MSE 226: Project Part 1 - Data Exploration and Prediction

1a. Identify a Realistic Stakeholder

Our stakeholder is Julie Balutis, the Director of Grants Policy and Management at the Institute of Museum and Library Services (IMLS).[1] As the federal agency funding the nation's public library infrastructure, this office is charged with strategic grant planning to ensure congressionally appropriated funds are used efficiently. Their mission is to guarantee every federal investment delivers measurable community benefits and long-term sustainability. The core challenge for this stakeholder is accurately anticipating patterns of library use before committing to major capital expenditures, essential for equitably allocating limited federal resources and justifying funding decisions amid increasingly constrained public investment.

1b. Stakeholder Questions and Decision Use

The model supports both short- and long-term decision-making for IMLS. In the short term, it helps determine whether to proceed with a proposed library construction project by answering: *"Is the projected level of community use sufficient to justify moving forward with this investment?"* This allows IMLS to compare projects based on their anticipated community impact. In the long term, the model provides data-driven forecasts that help the agency plan future grant programs, assess how proposed projects align with broader trends in library use, and monitor evolving visitation patterns to refine funding priorities.

2. Outcome Variable

The outcome variable for this analysis is "annual library visits", a continuous measure representing the total number of in-person entries reported by each library in fiscal year 2023. This metric aligns directly with the stakeholders' core interest: forecasting community engagement with libraries. For our predictions, we take the log of this variable as it is heavily right-skewed, so it can be modelled better by linear models. For comparison purposes, non-linear models also used the log-transformed outcome.

3a. Criteria Confirmation

The dataset meets all project requirements. After cleaning, it includes 6,988 libraries (rows) and 11 predictors (columns), drawn from the 2023 Public Libraries Survey (PLS) compiled by IMLS.[2] It contains both categorical and numerical features, represents real-world administrative data, and excludes temporal components. Missing data is removed to ensure completeness (described more below).

3b. Dataset Description and Appropriateness

Each observation represents one U.S. public library in FY2023. The original dataset included 9,252 observations, but 2,264 were removed due to imputed, missing, or closed libraries (given by zero reported visits). This ensures data integrity while maintaining a representative sample.

The dataset includes five categorical and seven numerical variables (including a numerical outcome variable) relevant to predicting library visitation. Categorical variables include: interlibrary relationship, which reflects whether the library is independent or part of a broader network; FSCS definition status, which identifies libraries that meet federal library criteria; and the overdue fines policy, which indicates potential barriers to patron access. Additional categorical features include region (e.g., New England,

[1] Julie Balutis, Director of Grants Policy and Management. (2024). Imls.gov. https://www.imls.gov/about/contact/staff-directory/leadership/julie-balutis-director-grants-policy-and-management

[2] "Public Libraries Survey (PLS)." Imls.gov, 2023, www.imls.gov/research-evaluation/surveys/public-libraries-survey-pls.

Plains, Far West) and community type (City, Suburb, Town, Rural), representing differences in population density and geography. Numerical variables describe scale and capacity factors. These include the Legal Service Area (LSA) population and county population, which measure the potential user base; print and e-book volumes, which indicate the depth and accessibility of library collections; and the number of branches and bookmobiles, which represent the library system's physical reach and service availability.

4. Holdout Set

To ensure unbiased evaluation, 20% of the dataset (1,398 libraries) was reserved as a holdout set prior to any analysis.

5a. Data Collection and Issues

The Public Libraries Survey (PLS) is an annual national census administered by IMLS through state library agencies. Each agency collects data directly from local libraries via a standardized reporting portal. The FY2023 cycle achieved a 96.4% response rate across all 50 states and the District of Columbia, though several U.S. territories did not participate. As with most administrative datasets, some records may include measurement or reporting errors, and IMLS performs statistical imputation for non-responses. To maximize data reliability, we excluded all imputed entries and retained only directly reported observations, ensuring the final sample reflects observed data. Finally, while the FY2023 dataset provides a robust snapshot of national library activity, its predictive accuracy may not accurately reflect recent library trends.

