**Funding by Keywords and Program Types**

**for IMLS Discretionary Funds (1996-2014)**

Amanda Lee

27 October 2015

The purpose of this project is to, by matching funding amounts granted by the Institute of Museum and Library Services (IMLS) with program types and keywords, identify and possibly predict priorities for this independent US federal agency. While IMLS sets its agenda and presents its goals to the public, I purports that funding amounts are a truer reflection of what issues take precedence for any organization.

**Background**

Established in 1996 by the Museum and Library Services Act (MLSA), the Institute of Museum and Library Services is the main source of federal support for libraries and museums in the United States. In fiscal year 2015, the agency’s budget was $227,860,000. From this amount, $28,724,000 was allotted to museums through the Museum Services Act and $180,909,000 was allotted through the Library Services Technology Act, which includes $154,848,000 for grants to states. (See Appendix A). The President has request $237,428,000 from Congress for 2016.

The data used in this endeavor is titled “Administrative Discretionary Grants FY 1996-2014” and the result of the agency’s authorization to carry out and publish analyses of the impact of museum and library services. The data is sourced from and produced by IMLS itself. This dataset is produced each year, so there is another dataset that ends in 2013 available for the public to download and one ending in 2015 can be expected as well. For the 13,594 records, there are 36 features which contain information about the organization and project that was awarded the discretionary funds, and the program through which the funds were granted. (See Appendix B).

**Analysis**

The rest of this paper will be between the two exercises which are:

1. Funding Amounts for Program Types
2. Funding Amounts for Keywords

The processes are ultimately very similar and I will try to avoid unnecessarily repeating explanations.

1. **Funding Amounts for Program Types**

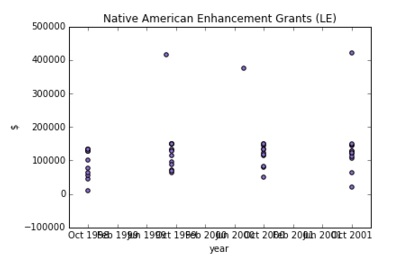
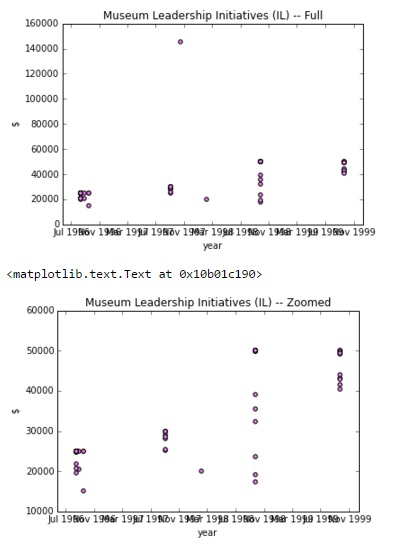
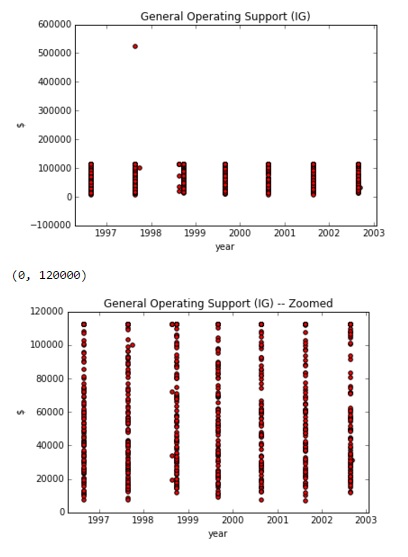
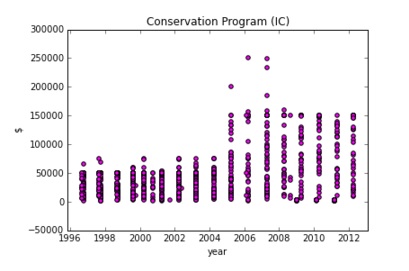
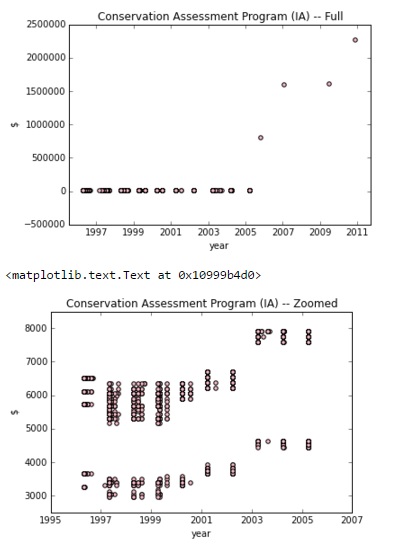
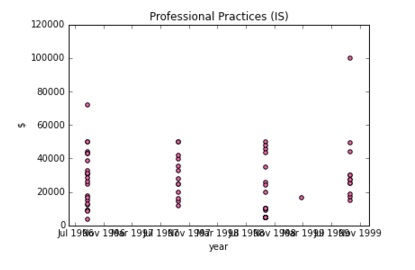
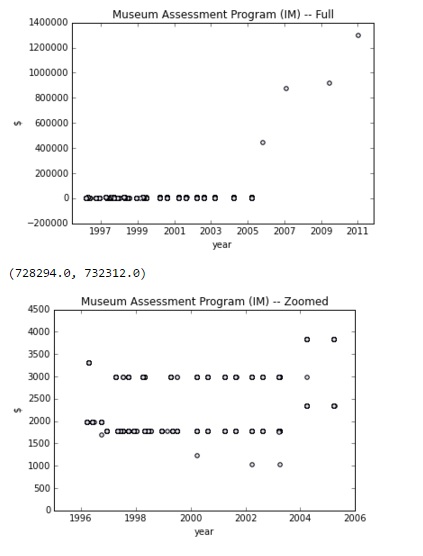
This exercise plots dollar amounts over time for each of the 18 program types.

