable correspond to changes in another, and to determine whether there is any association between two variables. The Pearson correlation coefficient, denoted by "r", is a commonly used metric that ranges from -1 to 1. A calculated p-value indicates the confidence in the calculated correlation value.

3.2 Regression Analysis

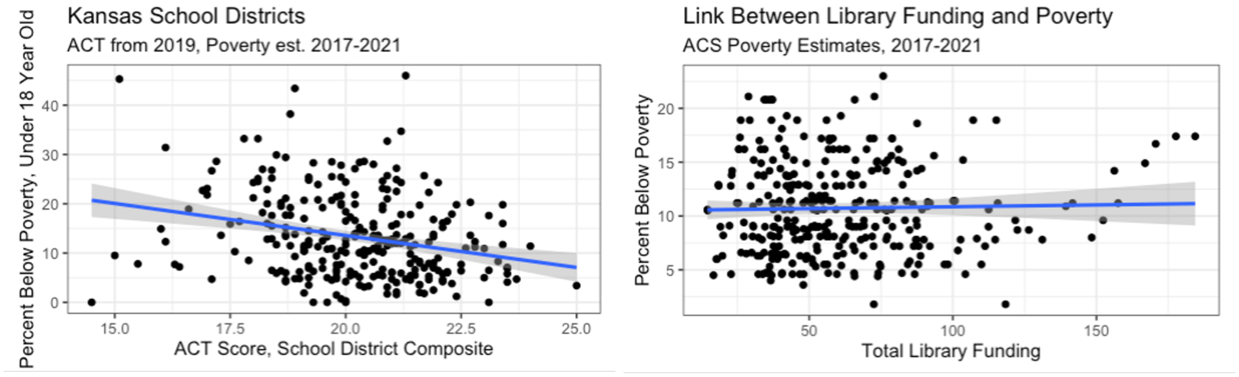
Regression analysis is a powerful tool for modeling and analyzing the relationships between variables. It is particularly useful for predicting the value of a dependent variable based on one or more independent variables. However, for our project, the purpose of the regression analysis was not prediction but rather to investigate the relative impact of library resource(s) on student test scores.

3.3 Classification Methods

In the realm of machine learning, classification methods are crucial for assigning observations to predefined categories or classes. Although our our project does not pose a classification problem, we wished to explore these methods nonetheless, by dividing test score outcomes into groups, or "classes." The methods we used are the following:

- **Random Forest** is an ensemble learning method that builds multiple decision trees during training and merges their predictions for more accurate and robust results.
- **Gradient Boosting Decision Tree** (GBDT) is another ensemble technique that combines the predictive power of multiple weak learners, usually decision trees.
- **Support Vector Machines** (SVM) are a class of supervised learning algorithms used for classification and regression tasks. SVMs find a hyperplane in a high-dimensional space that best separates data points belonging to different classes.