5b. Covariate Correlations with the Outcome

Through data exploration, we found that all numeric variables, besides the number of book-mobiles and the number of branches, are heavily right-skewed. We found this by creating histograms and QQ plots of the variables. As a result, we can apply a log(x+1) transformation to normalize the distributions (some variables, such as the number of ebooks, are 0 at some libraries).

The variables most strongly correlated with annual library visits are the Legal Service Area population (Log transformed: 0.87, Not log transformed: 0.90), print volumes (0.87, 0.78), number of library branches (0.53, 0.84), county population (0.51, 0.18), eBook volumes (0.35, 0.14), and number of bookmobiles (0.29, 0.32). Notably, county population and the number of library branches, factors one might expect to be highly predictive, show only moderate correlations with annual visits.

5c. Inter-Covariate Correlations

Strong interrelationships appear among the Legal Service Area population and print volume (0.87, 0.70), print volume and number of branches (0.65, 0.69), and Legal Service Area population and number of branches (0.60, 0.89). This reflects the natural co-scaling between larger communities and more resource-intensive library systems. Interestingly, county population and local service area population are not highly correlated with each other (0.52, 0.18), indicating that library service jurisdictions often extend across or diverge from county boundaries.

5d. Natural Subgroups

Several natural subgroups emerge within the data. These include a library's geographic region, community type (urban, suburban, town, rural), and system structure (e.g., federated versus independent systems). We found that libraries meeting the FSCS criteria also exhibit substantially higher average *log(visits+1)* than those that do not. Across community types, city libraries record the highest average *log(visits+1)*, followed by suburban, town, and rural libraries, consistent with our expectations.

5e. Missing Data Handling

The original dataset used coded values: (–1, –3, –4, and –9) to indicate missing or inapplicable responses and included an RSTATUS field identifying whether each record was directly reported or statistically imputed by IMLS.[3] To ensure data accuracy, we retained only directly reported observations, those with RSTATUS values of 1 or 3, and replaced all coded missing values with standard "NA" entries.[4] We then removed libraries that reported missing or zero annual visits (this meant that the library was closed in 2023) and any remaining incomplete records. These cleaning steps reduced the sample size to 6,988 complete library systems, which remains a large sample for statistical analysis and avoids the potential biases introduced by imputed data. However, by removing libraries that did not report or impute data, we may have introduced bias into our estimate if these data are not missing at random.

6a. Prediction Error Metrics

Our evaluation metric is the Root Mean Squared Error (RMSE), computed on the log-transformed outcome, *log(visits+1)*. RMSE is sensitive to large prediction errors and provides an interpretable measure of overall predictive deviation. Note that even though performance is evaluated on the log scale, the predictions can be easily converted back to the original scale (annual visits), using $e^{prediction} - 1$.

6b. Baseline Models

(1) Sample mean: always predicts the sample mean of *log(visits+1)*: **10.011** (note that this is *not* the log of mean visits), representing a no-information predictor. This model gives a 10-fold CV RMSE of **1.677**.

(2) Univariate population model: an OLS regression using only *population_lsa* and an OLS regression using only *log(population_lsa+1)*, capturing how well population size alone explains visitation. This model gives 10-fold CV RMSEs of **1.510** and **0.817**, respectively.

These establish a threshold of **0.817** for evaluating whether more complex models meaningfully improve predictive accuracy beyond the basic demographic scale.

6c. Training and Evaluation Procedure

All models were trained and evaluated using 10-fold cross-validation (CV) on the training data. To test which covariates would enhance performance if they were log(x+1)-transformed, we conducted a 10-fold CV across all 16 possible combinations of log-transformed versus not-transformed covariates for each model (except for the neural network) to select which covariates to transform. For LASSO, Ridge and Random Forests, we also employed randomized search to test 25 possible variations of each model's hyperparameters, and identify the optimal hyperparameters for each. The final model was selected as the one achieving the lowest cross-validation RMSE. We also considered further feature engineering, such as computing Library Volume Size / Legal Service Area Population, but decided to leave those as future optimizations.

