**DATA SCIENCE**

**Report File**

(Project Semester January-April 2025)

# Exploratory Data Analysis using Python

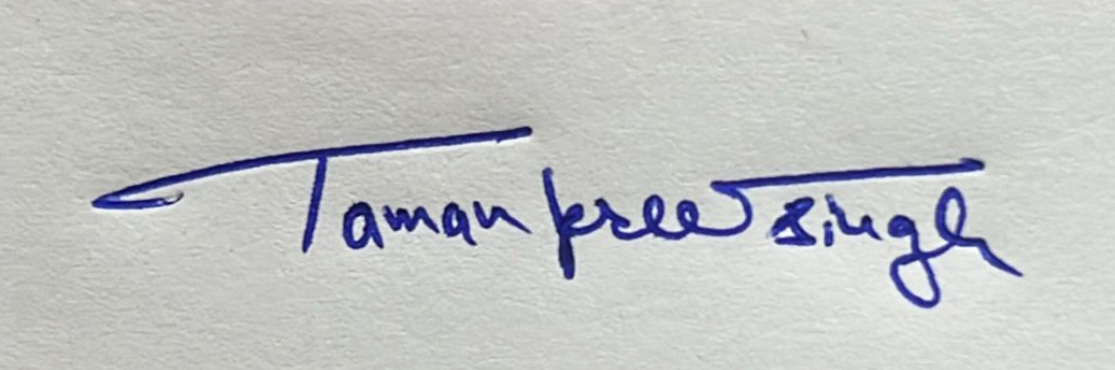
Submitted by  
 Taman Preet Singh  
Registration No: 12309438  
Program: Bachelor of Technology Section: K23GD  
Course Code: INT375

Under the Guidance of  
Mrs. Baljinder Kaur (U. Id: 27952) Assistant Professor  
Discipline of CSE/IT  
Lovely School of Computer Science and Engineering  
Lovely Professional University, Phagwara

CERTIFICATE:-

This is to certify that Taman Preet Singh bearing Registration no. 12309438 has completed INT375 project titled, “Exploratory Data Analysis of Public Libraries using Python” under my guidance and supervision. To the best of my knowledge, the present work is the result of his 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: 13 April, 2025

## DECLARATION:-

I, Aman, student of Bachelor of Technology under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.  
  
Date: 13 April, 2025  
  
Signature:   
  
Registration No.: 12309438  
  
Name: Taman Preet Singh

## ACKNOWLEDGEMENT:-

I would like to express my deep gratitude to Mrs. Baljinder Kaur, my project guide, for her valuable guidance, support, and encouragement throughout the duration of this project. I am also thankful to Lovely Professional University for providing the resources and environment to carry out this work successfully.

## TABLE OF CONTENTS:-

1. Introduction

2. Source of Dataset

3. EDA Process

4. Analysis on Dataset

5. Conclusion

6. Future Scope

7. References

## 1. INTRODUCTION:-

The aim of this project is to analyze patterns and trends in public library usage using Exploratory Data Analysis (EDA) techniques in Python. Public libraries play a vital role in communities, providing access to knowledge, culture, and essential services. Understanding the operational and usage metrics of libraries can guide policymakers and stakeholders in making data-driven decisions. This project explores the structure, services, and user engagement patterns of public libraries in the U.S. using real-world data from Data.gov.

## 2. SOURCE OF DATASET:-

The dataset used in this project, titled “Public Libraries”, was obtained from the U.S. Government’s official open data platform, Data.gov. The dataset includes operational statistics like population served, total visits, circulation, program attendance, and financials. It provides valuable insights into how public libraries operate across different counties and years. The dataset was accessed in CSV format and used for exploratory analysis.  
  
Dataset Link: <https://catalog.data.gov/dataset/public-libraries-b1aaf>

Linkedin post link: <https://www.linkedin.com/posts/taman5_datascience-eda-python-activity-7316766385047576576-1hFn?utm_medium=ios_app&rcm=ACoAAEf0NGIBP_YJj3uV4dL1xh4xgHKw7-I5jlc&utm_source=social_share_send&utm_campaign=whatsapp>

Github repository link: <https://github.com/Tamanpreet84/public_libraries_EDA>

3. EDA PROCESS:-

The Exploratory Data Analysis (EDA) process involved the following steps:

- Loading the dataset using pandas.

- Identifying and handling missing values by replacing them with column means.

- Removing duplicate entries from the dataset.

- Data type inspection and conversion as needed.

- Generating summary statistics to understand data distribution.

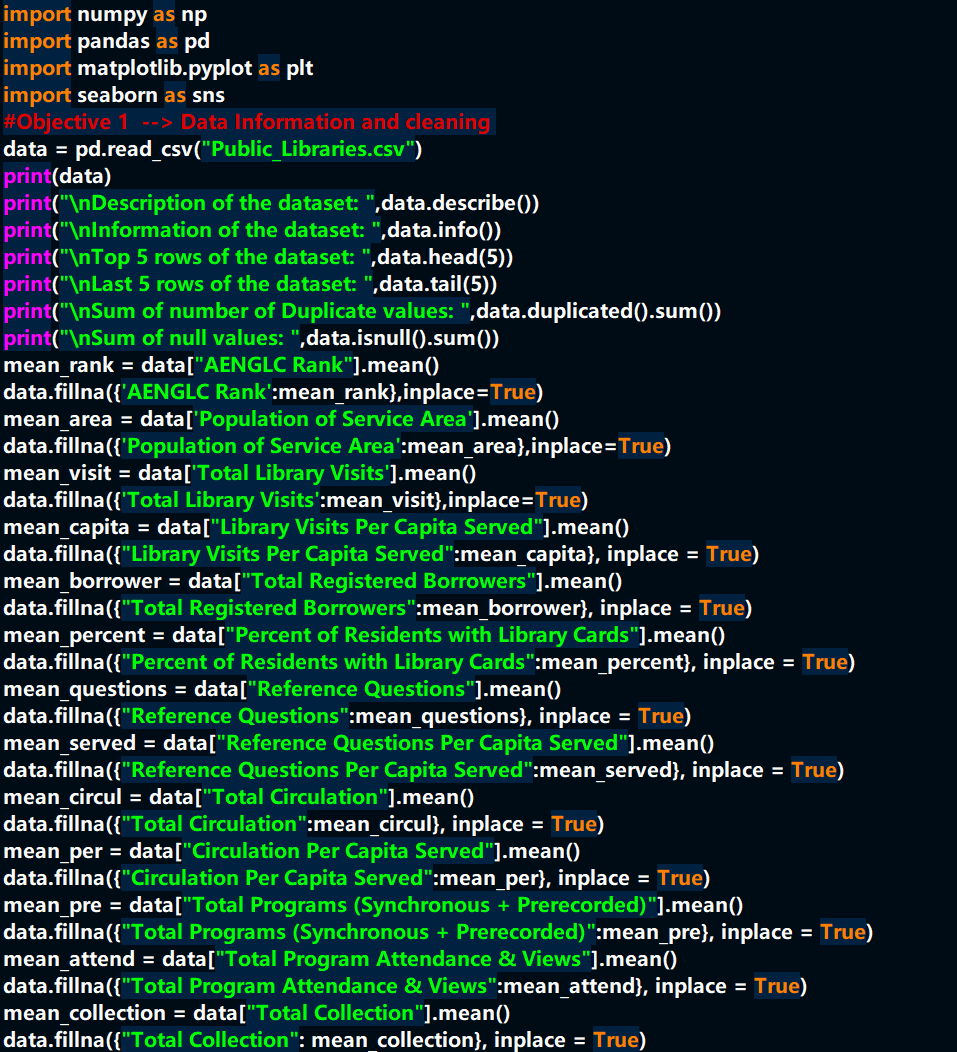
- Visualizing data using matplotlib and seaborn to uncover trends, patterns, and relationships between key attributes.

