Kathy Trieu

Texas A&M University
Department of Statistics
STAT 692 Statistical Consulting
November 30, 2023

Research Question

How do the characteristics of libraries in California contribute to variations in visitor attendance?

Table of Contents

Background

Libraries are vital pillars in our democratic society, as sociologist and New York University professor Eric Klinenberg emphasized in "Palaces for the People: How social infrastructure can help fight inequality, polarization, and decline of civic life." Klinenberg contends that public infrastructures, including libraries, play a pivotal role in fostering thriving communities, particularly when it comes to economically challenged neighborhoods. These social infrastructures serve as spaces that draw individuals out of their homes, promoting connections and mutual support during challenging times. Modern public libraries have evolved beyond mere repositories of books, now offering diverse services such as tool rentals, internet access, resume assistance, language learning, and even relief from extreme weather conditions. Despite this expanded utility,

libraries have experienced a decline in visitor numbers, possibly influenced by shifts in public interest or increased online resources.

The relevance of this decline is underscored by Klinenberg's assertion that community vitality thrives on social interaction, a necessity in an era marked by growing polarization and increased reliance on the internet. Librarians, faced with new challenges such as book banning and ideological policing, found their institutions further tested by the COVID-19 pandemic in 2020. While some libraries temporarily closed, others adapted by offering online book rentals and contactless services. As quarantine measures eased in 2021, library visits began to rebound, yet they have not fully recovered to pre-pandemic levels. This raises a pressing question: will libraries endure in the face of evolving challenges that strain the symbiotic relationship between public infrastructure and a healthy democracy?

Examining library metrics empowers librarians and management to optimize their resource allocation, enhancing the frequency of community members' visits to the library.

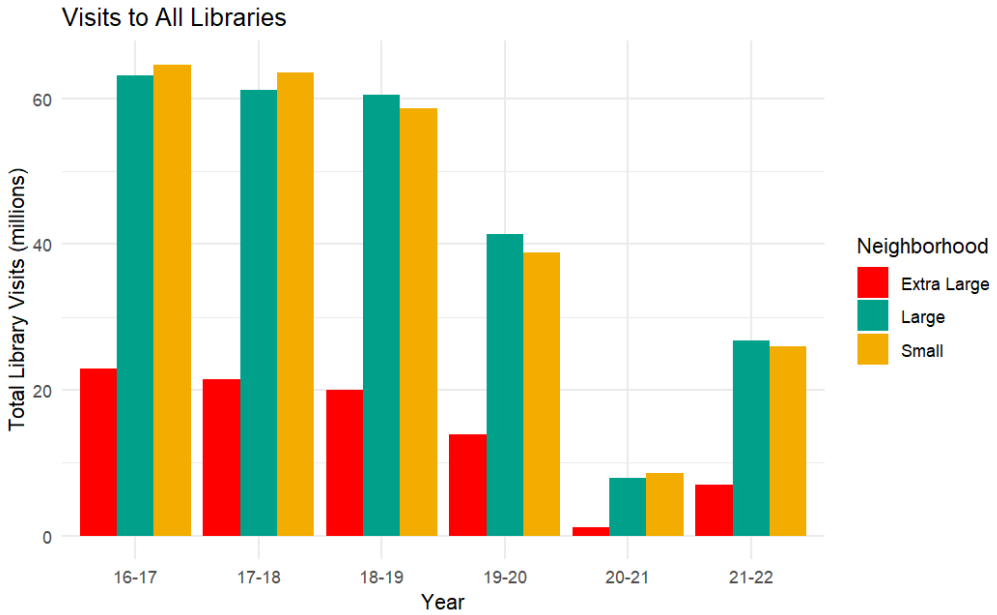

Figure 1. Total visits to all libraries in California from 2016 to 2022 separated by size of the population served by the library

Data Description

Recognizing that libraries cater to diverse populations with varying needs, our analysis focused solely on California libraries and considered the nuanced characteristics of

libraries based on the size of the communities they serve. Categorizing libraries into small, large, and extra-large based on population size revealed significant differences in their attributes. Small libraries predominantly serve populations of under 300,000, while large libraries cater to communities ranging from 300,000 to 2 million. Extra-large libraries, encompassing systems serving over 3 million people, present a distinct subset.

Although the database aggregates counts for library systems, precluding an analysis of individual libraries within a system, these systems were treated as individual libraries because community members with access to such systems can visit any library within the network. This interconnectedness is exemplified in regions like Los Angeles County, where patrons seamlessly borrow and return books across multiple library locations within the system. However, this report uses data from libraries categorized as “Small” in the model building and subsequent analysis to improve analysis results. Some data visualizations in the descriptive analysis section show information from all neighborhoods to illustrate their differences and instantiate the need to treat them separately.

The data utilized in this project was sourced from the state of California and was publicly accessible online. This analysis constitutes an observational study, focusing on correlation and prediction rather than establishing causation, which remains elusive. Each year's dataset provided information on various aspects, including library visits, computer usage, materials available, program numbers, expenditure, income, and more. While financial data and a few binary variables were part of the dataset, most values represented counts. A minor challenge arose during the data compilation due to variations in data collection and presentation from year to year. Despite certain variables capturing consistent information across years, discrepancies in variable names existed. Over time, new variables were introduced, holding potential for future analyses, but were excluded from the current study due to limited availability.

Several notable weaknesses are inherent in this analysis. Firstly, although libraries receive detailed instructions on measuring variables, they are not mandated to furnish actual counts; they can take counts for a single, self-determined "regular" week, excluding major holidays. These counts are then extrapolated to generate an annual estimate. While datasets starting with the 2019-20 dataset include a variable indicating whether the provided information is an actual count or an estimate, earlier datasets lack this distinction, making it impossible to discern which counts are precise or estimated. Despite this, the analysis proceeded without differentiation. Another weakness lies in the absence of census data detailing the demographics of the populations served by the library. Consequently, the model does not account for variations in libraries serving

populations with differing income or education levels or those catering to racially homogeneous versus diverse populations. Including such information could significantly enhance the model, addressing potential confounding factors. Therefore, interpreting the results from this study necessitates an awareness that the modeled relationships are not exhaustive of all relevant information and do not imply causality.