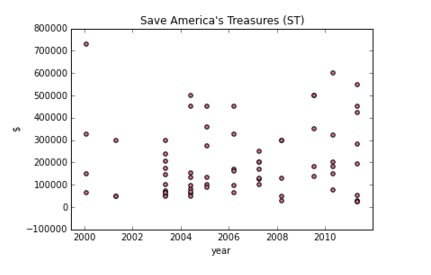
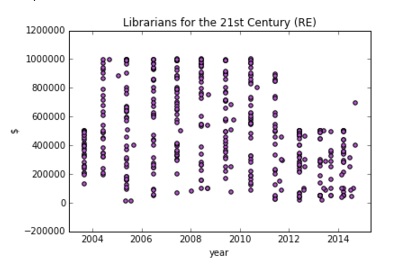
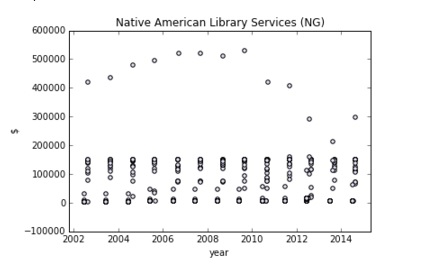
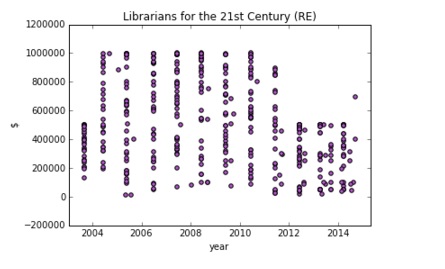
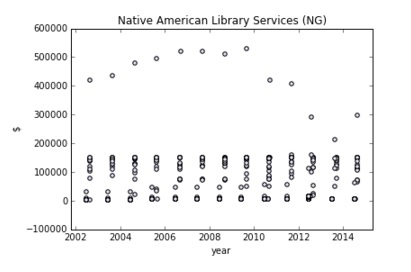
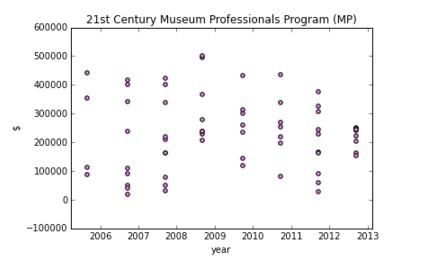
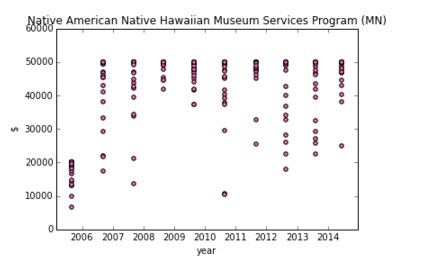
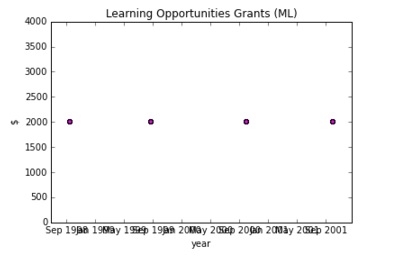
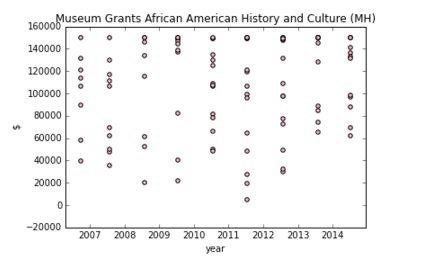
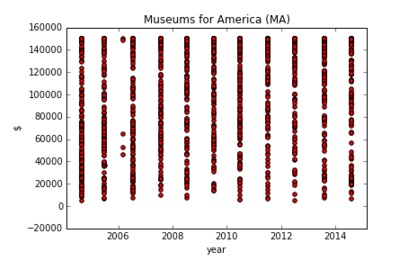
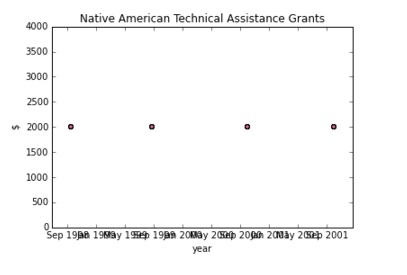
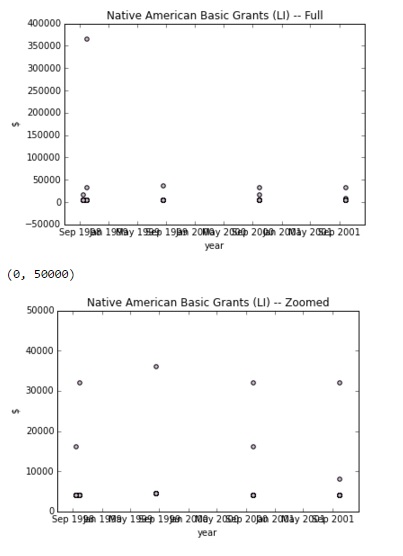
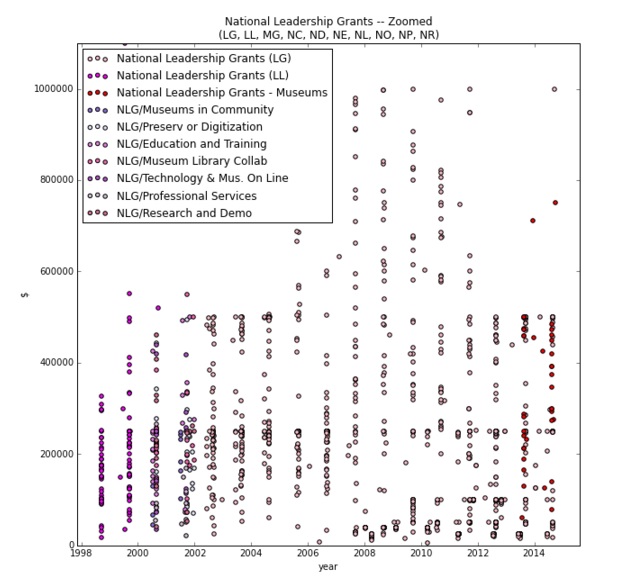
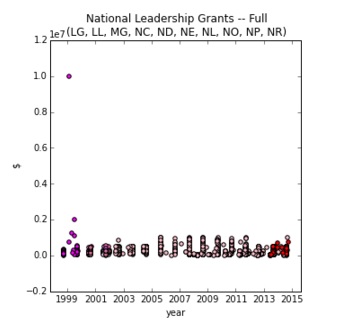
There are 18 programs through which awardees receive funding and each program name has a matching two-letter designation called the program type. “National Leadership Grants” is an exception in that it encompasses 10 program types (different National Leadership Grants programs). Program types specify the subject matter somewhat. For instance, programs could pertain to library services, museum conservation or the administration of either, or community outreach or exposure to minority cultures. Two-letter program types were preferred in this exercise over the lengthier and spelling-error-prone program name.

