Classifying Neighborhoods

CS591 L1: Data Mechanics for Pervasive Urban Systems

Kyle Mann and Jonathan Liu
Professor Andrei Lapets

May 3, 2016

Narrative

Our goal in this project was to originally find dangerous areas in Boston, and possibly reason as to why they are dangerous; however, due to complications and a change in interest it has change. Our current goal is to classify a zip code. More specifically, we wish to allow a user to specify some constraints and have them receive a zip code that satisfies those constraints. Something like this would include: "I want the safest neighborhood with the lowest tax rates, lowest crime rates, with 2 schools, and 1 hospital". We could still come up with a way to have a specific metric rating (1 - 10 per say) of a zip code. From this idea and feature, we sought out to solve some optimization problems that involve a variable number of zipcodes, ie: “If I want to invest in a maximum of 6 zip codes and affect the most schools with the lowest overall tax rates, which zip codes should I invest in?”

In order to try and solve the problem we have done the following. We have selected some data from the city of Boston (see source sets for details); these include, property information, crime rates, liquor license, public schools, and hospitals. We next combined data sets to get a "profile" for a zip code. We attempted to correlate crimes to alcohol by the street address and proximity (approx. walking distance) between the crime and liquor source. Additionally, we calculate the average tax per square foot in USD. From this we also see how many hospitals and schools are in each zip code. Using all this, we can create a profile for the neighborhood and begin to do some optimization and satisfaction problems. Additionally, we wanted to investigate if there is any correlation between liquor sales availability and crime - particularly crime involving alcohol.

As you proceed it is important to note that zip codes are a fairly variable geospatial area (most of them are very large, which makes categorizations of them less accurate, as there are certainly worse areas than others in a zip code).

Source Datasets

1. Crime data from city of Boston:
   https://data.cityofboston.gov/Public-Safety/Crime-Incident-Reports/7cdf-6fgx
   This dataset was chosen because we believe it represents a pretty strong correlation to the dangerousness of a neighborhood - high violent crime is probably a dangerous area.

2. Liquor license data from city of Boston:
   https://data.cityofboston.gov/dataset/Liquor-Licenses/hda6-fnsh
   This dataset is a bit of an experiment. Perhaps it is the case that in areas with easily available alcohol, there is high crime rate. In particular, violent crime related to alcohol - this would be a cause of a neighborhood becoming more dangerous potentially.

3. Property Assessment: https://data.cityofboston.gov/resource/qz7u-kb7x
   This dataset was chosen to see if we could learn anything about the value of an area based on the assessment from the census. We calculate the average tax per square foot in properties.

4. Public Schools: https://data.cityofboston.gov/resource/e29s-ympv
   This is used to rate a zipcode - how many schools does it have?

5. Hospital Locations: https://data.cityofboston.gov/resource/46f7-2snz
   Also used to rate a zipcode - how many hospitals does it have?

Methods

From these 5 source data sets we constructed a few data sets, as this was done in steps, we have some intermediate sets, with the final set (zip code profiles) being the most important. Using the zip code profiles we were able to solve some optimization problems and examine possible correlations.

Data Transformations

1. Zip codes with property and liquor information - this dataset combined sources (2) and (3). Each field in the sources had a zip code, which was used as the key. To get something meaningful out of the property data set, we calculated the average dollar per square foot of taxes paid on the properties in that zip code. We simply combined this with an aggregation of how many liquor locations are in that zipcode. We also counted how many properties were in the zip code. Here is a sample of what that looks like:

   ```
   {
           "zipcode" : "02215",
           "number_properties" : 50850,
           "avg_tax_per_sf" : 5.667375356194619,
           "liquor_locations" : 36
   }
   ```