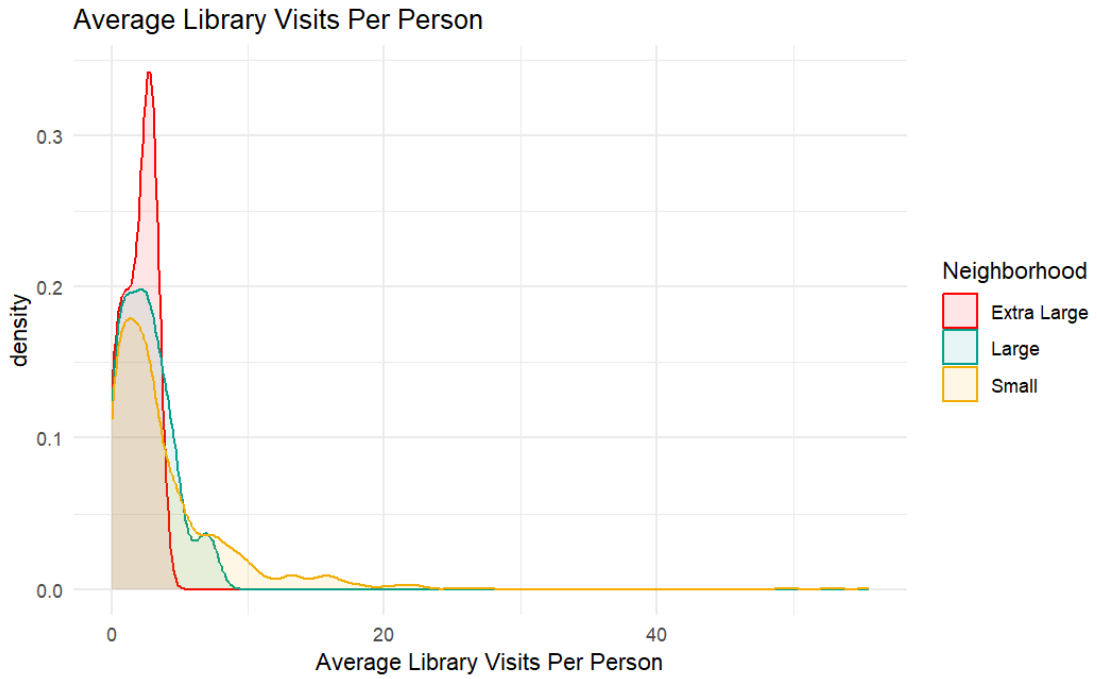

Figure 2. Average library visits per person are calculated by dividing the total number of library visits by the number of people in the population served by the library, separated by the size of the library's neighborhood.

The initial stages of data management involved handling data in segments, necessitating the removal of entire features due to sparse information and the renaming and combining of features. Subsequently, observations missing recorded supervisors (Library Visits) were eliminated. Evaluating the data type of each feature revealed they were all character strings, prompting the need to cleanse numeric data of symbols like commas and dollar signs. Missing values in each feature were then imputed with the feature median. The subsequent assessment of skewness levels for each feature unveiled a consistent trend of high right skewness, a characteristic attributed to a substantial proportion of true zero values.

Skewness was reduced only slightly when excluding libraries categorized as "extra large" and "large" from the dataset. Subsequently, the remaining extreme values underwent a thorough examination for accuracy. It became evident that certain values had been erroneously entered into the database, some displaying discrepancies of up to

100 times larger than values recorded at the same library in the preceding and subsequent years. Additionally, some libraries consistently reported identical numbers year after year, an improbable scenario. Consequently, data from these libraries was deemed unreliable and omitted from the dataset.

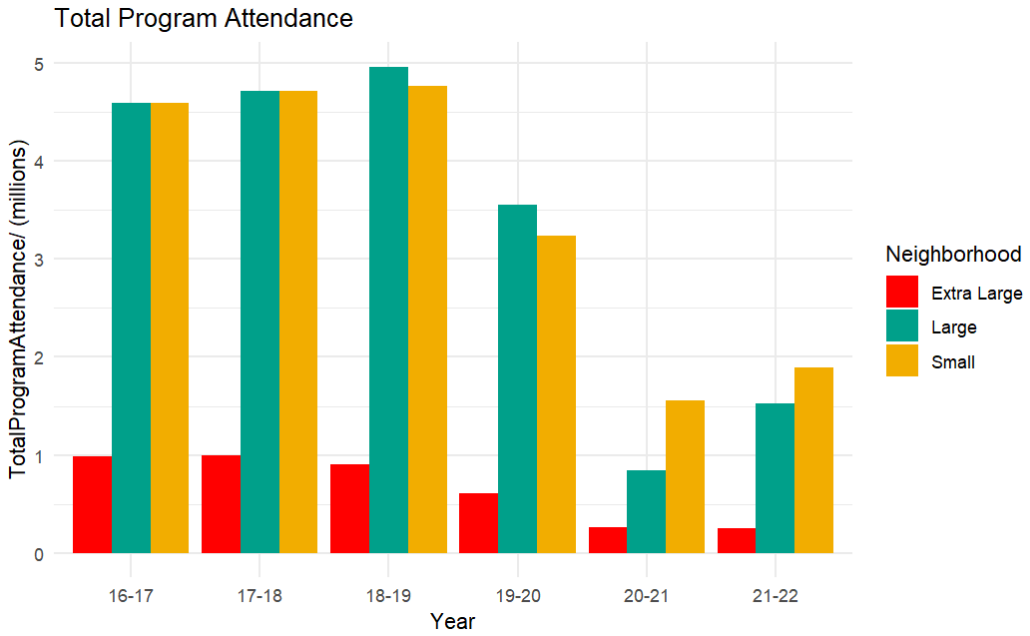

Figure 3. Total rate of program attendance from 2016 to 2022, separated by size of the population served by the library