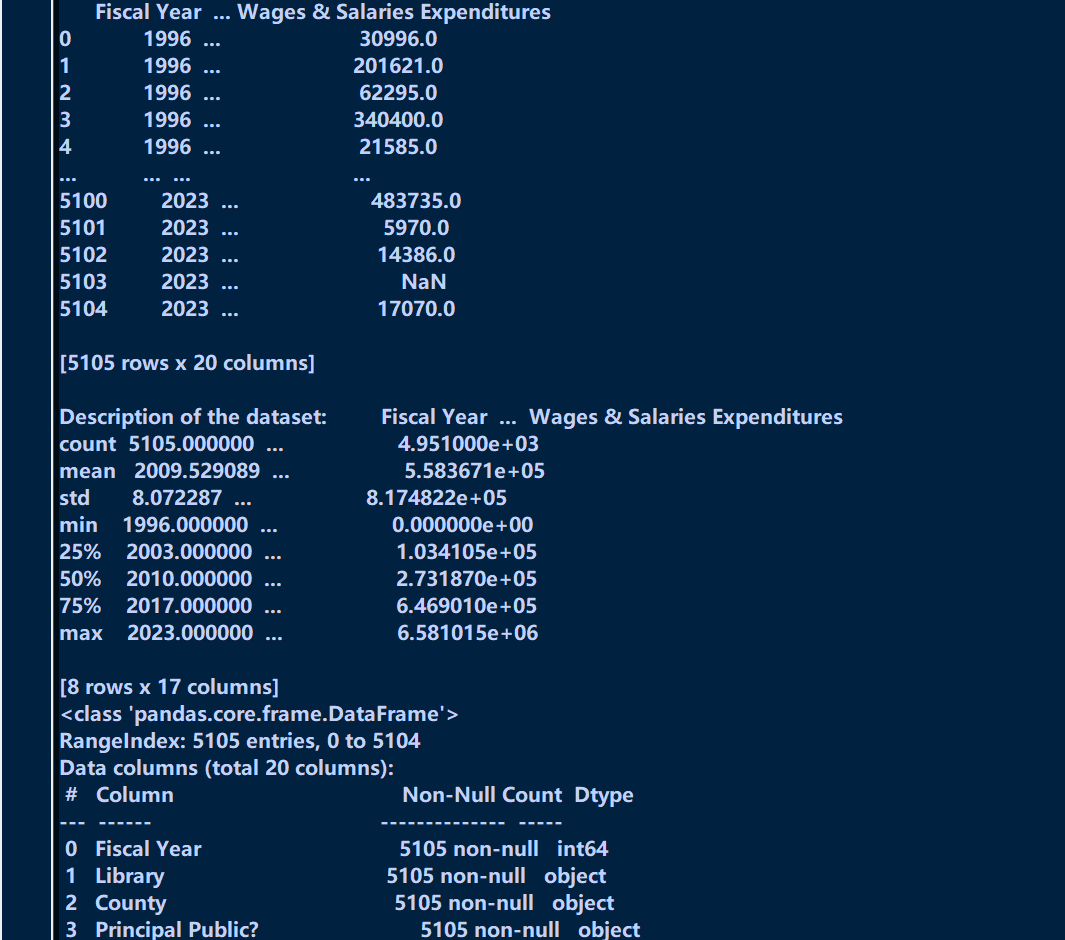
- Saving the cleaned dataset for future use.