2. Another dataset we attempted to craft was the crimes per zip code; however, if you read into our challenges section, you will see why this failed. We tried to do this by reverse geocoding the zip code from a crime. The crime incident reports (1) did not contain a zip code field, but due to copyrights, we could not actually obtain these results without purchasing a license. This data set is never used, since it would cost money to obtain.
3. Another dataset we created was to attribute crime to a zip code and potentially a liquor license. The source data set (1) for crime reports does not contain zip code information, so we had to look at what it had in common with the liquor data. They both had a street,

and coordinates. So, if a crime happened on the same street as a liquor license location, we were able to correlate it by street (using regular expressions to make sure the street was the same except for formatting), and then we computed the euclidean distance. (Although the earth is not flat, since Boston is relatively small, euclidean distance is accurate enough for our data). If a crime was within walking distance (~1.5 miles) we attributed it to the liquor store and the zip code. From this, we can now produce aggregations on how many crimes (which are potentially liquor related) occurred in a zipcode, or even on street. Here is a potentially daunting sample (feel free to ignore it) we decided to store the entire liquor location into the crime.

```
{
        "_id" : ObjectId("570f02c58b11311664e826bd"),
        "reptdistrict" : "D4",
        "streetname" : "WARREN AV",
        "ucrpart" : "Part Two",
        "y" : "2950408.285",
        "year" : "2013",
        "domestic" : "No",
        "reportingarea" : "152",
        "shooting" : "No",
        "day_week" : "Tuesday",
        "shift" : "First",
        "weapontype" : "Other",
        "location" : {
                "type" : "Point",
                "coordinates" : [
                        -71.077917,
                        42.343296
                ]
        },
        "x" : "770274.0752",
        "month" : "8",
        "liquor_license" : {
                "issdttm" : "2013-11-25T12:44:00",
                "primapplicant" : "FACILITY CONCESSION SERVICES, INC.",
                "opening" : "NULL",
                "comments" : "NONE",
                "licstatus" : "Active",
                "locationcomments" : "In whole of said building.",
                "liccatdesc" : "CV7 All Alc.",
                "_id" : ObjectId("570f02028b113116559ab4e2"),
                "city" : "Boston",
                "state" : "MA",
                "closing" : "2:00 AM",
                "businessname" : "FACILITY CONCESSION SERVICES, INC.",
                "location" : {
                        "needs_recoding" : false,
                        "latitude" : "42.34996",
                        "longitude" : "-71.06617"
                },
                "dbaname" : "CHARLES PLAYHOUSE",
                "zip" : "02116",
                "expdttm" : "2015-12-31T00:00:00",
                "address" : "WARRENTON ST",
                "stno" : "74",
                        "licenseno" : "LB-99272",
                        "liccat" : "CV7AL",
                        "capacity" : "871",
                        "patronsout" : "2:30 AM",
                        "stnohi" : "78",
                        "phone" : "NULL"
                },
                "zip" : "02116",
                "fromdate" : "2013-08-13T16:00:00.000",
                "incident_type_description" : "VAL",
                "naturecode" : "HITRUN",
                "compnos" : "130509531",
                "main_crimecode" : "VAL"
        }
```

4. From the (1) and (3) data sets we created, we then combined the school and hospital data. Using aggregations, we created what we call a "zip code profile" which can be seen below:

```
{
        "_id" : "02215",
        "avg_tax_per_sf" : 5.667375356194619,
        "num_schools" : 4,
        "num_hospitals" : 1,
        "liquor_locations" : 36,
        "number_properties" : 50850,
        "num_crimes" : 5031
}
```

The zip code profile seen above is the most interesting data set we constructed, and the basis for the following optimizations and correlations. It is important to note that we have yet to do any filtering of the crimes, so they may not be related to alcohol really - but this is a slight change to keep in mind.

Optimizations, Satisfactions, and Correlations

Using our zip code profile we came up with the following problem solutions:

Correlations

We wanted to see if there were various correlations between our data. We attempted to correlate each of the various attributes to the others:

1. Liquor locations to number of crimes - this had a very positive correlation, and an extremely low p-value. However, based on how we got the crimes from the liquor locations, this is not surprising; perhaps if we filter the crimes more, we can find a more meaningful correlation.

2. Number of crimes to average tax rate - A slightly negative correlation, with a high p-value, so not much we can conclude.
3. Number of crimes to percentage of properties that are liquor locations - slightly negative correlation, with high p-value, so not much to conclude.

Overall, our correlations were fairly inconclusive, as we had pretty high p-values for most of them; however, the (1) correlation was pretty successful, but unsurprising.