Descriptive Analytics

As depicted in Figure 1, the total annual library visits have experienced a consistent decline each year from 2016 to 2020. The notable drop in 2020 can be attributed to quarantine restrictions imposed during the COVID-19 pandemic. While visitor numbers began to rise after 2020, the rebound did not reach the levels observed in previous years. Although data from 2023 is not yet available online, it would be intriguing to analyze it once published. Considering the diverse population sizes served by each library, library visits per person were computed using population data. Figure 2 illustrates the variation in average library visits per person across different library sizes. Despite their differences, the distributions exhibit a similar pattern, with a median average visit of around 3 per year. Notably, the distributions are positively skewed, indicating that more libraries have fewer than 5 visits per person. It is important to acknowledge that the distribution for extra-large libraries, with only 12 data points, appears less smooth than small libraries, which encompass over 700 libraries.

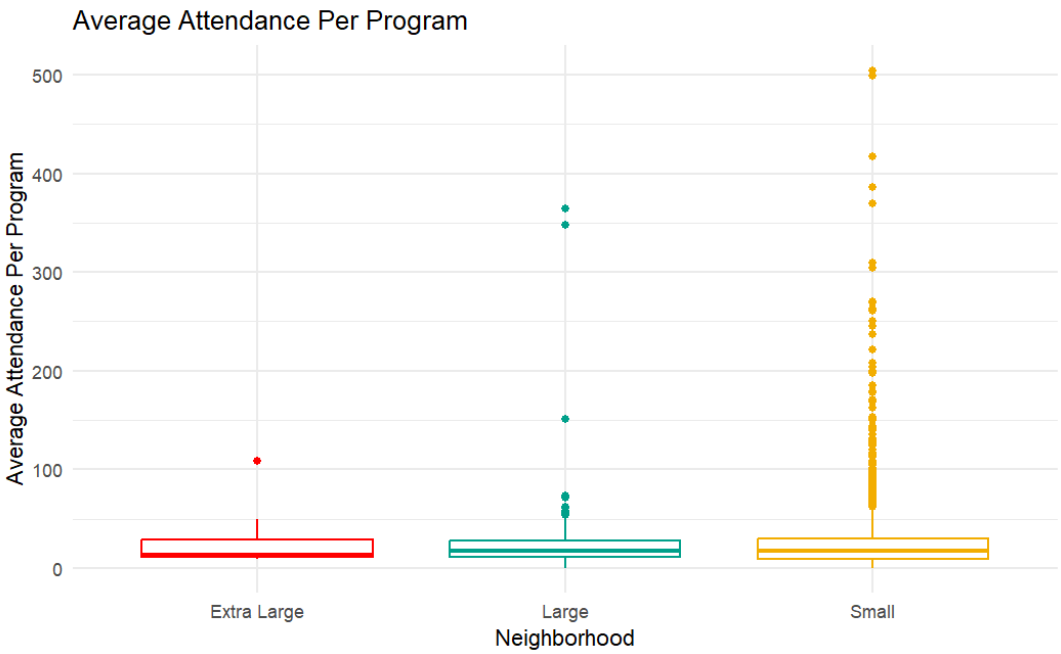

Figure 4. Average program attendance per available program per person grouped by neighborhood size

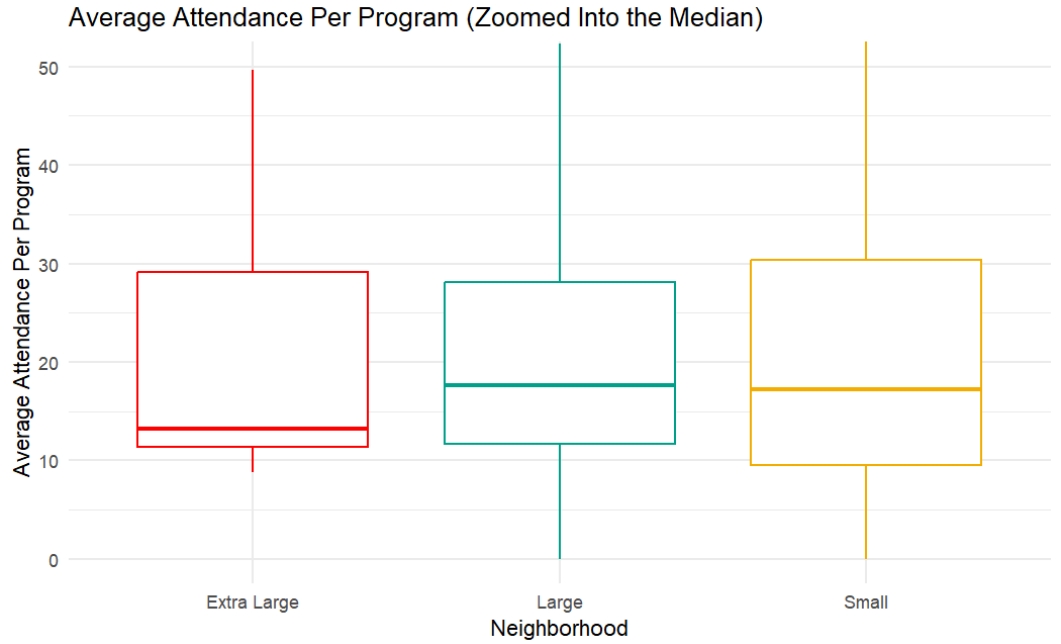

Figure 5. Average program attendance per available program per person grouped by neighborhood size and zoomed into the median

Another noteworthy aspect for libraries to explore is the average program attendance per person, depicted in Figure 3. Computed by dividing the number of program attendance by the population served by the library, this metric does not account for the number of programs available. Therefore, for example, this number cannot capture

whether libraries are hosting many programs with very little attendance or few programs with high attendance. Surprisingly, from 2016 to 2020, the average program attendance per person increased for small and large libraries, despite decreased library visits during the same period. Small libraries demonstrated a better ability to rebound than large ones, possibly attributed to their more manageable compliance with distance requirements and safety measures. However, the lack of detailed information necessitates further research to validate these observations.