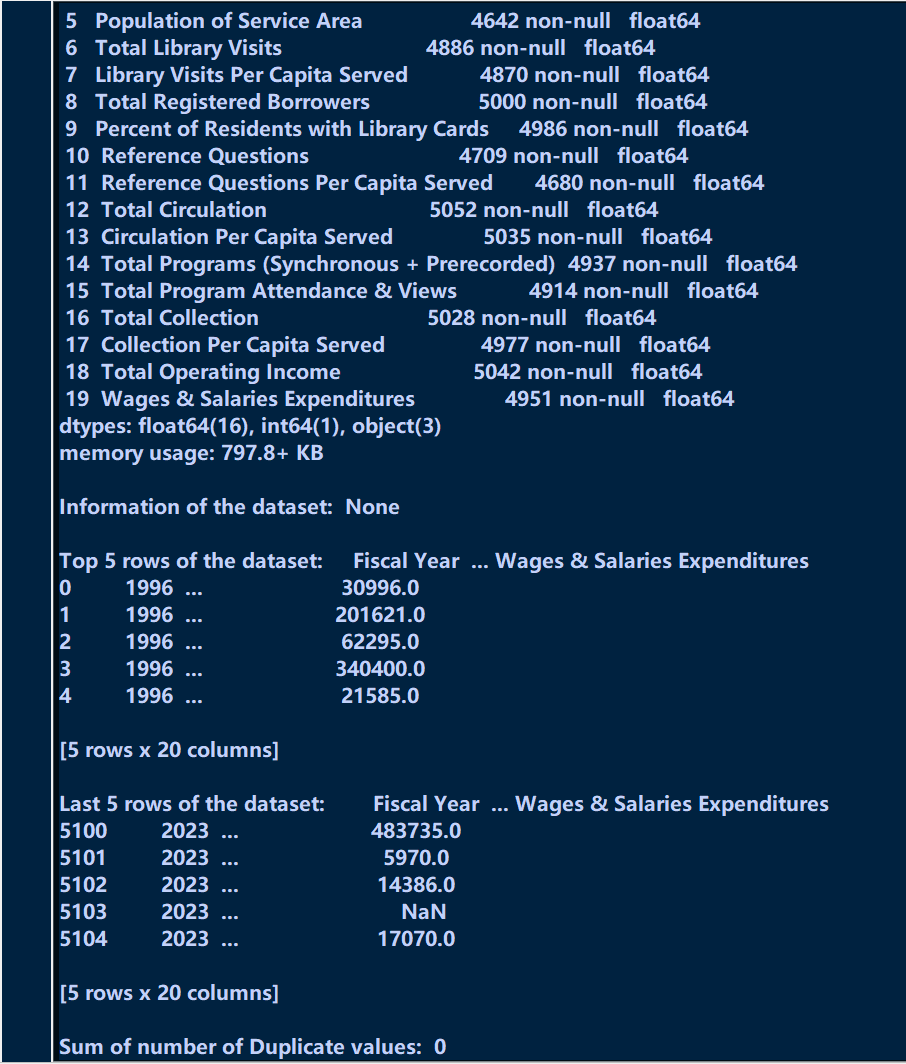
4. ANALYSIS ON DATASET:-

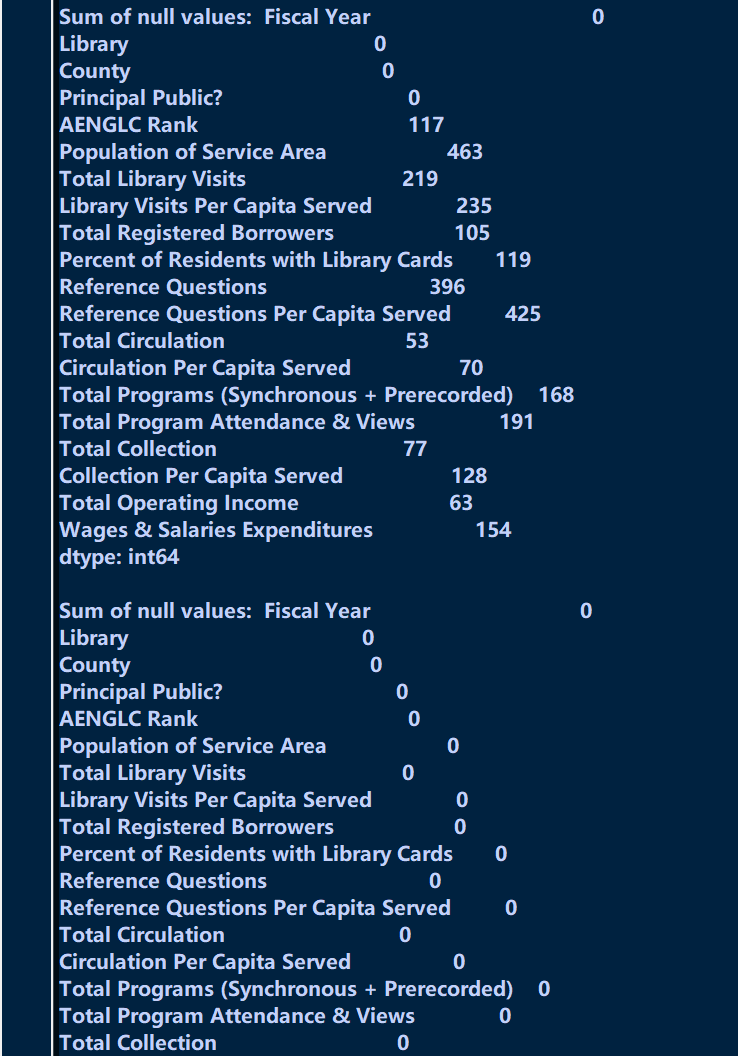
4.1 Data Overview and Cleaning

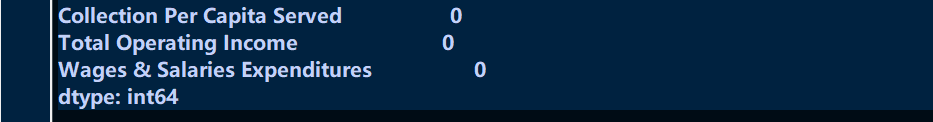
Initial inspection included printing dataset information, identifying duplicate and null values. Columns with missing data were imputed using their respective mean values. Data types were reviewed to ensure consistency for analysis.





