Exploratory Data Analysis on Public Libraries

By Sabrina Alves
Malachai Cravens
Enoc Serge Kouegbe
Marco Lopez

Table of Contents

Definition of U.S. Public Libraries

- Funded for the public by federal, state and local taxes as well as grants
- Serve the general public
- The U.S. is home to approximately 9,170 libraries

Why Libraries?

Is there correlation between demographics of area and use of library?

Define the Ask

Is there correlation between demographics of area and use of library?

Demographics of Area:

- Education Level of Population
- Median Household Income

vs

"Use of Library"

- Registered Users
- Total Collection Expenditures

Common Population (for data merge):	Data Collected by: ● Zip code ● state

Research Questions

Is there correlation between demographics of area and use of library?

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

Institute of Museum and Library Services

- Public Libraries Surveys
  - Data collected from over 9,000 public libraries since 1988.
  - Captures information regarding:
    - Registered Users (library card holders)
    - Expenditure on Print and Digital Collections
    - Library Visits

"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."

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

Census API

Using the Census Wrapper:

- Data collected from the American Community Survey (5-year data)
  - Benefit: increased statistical reliability for less populated areas or under reported groups
- Data collected by zip code

Null Values

In Census Data:

Rows removed when null values were present in columns of interest.

In ILMS Data:

Rows removed when null values were present in columns of interest.

Determining Null Values:

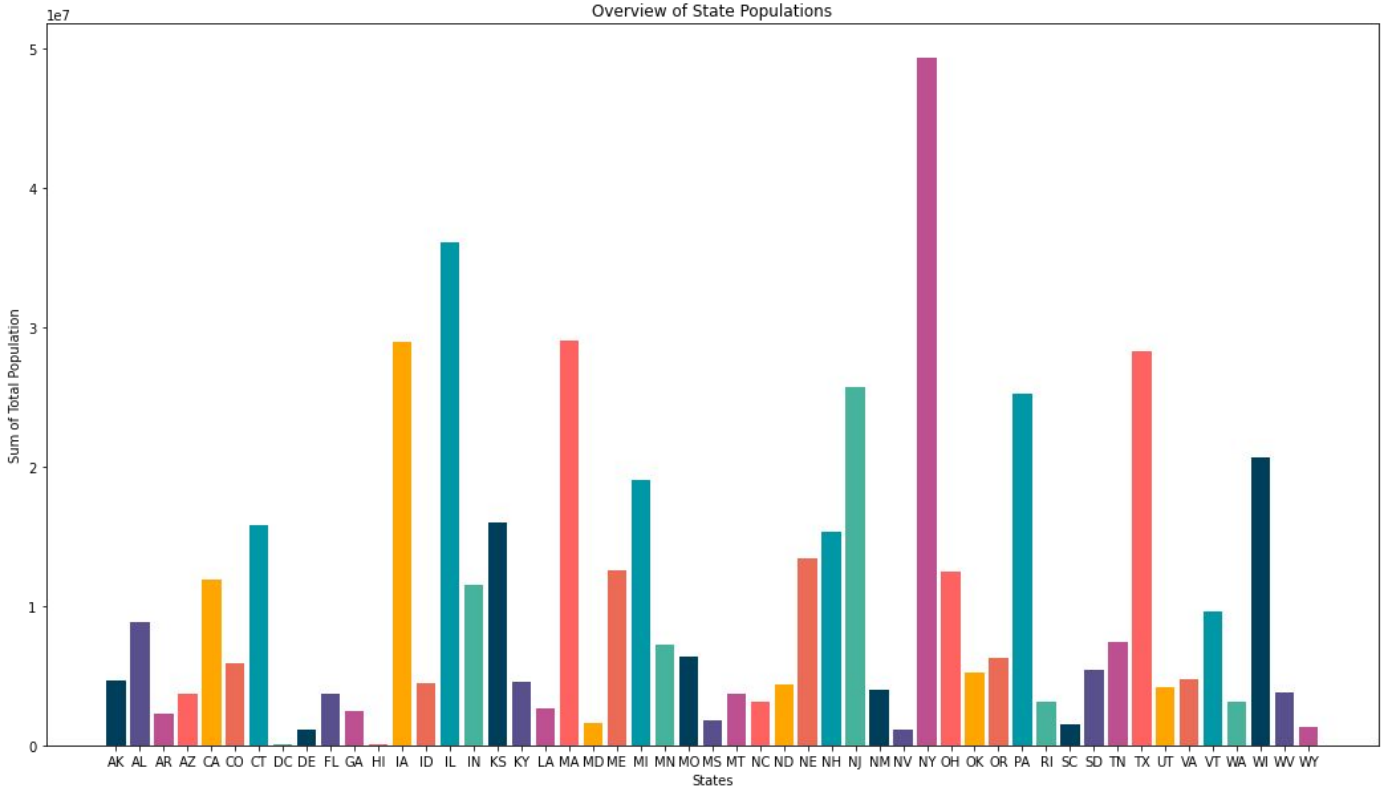
Cell values of: -666666666.0

Merge Files

- Library Data and Census Data merged on zip code columns
- Census zip codes that did not pair were dropped
- Libraries that did not pair were dropped

```
#merge data
df=pd.merge(library_data, census_data, left_on="Zip Code", right_on="zipcode",how= "left")
df
```

Census: Total Population



Outlier Footnote:
-No Outliers Removed from Upper or Lower Bounds
-No Outliers below 0