Figures 4 and 5 showcase the program attendance rate per available program per person, calculated by standardizing the total program attendance count by the number of available programs and dividing that by the population size. Figure 4 reveals a heavily positively skewed rate for small libraries, exhibiting large variance compared to others, partly due to the vast number of small libraries (over 700) as opposed to only 12 extra-large libraries. Figure 5 zooms in on the median to facilitate comparison, highlighting that extra-large libraries exhibit the lowest median attendance rate per program per person when standardized. This could be attributed to fewer data points being available or perhaps that these libraries produced more programs for a larger audience, resulting in lower attendance per program. A comprehensive investigation with more metrics than available is required to confirm this hypothesis.

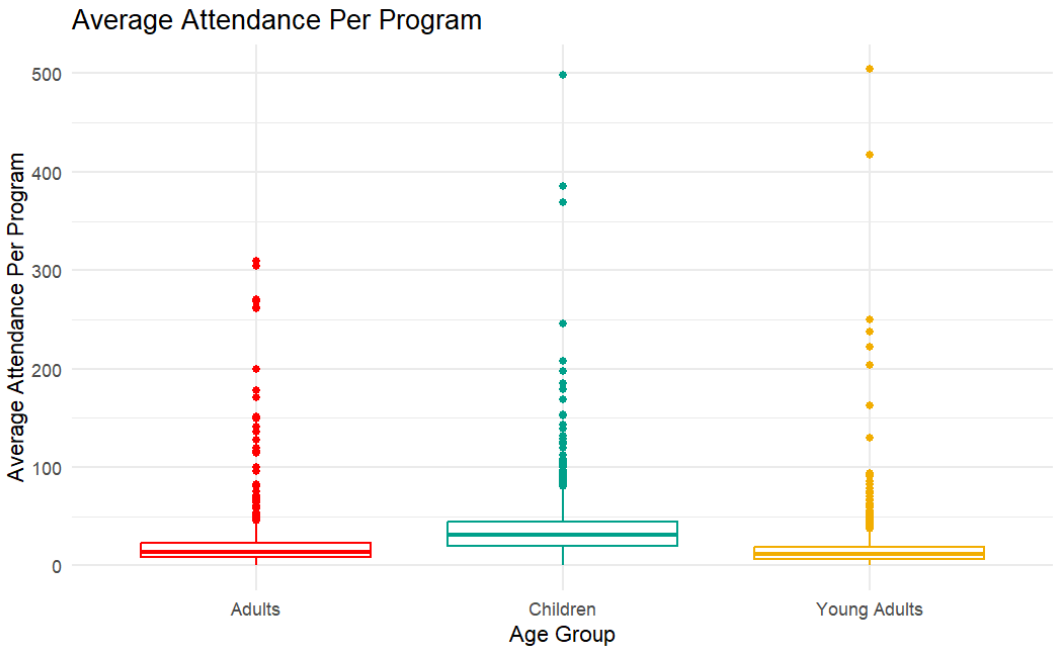

Figure 6. Program attendance per program divided by age group.

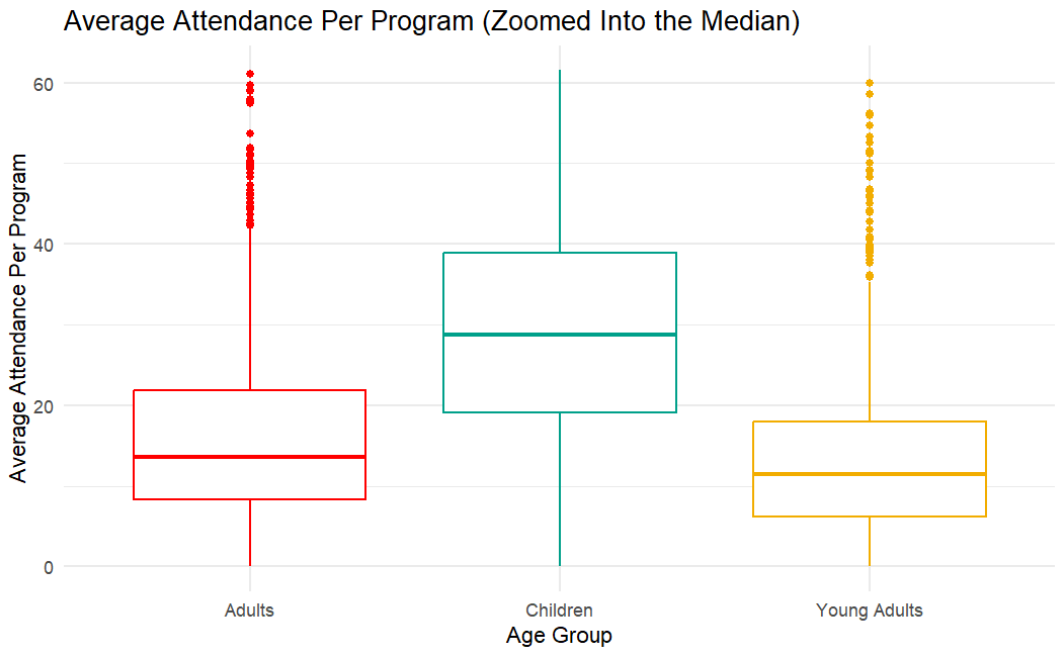

Figure 7. Program attendance per program divided by age group zoomed in to better show the median.