There are several features that contain funding amounts requested or originally awarded, but to determine how much a project is worth to IMLS, I chose the “Total Amount Awarded and Disbursed.” These amounts are stored as a string preceded by a dollar sign. I created a new list that contains the items in this series without the dollar sign and converted into a float. This new list was appended to the dataset and named “totAward\_float.”

The “Award Date” feature contained both strings and floats, which were NaN values. It had much fewer NaN values than either “Award Period from” and “Award Period To,” however. Otherwise it would have been interesting to measure how much an awarded amount is expected to last which speaks to the duration, scope, and perhaps impact of a project. Using the datetime module in a for loop, string values were converted to datetime values and NaN values were ignored. This was also appended to the dataset as “date\_asTime.”

Plotting NaN values in datetime produced points in the 17th Century and augumented the scale of some scatter plots so it became necessary to create a new dataset without NaN values in the “date\_asTime” column. This new dataset had 13 less records.



The data’s “stacked” appearance is probably due to IMLS program’s application periods and subsequent award dates. There also appears to be a range of funding amounts each award period, with some “grand prize” amounts for some of the programs (“Native American Library Services (NG),” “Native American Enhancement Grants (NE),” for example). Some programs grant awards sparsely, showing only 4 award dates with similar amounts. For these programs, using the “Award Period” start and end dates may have been the more demonstrative features.

1. **Funding Amounts for Keywords**

This exercise is a method to find keywords and plots funding amounts over time for those selected keywords.

The stop\_words module was used in the “Project Title” column. This feature describes the name of the project awarded discretionary funds. I split each “title” string and eliminated stop words provided by the module, returning a list in the string’s place. The result is a list of lists.

The list of lists was appended to the dataset as “titleList.” Additionaly, titleList is counted then sorted by counts.

The goal of having keywords is to identify trends in subject matter that does not duplicate the plotting by program type. Not surprisingly, words like “museum,” “library,” and “program” appear at the top and add little value. Although these and other words like “using,” “center,” “national,” “American,” “access,” and “strengthening” are not stop words, they also add little value but occur frequently. I also discovered that, for instance, “Art” might have 50 counts, but combined with “art,” “arts,” or “Arts,” I might find 250 records for projects about art.

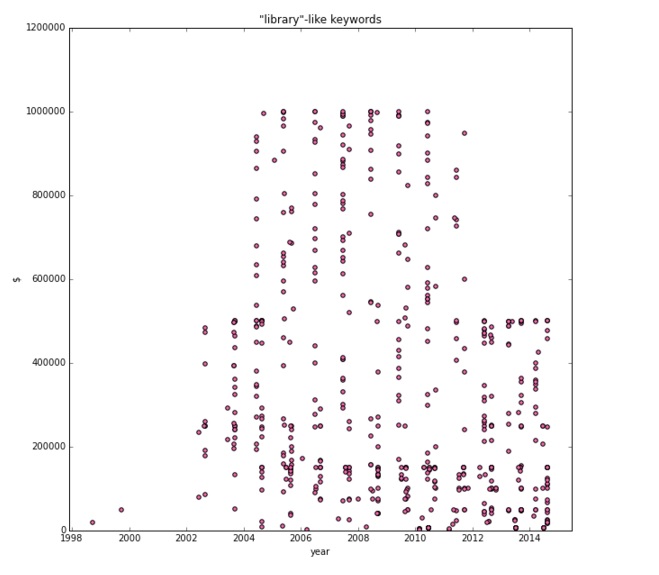
There were several missed opportunities as well. For one, word pairs could have identified a phrase like “national treasure” which is more specific than “national museum,” or “Native American” and “African American” rather than just “American.” I probably should have cleaned the lists up before sorting them by making all items lowercase and taking out punctuation marks. I also thought about taking words that occur in the program name out of the titleList for each record.