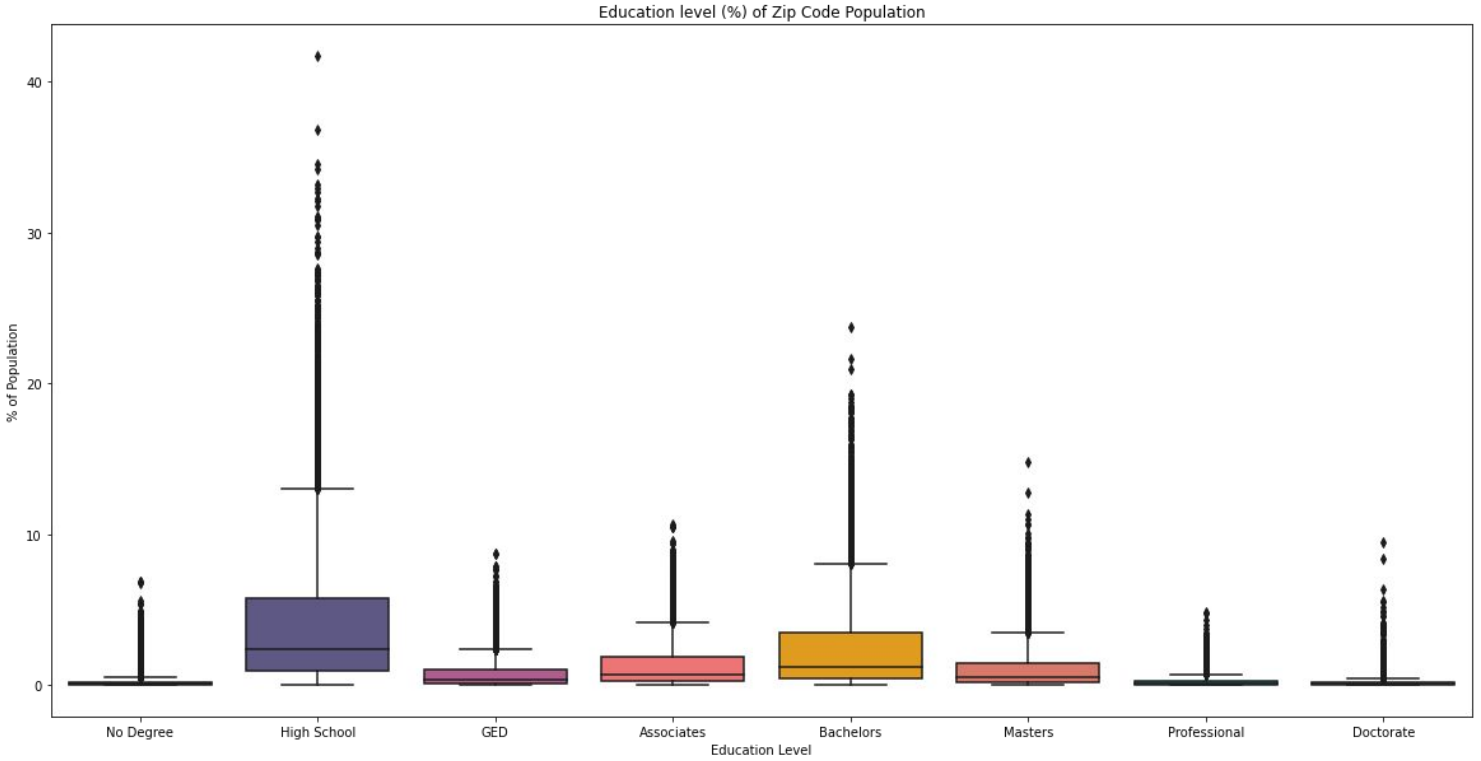
Total Population: The Great Equalizer

Uniformity between populations:

(metric) / (Total Population of Zip Code)

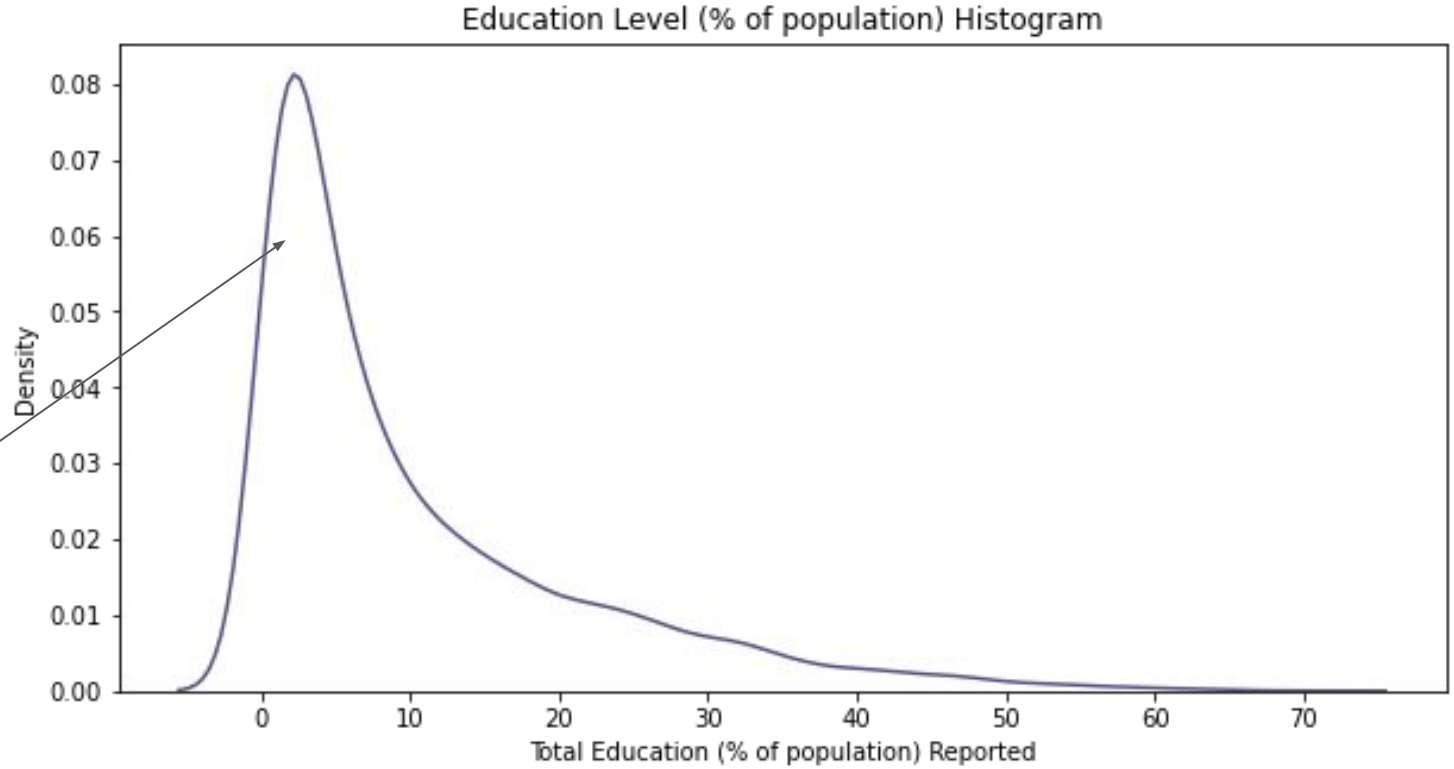
Columns Used in Percentages
No Education
High School Degree
GED
Associate’s Degree
Bachelor’s Degree
Master’s Degree
Professional Degree
Doctorate
Registered Users

