Exploratory Data Analysis on Use of Public Libraries by Local United States Populations

By Sabrina Alves

Malachai Cravens

Enoc Serge Kouegbe

Marco Lopez

Table of Contents

I. Decompose the Ask

Our main focus for this project was the question: Is there a correlation between use of a library and the demographics of the area? It was our intention to determine if demographic metrics recorded for an area can be used to predict use of local libraries. To narrow the scope of our question, we sought out two variables to define each of the parts in the question.

Public libraries are used by those residing in the same county (and sometimes even in relevant zones or in the general zip codes outside of the county) for various purposes, such as computer and internet use, research materials, and leisure reading, and through various forums, such as in person visitation and online apps. To determine use, we decided on registered users, or those that hold library cards and proved to be a member of the community the library supports, and total collection expenditures. Most programs require possession of a library card to use services from the library, so this metric was determined to be a good one to measure the percentage of the community who utilize the library to some degree. Total collection expenditure was chosen because of its two pronged focus: it exposes how much money is made available by local funding bodies, as well as shows the amount of money spent catering to the needs of the community. For example, when a new release is requested by multiple registered users, the library will purchase more copies of the release.

Determining the demographics of the area is a slightly more nebulous task. In order to retrieve the demographics of the area, a common identity needs to be shared between the library data and census data. Because requirements for membership to a public library differs between each library, it was decided that zip code was an appropriate population to “belong” to each library.

“Demographics” is the next term to be defined. Race, age, education and income levels are demographics captured by the American Community Survey that can be selected for demographics. Median household income and education level held by adults over the age of 25 were the compelling metrics for this project– does the income level of the community influence use of the library? Does a certain threshold (as percentage of population) of education mean more or less visits to the library?

II. Identify Data Sources

Data was retrieved from two sources: The Institute of Museum and Public Library Services[1] and the United States American Community Survey[2].

The Institute of Museum and Public Library Services (ILMS) is a government organization that provides information on public libraries in the United States through the Public Libraries Survey. ILMS describes the , “The Public Libraries Survey (PLS) examines when, where, and how library services are changing to meet the needs of the public. These data, supplied annually by public libraries across the country, provide information that policymakers and practitioners can use to make informed decisions about the support and strategic management of libraries.”[1] The survey is provided in CSV files.

The American Community Survey is an ongoing survey where data is collected each year to give “communities the current information they need to plan investments and services.”[2] The 5-year survey, as opposed to the 1-year survey, provides data by zip code and benefits from “the increased statistical reliability of the data for less populated areas and small population

[1] https://www.imls.gov/research-evaluation/data-collection/public-libraries-survey

[2] https://www.census.gov/data/developers/data-sets/acs-5year.html

subgroups," due to its multiyear estimates.[3] Data can be retrieved from the Survey using the Census API wrapper.[4]

III. Define Strategy and Metrics

The following programs are used for data retrieval, clean up, and analysis:

- Jupyter Notebook
- Pandas
- Census Wrapper
- Matplotlib
- Seaborn
- Numpy
- Statsmodel.api
- Sklearn.linear_model

To answer our research question, analysis was broken down into 4 sub-questions, to be distributed to each member of the team:

- Is the amount of money spent on the library collection affected by income in the area?
- Is the amount of money spent on the library collection correlated to the education level of the population?
- Is there a correlation between the education level of the population and registered users in the library?
- Is there a correlation between income level of the population and the amount of registered users?

[3] https://www.census.gov/data/developers/data-sets/acs-5year.html
[4] https://github.com/CommerceDataService/census-wrapper

IV. Build Data Retrieval Plan

Both ILMS and American Community Survey contained reliable data from 2015, so data will be retrieved from this year.

ILMS data was found in CSVs on their website. The CSV was pulled into Jupyter Notebook for cleaning.

In order to retrieve data from the American Community Survey, a census wrapper and an API key were used. One member will push the request and export data into a csv file, to be cleaned in a separate workbook. In order to pull the correct columns, IDs were retrieved from a wrapper resource.[5] Relevant data, such as income and education, and interesting auxiliary data, such as median age and median home value, were requested.

V. Retrieve the Data

We pulled ILMS data from the CSV file provided and kept most columns relevant to the library location. We also pulled money spent on collection, registered users, and library visits for analysis.

Our data contained certain columns we wished to use in our exploration, we kept all the columns related to education, from no high school diploma all the way to a Doctorates degree as long as it was an adult over the age of 25, so our data does not pertain to kids nor young adults. We also kept the Median Household Income column for our population demographic by zip code as we want to know if a library's collection or their registered users has any correlation to a population's income.

[5] https://gist.github.com/afhaque/60558290d6efd892351c4b64e5c01e9b

The American Community Survey request pulled the following columns for all zip codes in the United States:

Total Population
Median Household Income
Per Capita Income
Median Age
Median Gross Rent
Median Home Value
No Education Degree Held by Adults Over The Age of 25
High School Degree Held by Adults Over The Age of 25
GED Degree Held by Adults Over The Age of 25
Associates Degree Held by Adults Over The Age of 25
Bachelor’s Degree Held by Adults Over The Age of 25
Master’s Degree Held by Adults Over The Age of 25
Professional Degree Held by Adults Over The Age of 25
Doctorate Degree Held by Adults Over The Age of 25

The responded data was then converted to a CSV file for cleaning and analysis.

VI. Assemble and Clean

Null Values

Null values in ILMS columns of interest– total collection expenditures and registered users– were dropped. There was also one row in the data set that contained “R_14” and “R_18” string values that could not be converted into float values. We did a mask to filter out this row so we could use the float values for our statistical calculations.

