**“Analysis on Public Library Data: Python”**

Data Science Toolbox: Python Programming: INT375

**PROJECT REPORT**

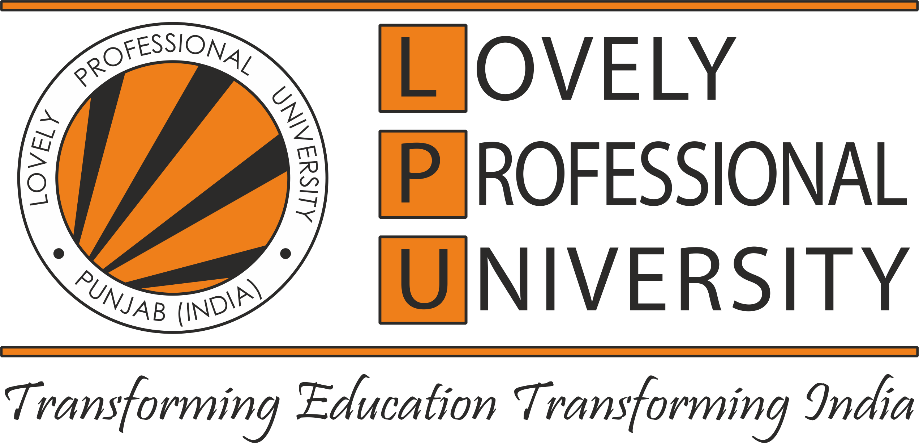
**Submitted By:** **Submitted To:**

Aditya Arora Mr. Vikas Mangotra

(12324327)

10

in partial fulfillment for the requirement of the award of degree of B. Tech. (Computer Science and Engineering)



LOVELY PROFESSIONAL UNIVERSITY

Phagwara, Punjab

**DECLARATION**

I, Aditya Arora, student of B. Tech. under CSE Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own honest work and is genuine.

Date: 12/04/2025

Registration No.: 12324327 Name: Aditya Arora

**CERTIFICATE**

This is to certify that Aditya Arora bearing Registration no. 12324327 has completed INT375 project titled, **“Analysis on Public Library Data”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Signature and Name of the Supervisor**

**Designation of the Supervisor**

**School of Computer Science and Engineering**

Lovely Professional University

Phagwara, Punjab.

Date:

**TABLE OF CONTENTS**

1. Project Overview: Introduction
2. Source of Dataset
3. Dataset Preprocessing:  
   i. General

ii. Evaluating Data completeness

iii. Analyse Per Capita Program Attendance by Library

iv. Analyse Total Program Attendance by Library

v. Investigate Relationship between Key Public Library Indicators

vi.Explore Link Between Per Capita Circulation Distribution by County

vii.Identify Graph of Population of Service Area vs. Total Library Visits

viii. EDA

viii. T-Test analysis

1. Conclusions
2. Future Scope
3. References

**PROJECT OVERVIEW**

This project involves analysis of public library data in the state of Connecticut, United States by utilizing Python as the main utility for data processing and analysis. The objective was to examine the relationship between multiple factors—library visits, circulation, expenditures, program attendance, and population size—and how each is connected with others, as well as what they indicate concerning library performance and community involvement.  
To accomplish this, the project utilized Exploratory Data Analysis (EDA), produced effective visualizations like bar charts, heatmaps, and boxplots, and performed statistical hypothesis testing like the t-test to confirm observed trends. A cleaned data set was employed to provide accuracy, and Python libraries like Pandas, Seaborn, Matplotlib, and SciPy were used across the workflow.  
Overall, this project not only shows the strength of data science in interpreting public service data, but also paves the way for real-world implementation in policy-making, library administration, and community building.

**SOURCE OF DATASET**

The data utilized in the current project has been meticulously sourced from publicly accessible records of U.S. public libraries, compiled to offer a comprehensive view of library performance across diverse communities. This dataset encompasses a wide range of detailed statistics related to library utilization, operational efficiency, financial management, and community engagement initiatives. It has been carefully curated to capture the multifaceted nature of library performance, incorporating both quantitative metrics and qualitative attributes to provide a holistic perspective. The dataset includes a variety of performance indicators for individual libraries, featuring both numeric data points—such as visit counts and budget allocations—and categorical features, such as geographic identifiers and library rankings. This rich collection enables researchers, policymakers, and library administrators to analyse trends, identify disparities, and develop strategies to enhance library services tailored to specific community needs.