Since neighborhood sizes led to greater variations and skew, large and extra-large libraries were excluded from further analysis. Figures 6 and 7 show which programs attracted more attendance per program per person among small libraries. Of the three age groups: adults, young adults, and children, the programs designed for children had a higher average attendance rate per program than the other groups. This might be because children need to be escorted by adults.

Model Building

Support Vector Machine Regression Model and Random Forest Model

The pronounced skewness in the distribution of both features and the target variable posed considerable challenges in selecting an appropriate model for this analysis. The primary focus was to discern the variables that significantly influenced the number of library visits and uncover any intriguing relationships between them. Consequently, choosing an appropriate model was crucial, aiming for one that could elucidate relationships rather than being a mere black box providing accurate predictions. While a multivariate linear regression model would have been optimal for interpretability, it was deemed unsuitable due to the skewed distribution of the data, which violates the assumption of normality inherent in regression models.

Two alternative models were considered: a support vector machine regression (SVMR) model and a random forest model. The SVMR model constructs hyperplanes that best

capture data trends, using a loss function that limits overall complexity and the impact of extreme values. This model is capable of illustrating relationships between variables and library visits. On the other hand, the random forest model, composed of a forest or group of decision trees, offers slightly superior prediction accuracy, detailed insight into the most influential variables, and the ability to capture more complex relationships. Decision trees, forming the basis of the random forest, recursively split the data into subsets based on selected features and thresholds, minimizing the variance of the target variable within each subset. The random forest aggregates results from each decision tree, highlighting features frequently deemed crucial.

To maximize the effectiveness of the analysis, results from both the SVMR and random forest models are included in this study. The SVMR model, with a significantly lower average error, is preferable for prediction. The SVMR feature importance graph aids in understanding which features contribute to the model's trend, while the random forest feature importance graph identifies key features explaining outcome differences. Despite differences in how feature importance is determined, there is an overlap between the most important features in both models, reinforcing their significance in influencing library visits.

Meticulous attention was given to ensuring accuracy while building both models. Hyperparameters play a pivotal role in model development, and their optimal values were selected through the cross-validation method. This involved computing the models with various hyperparameter values on subsets of the dataset. Following hyperparameter selection, the final models underwent testing on previously unseen data, a crucial step to preserve the integrity of performance metrics. Testing on data used during training could compromise the models' evaluation. Additionally, data were manipulated to satisfy the necessary assumptions for each model. For SVM, this involved scaling the data and removing highly correlated features. In contrast, random forest models have fewer assumptions; aside from excluding highly correlated and substantially missing features, no further feature removal was necessary. Both models demonstrate proficiency in handling high-dimensionality and numerous features. Random forests' natural ability to select the most important features mitigates the impact of less significant ones, contributing to their robust performance.

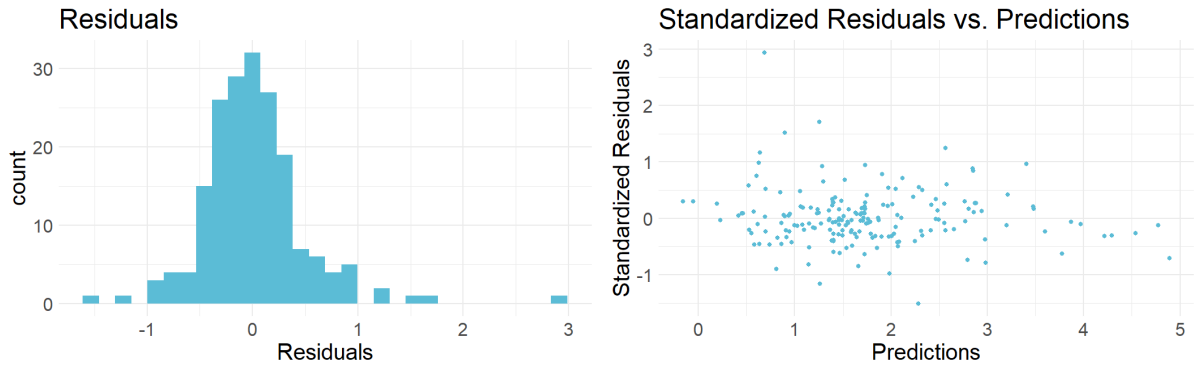

Figure 8. Residual plots from the support vector machine regression model

Following the testing of the models on previously unseen data, residuals were calculated, representing the discrepancies between actual and predicted values. In the context of SVMR models, assessing residuals is crucial to identify signs of heteroscedasticity, a problem that indicates inconsistent variance that changes as values increase or decrease. As demonstrated in Figure 8, the residuals exhibit a normal distribution, and upon examination, no discernible patterns emerge between the standardized residuals and the predicted values. This absence of patterns signifies the absence of heteroscedasticity.