Satisfactions and Optimizations

From the profile data set we can solve some satisfaction problems really easy by querying the data set. We could ask: “What is the zip code with the most crimes?”, “What is the zip code with the lowest tax rates?”, “What is the zip code with the most schools and lowest tax rate?”

These are satisfied by a simple query to our database.

However, what if you want something more complicated? Something like: “If I want to invest in a maximum of 6 zip codes, with the most schools and lowest tax rates, which ones should I choose?”

To solve this problem we used z3, which allows us to set up some constraints and optimize on them. Here is a sample run:

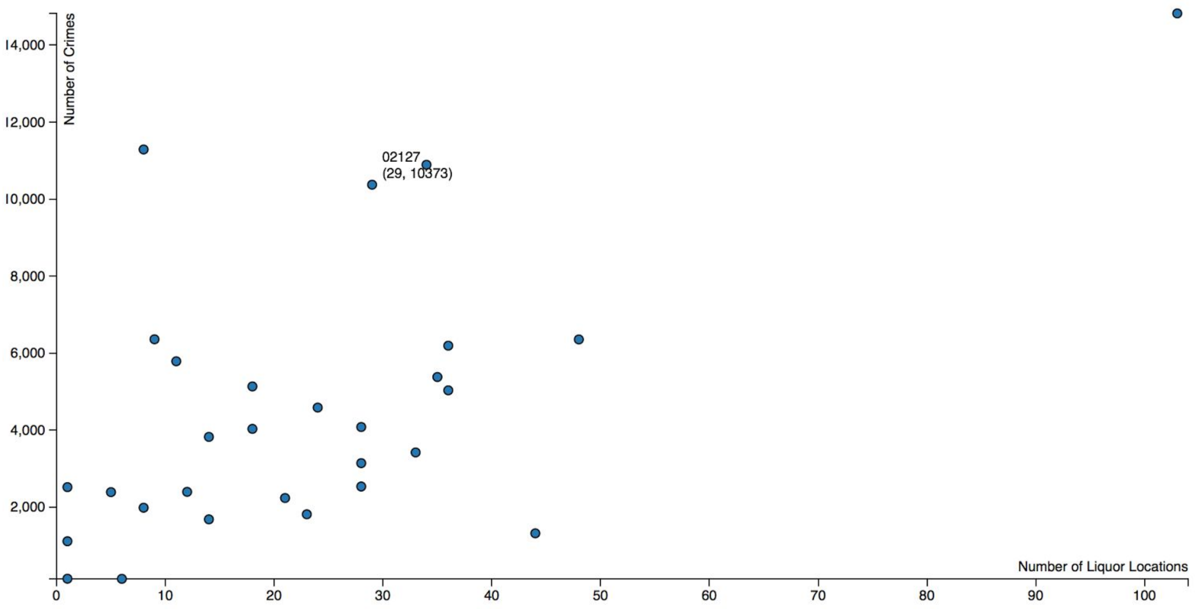
```
Enter a parameter to maximize (MAX) from the above choices: "num_schools"
Enter a parameter to minimize (MIN) from the above choices: "avg_tax_per_sf"

Here are the zipcodes you may be interested in:
02128
02124
02136
02119
02130
02127
```