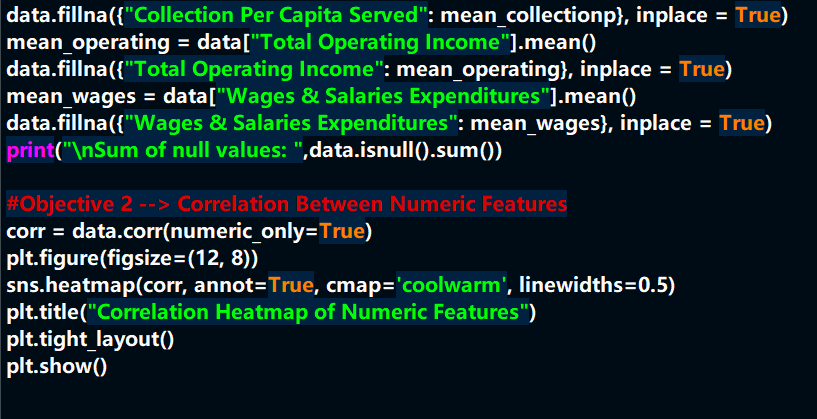


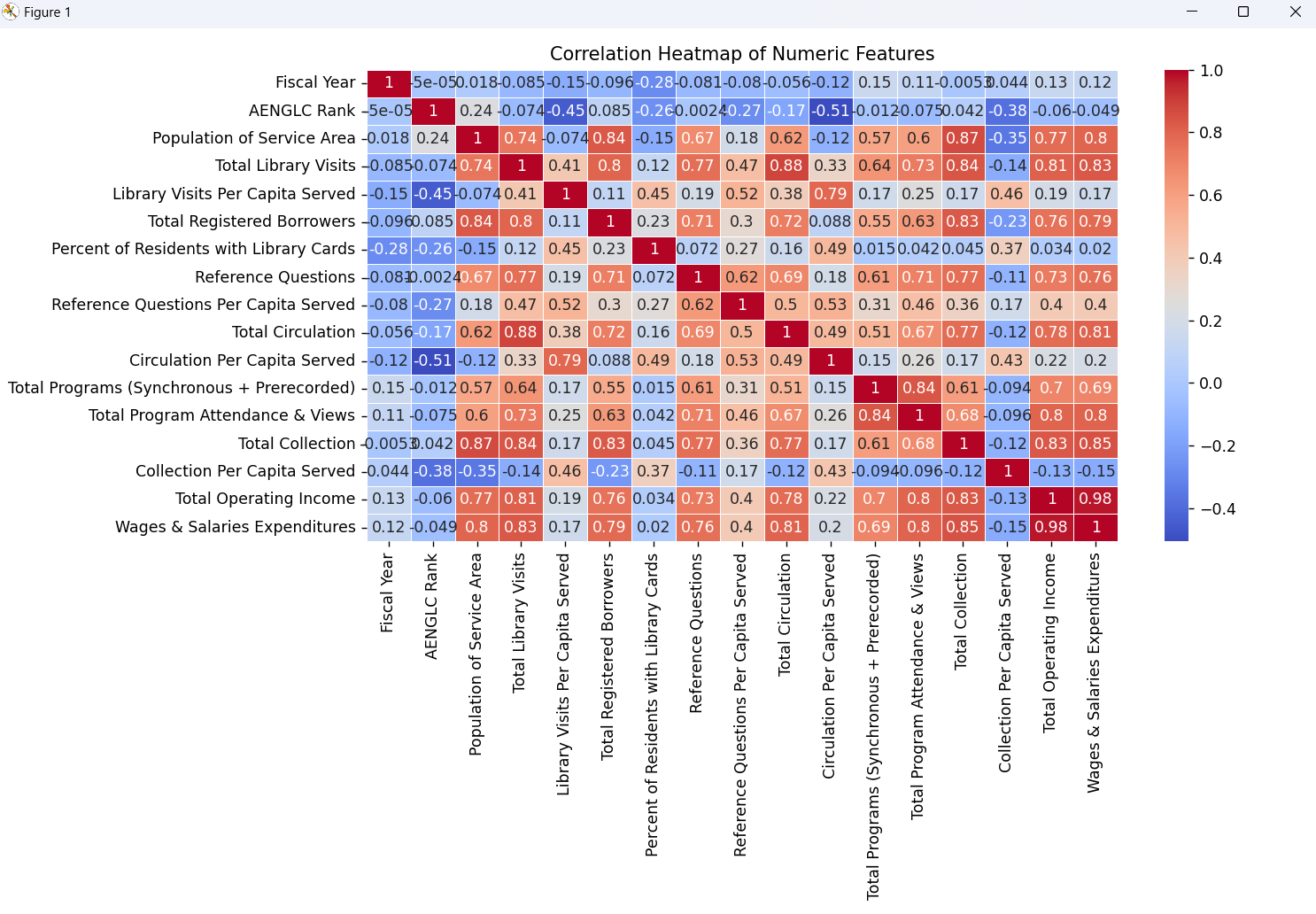




4.2 Correlation Between Numeric Features

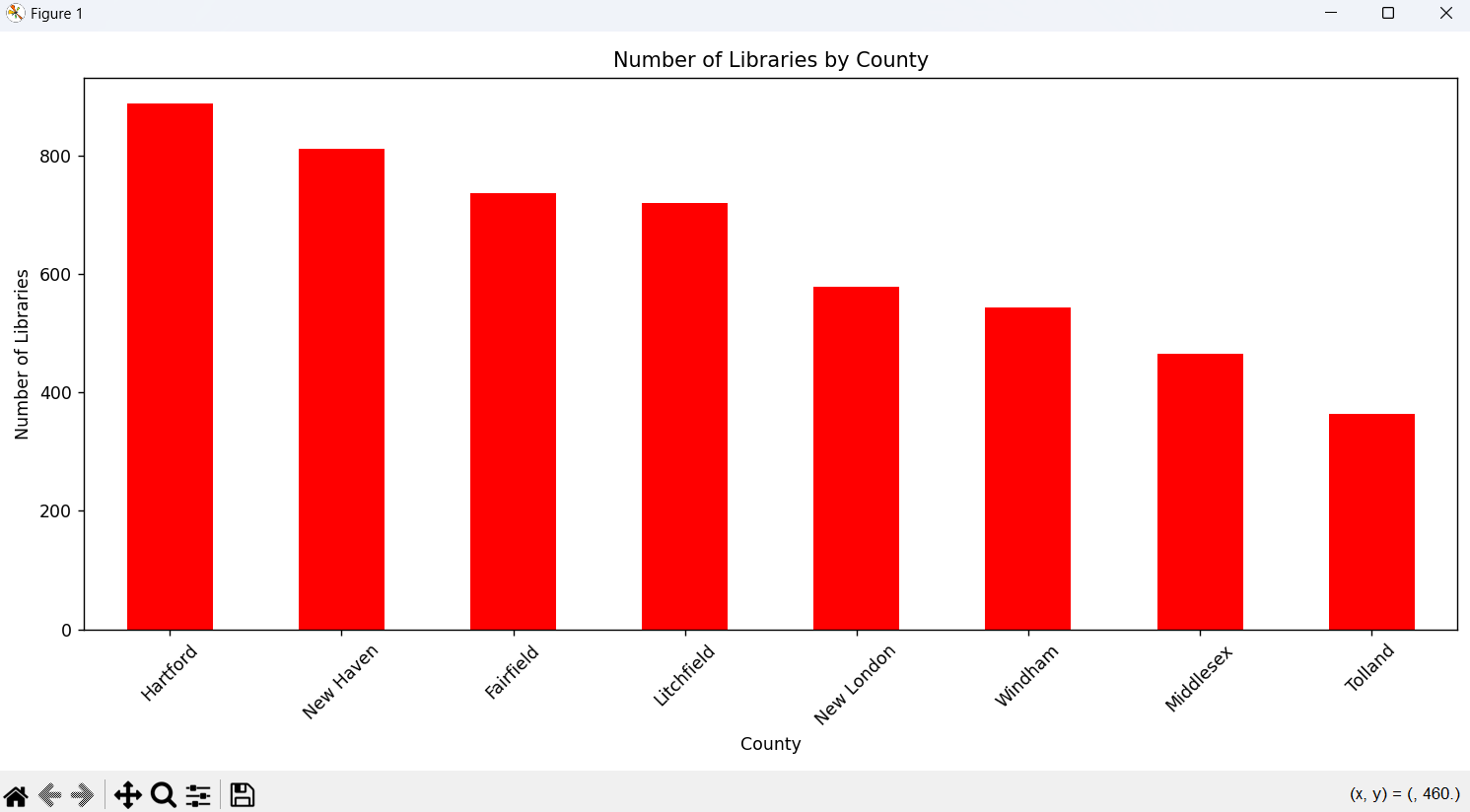
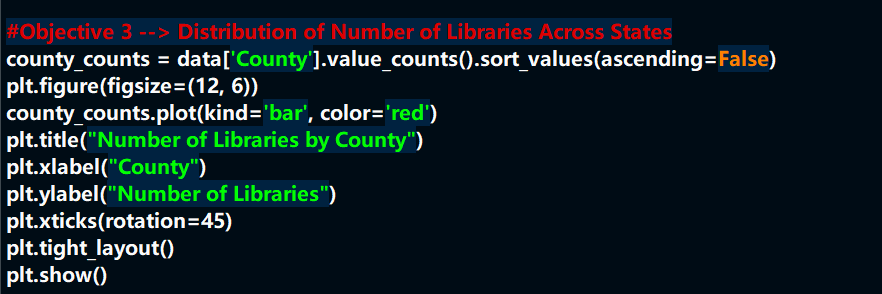
A heatmap was generated to show correlations between numeric variables like visits, collections, and income, revealing interdependencies among various library performance indicators.