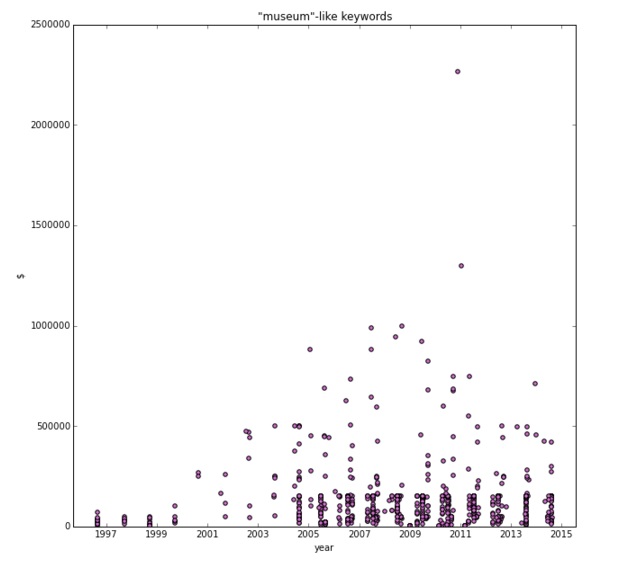
I had done the same for the “Description” features but found that there were many empty records. So although a description is longer and contains maybe a wider range of words to choose from than the “Title” feature, I chose to exclude the project description feature in order to be able to include more records. Every record contains a project title, so none are excluded through this criteria.

***“Museum” and “Library”-like keywords.*** Next, I created a dummy list for “museum”-words. Because “museum” may appear as “museum,” “Museum,” I searched for “useum” and created a new list that returns a 1 where such a word appears. For “library,” I used “ibrar.”

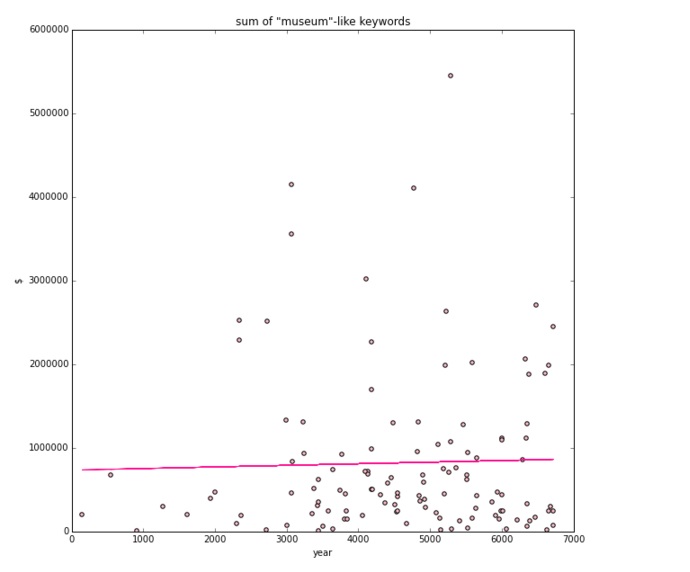
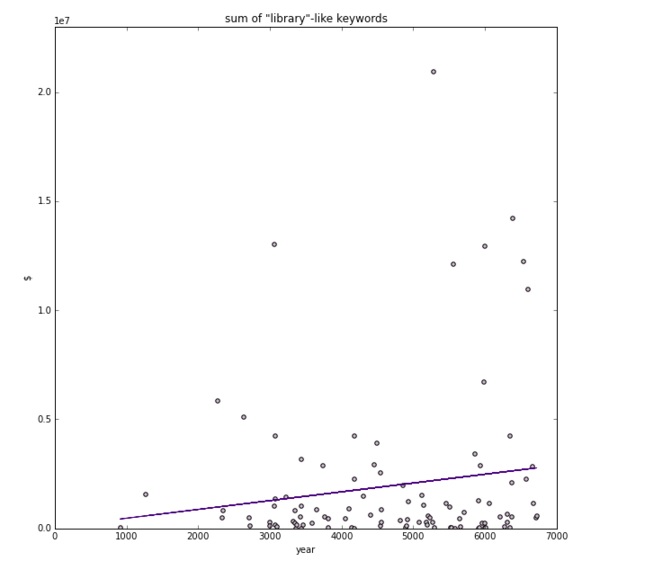
Similar to the above Project Type exercise, I turned award amounts into floats without a dollar-sign, and turned the strings in “Award Date” into datetime objects.

I also added another column that lists the date-time object as days as integers because of some issues I ran into further down the road. This ranges from 0 (first date in 1996) to 6757 (last date in 2014).

Using the dummy columns for “museum” and “library”-like keywords, I plotted award amounts over time. 