Census: Education Level Held By Adults 25 and Older



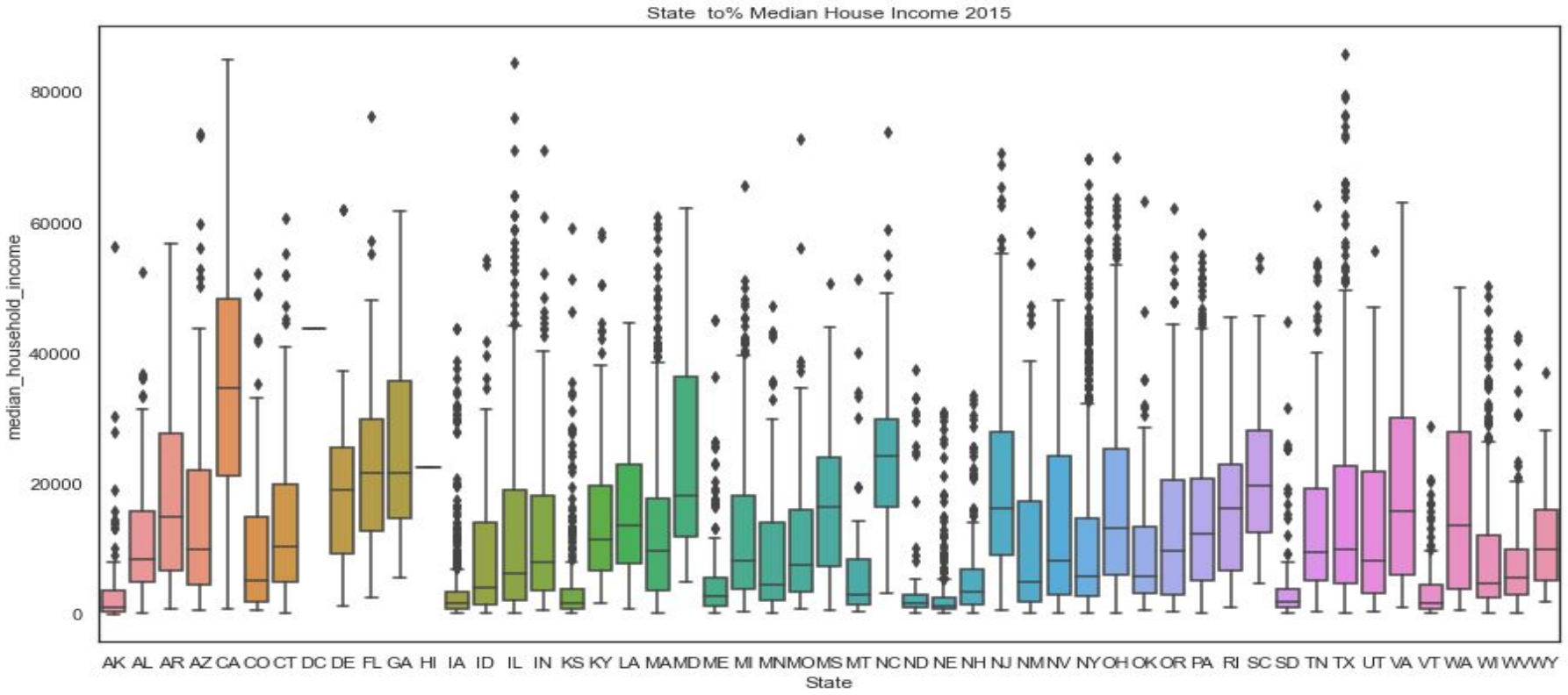
Outlier Footnote:
-No Outliers Removed from Upper or Lower Bounds
- Outliers removed when under 0