**Core fields:**   
1. Total Library Visits  
2. Population of Service Area  
3. Total Circulation  
4. Library Program Attendance & Views  
5. Public Internet Computer Usage  
6. Operating Expenses, Salaries, and Material Costs  
7. County, Library Name, and AENGLC Rank  
The dataset offers a valuable resource of real-world data to evaluate how library use, funding, and operations differ within various communities and population segments.

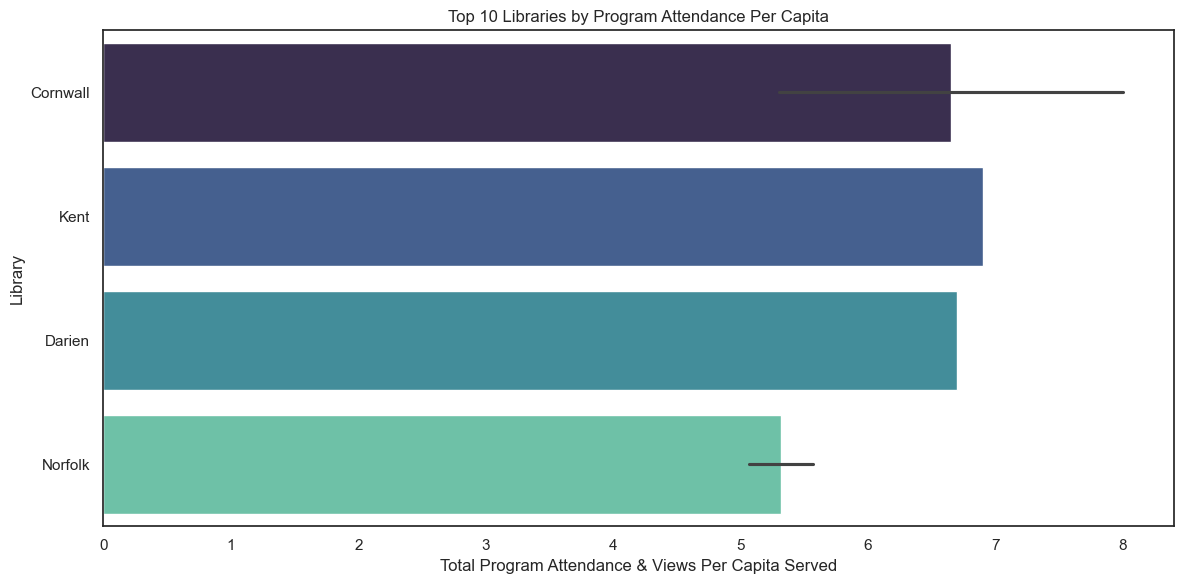
Link to the Dataset: [DATASET](https://catalog.data.gov/dataset/public-libraries-b1aaf)

Taken from Public Federal Data Website for America: [data.gov/](https://data.gov/)

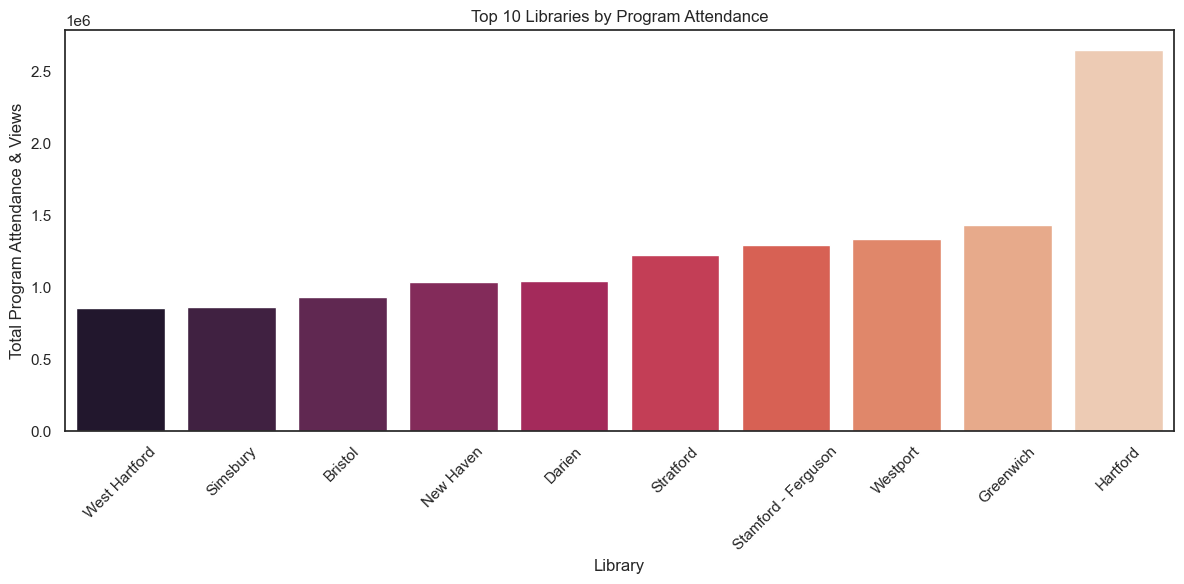
**DATASET PREPROCESSING**  
Prior to analysis, the dataset went through a series of cleaning and transformation processes to prepare for accuracy and use. The initial CSV had mixed data types, missing values, and some formatting inconsistencies.  
Preprocessing steps involved:  
  