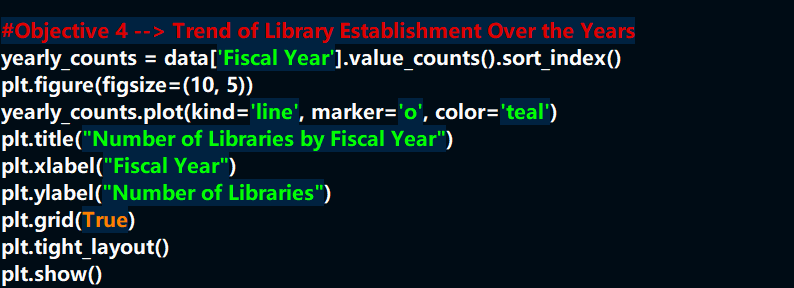
4.3 Distribution of Number of Libraries Across Counties

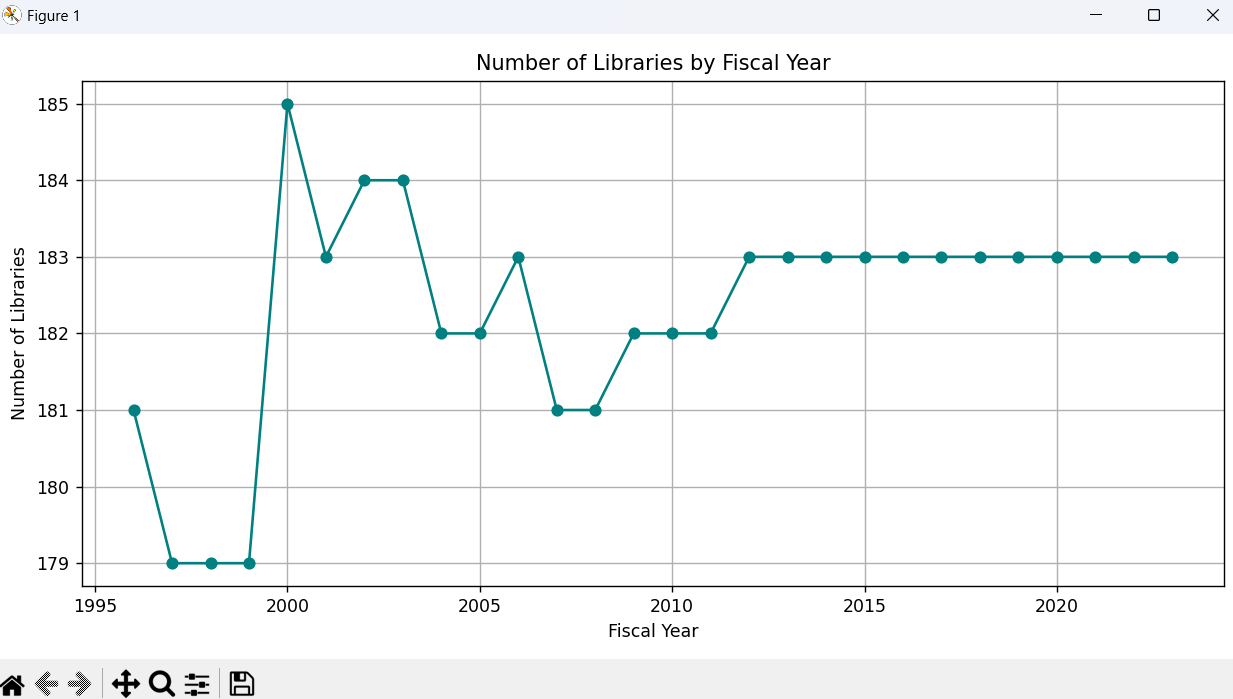
A bar plot displayed the count of libraries by county, highlighting regional distribution and operational density.



4.4 Trend of Library Establishment Over the Years

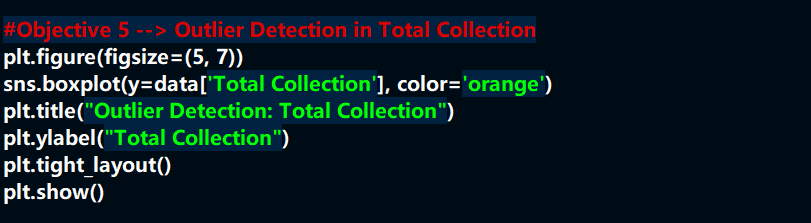
A line plot of fiscal years versus library count revealed trends in library establishment and data availability over time.

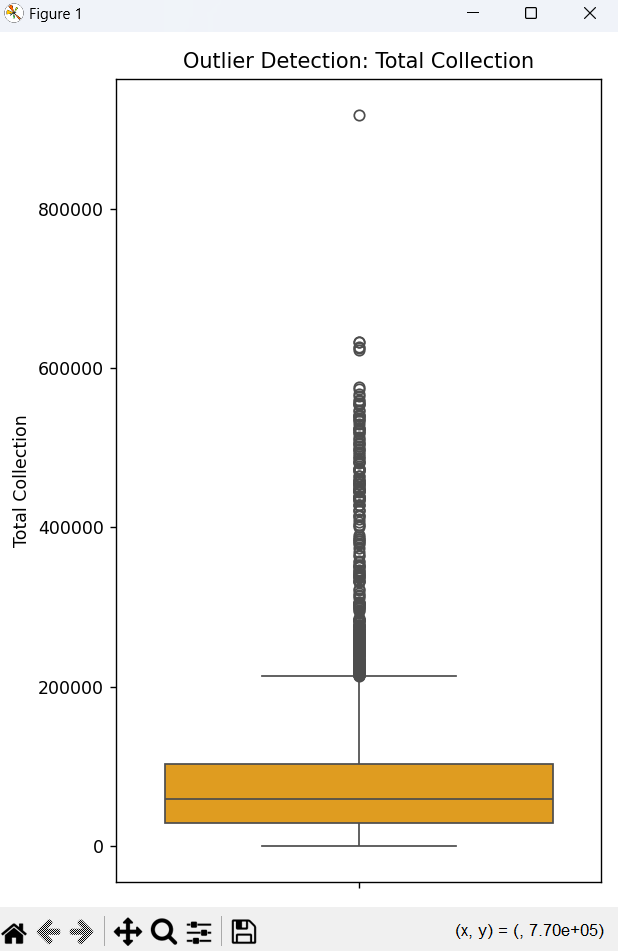




4.5 Outlier Detection in Total Collection

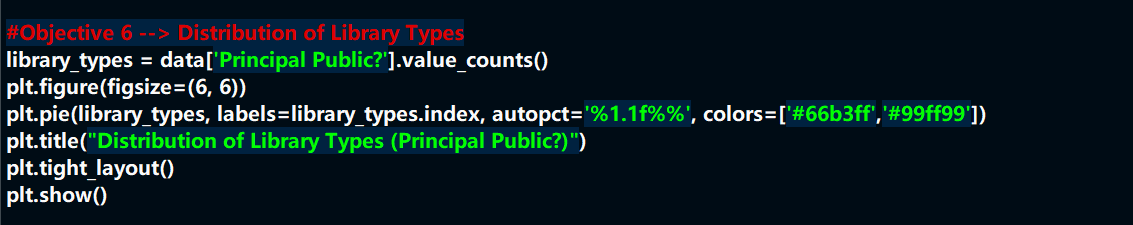
A box plot helped detect anomalies in the 'Total Collection' field, identifying libraries with exceptionally high or low collections.