Null values in census columns of interest– Total Population, Median Household Income, and Education Level Columns– in the census data were dropped, as analysis could not be performed on these rows without this information. The census data did not have null values but did have values such as -666666666.0 which were understood to be null values and dropped.

File Merge

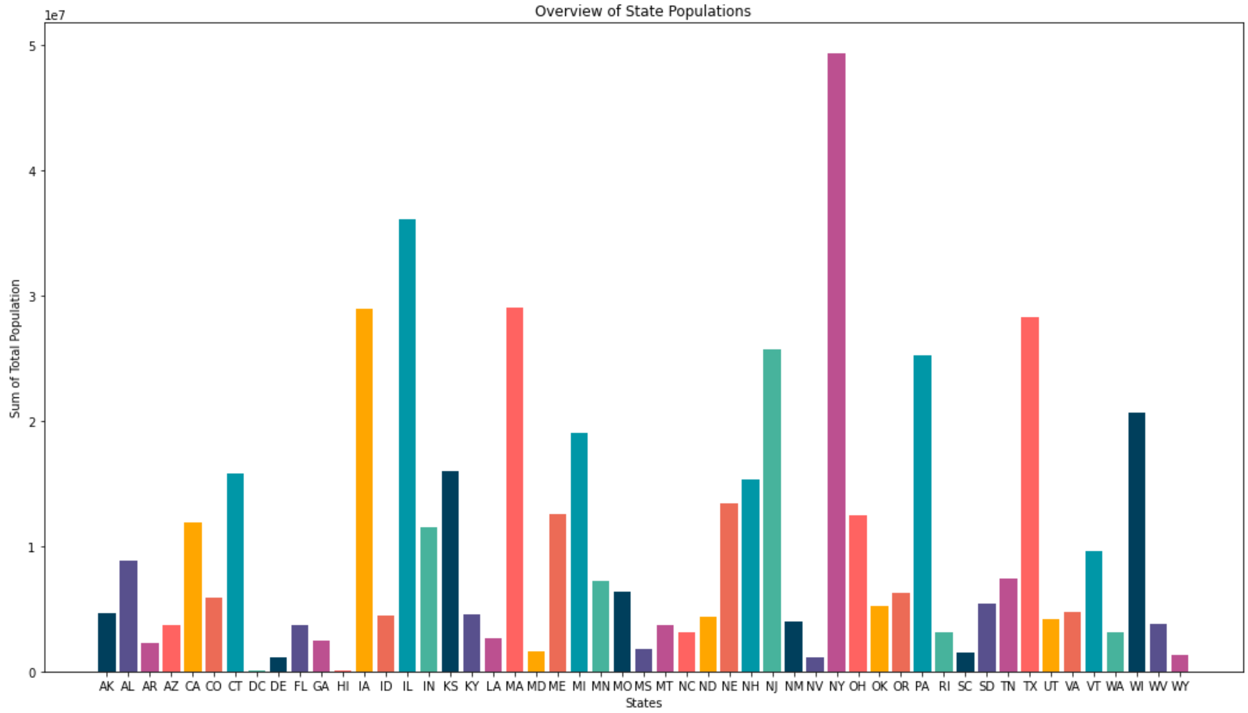
ILMS and census data were merged together on zip codes. Census rows that contained zip codes that did not pair were dropped. Likewise, libraries that had zip codes with no pair were also dropped.

Select columns were used from the census data for the research questions chosen for this wave of analysis. Those columns were:

Total Population
Median Household Income
No Education Degree Held by Adults Over The Age of 25
High School Degree Held by Adults Over The Age of 25
GED Degree Held by Adults Over The Age of 25
Associates Degree Held by Adults Over The Age of 25
Bachelor’s Degree Held by Adults Over The Age of 25
Master’s Degree Held by Adults Over The Age of 25
Professional Degree Held by Adults Over The Age of 25
Doctorate Degree Held by Adults Over The Age of 25

Total Population

Total population is needed for equalizing our variables for comparison. Below is a bar chart illustrating the distribution of populations by state (because zip codes, our population unit, can not be visualized easily as a categorical variable) in the areas that contain public libraries. New York, which is a leading state in terms of population count generally, has the largest population serviced by public libraries. Washington DC and Hawaii have the lowest counts of population serviced by public libraries.