**1. Loading and Cleaning**: The data was loaded with Pandas. Numeric missing values were replaced with 0, and categorical missing values were replaced with "Unknown" to ensure consistency.  
**2. Type Conversion**: Columns like Population of Service Area, Total Circulation, Total Library Visits, and expenditure values were converted to numeric types explicitly to allow for proper statistical analysis and plotting.  
**3. Column Renaming (where necessary):** Ambiguous or long column names were shortened for easier coding and interpretation (e.g., Total Program Attendance & Views was abbreviated in plots).  
**4. Dealing with Duplicates and Aggregates**: Libraries with similar names but dissimilar values were aggregated and their metrics combined for best-performing analyses.  
**5. Data Filtering**: R only kept rows containing meaningful values in key columns (e.g., non-zero attendance or circulation) for hypothesis testing and visualization to prevent the distortion of findings.  
  
The pre-processed, cleaned dataset was saved as data/cleaned\_libraries.csv and formed the foundation for all EDA, visualizations, and statistical tests conducted within the project, majorly using Jupyter Lab’s dashboard type layout.

**ANALYSIS ON THE DATASET**

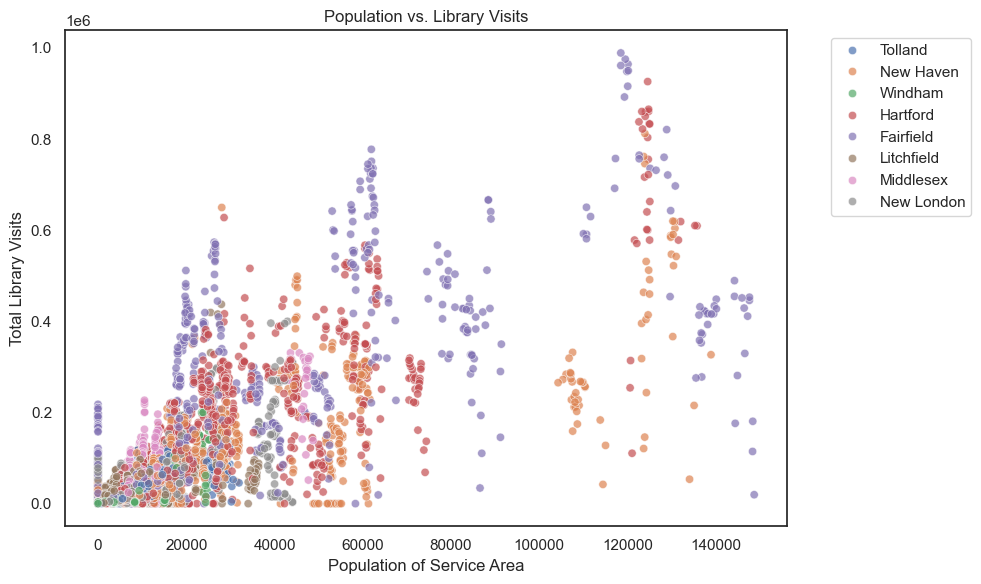
1. **Bar Chart: Per Capita Program Attendance by Library**  
   Smaller libraries such as Cornwall, Kent, and Darien are top in per capita program attendance, indicating high community participation despite smaller service populations.