Census: Education Level Held By Adults 25 and Older

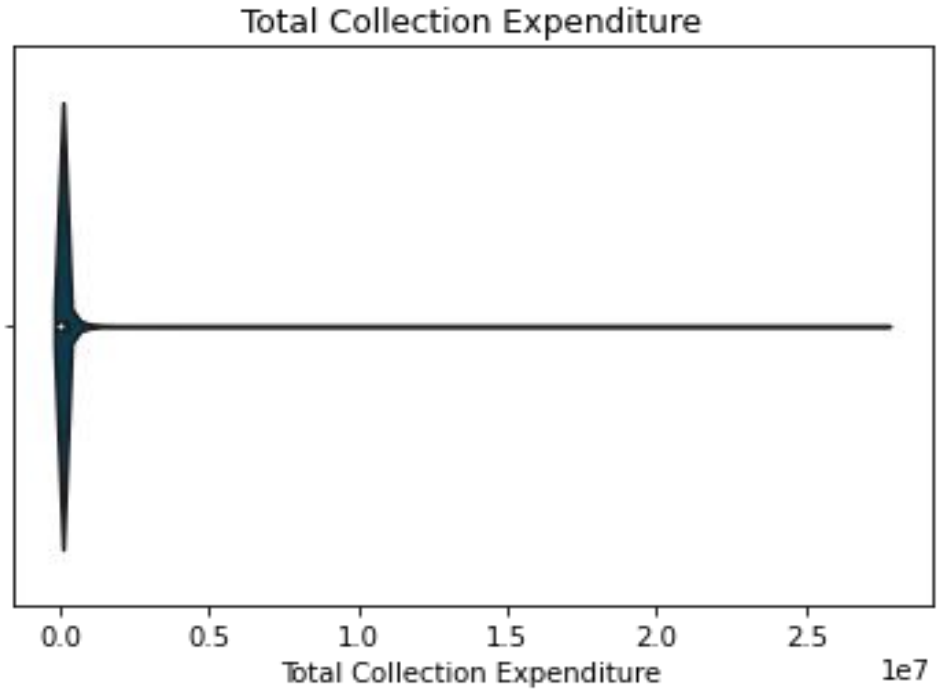


Data Under Reported

Census: Median Household Income

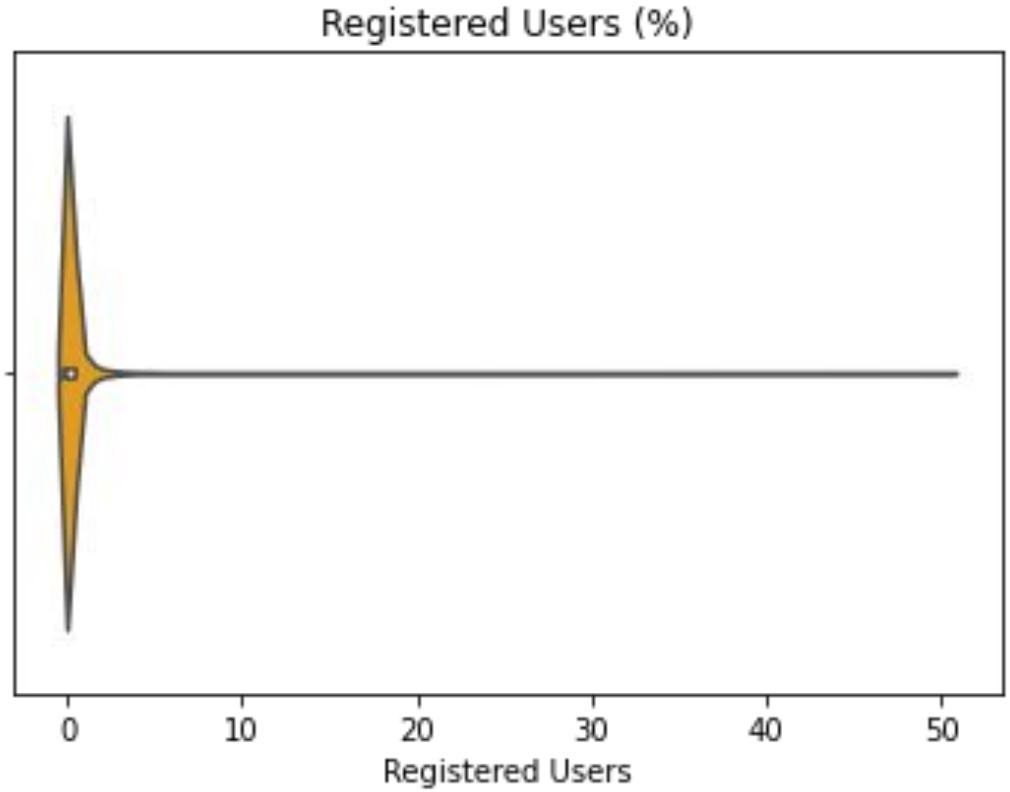


ILMS: Total Collection Expenditure



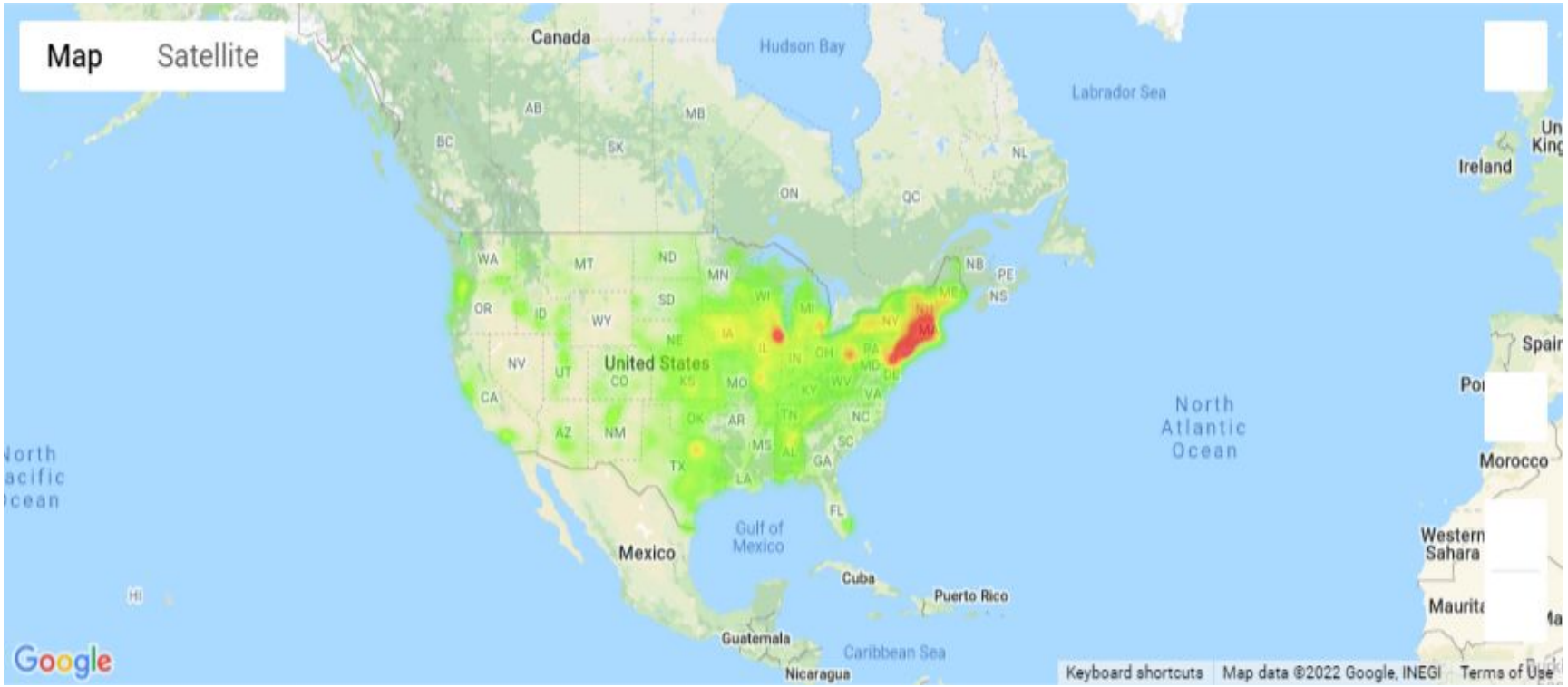
ILMS: Registered Users