[3] Institute of Museum and Library Services. Data File Documentation: Public Library Survey (PLS) — Fiscal Year 2023. Washington, D.C.: IMLS, 2025. Available at https://www.imls.gov/sites/default/files/2025-08/PLS-FY-2023-Data-Documentation-508.pdf. Page 26.
[4] Ibid, Page 49.

6d. Learning Algorithms for Comparison

To balance interpretability and predictive power, we compare three modeling approaches representing increasing levels of flexibility and complexity. **(1) Linear Regression:** Provides a transparent linear model for *log(visits+1)*. We also employ regularized regression (**LASSO** and **Ridge**) to extend OLS with penalties. Regularization strength (λ) was tested over 25 random log-spaced values (0.0001 - 10). **(2) Random Forest:** Averages many decorrelated decision trees trained on bootstrap samples to capture nonlinear relationships. We tuned the number of estimators using 25 random combinations between 200-800, the maximum tree depth between 5-30, the minimum samples required to split a node between 2-20, the minimum samples per leaf between 1-8, and the fraction of features considered at each split among ['sqrt', 'log2', 0.4, 0.6, 0.8]. For interpretability, we also trained a single **Decision Tree** with max depth (3-20), minimum samples split (2-30), and min samples per leaf (1-10). **(3)** For the **Neural Network**, we log-transformed all the numerical variables. Furthermore, we explored the use of embeddings instead of one-hot encoded categorical variables, which maps the categorical variables into an 8-dimensional vector space, and included two linear layers with 64 neurons and a ReLU layer in between. We then ran the training of the neural network with 5000 epochs, a learning rate of 0.001, and the Adam optimizer for the 10 folds of cross-validation, and averaged over the fold RMSE. We performed some manual hyperparameter tuning, such as adjusting the number of layers or embedding dimensions, but saw negligible performance gains. A future exercise could be to systematically optimize the hyperparameters.

6e. Best Model

Overall, all models substantially outperformed the baseline (**0.817**) in predictive accuracy. The Random Forest model achieved the lowest cross-validated RMSE (**0.657**), making it the best-performing approach. Its optimal configuration included 225 estimators, a maximum depth of 55, and a minimum sample split of 6. At each node, the model considered a random subset of features whose size equaled the square root of the total number of features. The model also utilized a log(x + 1) of the LSA population and county population, but not a transformation of the number of print and ebooks. This model also significantly outperformed the single Decision Tree (cross-validated RMSE = **0.719**).

Among the linear models, LASSO, Ridge, and OLS yielded almost identical cross-validated RMSE values (**0.673**), only slightly worse than Random Forests. LASSO ($\lambda = 0.0017$) and Ridge ($\lambda = 28.4850$) were only slightly better than OLS, suggesting that the dominant relationships between predictors and visitation are largely linear. For these models, the best models came from log(x+1) on all four variables.

Our neural network performed slightly better than OLS with a cross-validated RMSE of **0.668** but worse than Random Forests. Hyperparameter tuning might have potentially improved this value.

In summary, Random Forests achieved the best cross-validated RMSE, modestly outperforming the linear models. However, the close results among the linear models suggest that *log(visitation+1)* is primarily captured by broadly linear effects, with nonlinear interactions contributing only incremental gains in predictive accuracy.