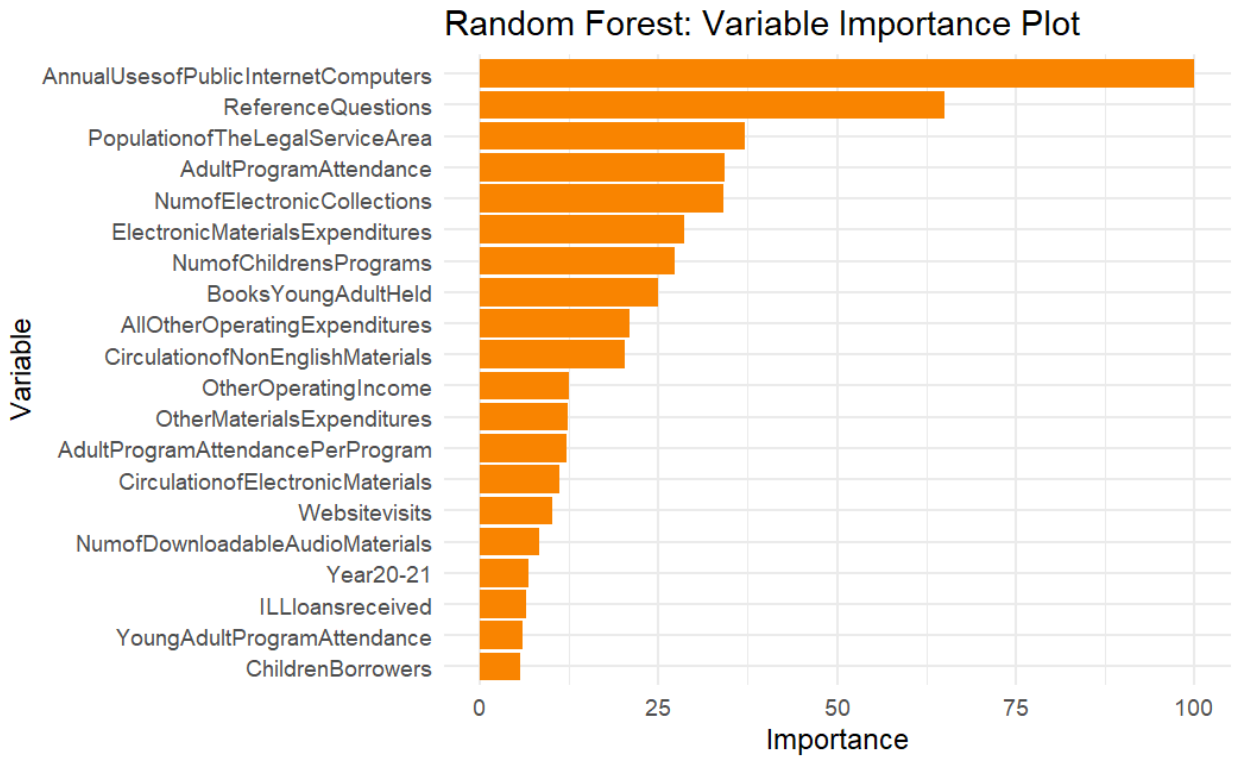

Figure 6. Top 20 most important variables in the random forest model

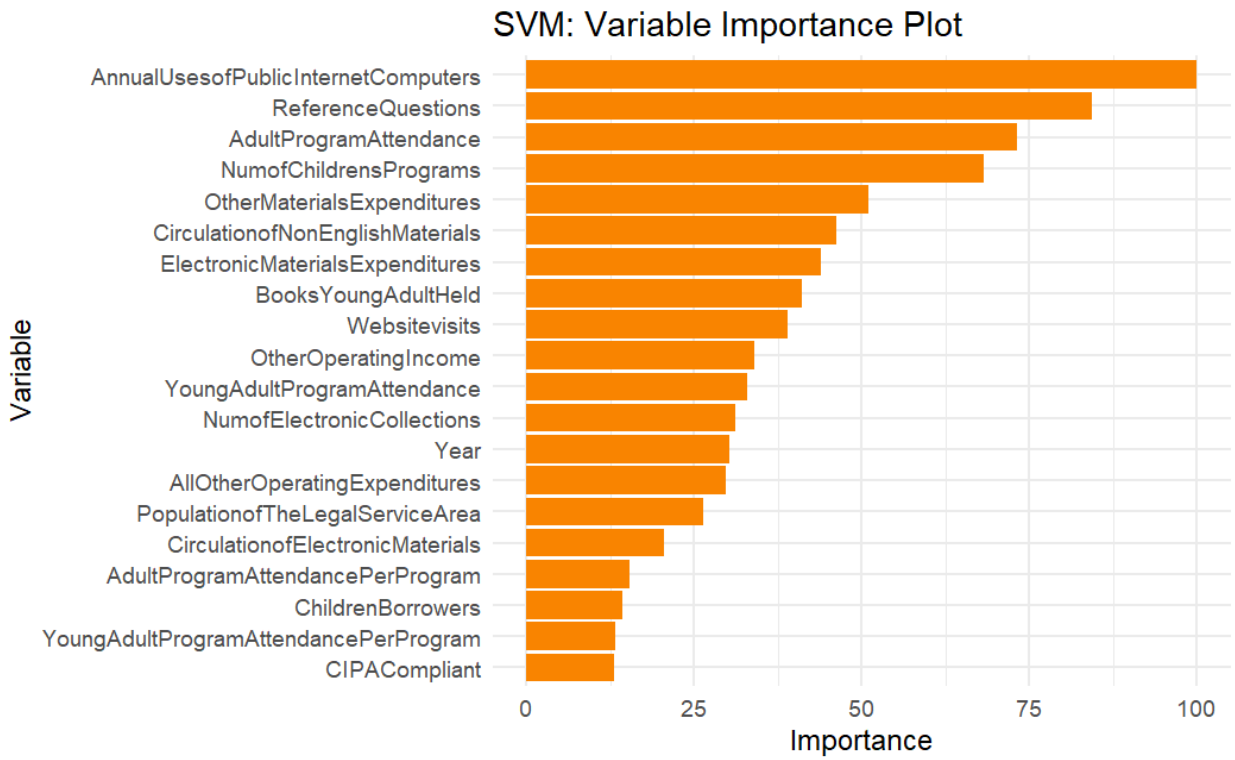

Figure 7. Top 20 most important variables in the support vector machine regression model

Model Analysis

The models indicate that certain variables exert a greater influence than others, signifying that the information within these variables has a more pronounced impact. While they don't precisely elucidate the relationship between these variables and the number of library visits, the models underscore that adjustments can lead to significant fluctuations in visitation numbers. Some of the variables identified as most crucial by both the support vector machine regression model and random forest model are intuitive, such as the rate of computer usage, which implies a connection to library visits. However, this relationship is not strictly linear, as a single library visitor can account for multiple computer uses or choose not to use the computer at all. Notably, as with all variables, the computer usage rate was calculated by dividing the counts of computer usage by the population served by the library. A higher rate for one library suggests more computer use per person in their community than another. Given the importance of computer usage rate in our model, its exclusion would considerably weaken predictions. A partial dependence plot not shown here was generated to delve deeper into the relationship, providing insights into how computer usage affects library visits when holding other variables constant. The plot revealed a generally positive relationship. However, it's essential to note that the nonlinear nature of our model

implies that the relationship may not be entirely accurate and may not capture more intricate interactions with other variables.