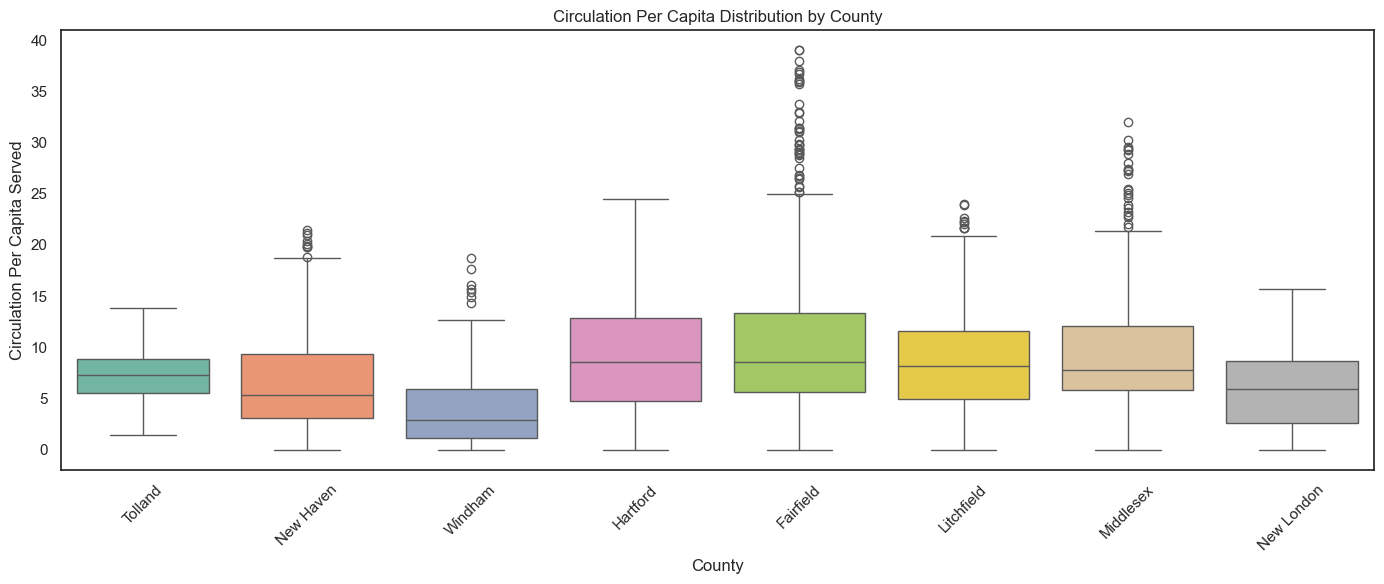
1. **Column Chart: Total Program Attendance by Library**  
   Larger urban libraries such as Hartford, Greenwich, and Stamford record the highest total program attendance, fuelled by their larger population and event programs.