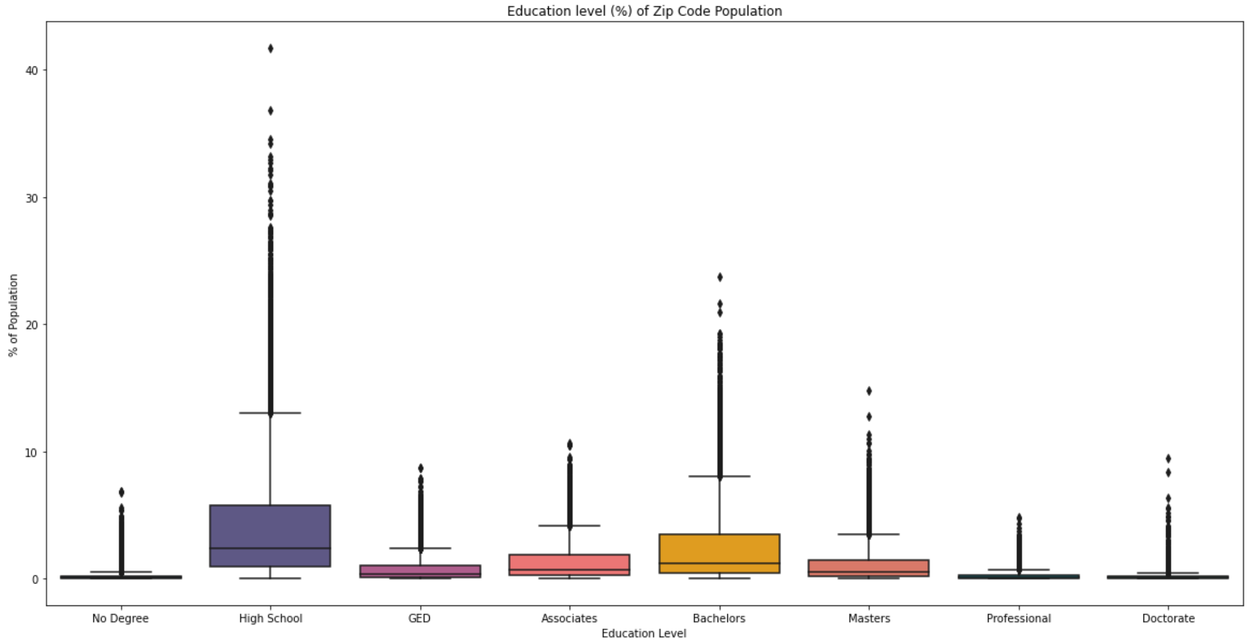


Education Level

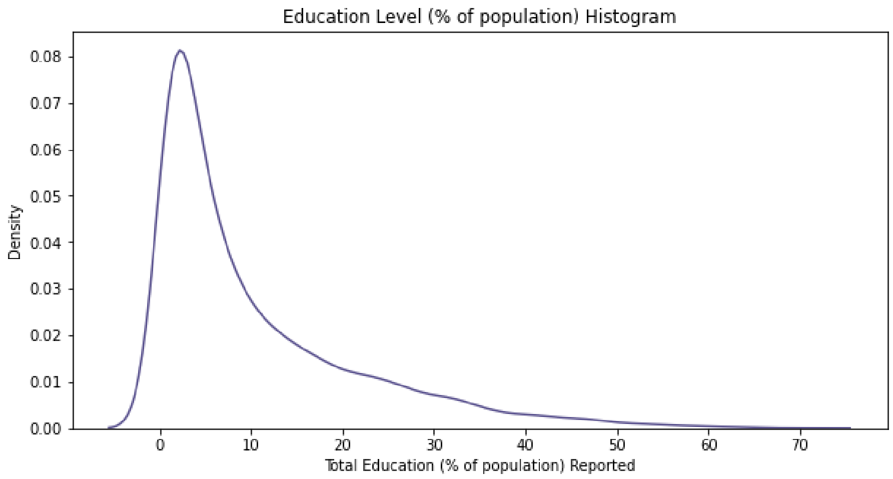
Education level of adults over the age of 25 is provided as a count of adults in the category by the American Community Survey. In order to accurately compare populations with each other, these counts were divided by population count and multiplied by 100 in order to get a percentage of education level in the zip code. The following visualizations and comments are based on these percentages.

Though the data has lower and upper bound outliers, outliers have not been removed from the data set. Our questions are general in nature and ask if utilization of public libraries can be predicted with recorded demographics of the area. The outliers are justified.

Below is a box plot visualizing the distribution of education data. Recorded education levels remain low– less than 10% of the population is recorded for each education level the majority of the time– and outliers are mainly in the upper bounds in the data.

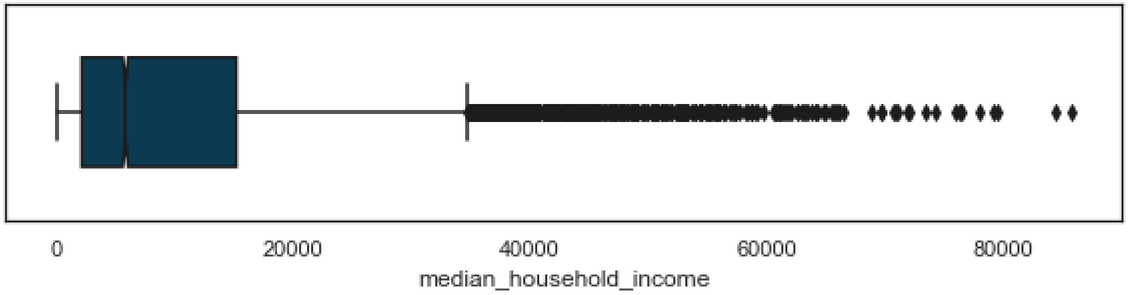


Below is a histogram of the total recorded data for each zip code. When the percentages of each education level is added together, most zip codes report 0-20% of the population. We would like to see this histogram left skewed (most of the data reporting in, at minimum, between 70 and 80%).



Median Household Income

We will keep our full range of median household income. We searched for any negative values reported and made sure they were removed. But we want to consider income in the wide range that is reported and we will then see if we can categorize it by amount further in our exploration. The box plot shows our wide range of our income; this may cause obscurity if left broken apart individually. As you can see our mean value is in our lower $100K but we didn't want to exclude the higher class as they are likely to be a part of a library's source of funding.