Gained from dividing the reported Registered Users, by the reported total Population per state

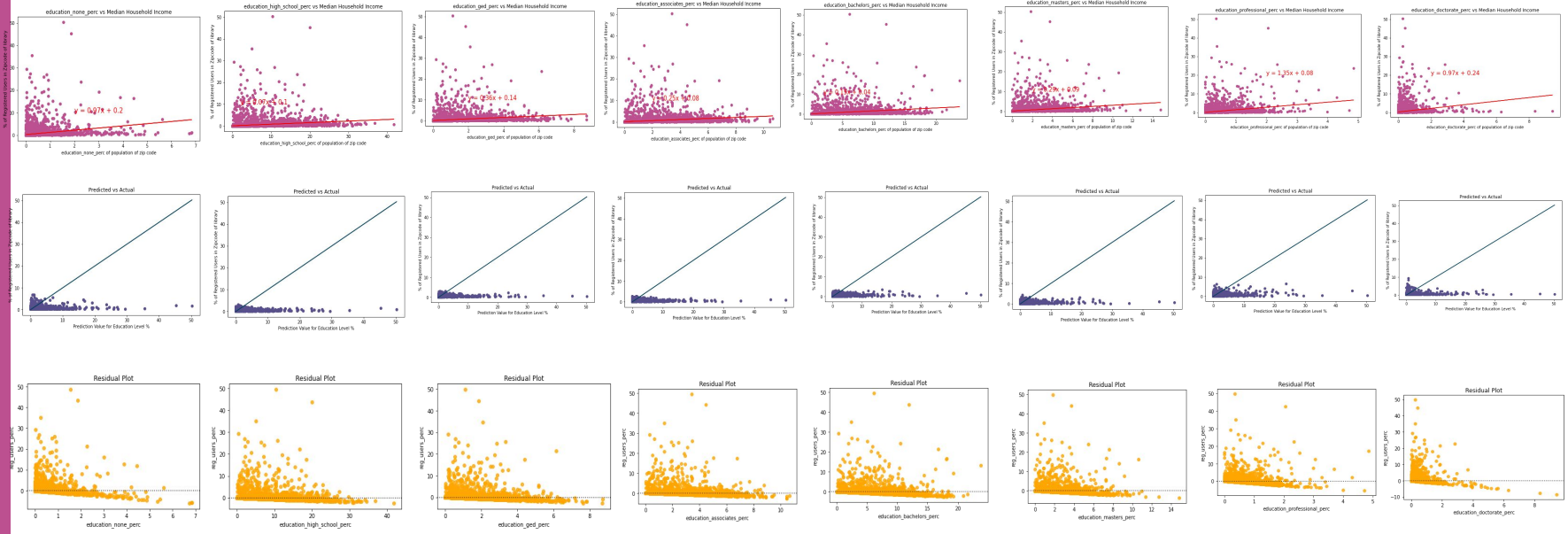


Outlier Footnote:
-No Outliers Removed from Upper or Lower Bounds
- Outliers removed when under 0
-Outliers removed when greater than population (>100%)

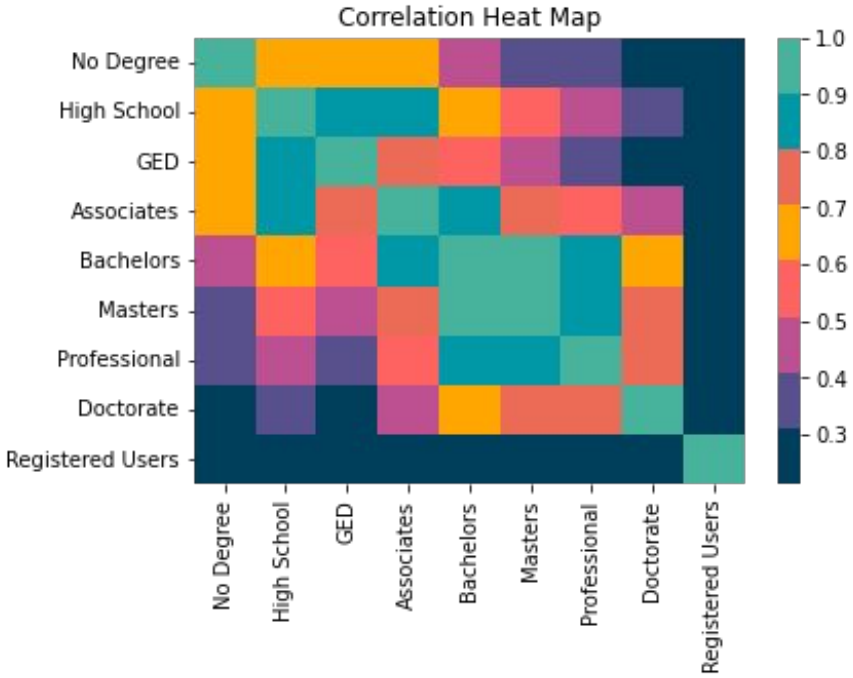
Registered Users (>50,000)