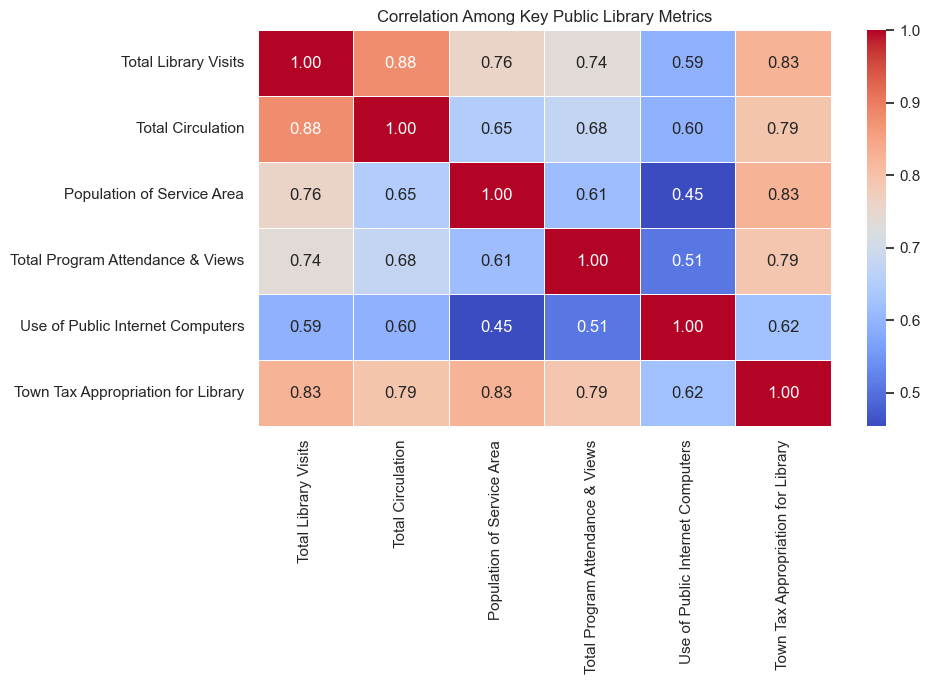
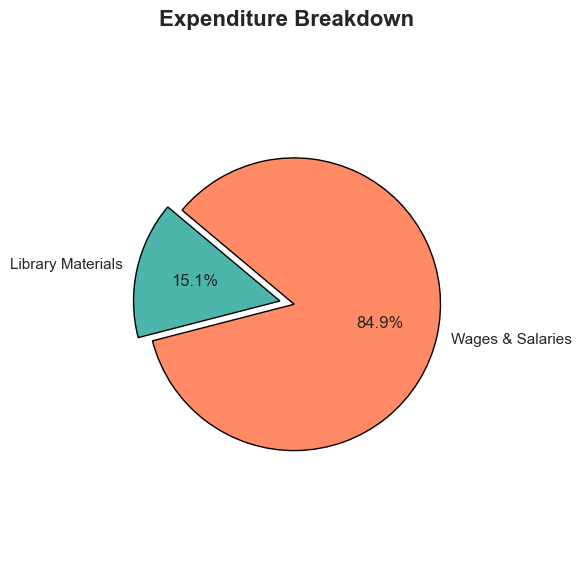


1. **Scatter Plot: Population of Service Area vs. Total Library Visits**  
   There is a positive correlation between population and visits to libraries, although some of the libraries with moderate populations have disproportionately high numbers of visits, which suggests excellent community integration.



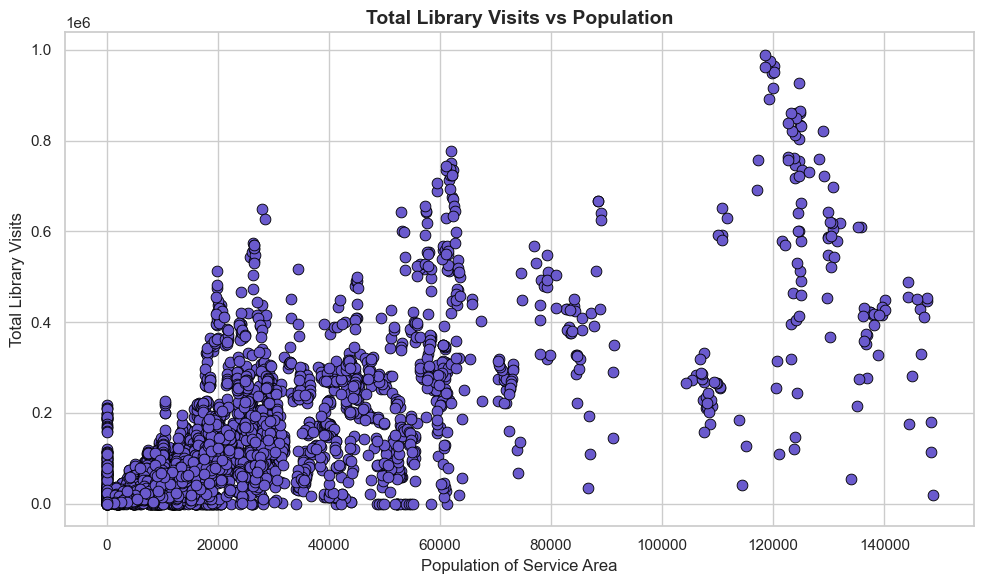
**4. Boxplot: Per Capita Circulation Distribution by County**  
Fairfield and Middlesex counties have higher median and outlier per capita circulation, which indicates intense borrowing activity and possibly wider collections or electronic lending services.