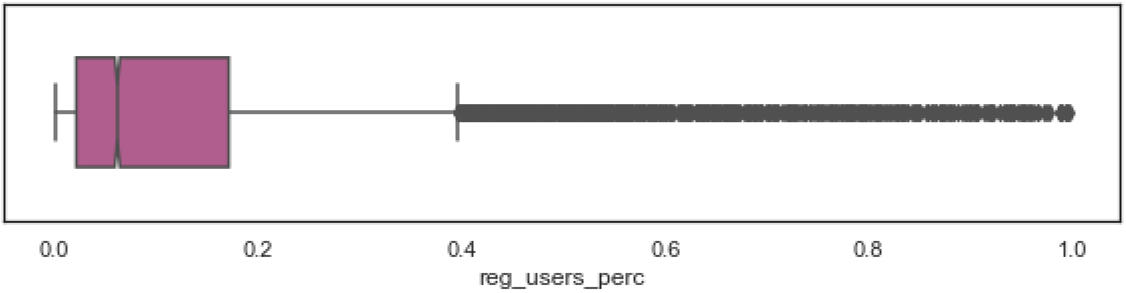


Total Collection Expenditure

One of our questions we wanted the data to answer was the amount of money spent on the library collection correlated to the education level of the population? On average, most of the linear regression analysis graphs in correlation of total collection expenditures vs education levels had the same representations. While we had positive linear regression directions, there were no strong correlations between total collection expenditures in relation to education levels.

Registered Users

We wanted to generate a percentage of the reported registered user by the reported population of the zip code. We had to watch out for some of the libraries that reported a negative number for the number of the registered users. As well as some zip codes reported a higher registered user count than what was counted in the total population. These high and low outliers were removed from our data and the following plot shows how we still have many outliers in our data as we are still comparing our entire US country of percent of registered users by population of each zip code. Again our percentage shows many outliers away from our mean line however we did not want to take away a population from our exploration so we will keep these outliers in our exploration.

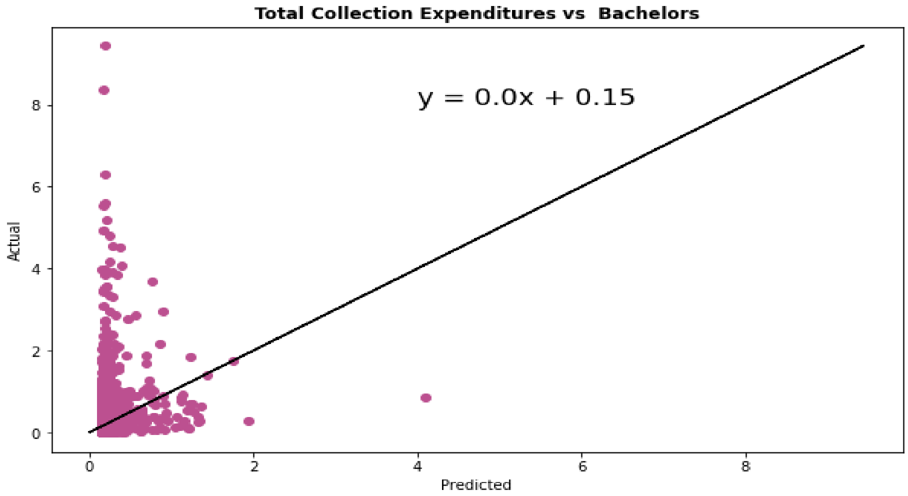


VII. Analyze for Trends

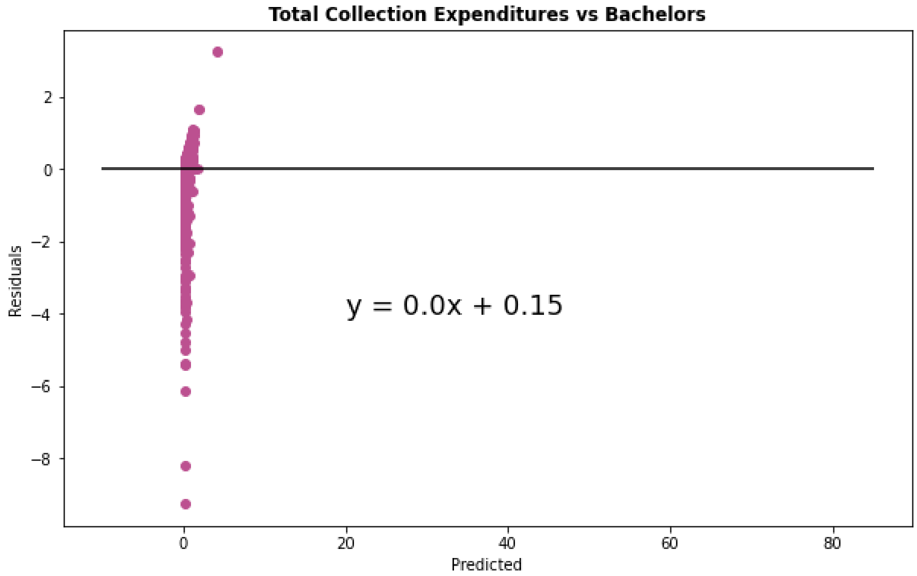
Education Level vs. Total Collection Expenditure

The trend we mostly observed on all education levels vs total collection expenditures was that there was not a strong correlation between two. As depicted in the graphs, they do not demonstrate any strong bearing resemblance to any correlations between the two data sets. In addition to support our findings of these trends, we found R squares were small in relation between the two data sets. The Kurtosis calculations were high in relation to the education levels in contrast to the total collection expenditures, which means there are no normal curves between the two data sets. Furthermore, in our predictions-actual and residual graphs, we clearly see there was no heteroscedasticity, we had the opposite, which was homoscedasticity where the variance error was the same across the values of any independent variable.

