

Group 5 Project 1 Proposal: Public Library EDA

Sabrina, Malachai, Marco, Enoc

For project 1, we intend to analyze public libraries in the United States. Our inspirations for this EDA are an analysis found on Kaggle, found here: <https://www.kaggle.com/imls/public-libraries?select=libraries.csv>, as well as a gallup article done on the visitation habits in 2019, found here: <https://news.gallup.com/poll/284009/library-visits-outpaced-trips-movies-2019.aspx>.

The resources where we can find data are The Institute for Library and Museum Services, the US Census, and potentially, Google places. The Institute for Library and Museum Services is a government organization that has the mission, “The mission of IMLS is to advance, support, and empower America’s museums, libraries, and related organizations through grantmaking, research, and policy development.” The Institute has compiled data for US public libraries by state, found here: <https://www.imls.gov/research-evaluation/data-collection/public-libraries-survey>. This resource includes funding, number of libraries per state and the address for each library.

Census data provides median income and population by zip code. We intend to use the Census API to obtain information, specifically the Poverty Statistics API, which can be found here: <https://www.census.gov/data/developers/data-sets/Poverty-Statistics.html>. Both resources so far use zip code, which will be useful for the merges done for analysis.

We also have the option of using the Google Maps API for the most current database of libraries in the United States. The Institute mentioned above only has information available until 2019.

Some of our initial thoughts on questions that we would like to ask of our data are:

- What is the median income for each zip code libraries are located in?
- What is the average number libraries are usually in each zip code?
- What is the average population a single library serves?
- What is the average funding for each library?

Some of our initial thoughts on questions for regression and correlation are:

- Is there a correlation between income and presence of libraries?
- Is there a correlation between population size and libraries?
- Is the correlation between availability of libraries and state?
- Is there a correlation between expenditure on physical books and e-books?

Our use of color to display our data will be for contrast on our variables. We have decided to use the following color palette (also available below): #003f5c, #58508d, #bc5090, #ff6361, #ffa600. We are focused on contrast and correlation, rather than time trends, so using contrasting colors will be beneficial, as opposed to a color gradient. Additionally, we do not have any increase/decrease or “good” / “bad” data, which would be more typical of financial analysis, so we decided to stay from the green/red color paradigm, so we don’t impose any moral weight to our data visualizations.