4 Results and Discussion



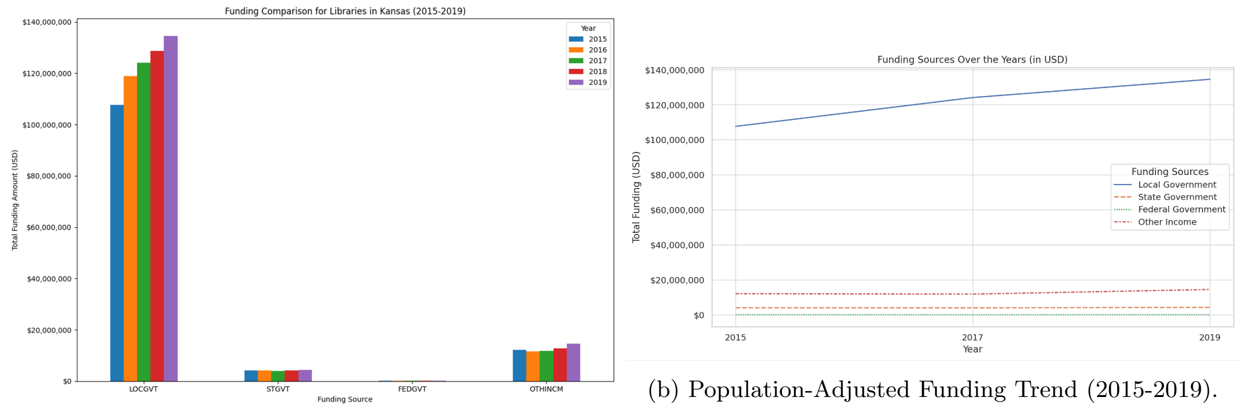
(a) Library Funding and Poverty

(b) ACT Scores and Poverty in Kansas School Districts

Figure 1: Interplay of ACT Scores and Poverty: Scatterplot of Kansas School Districts Reflecting Poverty Estimates with ACT Performance from 2019

While planning our analysis of the library and test score data sets, we were concerned about the impact of family incomes as a possible confounding factor, because income is known to have a strong association with educational outcomes (Ref. 1). Indeed, we found that as expected, test scores are negatively correlated with poverty rates (Figure 1 a; see also similar results for 2015 and RLA scores). This was a concern because we suspected that most library funding is from local sources, which in turn would be impacted by local incomes. Indeed, we found that by far the biggest portion of library funds come form local sources (Figure 2). However, total funding for libraries was not significantly correlated with poverty rates (Figure 1 b), so we can do some of our analyses without accounting for incomes.

Also of potential significance is that we observed a consistent increase in funding across all sources (Figure 2 b). Notably, this upward trend persisted even when we adjusted for population growth, suggesting a genuine escalation in investment levels towards public libraries. This finding is important because it means we will need to eventually analyze more years in order to follow a potential long-term impact of library resources.



(a) Annual Public Library Funding (2015-2019).

(b) Population-Adjusted Funding Trend (2015-2019).

Figure 2: Comparative Analysis of Public Library Funding (2015-2019): Annual Trends and Population-Adjusted Perspectives.