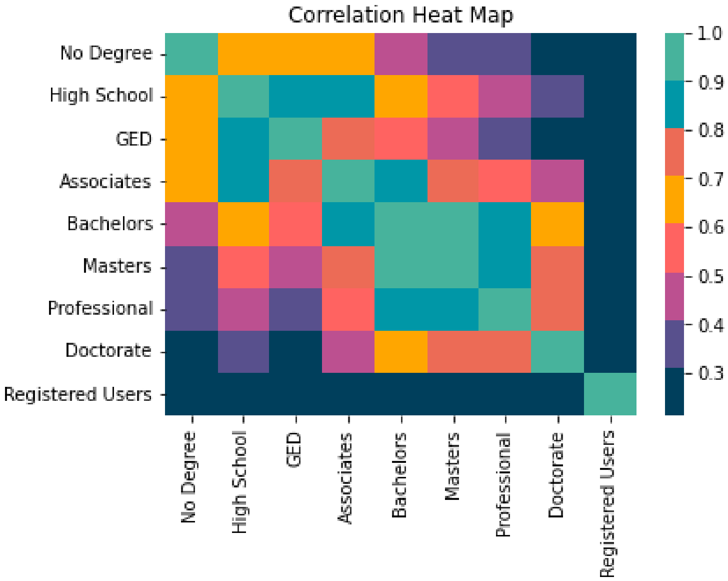
Example of an average linear regression education levels (Bachelors in this case) in relation to total collection expenditures.



Example of an average predictions-actual, residual graph education levels (Bachelors in this case) vs total expenditures, where homoscedasticity is depicted

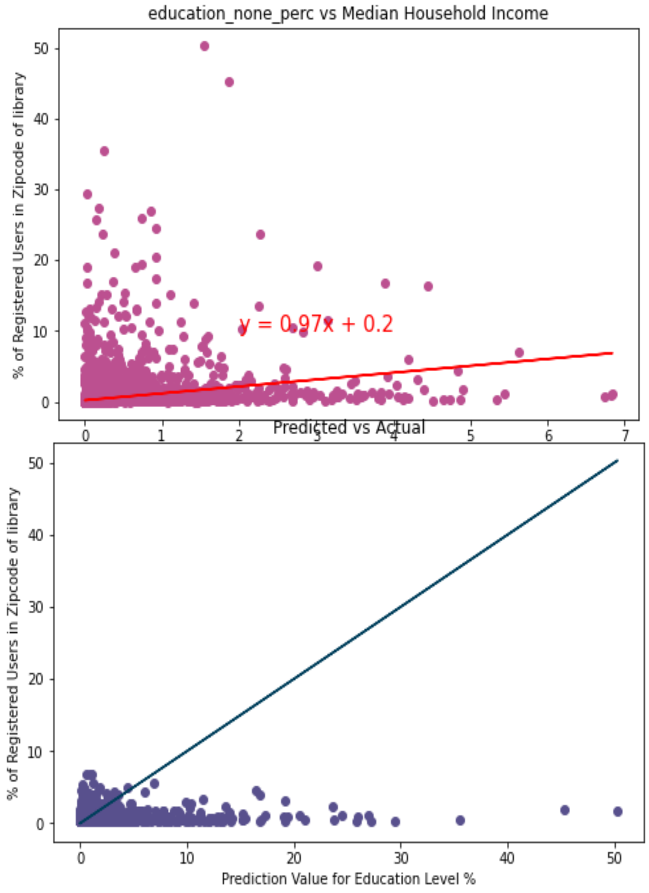


Education Level vs. Registered Users

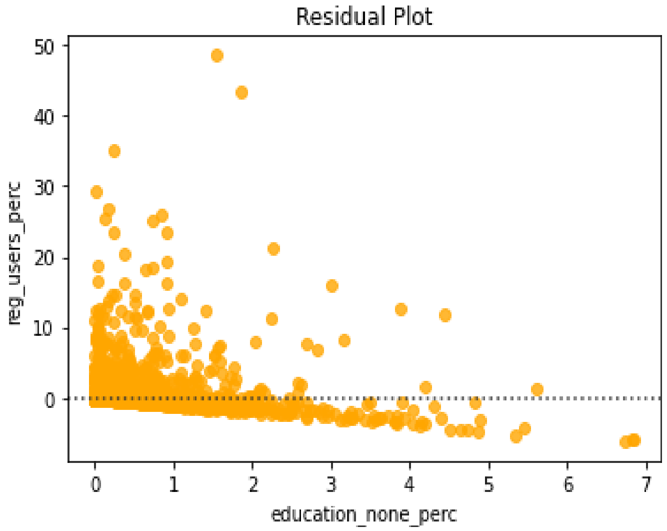


Both variables analyzed are percentages of the population of their zip codes. First, after outlier analysis was performed, linear regression and t-tests were performed to see if there was a correlation between the two variables. The heat map to the right summarizes our findings well– there is very little correlation between the percentage of the population at a certain education level and the percentage of the population with library cards. For example, the correlation of percentage of population that holds a bachelor's degree with percentage of registered users has an r-squared coefficient of less than 0.3, which implies weak or no correlation.