6f. Process

We followed an iterative process combining data exploration, feature engineering, model comparison, and automated hyperparameter tuning to identify the best model. Exploratory analysis showed that both the outcome and most numeric predictors were heavily right-skewed, so we applied log(x+1) transformations to stabilize variance and tested all 16 possible combinations of transformed and untransformed predictors to evaluate their effect on predictive accuracy. Categorical variables, including region, community type, and policy indicators, were one-hot encoded, and continuous predictors were standardized for use in regularized models. We first trained linear models (OLS, Ridge, and LASSO) to establish a strong baseline, which revealed that visitation was largely explained by linear relationships. To capture potential nonlinear effects, we next trained Decision Tree, Random Forest, and Neural Network models. For models with tunable parameters (LASSO, Ridge, Random Forest, and Neural Network), we used Randomized Search Cross-Validation, evaluating 25 parameter combinations across 10 folds for each of the 16 transformation sets to efficiently optimize performance. The Random Forest achieved the lowest cross-validated RMSE (0.655), outperforming both linear and neural models.

7a. Stakeholder Recommendations and Model Use

Based on our evaluation criteria, we believe this model is a strong predictor of annual visits to a library. While no model is perfect, this approach provides a strong estimate that can inform several key decisions:

- **Investment decisions:** Support go/no-go decisions for proposed funding initiatives by estimating the likely visitations at libraries.
- **Funding allocation:** Guide long-term IMLS grant decisions toward libraries projected to attract the most visitors.
- **Data completion:** Estimate visitation numbers for libraries with missing or incomplete records, improving data consistency across the system.

The stakeholder can use the model provided in the .zip file to make a prediction for log(x+1) annual visits, and then convert that to annual visits by exponentiating the outcome minus 1.

7b. Practical Limitations and Reliability

While the model performs strongly overall, several limitations should temper its use in decision-making:

1. **Prediction uncertainty at local scales**: Although aggregate accuracy is high, predictions for small or atypical libraries may be less reliable due to limited representation in the data.
2. **Static snapshot**: The model is based on FY2023 data and does not incorporate temporal dynamics. Predictions assume relationships between visitation and inputs remain stable over time.
3. **Correlation, not causation**: The model identifies statistical associations, but cannot confirm causal effects. Stakeholders should avoid interpreting model OLS coefficients as proof that changing one factor will directly increase visits. Additionally, in practice, locations with higher populations may have higher visitation. This should not exclude low-population libraries from being built, but instead should be used to predict which low-population libraries will have maximal visitation.
4. **Potential bias from data cleaning**: Because we excluded imputed or missing records, the sample may overrepresent larger or better-reported library systems, potentially biasing predictions upward for under-resourced libraries.
5. **Unmeasured influences**: The model does not capture other relevant factors like programming quality, staffing levels, or local political climate, which may affect community engagement.

8. AI Tools

ChatGPT was used to think of more complex learning algorithms that we could utilize and understand the concepts in more depth. We also asked ChatGPT for syntax questions in different coding languages and about commonly used default hyperparameters.

MSE 226: Project Part 1 - Data Exploration and Prediction

9. Appendix

Table 1: Normalizing Numeric Features.