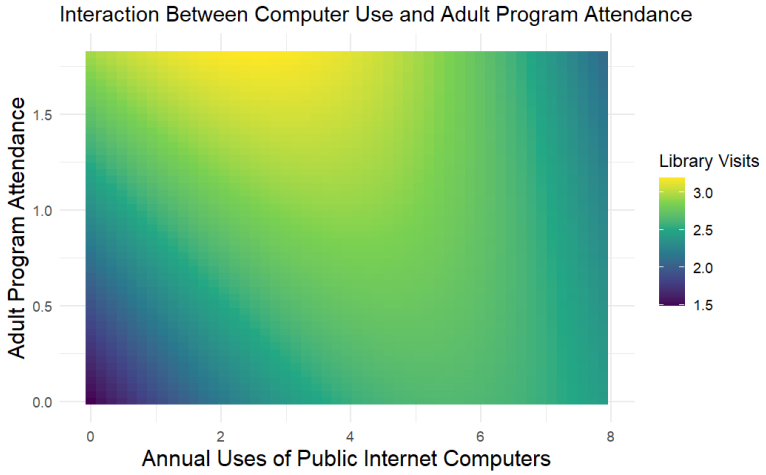

Figure 8. Interaction between computer use and adult program attendance.

Another less apparent yet impactful variable in both models is the circulation of non-English materials. Circulation, defined as the loans of materials, including renewals, significantly influences the prediction of library visits. Similar to computer usage, this variable was computed by dividing circulation counts by the population served by the library. Some populations circulate non-English materials more than others, contributing to the model's predictive accuracy. A partial independence plot for this variable revealed a generally positive relationship between the circulation of non-English materials and library visits.

While the average attendance per program is higher for children's programs compared to adult and young adult programs, both models identified adult program attendance as slightly more important than the number of children's programs. While this doesn't provide an exact relationship between adult attendance and library visits, it introduces a consideration for libraries that may have been overlooked–adult programming changes can influence library visits. Therefore, it might be worthwhile for libraries to experiment locally with adjustments to adult programming. As expected, children's programming also emerged as a significant feature in the model.

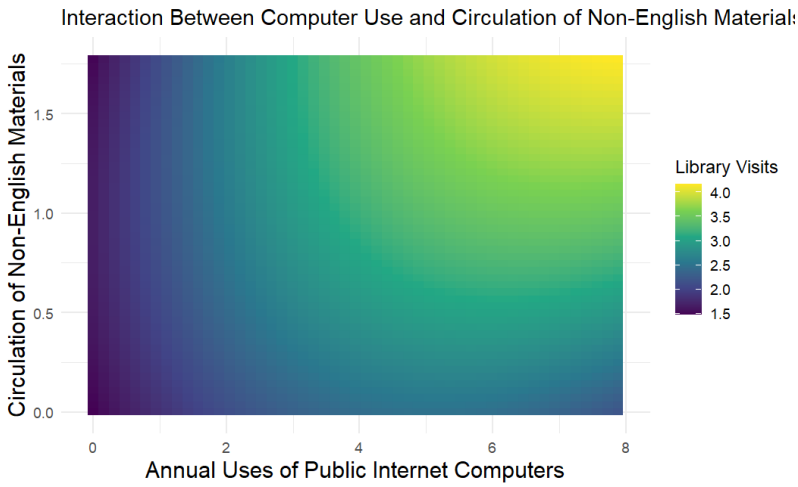

Figure 9. Interaction between computer use and circulation of non-English material. Yellow depicts higher rates of library visits. This plot shows that both variables must increase in order for library visits to increase.

When the presence of another variable influences the impact of one feature on the target variable, it is termed an interaction, a crucial aspect to investigate due to its real-world significance. Interactions cannot be explicitly examined despite being an inherent part of the random forest model. In random forest models, interactions are naturally captured through the amalgamation of decision trees. Each tree employs a random subset of variables and is constructed independently, resulting in some trees excluding certain variables while others include them. The comparison between these trees can unveil insights into interactions and other intricate relationships.

Table 1. Model Performance Metrics

Metric	**SVMR**	**Random Forest**
Mean Squared Error (MSE)	0.20	3.22
Root Mean Squared Error (RMSE)	0.44	1.79
Mean Absolute Error	**0.30**	**1.18**
R-Squared	**0.79**	**0.82**

Conversely, support vector machine regression (SVMR) models do not inherently capture interactions unless certain kernels are used. In constructing SVMR models, researchers must choose a kernel type. A polynomial kernel was selected for this report which implies that the model considered interaction terms of the second degree, involving interactions between pairs of variables. Unfortunately, extracting these

interactions from the model was hindered by limitations in the software package's architecture. However, some interaction plots of important features were generated to determine which interactions exist and examine the nature of their relationship. Figure 8 shows the Interaction between computer use and adult program attendance. This interaction is oddly shaped. Holding computer use constant, increases in adult program attendance will still increase the target variable value. However, the same is not true if adult program attendance is constant and computer use increases. Figure 9 shows how the circulation of non-English materials and computer use positively interact. This means that as each increases in value, their predicted value increases. Additionally, a higher value of library visits necessitates that both the circulation of non-English materials and computer use are high.