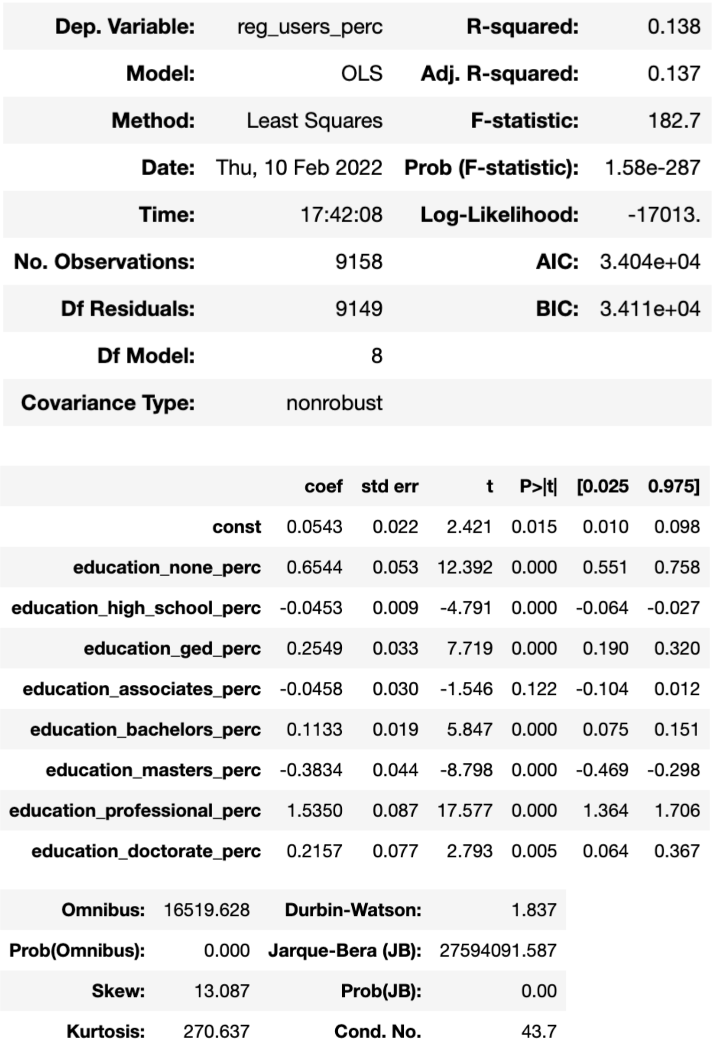
Our linear regressions corroborate the heatmap. To the right, the linear regression done for No Degree held percentages and registered users is displayed. The line of best fit in the top most chart does not obviously capture the points of data. This is further proven in our "Predicted vs. Actual" plot (middle chart, in purple), which shows the predicted data clustered in the left most quadrant of the trace and plotted actual data far from the predictions. The residual plot (in yellow, right) is further proof. The data lacks a symmetrical distribution. This means that the line of best fit can not be used for predicting the percent of adults over the age of 25 with no education within our data



The r-squared value from the statsmodel below shows the weak correlation with an r-squared value of 0.138.



The t-tests performed reject our null hypothesis but were performed on skewed data. One of the assumption t-tests are performed under is normality, and this data has significantly right skewed. In the image below, the skew and kurtosis are significantly large.

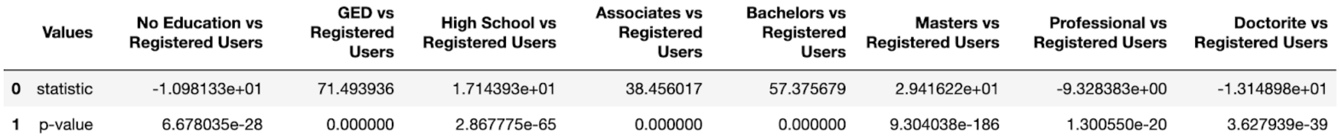


Dep. Variable:	reg_users_perc	**R-squared:**	0.138
Model:	OLS	**Adj. R-squared:**	0.137
Method:	Least Squares	**F-statistic:**	182.7
Date:	Thu, 10 Feb 2022	**Prob (F-statistic):**	1.58e-287
Time:	17:42:08	**Log-Likelihood:**	-17013.
No. Observations:	9158	**AIC:**	3.404e+04
Df Residuals:	9149	**BIC:**	3.411e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	0.0543	0.022	2.421	0.015	0.010	0.098
education_none_perc	0.6544	0.053	12.392	0.000	0.551	0.758
education_high_school_perc	-0.0453	0.009	-4.791	0.000	-0.064	-0.027
education_ged_perc	0.2549	0.033	7.719	0.000	0.190	0.320
education_associates_perc	-0.0458	0.030	-1.546	0.122	-0.104	0.012
education_bachelors_perc	0.1133	0.019	5.847	0.000	0.075	0.151
education_masters_perc	-0.3834	0.044	-8.798	0.000	-0.469	-0.298
education_professional_perc	1.5350	0.087	17.577	0.000	1.364	1.706
education_doctorate_perc	0.2157	0.077	2.793	0.005	0.064	0.367

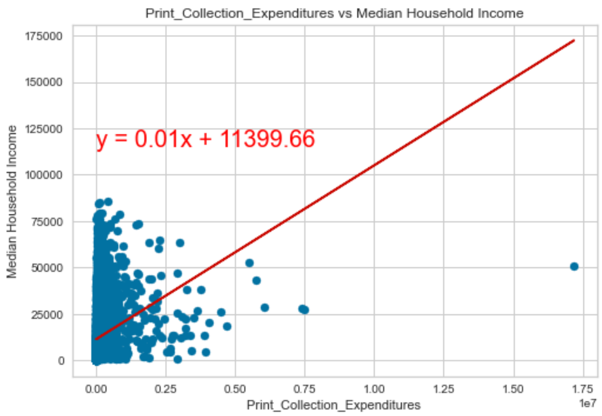
Omnibus:	16519.628	**Durbin-Watson:**	1.837
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	27594091.587
Skew:	13.087	**Prob(JB):**	0.00
Kurtosis:	270.637	**Cond. No.**	43.7