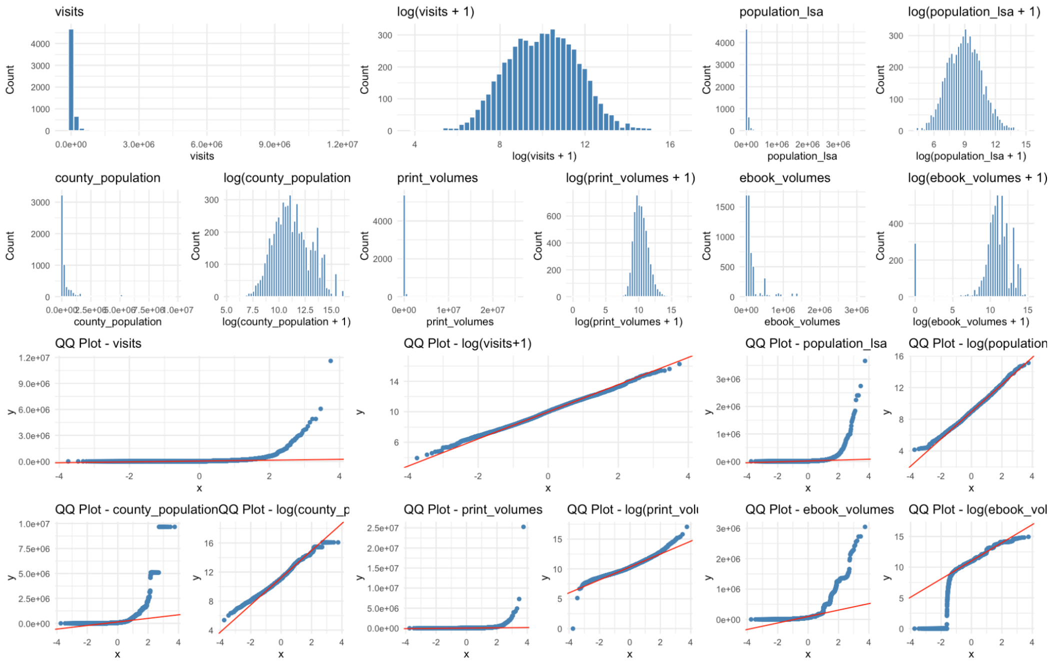
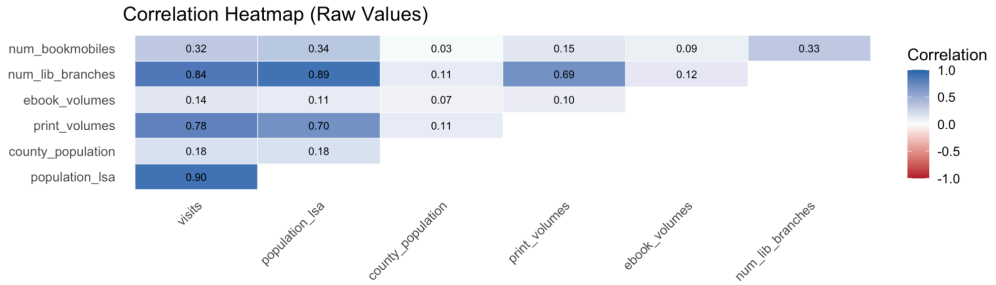
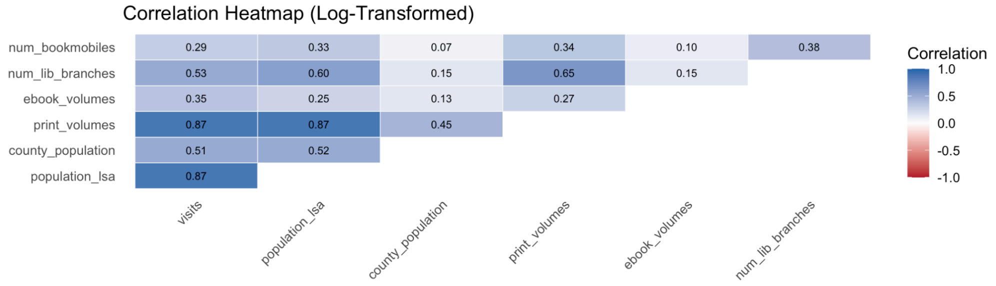


Table 2: Correlation Between Variables


Correlation Heatmap (Raw Values)
num_bookmobiles 0.32 0.34 0.03 0.15 0.09 0.33
num_lib_branches 0.84 0.89 0.11 0.69 0.12
ebook_volumes 0.14 0.11 0.07 0.10
print_volumes 0.78 0.70 0.11
county_population 0.18 0.18
population_lsa 0.90
visits
population_lsa
county_population
print_volumes
ebook_volumes
num_lib_branches
Correlation


Correlation Heatmap (Log-Transformed)
num_bookmobiles 0.29 0.33 0.07 0.34 0.10 0.38
num_lib_branches 0.53 0.60 0.15 0.65 0.15
ebook_volumes 0.35 0.25 0.13 0.27
print_volumes 0.87 0.87 0.45
county_population 0.51 0.52
population_lsa 0.87
visits
population_lsa
county_population
print_volumes
ebook_volumes
num_lib_branches
Correlation