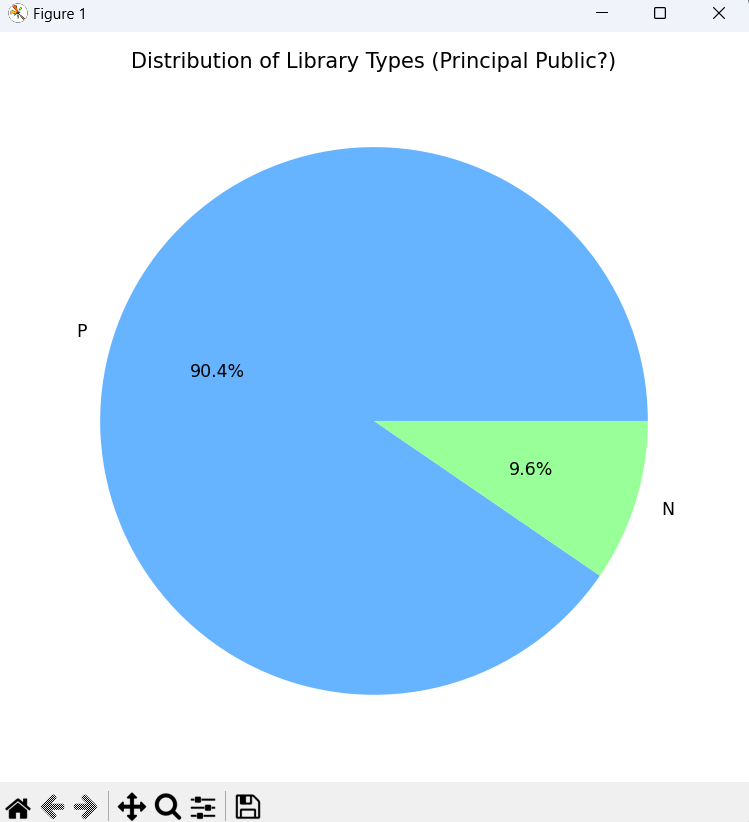




4.6 Distribution of Library Types

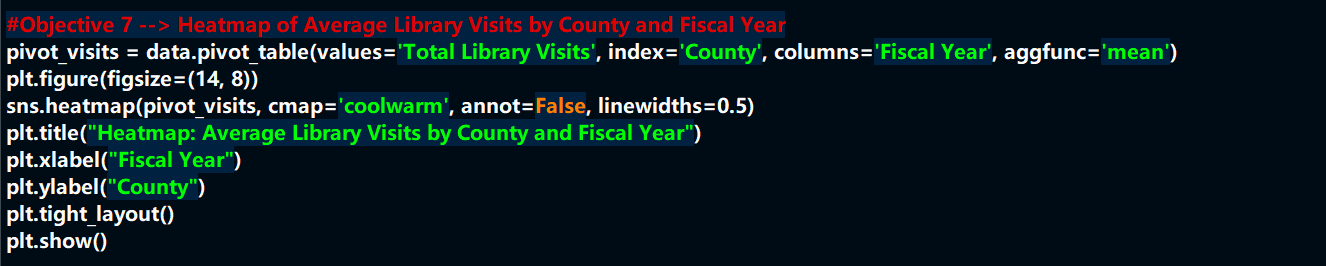
A pie chart showed the proportion of libraries marked as 'Principal Public', indicating the majority classification of libraries.