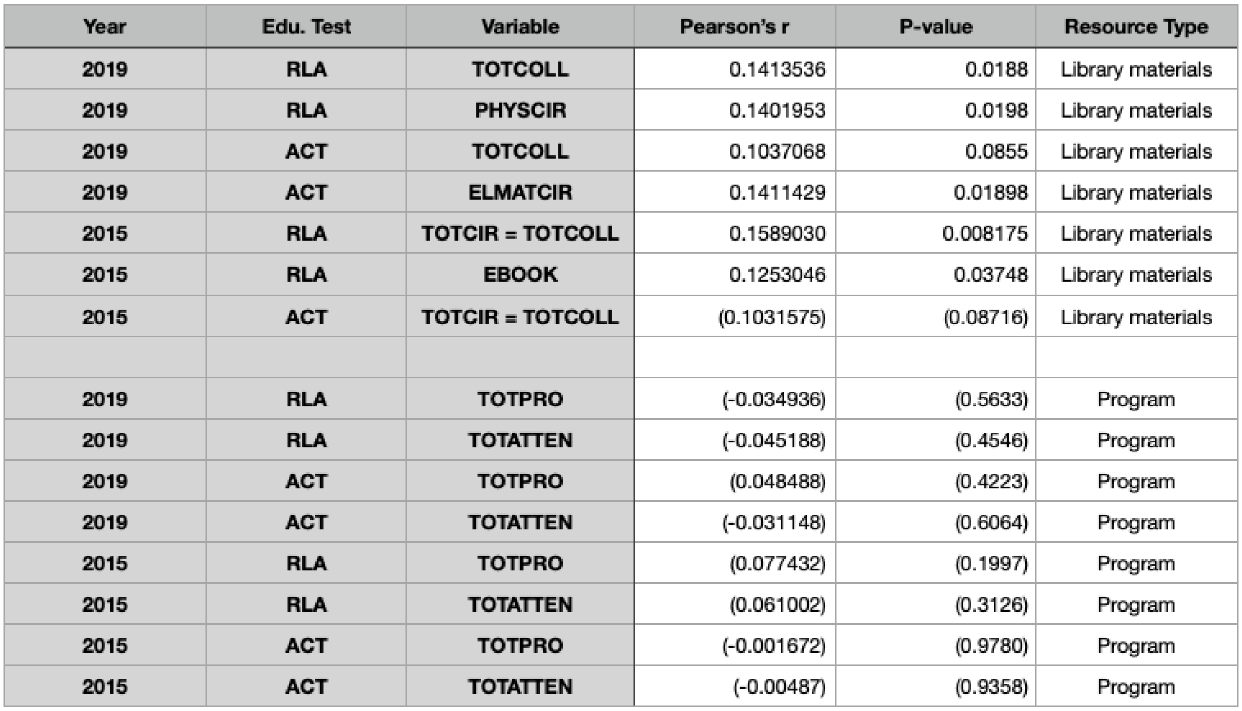
4.1 Correlation analysis

We calculated Pearson's Correlation Coefficients using the cor.test function in R, which also generated a 95% confidence interval for the coefficient, and a p-value (see Appendix B and C which are code files for 2019 and 2015). We performed this analysis for all of our selected library resource variables, analyzing correlation with both ACT and RLA student scores and for both 2019 and 2015. Our highest-confident correlations are shown in our Colab file (Appendix A) and in Table 1 (values in parentheses had p-value above a 0.05 confidence cutoff). Although all correlations were weak, several variables that related to circulating library materials yielded the highest-confidence correlations, in particular TOTCOLL for 2019 (total circulating materials, called TOTCIR in in the 2015 data set). Electronic materials (ELMAT), ebooks (EBOOK), and physical materials (PHYCIR) yielded the highest-confidence correlations. This pattern of library materials yielding the best results held for both ACT and RLA scores, and both 2019 and 2015. In contrast, variables that represented the number of programs (activities) and attendance consistently had no significant correlation.

4.2 Regression Analysis

We used regression analysis to investigate the relative impact of library resources on student test scores. We focused primarily on the variable with the highest-confident correlation with test scores, which is total circulating library materials (TOTCIR in 2015, renamed TOTCOLL in 2019, likely due to slightly different makeup of the "total" materials considered). Furthermore, we focused primarily on RLA scores, which yielded more promising correlation results than ACT scores. We did this by comparing R-Squared, Adjusted R-Squared, AIC, and BIC calculations when testing library variables in regression models. Adjusted R-

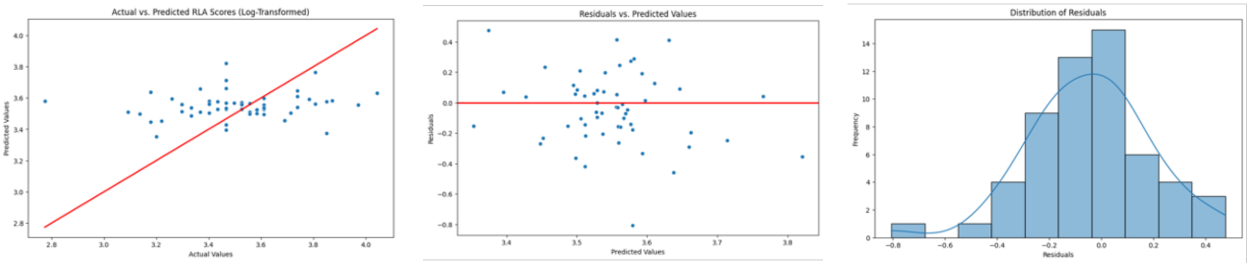
Table 1: Correlation table



Year	Edu. Test	Variable	Pearson's r	P-value	Resource Type
2019	RLA	TOTCOLL	0.1413536	0.0188	Library materials
2019	RLA	PHYSCIR	0.1401953	0.0198	Library materials
2019	ACT	TOTCOLL	0.1037068	0.0855	Library materials
2019	ACT	ELMATCIR	0.1411429	0.01898	Library materials
2015	RLA	TOTCIR = TOTCOLL	0.1589030	0.008175	Library materials
2015	RLA	EBOOK	0.1253046	0.03748	Library materials
2015	ACT	TOTCIR = TOTCOLL	(0.1031575)	(0.08716)	Library materials
2019	RLA	TOTPRO	(-0.034936)	(0.5633)	Program
2019	RLA	TOTATTEN	(-0.045188)	(0.4546)	Program
2019	ACT	TOTPRO	(0.048488)	(0.4223)	Program
2019	ACT	TOTATTEN	(-0.031148)	(0.6064)	Program
2015	RLA	TOTPRO	(0.077432)	(0.1997)	Program
2015	RLA	TOTATTEN	(0.061002)	(0.3126)	Program
2015	ACT	TOTPRO	(-0.001672)	(0.9780)	Program
2015	ACT	TOTATTEN	(-0.00487)	(0.9358)	Program