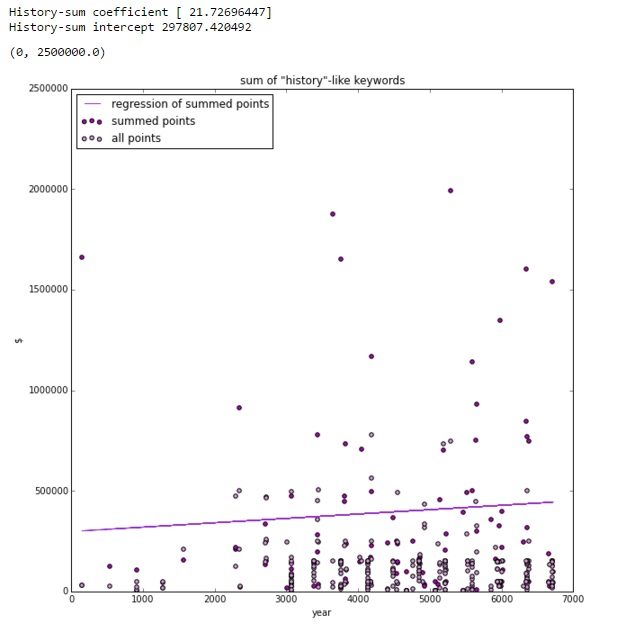


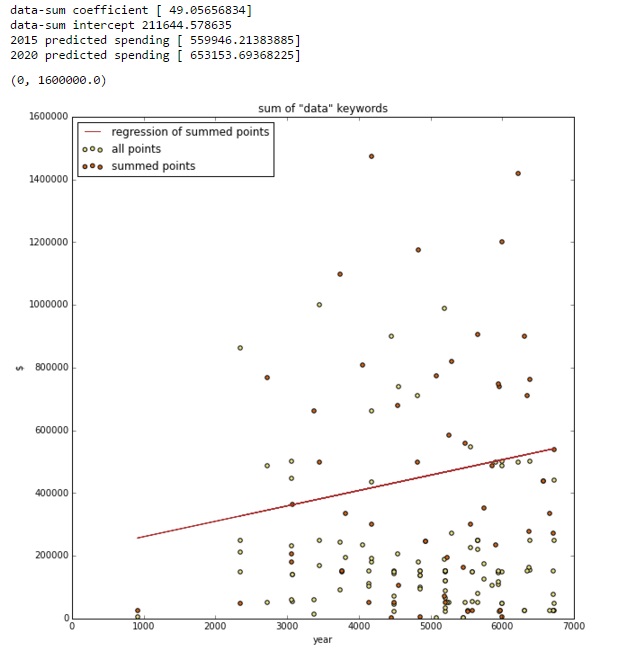
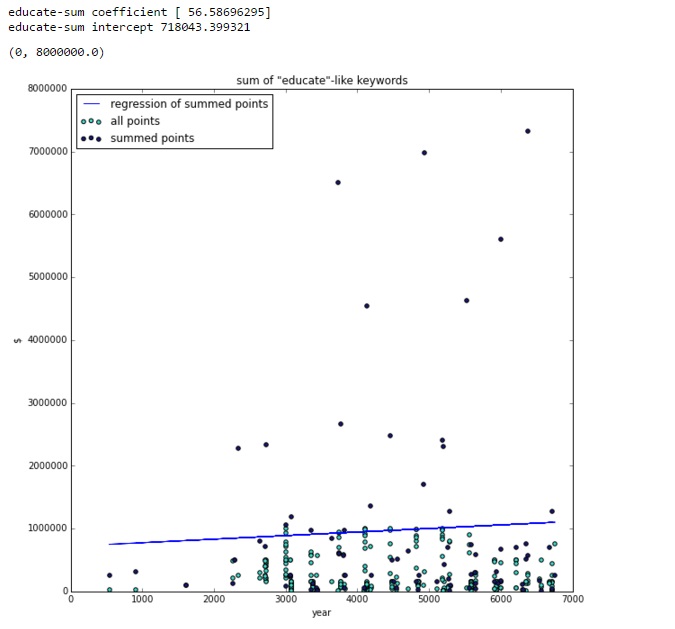
I then summed award amounts over time and created a linear regression model for both “museum” and “library.” This exercise was inspired by the annual budget which allocates a $150 million to Grants to States through the Library Services Technology Act (LSTA), versus the $28 million through the Museum Services Act. However, it is not known whether those grants to states are included in this dataset, and if they are, under which program, so no such conclusion could be drawn at this time.