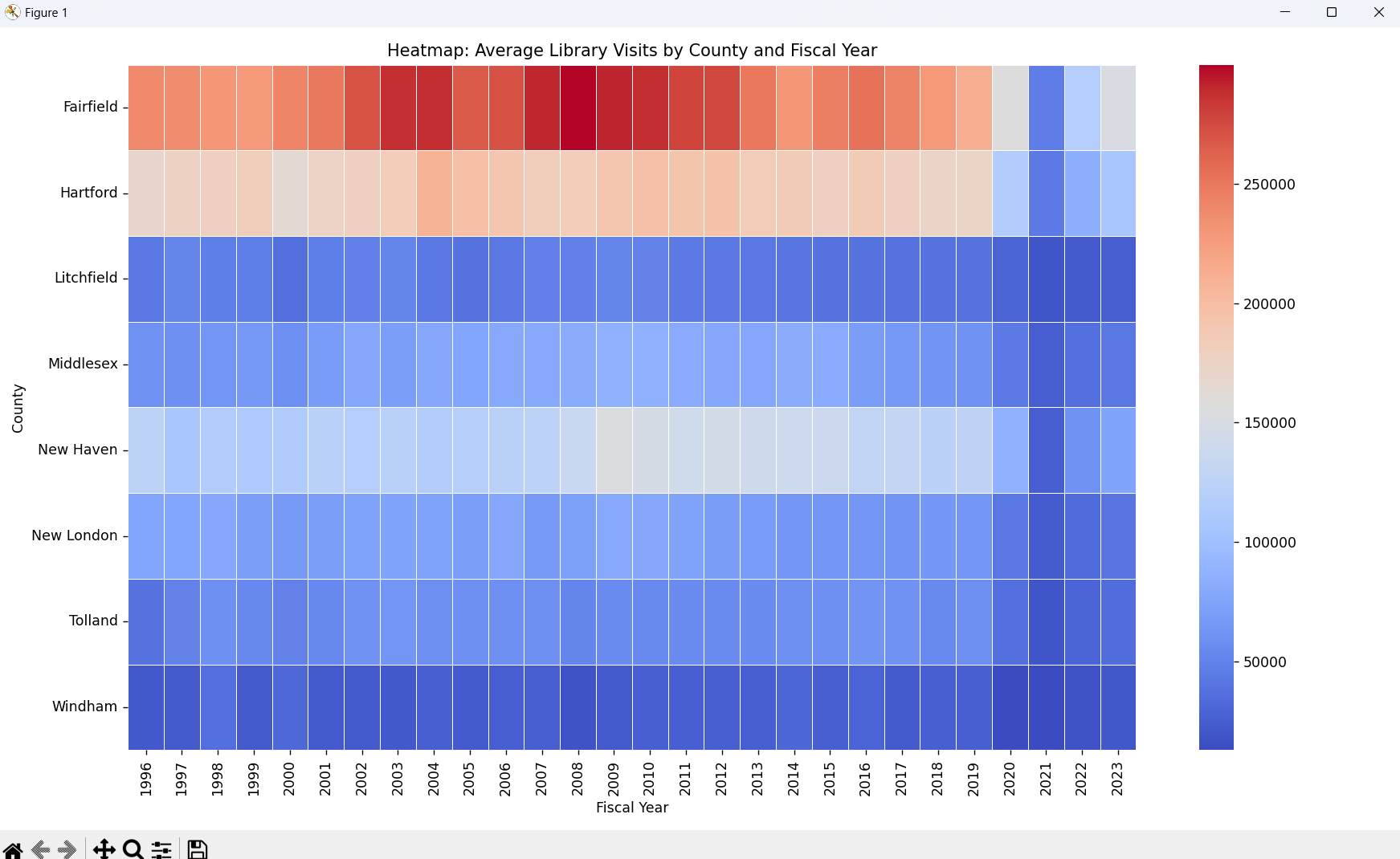




4.7 Heatmap of Average Library Visits by County and Fiscal Year

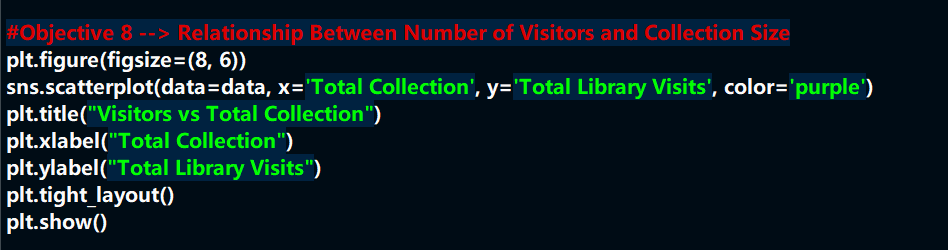
This visual compared the average number of visits across different counties and years, revealing usage trends over time.





4.8 Relationship Between Visitors and Collection Size

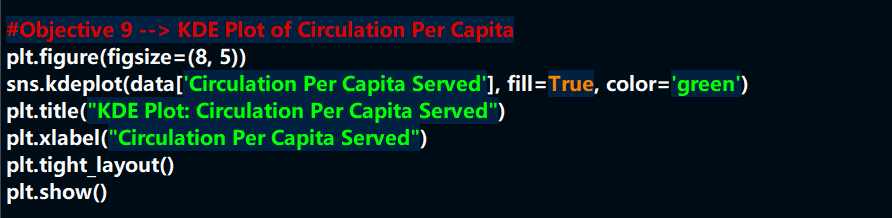
A scatter plot showed a positive correlation between the size of collections and number of visits, suggesting content availability affects footfall.

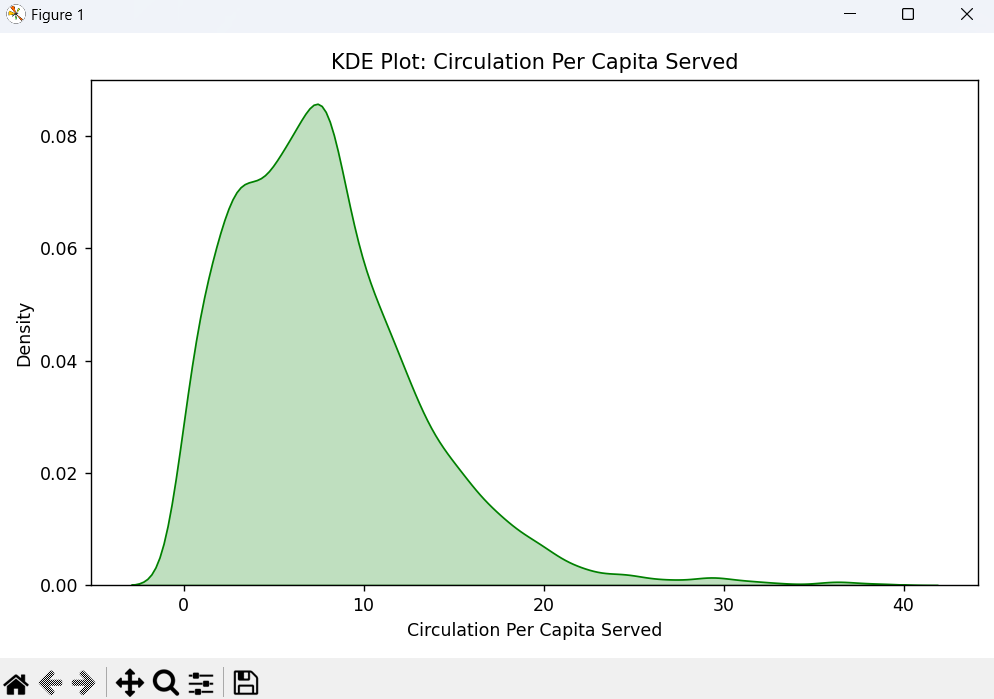




4.9 KDE Plot of Circulation Per Capita

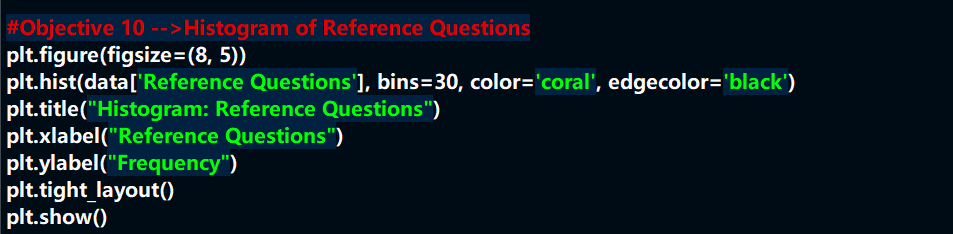
A KDE plot illustrated the distribution of circulation per capita, helping identify the most frequent usage patterns.