Education Level vs Registered Users



Education vs Registered Users



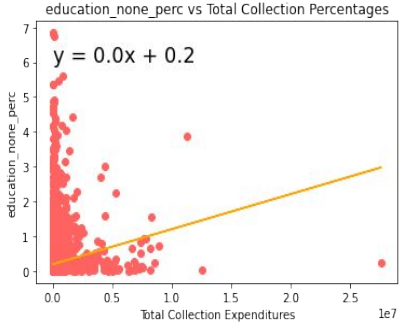
Education vs Registered Users

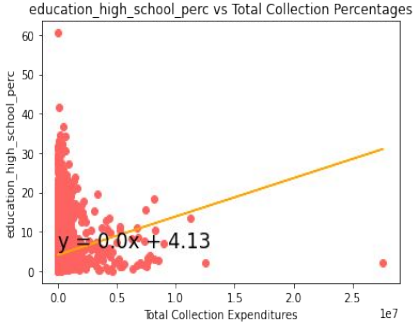
Dep. Variable:	reg_users_perc	**R-squared:**	0.138
Model:	OLS	**Adj. R-squared:**	0.137
Method:	Least Squares	**F-statistic:**	182.7
Date:	Thu, 10 Feb 2022	**Prob (F-statistic):**	1.58e-287
Time:	17:42:08	**Log-Likelihood:**	-17013.
No. Observations:	9158	**AIC:**	3.404e+04
Df Residuals:	9149	**BIC:**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

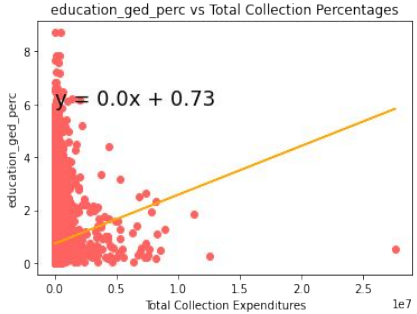
	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0543	0.022	2.421	0.015	0.010	0.098
education_none_perc	0.6544	0.053	12.392	0.000	0.551	0.758
education_high_school_perc	-0.0453	0.009	-4.791	0.000	-0.064	-0.027
education_ged_perc	0.2549	0.033	7.719	0.000	0.190	0.320
education_associates_perc	-0.0458	0.030	-1.546	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

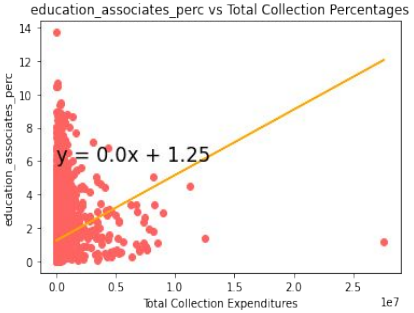
Omnibus:	16519.628	**Durbin-Watson:**	1.837
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	27594091.587
Skew:	13.087	**Prob(JB):**	0.00
Kurtosis:	270.637	**Cond. No.**	43.7

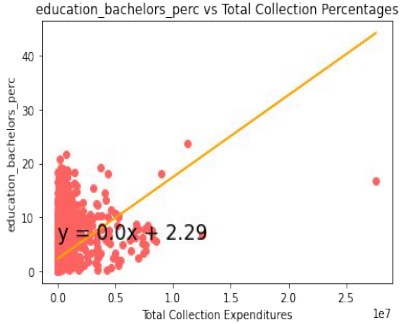
Education Levels vs Total Collection Expenditures