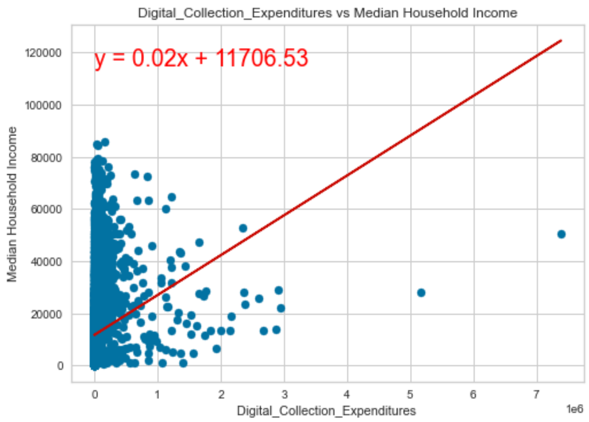
When the null hypothesis is: "Local education level has no effect on registered users (% of population)." and the alternative hypothesis is: "Local education level has an effect on registered users (% of population)." in our paired t-test, we are forced to reject the null hypothesis despite not being able to say there is a correlation between the two variables. Below is a table of the p-values generated from the paired t-tests, with values in each of the tests less than 0.01.

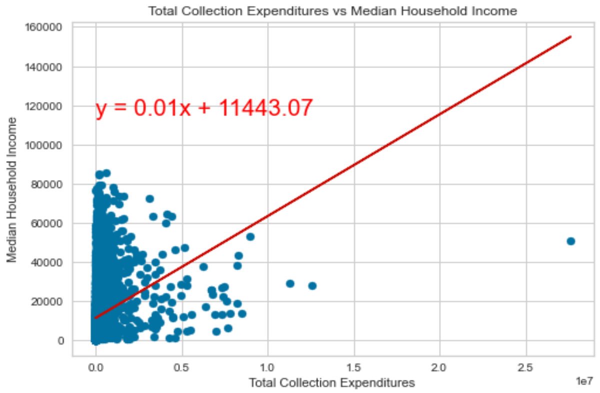


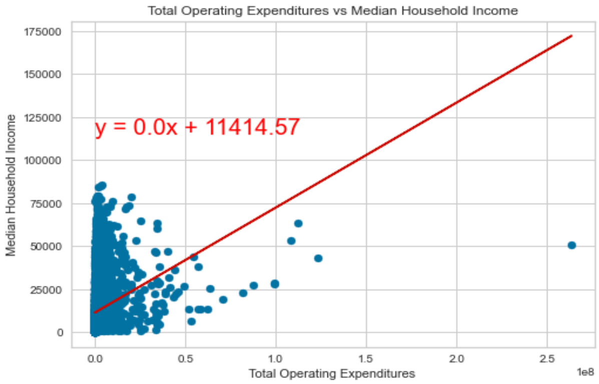
	Values	No Education vs Registered Users	GED vs Registered Users	High School vs Registered Users	Associates vs Registered Users	Bachelors vs Registered Users	Masters vs Registered Users	Professional vs Registered Users	Doctorite vs Registered Users
0	statistic	-1.098133e+01	71.493936	1.714393e+01	38.456017	57.375679	2.941622e+01	-9.328383e+00	-1.314898e+01
1	p-value	6.678035e-28	0.000000	2.867775e-65	0.000000	0.000000	9.304038e-186	1.300550e-20	3.627939e-39