Squared accounts for the increase in the number of independent variables between models (which is expected to increase ”un-adjusted” R-Squared). AIC is Akaike information criterion, and BIC is Bayesian information criterion; these are additional metrics that assist in comparing regression models (Ref. 9). Detailed results are in our Colab file (Appendix A); the results are briefly summarized here.

We examined the impact of total circulating materials alone or in combination with other variables; we tested both first order and second order regression as well as logistic transformation. As indicated by our low R-Squared and Adjusted R-Squared values, even our most confidently-correlating variables contribute only a very small component of the many factors that impact test scores. Among our best results was a first-order multi-variable model that included total circulating materials and ebooks, for 2015 RLA test results (R-squared 0.039, Adj. R-Squared 0.032). In contrast, a model with poverty rates for 18-and-under yielded an adjusted R-Squared of 0.157, sonsisent with the known impact of poverty on test scores. When added to poverty rates in a multivariable model, total circulating materials improved the adjusted R-Squared to 0.171.



(a) Actual vs. Predicted RLA Scores (b) Residuals vs. Predicted Values (c) Histogram of Residuals

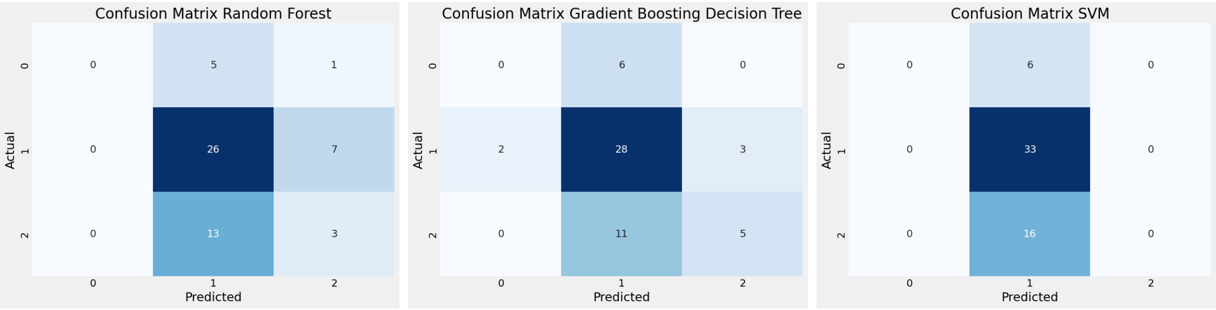
Figure 3: Assessment of Regression Model Performance through Residual Analysis and Predictive Accuracy

Table 2: Accuracy results for different classification methods for ACT and RLA scores

Classifier	**Accuracy for ACT scores**	**Accuracy for RLA scores**
Random Forest	0.53	0.53
GB Decision Tree	0.6	0.47
SVM	0.6	0.6