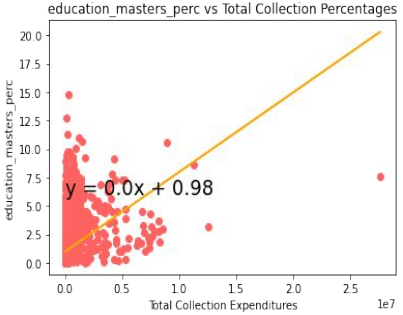


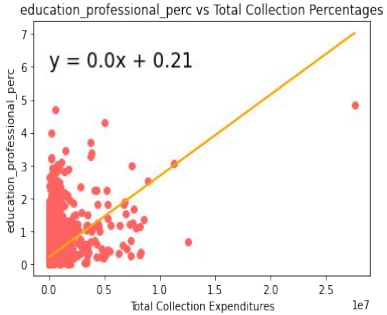


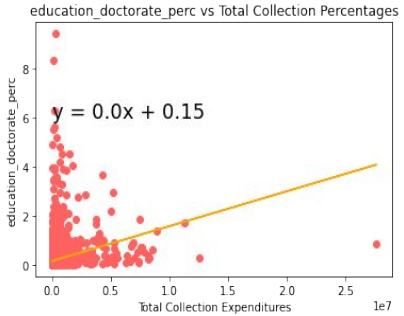




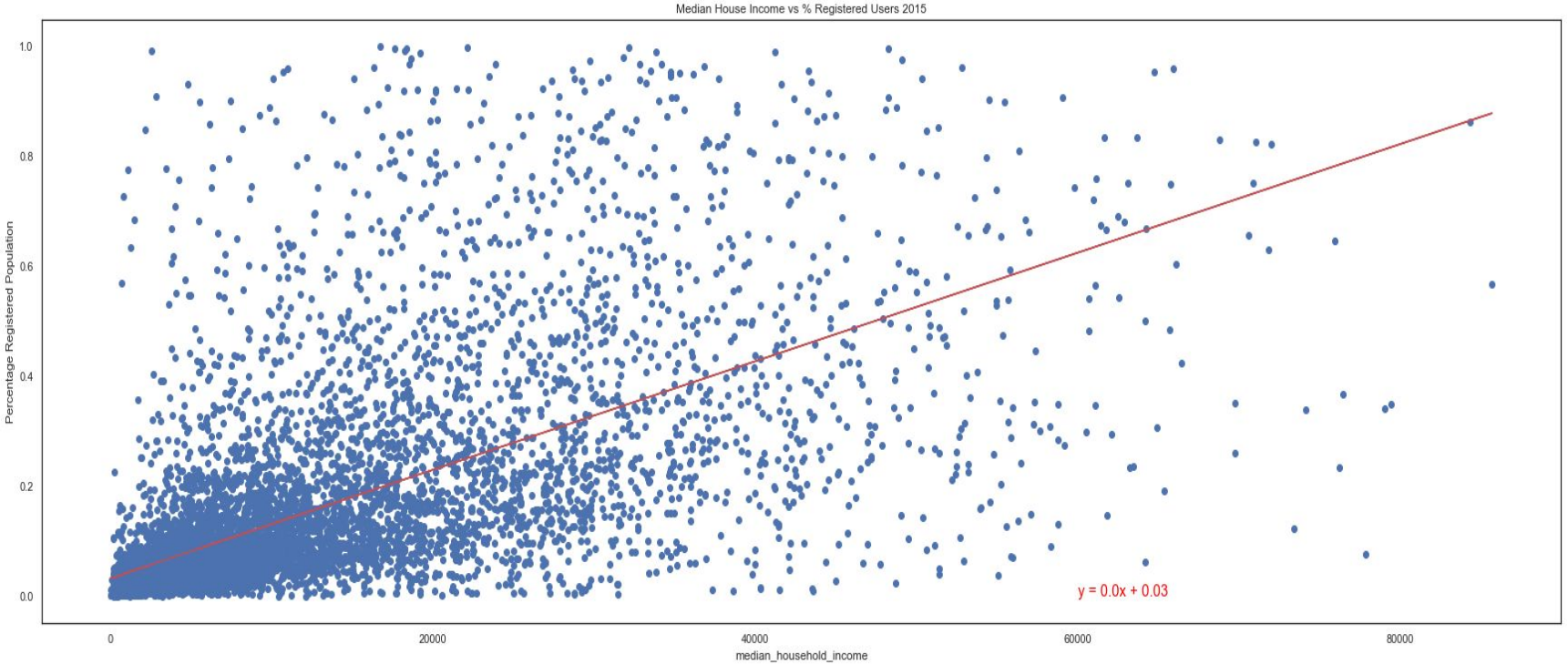






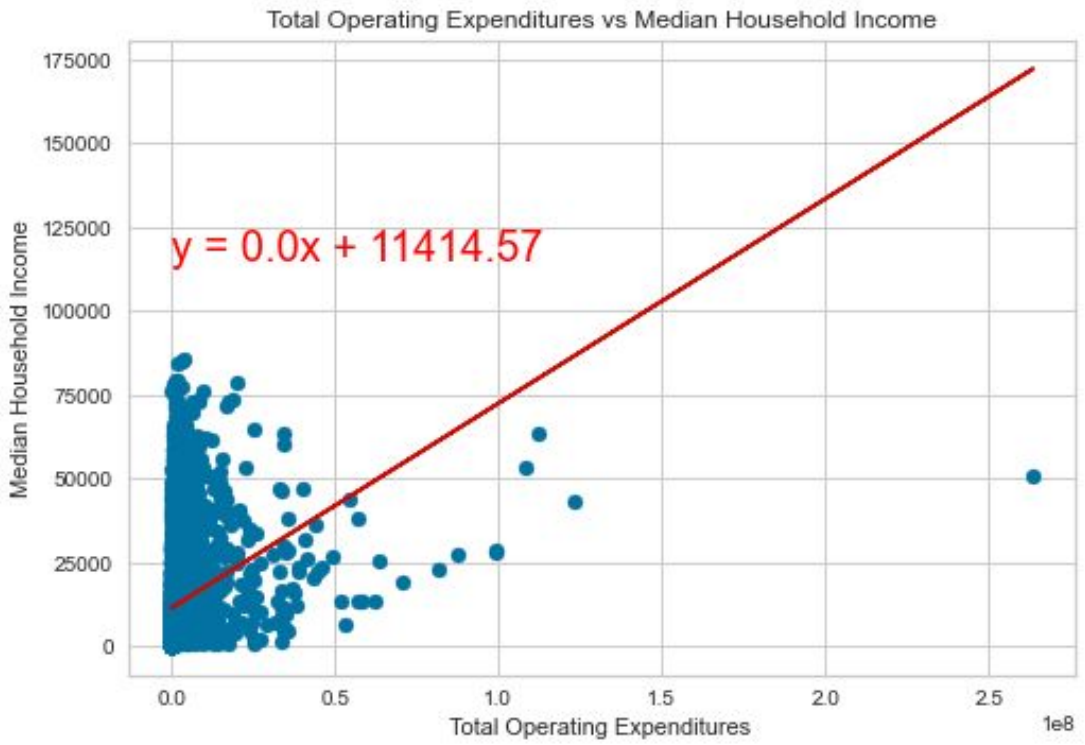


Median Household Income vs Registered Users

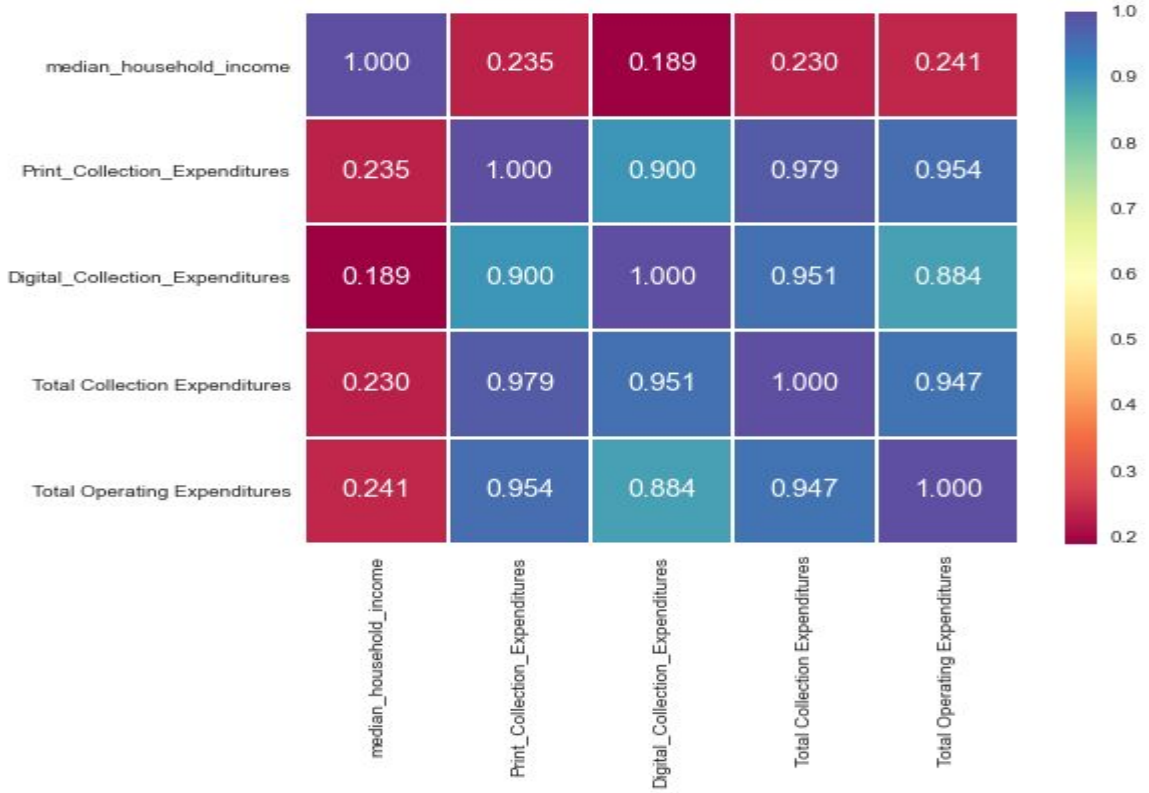


Median Household Income vs Total Collection Expenditures

The following example illustrated a scenario where two variables have no correlation.



Median Household Income vs All Expenditures



Education vs Registered Users

	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

NULL: Local education level has **no** effect on registered users (% of population).

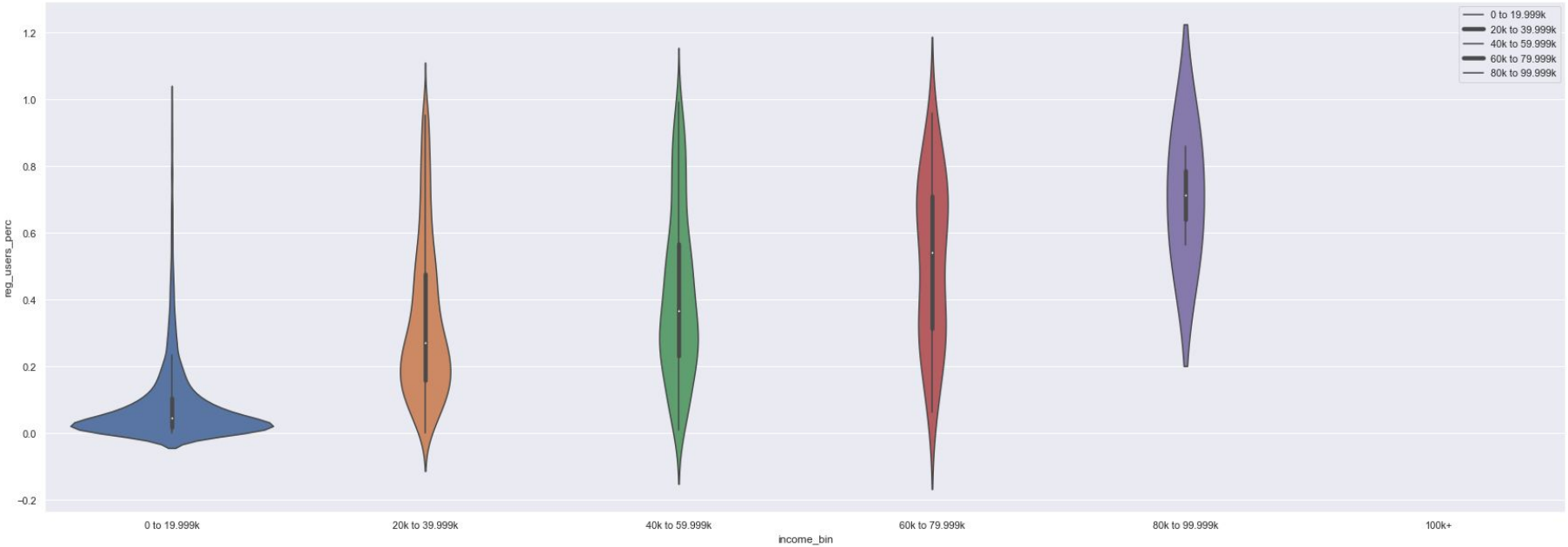
ALTERNATIVE: Local education level has **an** effect on registered users (% of population).

RESULT: REJECT THE NULL

*skewed data

Education vs Total Collection Expenditure

Median Household Income Bins vs Registered Users



Independent test our group 3 and group 4 median income groups cause they have the closest variance.

```
: print(stats.ttest_ind(group3, group4, equal_var=True))

Ttest_indResult(statistic=-2.581670063687598, pvalue=0.010272617612363042)
```

*skewed data

Conclusion

- Our data suggest we should reject our Nulls of our Hypothesis
- We can not draw a complete correlation between our variables

Limitations

- Errors in data– counts misreported in census
- Wide range of country reviewed– breakdown by state; income level
- Median Household Income – outliers in upper and lower bounds
- The variance in our sample sizes per some variable
    - Income Bins

Future Work

- Review Programs offered by ILMS
- Review Library Visits information offered by ILMS
- Break Down further what we group by when it comes to our data
  - By state, city
- Can we determine another variable we could have chosen to explore?

Q & A