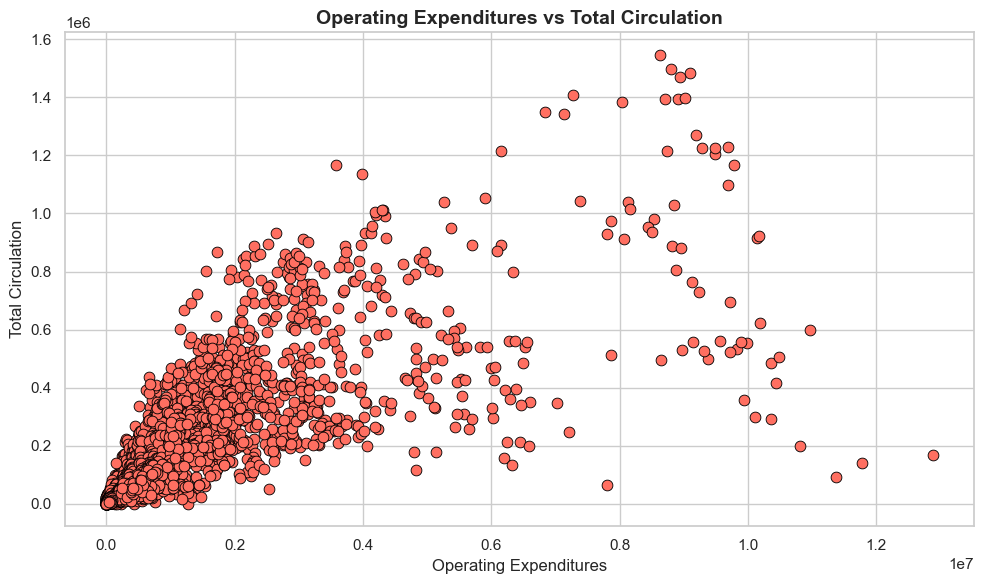
1. **Heatmap: Correlation Between Key Public Library Indicators**  
   Total visits to libraries, circulation, and tax allocation are highly correlated with one another, noting that better-funded libraries tend to have more visitors and circulate more books.
2. **Pie Chart: Expenditure Distribution of Libraries (Wages vs Materials)**More than 84% of library expenditures are used for wages and salaries, pointing to the high dependency of library functions on personnel with only 15% used for materials.

**EXPLORATORY DATA ANALYSIS AND T-TEST**

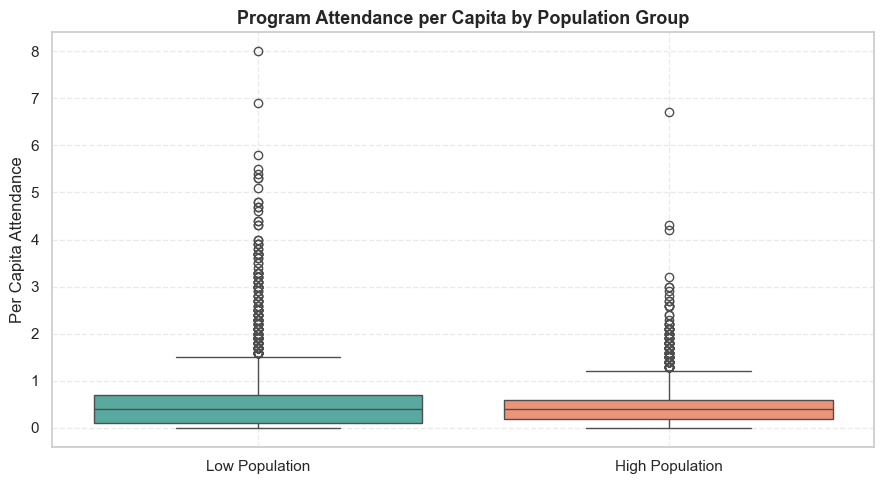
**Chart 1: Total Visits to the Libraries vs Population**  
X-axis: Population in Service Area  
Y-axis: Total Visits to Libraries  
  
Overall increasing trend, indicating that higher populations have more visits to libraries, but not entirely linear. There's extensive variation, particularly in the middle-range population categories — might be due to local activity, services, or library size.



**Chart 2: Operating Expenses vs Total Circulation**  
X-axis: Operating Expenses  
Y-axis: Total Circulation  
  
Very strong positive correlation. Libraries that spend more circulate more items — an indication that spending translates directly into use. There's a greater spread at higher spending levels, indicating decreasing returns or differing efficiency.  
The chart below is a program attendance per capita for two different population groups.