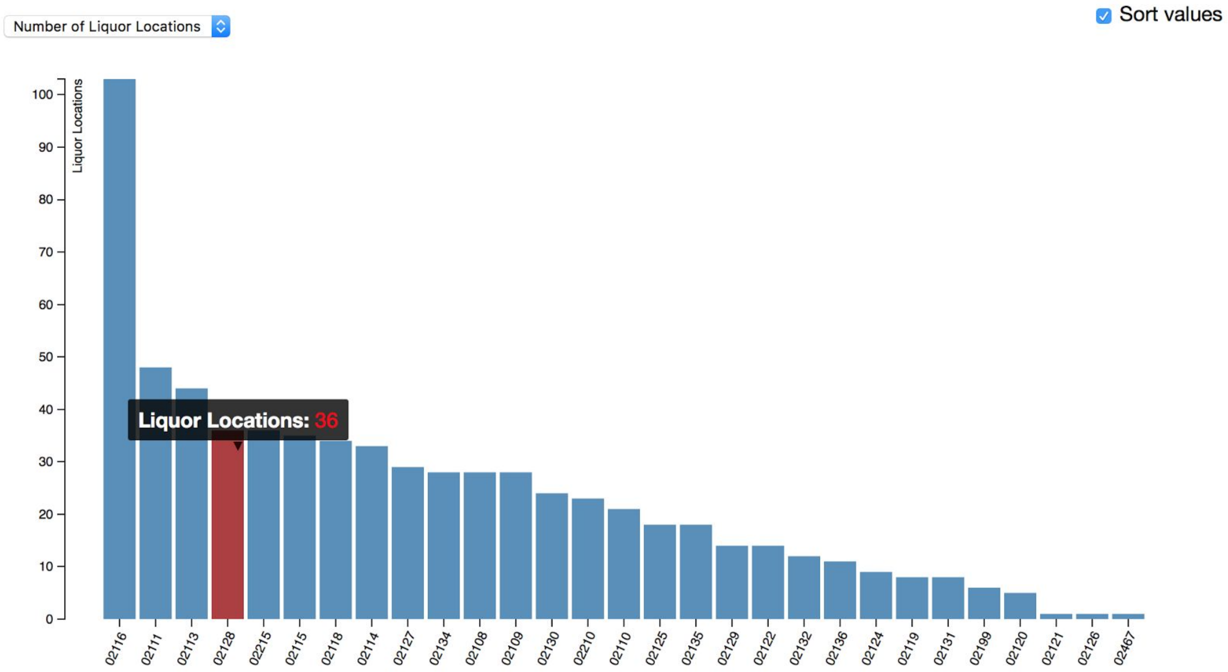
It would then display the profiles for those zip codes.

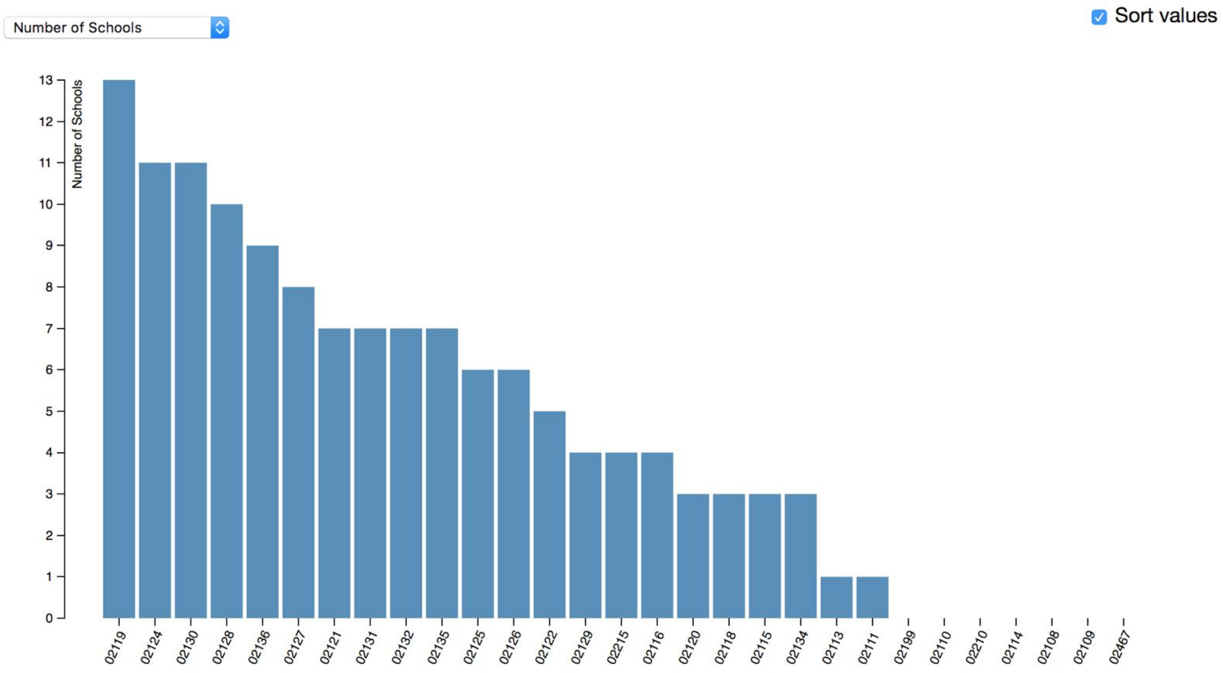
Results and Conclusions

Now we present some visualizations of our data we constructed:

Visualizations

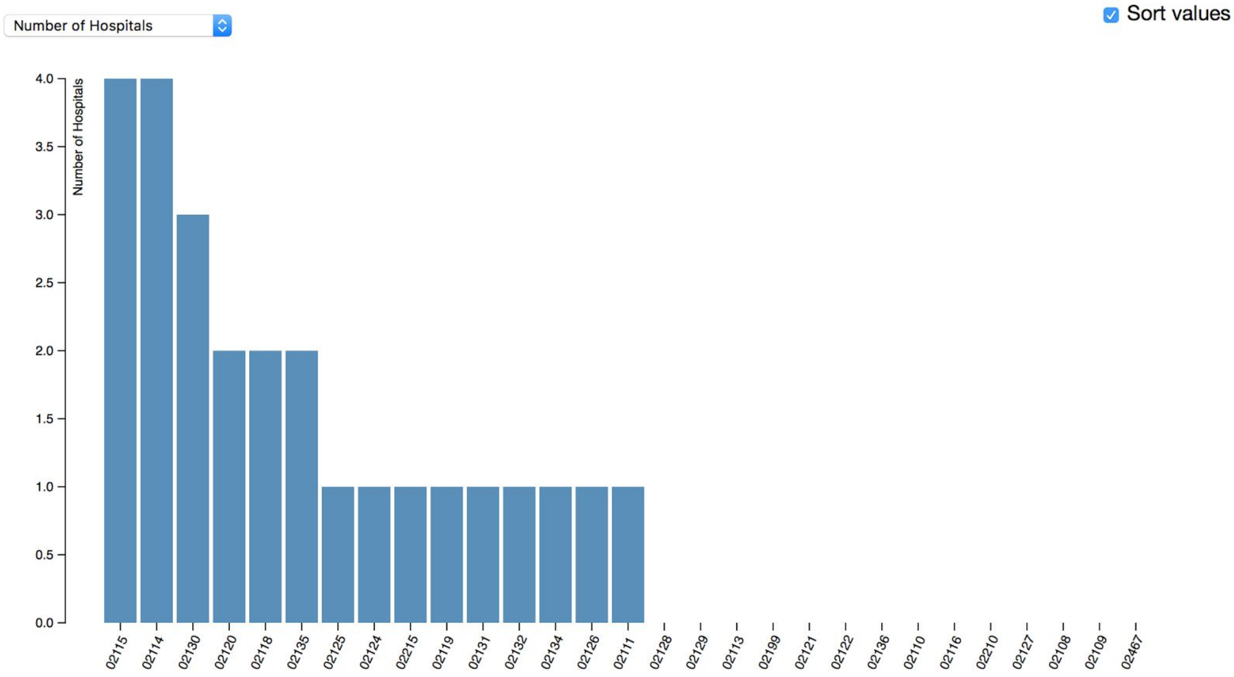






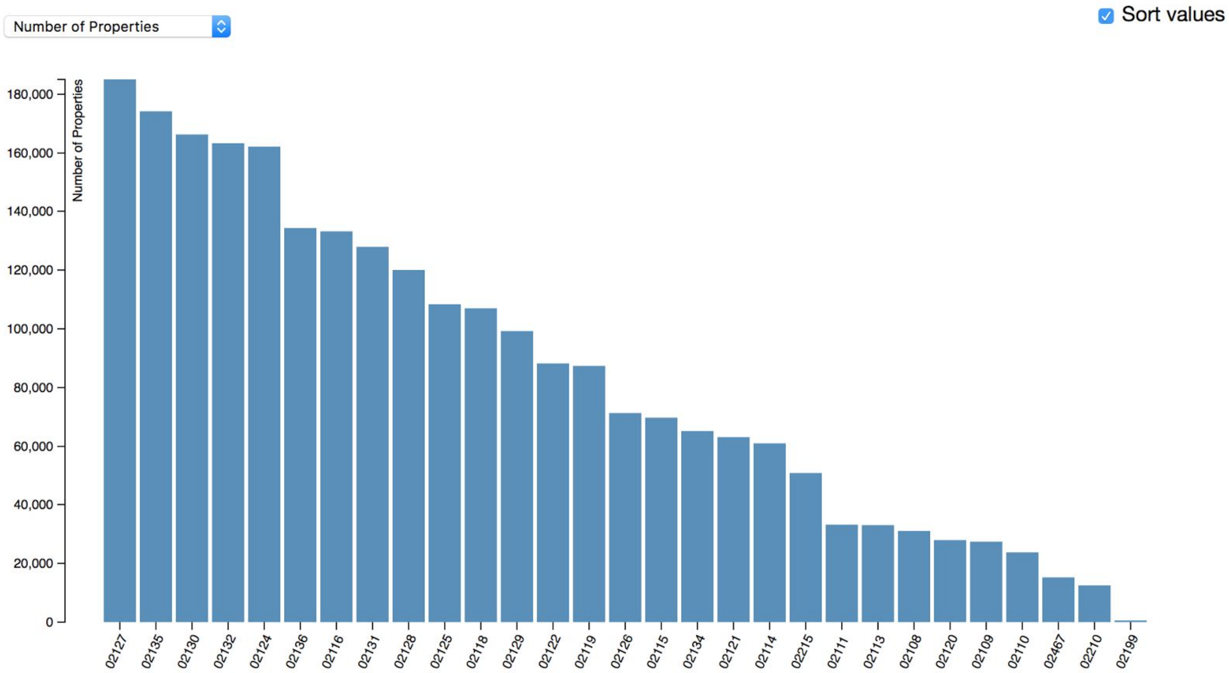
Number of Schools

☑ Sort values



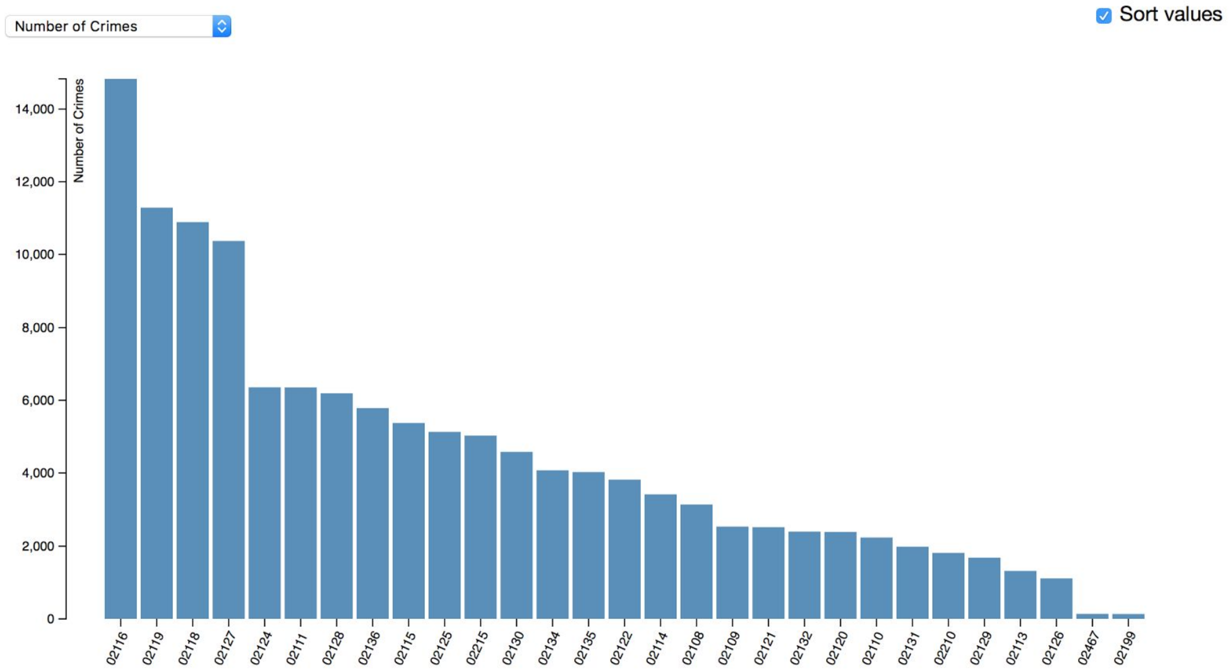
Number of Hospitals

☑ Sort values



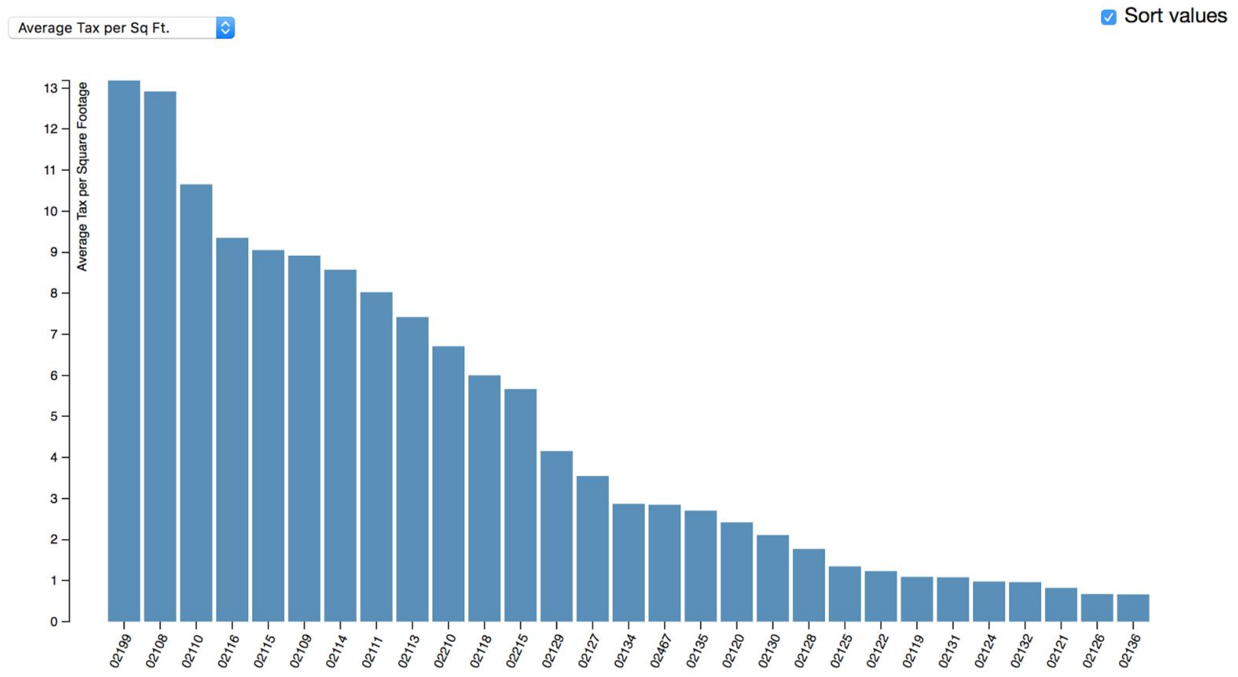
Number of Properties

☑ Sort values



Number of Crimes

☑ Sort values



Average Tax per Sq Ft.

☑ Sort values

Conclusions

We did find some useful results with our optimization feature, and believe that with a little refinement on how to correlate crimes to liquor locations, it will be even better. Additionally, we found some interesting information such as: The Prudential's zip code has the highest tax per square foot!

Problems and Looking Forward

The biggest problem we encountered was the zip code information being copyrighted, which set us back a lot. One thing we want to improve upon would be the way we correlate crime to a liquor location. Using this to obtain zip code is probably fine (we would probably hit nearly 80-100% of the crimes by extended our range of correlation), but this is not enough to deem a crime alcohol related, and we should do some filtering to secure that. Furthermore, we can separate crime types in the zip code profile to make it more robust and informative. Additionally, we would like to add even more details and information to the profile dataset. I would say our correlations remain an open problem, but the optimizations, with the current data are solved, and could be improved upon with more data.

Other problems we encountered were with data formats, as different datasets stored zip codes as integers and others as strings; some datasets stored the street address differently; some datasets also saved coordinates as strings.