Median Household Income vs. Total Collection Expenditure





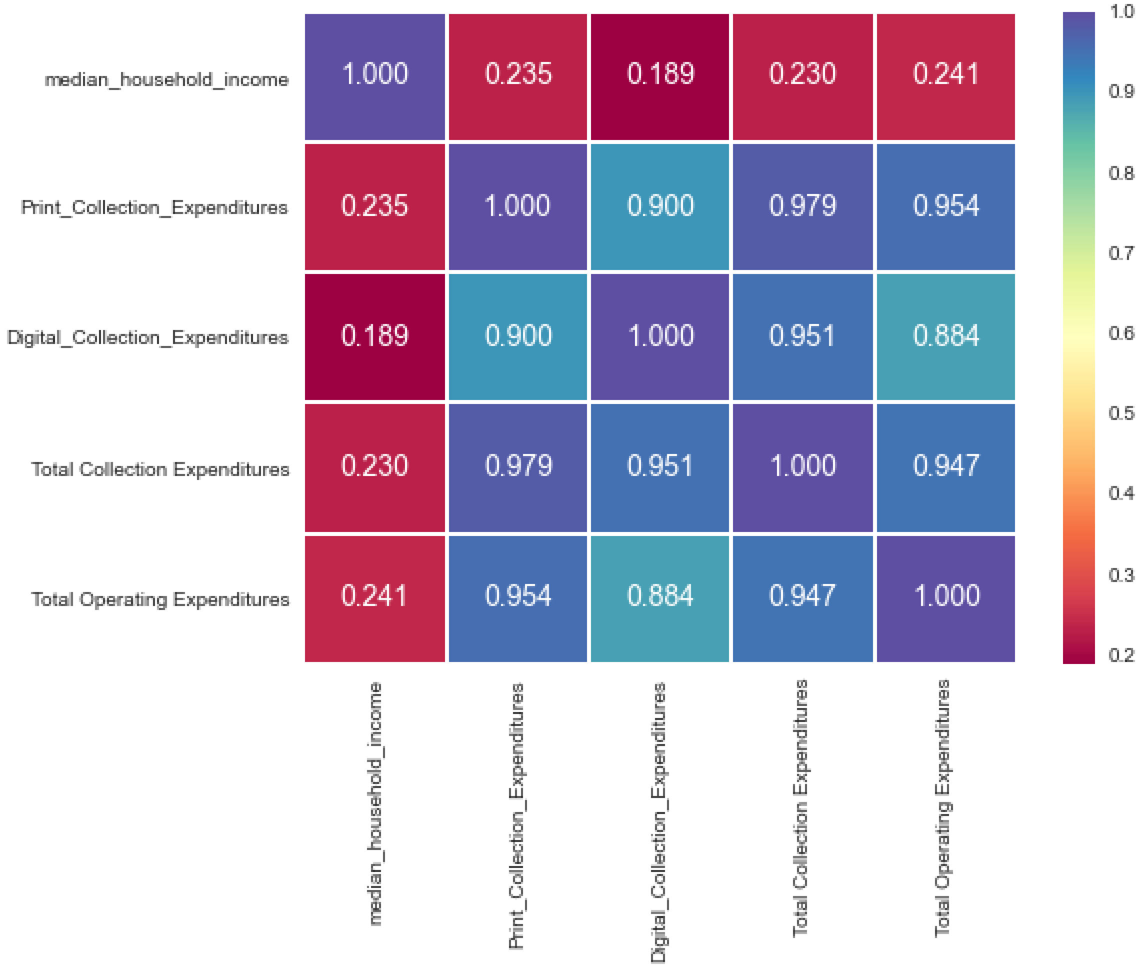




The linear regression lines between median household income and all the different expenditures suggested that the variables are not correlated.

Correlation

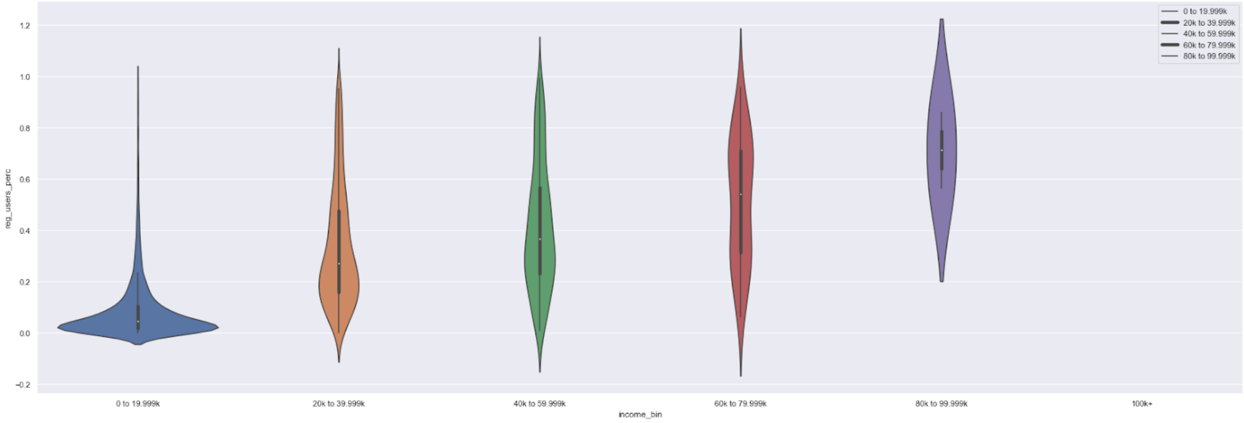
The heat map below shows the correlation between the median income and all the expenditures in our dataset. This validates that there is not a strong correlation between any of these variables. We can interpret that the median household income does not affect the expenditures of the libraries in a specific area.



	median_household_income	Print_Collection_Expenditures	Digital_Collection_Expenditures	Total Collection Expenditures	Total Operating Expenditures
median_household_income	1.000	0.235	0.189	0.230	0.241
Print_Collection_Expenditures	0.235	1.000	0.900	0.979	0.954
Digital_Collection_Expenditures	0.189	0.900	1.000	0.951	0.884
Total Collection Expenditures	0.230	0.979	0.951	1.000	0.947
Total Operating Expenditures	0.241	0.954	0.884	0.947	1.000

Median Household Income vs. Registered Users

Our data exploration resulted in us finding that there is no direct correlation between the median household income by zip code and the percentage of the reported registered users. Our data sample was too wide to draw a direct higher p value during our test. And even when we broke our income into bins ranging 0 to 20k, then raising in increments by 10k and performed our anova test we found that not one showed a strong correlation due the outlier represented in our violin plot below.



VIII. Acknowledge Limitations

Our exploration was limited in many areas, for example, when using census data there is a degree to the data that gets miss reported or not responded to which can account for a percentage of our population to be unaccounted for. We were also limited in exploring such a wide range of data given we looked at pour variables per zip code. We understand we would want to break it down to a specific state or even a city to help limit the outliers in our data and then further explore our variables. A major limiting factor was that we took only the year from 2015. To better explore our data is seeing how our variables have progressed over time. Population demographic is a constantly evolving variable as population is constantly growing. And we would want to take how much a library's usage and collection would be used over time.

IX. Call to Action

We encourage that we continue exploring in a more detailed scope. Meaning we focus down to a city, a state, or general area, as well as we need to collect more data through the years instead of one sample year of 2015. We can not use these demographics of population (median household income, and population count) versus the total collections stored within a library or the number of reported registered users on a country wide basis due to the obscurity. We could not generate a high correlation coefficient that suggest our variables were contagent with one another on a country scale but it is possible with more samples is a smaller regional area we can then possibly find that the data does support correlation between our variables and we could then use that factor in our pursuits in aiding libraries to more visitations or ways to draw means to add to their collection.