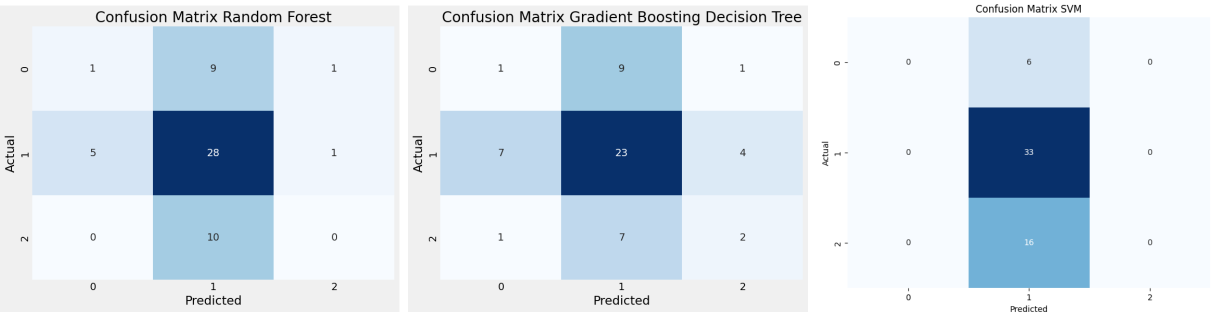
4.3 Classification Methods

Attempts were made to classify ACT scores and RLA scores using hyperparameter-tuned Random Forest, Gradient Boosting Decision Tree, and Support Vector Machines (SVM). The scores were categorized into three classes for this: Low, mediocre and high scores. However, the outcomes were discouraging, as seen in Table 2 and Figure 4: The accuracy showed an almost random behavior, or even slightly worse than random behavior, of the classifiers, which leads to the rejection of the hypothesis that scores can be predicted by library values using classification methods.



(a) Random Forest (b) GB Decision Tree (c) SVM

Figure 4: Confusion matrices for different classification methods for ACT scores



(a) Random Forest (b) GB Decision Tree (c) SVM

Figure 5: Confusion matrices for different classification methods for RLA scores

5 Conclusion

Our key conclusion from our investigation is that it is worthwhile to pursue further analysis, primarily by including data from more years. In addition, one could compare results from Kansas to a similar analysis for other states, and also include more types of student assessments. Although most of the library resource variables had either weak or no correlation with student test scores, our results followed a pattern that was consistent for the two years we analyzed. Furthermore, the results were consistent for two very different types of student tests. The pattern we saw also makes logical sense: any impact of library resources would accumulate over a number of years. The quantity of library materials (and the variable of total circulation) should be fairly steady over time, since it takes years to build up collections. Therefore a survey of library materials from 2019 would reflect also the years that lead up to the 2019 (or 2015) student tests. In contrast, the programming/activities available for 2019 is not expected to have any impact on test scores for that particular year, and programming is expected to be more variable. This does not mean that programming is not significant, but that it would need to be analyzed over a long period of time.

It is expected that library resources are only a small component of many factors that impact educational outcomes — and not all important outcomes are readily measured with standardized tests. Our results suggest that we may be able to analyze in more detail which library resources have the most positive impact on student achievement and perhaps other aspects of life, for young and old Kansans alike.

5.1 Division of Labor

Table 3: Group members' contribution to the project

Name	**Contribution**
Ina	Classification, Regression, Report writing
Bakhbyergyen	Funding trends, Regression, Report writing
Edina	Data search, Data wrangling, Correlation analysis, Regression, Report writing
Paul	Report writing

References

[1] Miller, C. C., and F. Paris. The Upshot: New SAT Data Highlights the Deep Inequality at the Heart of American Education. The New York Times, Oct. 23, 2023. https://www.nytimes.com/interactive/2023/10/23/upshot/sat-inequality.html

[2] Libraries Survey data: https://www.imls.gov/research-evaluation/data-collection/public-libraries-survey

[3] Pelczar, M., Frehill, L. M., Nielsen, E., Kaiser, A.& amp; Li, J. (2021). Data File Documentation: Public Libraries in the United States Fiscal Year 2019. Institute of Museum and Library Services: Washington, D.C.

[4] The Institute of Museum and Library Services. 2017. Public Libraries in the United States Fiscal Year 2015. Washington, DC: The Institute.

[5] Census data (select American Community Surveys, 5-year estimates and year, the state, school district-level, and poverty rates) https://data.census.gov/table?g=040XX00US20&y=2021

[6] ACT scores from KSDE: https://ksreportcard.ksde.org/act_scores.aspx?org_no=State&rptType=3

[7] U.S. Department of Education EdFacts data: https://www2.ed.gov/about/inits/ed/edfacts/data-files/index.html

[8] Geographic Reference Files: https://www.census.gov/geographies/reference-files/2010/geo/relationship-files.html#par_textimage_19960473

[9] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). *An Introduction to Statistical Learning* (2nd ed.). Springer. https://www.statlearning.com/