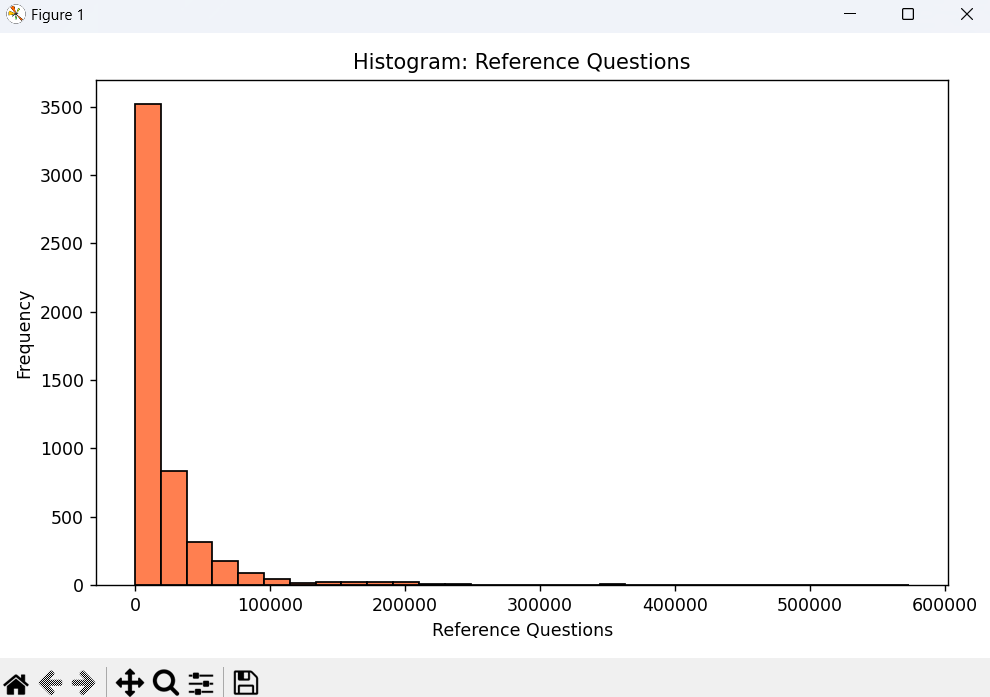




4.10 Histogram of Reference Questions

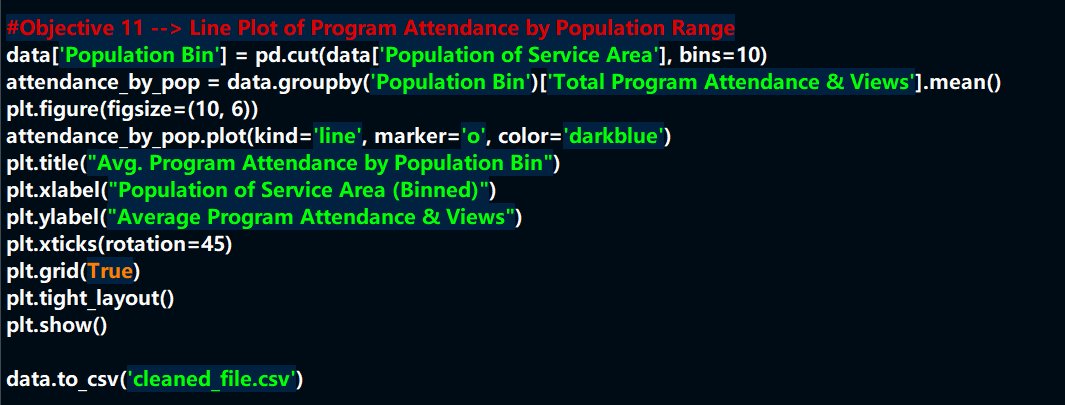
A histogram showed frequency distribution of reference questions, helping visualize user engagement in terms of inquiries.

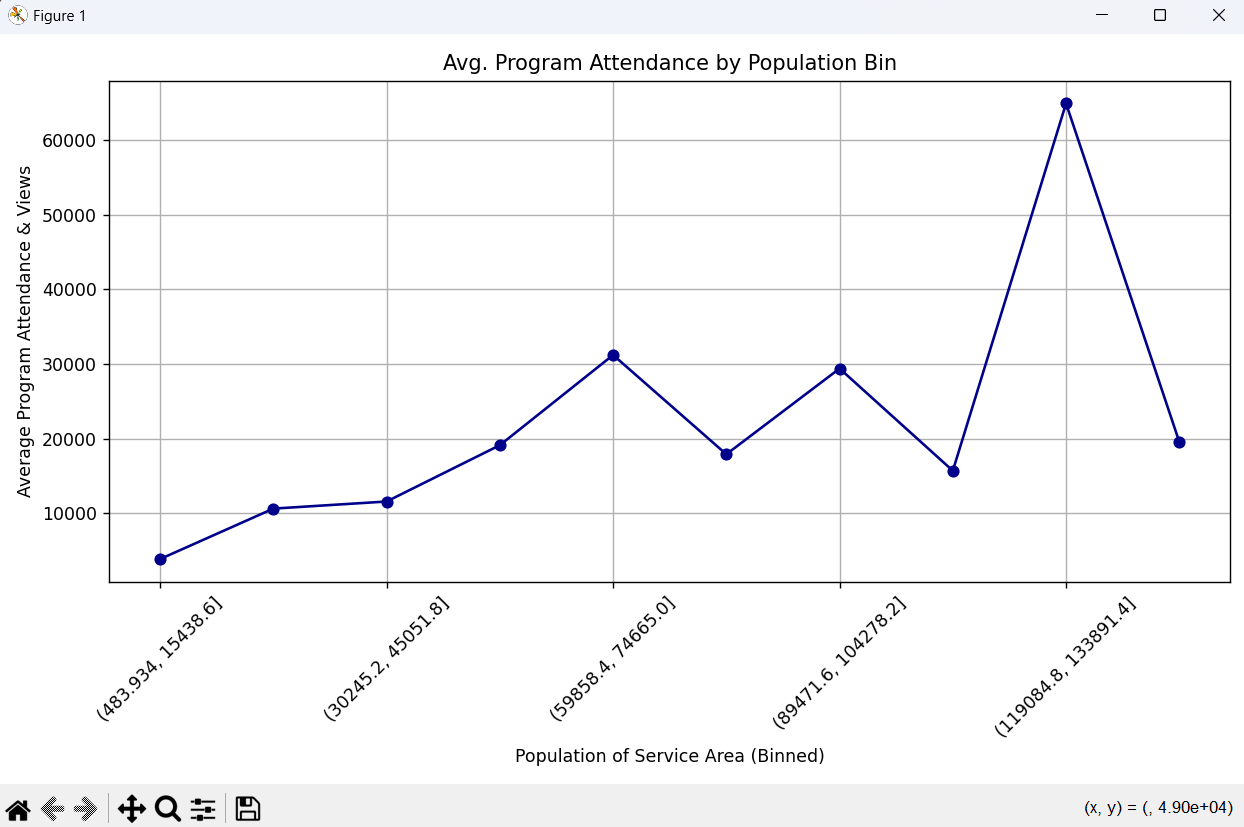




4.11 Program Attendance by Population Range

Libraries were grouped by population served. A line plot revealed how program attendance varied with community size.





5. CONCLUSION:-

This project effectively demonstrated the power of Exploratory Data Analysis using Python to gain insights into public library operations. Through visualization and statistical analysis, patterns such as regional disparities, operational efficiencies, user engagement, and service reach were uncovered. These insights can help policymakers, library administrators, and researchers to evaluate library performance and plan better resource distribution.

6. FUTURE SCOPE:-

Potential future enhancements to this project include:

- Incorporating geospatial data to analyze libraries based on geographical distribution.

- Developing an interactive dashboard using Plotly Dash or Tableau for real-time insights.

- Integrating with demographic datasets to assess service effectiveness by community type.

- Applying predictive modeling to forecast future library visits or program participation.

7. REFERENCES:-

[1] Data.gov, “Public Libraries Dataset,” [Online]. Available: <https://www.data.gov/>

[2] W. McKinney, Python for Data Analysis, 2nd ed. O’Reilly Media, 2017.

[3] M. L. Waskom, “Seaborn: Statistical Data Visualization,” Journal of Open Source Software, vol. 6, no. 60, pp. 3021, 2021.

[4] J. D. Hunter, “Matplotlib: A 2D Graphics Environment,” Computing in Science & Engineering, vol. 9, no. 3, pp. 90–95, 2007.

[5] Lovely Professional University, “Minor Project Guidelines – INT375,” 2025.