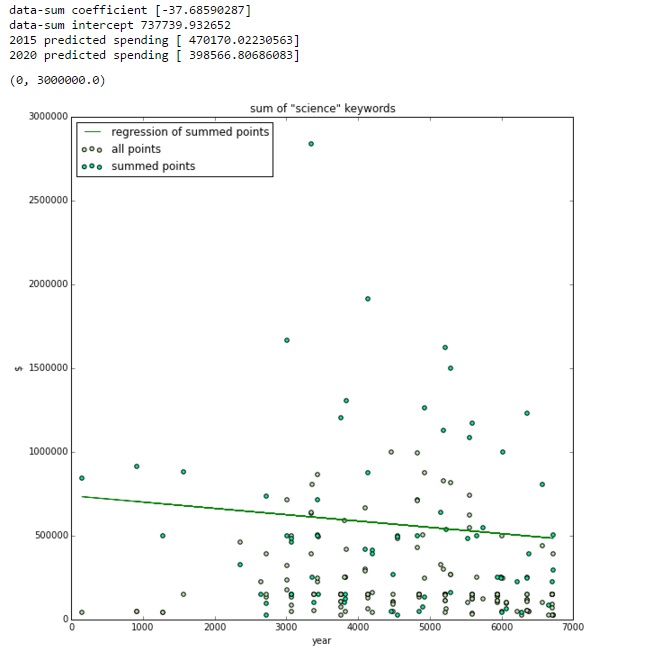
What was interesting is that the first time I ran the regression model for summed “museum”-like keywords, its coefficient was -30, but after running it a few more times while testing my options, it settled on +19.

***Other Keywords.*** I did the same for words that begin with “histor-,” “educat-,” “data,” and “scien-.”





Starting from “data”-like keywords, I added in two calculations: an approximate spending amount for a day in 2015, and in 2020. Because I have not used any other features besides keywords however, these calculated values are hardly predicted values.



It was interesting to note that total funding for “science,” based on the key-word search, seems to have decreased over time. This observation leads me to believe that for some reason or another, “science” became less of a priority for the IMLS. This could be for funding constraints, an administrative wish to focus on issues set by its policy, or perhaps another agency or organization, such as the National Science Foundation, took on priorities that would overlap with IMLS’s.

**Conclusion**

***Challenges.*** My biggest challenge was probably the source of all my challenges, and that was my unfamiliarity with Python. I probably spent more time getting familiar with the data and programming concepts than I did running actual analysis.

Much of the data I wanted to use was stored as strings—descriptions, dollar amounts, ZIP codes, dates. In order to use these, I realized (maybe too late) that I needed to convert these strings into other objects. Working with the data, then, usually meant trying to find an the right type of object that would fit the code’s input args.

For instance, ZIP codes were stored as string, but could be five characters or ten characters (XXXXX-XXXX). I thought it would be easy enough to return each one as just the first five and convert that string object into an integer or float, but I ran into errors and ran out of time.

***Success.*** On the flipside, however, overcoming these challenges in regards to making the data work with the problem I was trying to solve was my biggest success. It bred a level of comfort with the problem solving and exploring possible solutions.

If compared to the Data Science course lessons, however, I probably only used the skills learned from the first two classes. Therefore the outcome of this project probably produced more missed opportunities than successes.

***Missed Opportunities.*** I ran out of time (and patience), unfortunately, and missed some opportunities. For one, I would have liked to actually try to train/fit and predict a model with multiple features, especially with ZIP Code and Organization Type. The problems I had with these two were massaging the data into features I could run regression-type problems on.

Even with what I had accomplished, I think I could have turned much of these processes into functions through which I could run many, many more keywords.

I think if I had managed to turn many of the string-features into float-features, I could have tried to predict something. And to this point, I lost any subsequent opportunities.

***Next Steps.*** Knowing what I know now about the dataset, about the features, and with a greater familiarity with the Python language, the next step would be to explore these missed opportunities, especially with more features than just date and award amount. I think the first step before anything, however, would be to rebuild this dataset as float-objects, and clean up much of the processes made and coding syntax used.

**Appendix A**



**Appendix B**