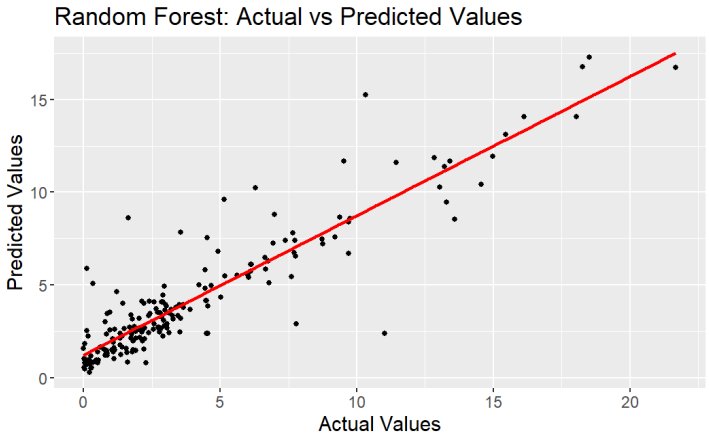

Figure 10. Predicted values compared to actual values in the random forest model

Both models exhibit moderate predictive accuracy, with sufficiently low errors between predictions and actual values, providing reasonably accurate forecasts of library visits. In Figures 10 and 11, the relationship between actual and predicted values is depicted for both models. Data points closer to the red line are more accurate. Paying close attention to the scales of both graphs reveals that the Support Vector Machine Regression model performed significantly better in terms of prediction.

In Table 1, performance metrics are provided which support that the support vector machine regression (SVMR) model outperformed the random forest model in terms of prediction quality. While all three values, the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), are lower for SVMR than random forest, the Mean Absolute Error is a suitable metric for this scenario due to its insensitivity to extreme values. This underscores the superior performance of the SVMR model with an MAE of 0.3, compared to the random forest model's MAE of 1.18.

Despite the SVMR's stronger predictive capabilities, the random forest model captures more variation, as indicated by its higher R-squared value of 0.82 (82%) versus 0.79 (79%) in the SVMR model. This suggests that the random forest model provides a more comprehensive understanding of how the variables relate to changes in library visits. An intriguing observation is the distinct ranking of one variable, the population of the legal service area, in both models. While ranked as the third most important feature in the random forest model, it drops to the 15th position in the SVMR model. This discrepancy implies that the relationship between population and library visits is too complex to be effectively captured by the SVMR model. This speaks to the random forest model's ability to capture more variance.

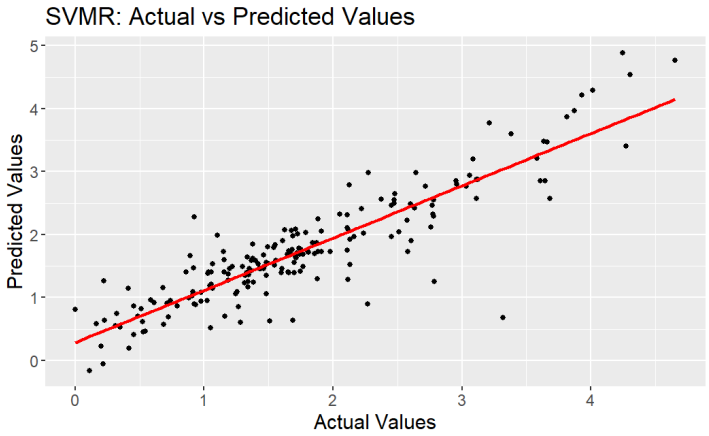

Figure 11. Predicted values compared to actual values in the support vector machine regression model

Predicting library visits based on specific predictors, such as computer usage, can offer valuable insights to librarians seeking to assess the potential impact of their current performance metrics on annual visit counts. For instance, leveraging metrics collected throughout the year allows librarians to extrapolate for the entire year and input this into the model for a year-end estimate. Librarians can also experiment with metric adjustments to obtain predictions, aiding decision-making regarding resource allocation for the populations they serve. This unconventional approach is particularly applicable in this context, where precision requirements are less stringent than in fields like cancer research or engineering. The SVMR model serves as a simple prediction generator, providing librarians with a tool to assess their performance throughout the year.

Conclusion

The significance of libraries as essential social infrastructures, underscored by sociologist Eric Klinenberg, emphasizes the imperative to comprehend the intricacies influencing their visitor attendance. This research extensively explored the attributes of

libraries catering to populations under 300,000, revealing their substantial impact on variations in library visits. The findings from this study offer valuable insights for librarians and management teams, enabling them to elevate community engagement and optimize resource allocation effectively.

The Support Vector Machine Regression (SVMR) model demonstrated better accuracy in predicting library visits, while the Random Forest model captured a broader spectrum of variation, highlighting the nuanced relationship between variables and library visits. Librarians can harness the SVMR model as a predictive tool, facilitating experimentation with adjustments and empowering data-driven decisions that deepen their understanding of community needs. The feature importance graphs generated by both models provide libraries with a way to identify key operational aspects significantly influencing annual library visit counts.

Noteworthy discoveries encompass the roles of computer usage, circulation of non-English materials, and programs for both adults and children. These variables, corroborated by both models, are recommended as variables to start with in localized experimentation, which hold the potential to potentially uncover causal relationships. While this study lays the foundation for further research into these variables, demographic data should be added to more comprehensively capture each library's distinctive role within its community. Additionally, improvements in data collection methods, such as providing more detailed instructions and encouraging precise counts rather than estimates, could enhance the robustness of future analyses.

In summary, this study sheds light on the multifaceted factors influencing library visits, providing librarians the chance to not only adapt but thrive in an evolving landscape. Hopefully, this study may immediately empower libraries with actionable information to fortify their position as indispensable community hubs, fostering meaningful social connections and sustained engagement.