**Chart 3: Program Attendance per Capita (Low vs High Population)**

**Hypothesis Test: Does population affect per capita program attendance?**  
Test: T-Test between two population groups  
  
Insight: The distributions are comparable, but small-population libraries may have a bit more variability (greater spread and more outliers). Depending on your p-value, this could imply per capita participation is pretty uniform regardless of population — which is a compelling result.

**Results**  
t-statistic = -6.351  
  
This is the calculated test statistic from the t-test. The negative sign simply tells you which group had the higher mean (likely the low population group in this example).  
The bigger the absolute value, the stronger evidence you have of a genuine difference.  
p-value = 0.0000  
The p-value gives you the probability that the difference you observed occurred by chance.  
A p-value this low (basically zero) indicates that the difference is extremely unlikely to be random.  
  
Usually, if p < 0.05, we say it's statistically significant — your value is way lower than that.  
  
**Interpretation**  
There is a statistically significant difference in per capita program attendance between low and high population libraries.  
So yes, population size does seem to impact per capita program attendance.

**Conclusion:**

The hypothesis is supported

**REASONS**

1.\_Tighter Community Bonds in Smaller Communities  
2. Reduced Competition for Attention  
3. Better Targeting and Accessibility  
4. Increased Attendance by Smaller Denominators  
5. Feeling of Ownership and Loyalty  
6. Program Prioritization

**CONCLUSIONS**

1. **Library Visits Increase with Population – But Not Linearly**

* Larger service area populations generally lead to more library visits.
* However, visit counts vary widely within similar population groups, suggesting other factors (like programming, outreach, or location) may also play a big role.

1. **Higher Operating Expenditures Correlate with Higher Circulation**

* Libraries that invest more money typically see greater circulation numbers.
* This highlights the positive impact of funding on usage and suggests that investment in library services yields measurable returns in engagement.

1. **Population Size Impacts Per Capita Program Attendance**

* A t-test comparing low and high population groups shows a **statistically significant difference** in per capita program attendance (p < 0.001).
* Libraries in **lower population areas tend to have higher attendance per capita**, likely due to stronger community ties, fewer competing services, and more personalized programming.

1. **Smaller Libraries Can Achieve High Engagement**

• Libraries serving smaller populations (e.g., Kent, Cornwall, Darien) report high program attendance per capita.

• This shows that impactful community engagement doesn’t require a large population base—it depends more on programming and outreach.

1. **Higher Operating Expenditures Correlate with Higher Circulation**

• Libraries that invest more money typically see greater circulation numbers.

• This highlights the positive impact of funding on usage and suggests that investment in library services yields measurable returns in engagement.

1. **Library Spending Is Primarily Staff-Centred**

• Over 84% of library budgets are spent on staff wages and benefits.

• This underscores that libraries are people-driven institutions, and their success heavily relies on skilled and dedicated staff.

**FUTURE SCOPE**

**1.Scale to Statewide or National Benchmarks**  
• Integrate additional counties or states' data to compare performance and share best practices regionally.  
• Facilitate comparative review of funding mechanisms, programming effectiveness, and user engagement.  
  
**2.Add User Feedback and Surveys**  
• Integrate qualitative data such as patron satisfaction, user input, and demographic data.  
• Helps place quantitative data in context and informs improvement based on actual community needs.  
  
**3.Predictive Analytics for Resource Planning**  
• Utilize machine learning to predict circulation trends, program visitation, or use of digital resources.  
• Facilitates anticipatory staffing, collections, and budgeting planning.  
  
**4.Digital Services and Technology Impact Study**  
• Determine how digital resource investment (e-books, online databases, virtual programs) influences use.  
• Practical for informing future tech spending and hybrid service approaches.  
  
**5.Create an Interactive Dashboard**  
• Develop a dashboard accessible by library directors or the general public to browse in real-time trends.  
• Encourages transparency and data-driven decision-making at every level.  
  
**6.Library Equity and Access Study**  
• Examine how funding and services differ across socioeconomic regions.  
• Advocates for equitable distribution of library resources and outreach.

**REFERENCES**

[Public Libraries - Catalog](https://catalog.data.gov/dataset/public-libraries-b1aaf)

[Full article: Library analytics and metrics: Using data to drive decisions and services](https://www.tandfonline.com/doi/full/10.1080/00049670.2015.1100267)

[Library assessment and data analytics in the big data era: Practice and policies - Chen - 2015 - Proceedings of the Association for Information Science and Technology - Wiley Online Library](https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/pra2.2015.14505201002)