Project 3

NYC 311

Revealing Race and Income in 311 Calls

Jonathan Yip and Tommy Heffernan
TF: Chelsea Yeh

In Project 3, we decided to build on the NYC 311 call data we used in Project 2. One of the most intriguing insights we saw while exploring the data was that some complaints seemed to correlate with income and ethnicity. In order to further explore this possibility, we used data from the US Census Bureau's American Community Survey to show users the correlations between race/income and complaints as well as the different quality of life issues that are most prevalent for different socioeconomic groups.

I. Project 3 Proposal

1. Increased Data and Views

Our initial avenue for expansion and improvement is to use more of the data in the 311 dataset. As we discussed earlier, the 311 data is absolutely staggering in size and comprehensiveness. Given the trade-offs between the quantity of data and user experience/computational resources, we do not think that expanding the timeframe would be prudent (each additional year would probably add tens of MBs). But, we have only grazed the surface in the complaint descriptors/types, and their computational overhead will be relatively smaller.

In thinking about our storytelling, our first thoughts revolve around the powerful message of socio-economic discrepancy in complaints. As we noted in Project 2, there appear to be some complaints that are strongly correlated with socio-economic status of neighborhoods, including heating and broken elevators. As such, we are considering building Project 3 around investigating this trend further in depth. Other possible complaints that may hew to this insight include "Homeless Encampments," "Benefit Card Replacements," "Asbestos," and others.

Furthermore, in order to make these trends more visible and quantifiable (especially to viewers not from New York), we propose including demographic data not in the 311 dataset, particularly median income, perhaps from census data.

To expand the actual visualization, our initial proposal is to add more views. One such idea we have discussed is to add a scatterplot that will allow for brushing with the map and show correlations with census/median income data. This will allow users to more fully explore our storytelling about income/311 trends.

2. Storytelling

In thinking about how to include more storytelling in our Project 2 exploratory data visualization, we focused on the question that Alberto Cairo posed: "So what?" One of the most interesting stories that emerged from Project 2 was the discrepancy in complaints by socio-economic neighborhood. While there were plenty of amusing or interesting trends (loud party complaints are highest in the Lower East Side!), we thought that the socio-economic ones were the most insightful and revealing. That said, the success of this storytelling element is dependent on the patterns that emerge in other complaint types. If it turns out that few other intriguing socio-economic trends emerge, then we will have to reconsider the story.

3. Implementation Plan

The rest of our project will likely also be implemented in D3 and Javascript, so no new technologies are expected.

4. Timeline

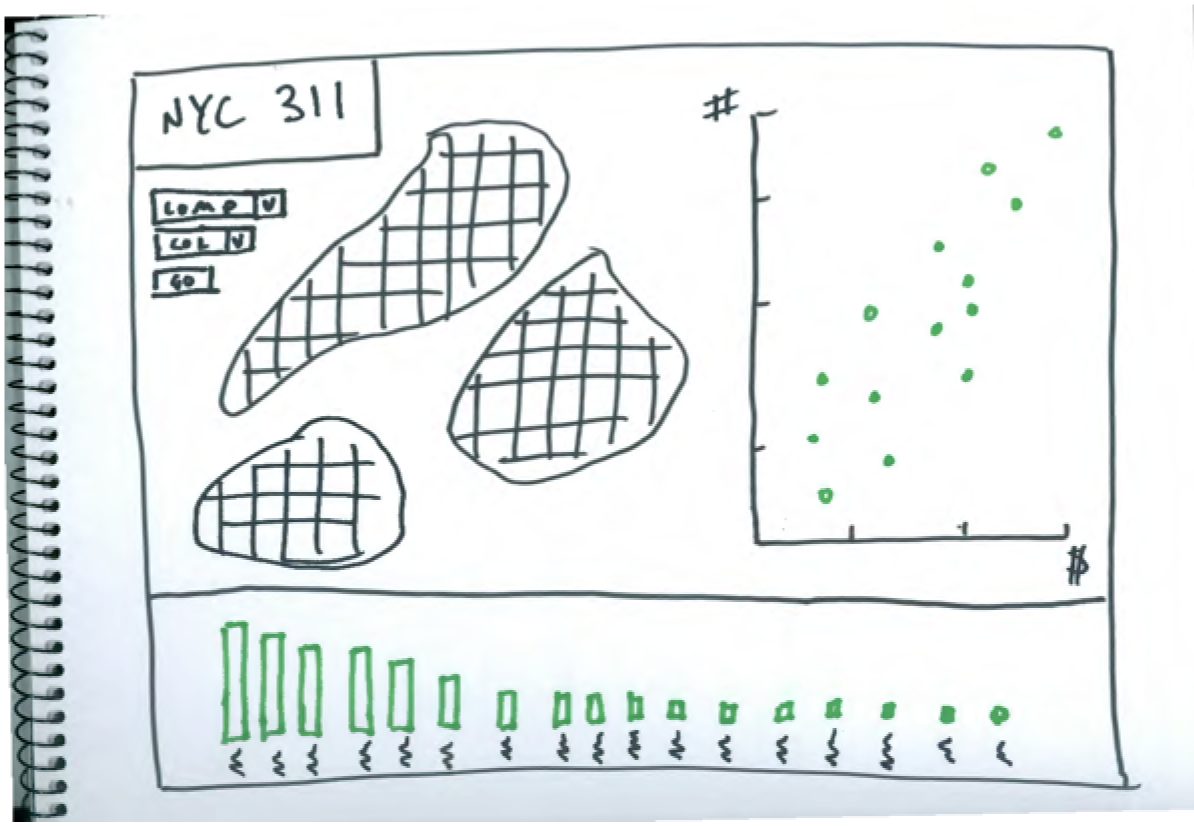
Tasks:

- Pull more data from NYC Open Data
- Pull income/demographic data from Census or otherwise
- Pre-filter/clean data
- Preliminary mapping to investigate trends in data
- Finish sketching/designing end layout
- Implement new scatterplot view
- Implement brushing with map and scatterplot
- Layout adjustments
- Finalize screencast, processbook, readme, etc.

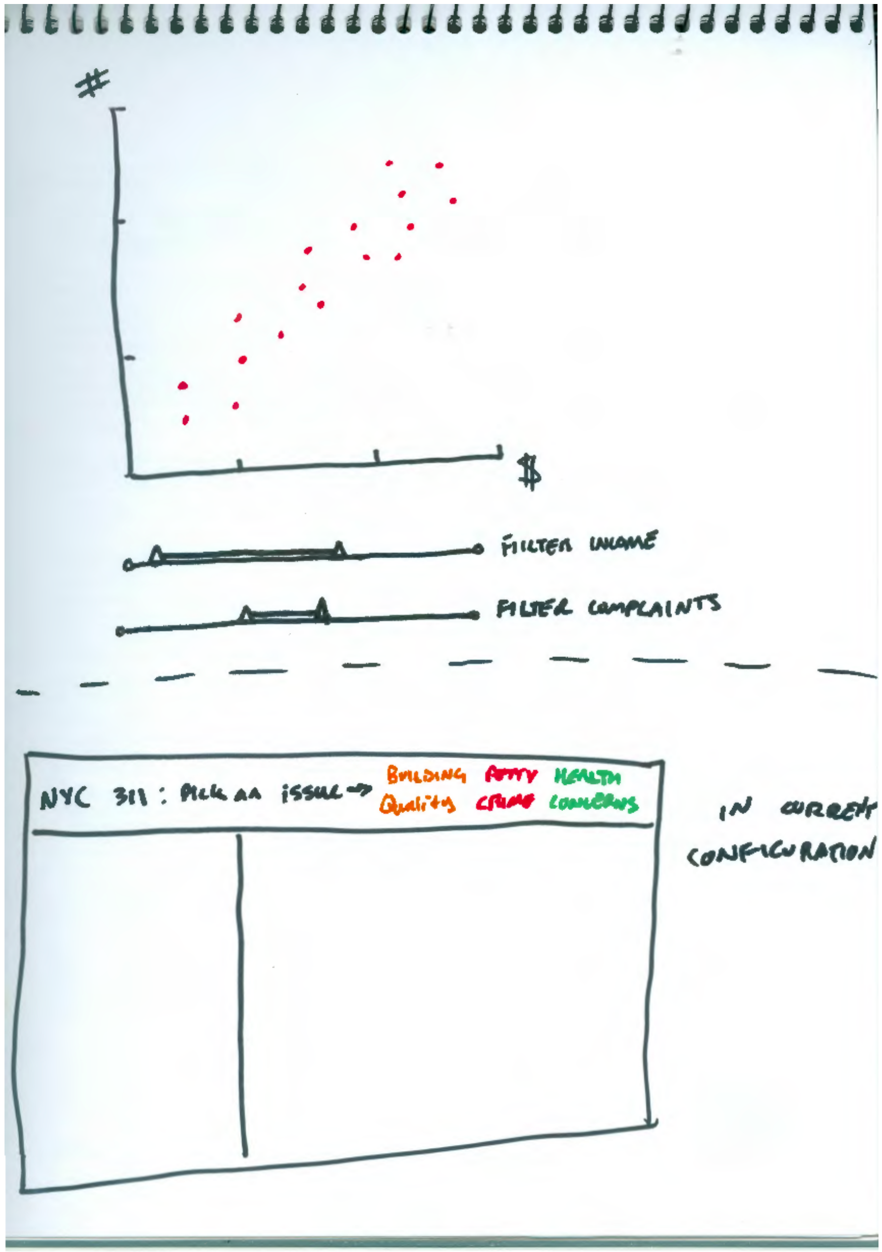
Our sense is that we would like to finish collecting and refining data as well as have final sketches within the next week. After that, we will spend the following week actually implementing the new views and data.

Below are some sketches to illustrate our ideas for project 3.

This sketch depicts an alternate layout, which truncates the bar charts to make room for a scatterplot, which would plot complaints vs. socioeconomic variables.



This sketch depicts a similar layout, but in a different configuration. In this one, the bar chart is now vertical, so that the entirety of it can be shown. Additionally, the choropleth and scatterplot are in the same “window,” as they both focus on one issue, while the bar chart is separate, as it focuses on one zip code.



The first part of this sketch is a diagram of possible filtering for the scatterplot. Depicted are sliders to filter by number of complaints and by income level, but others could be added.

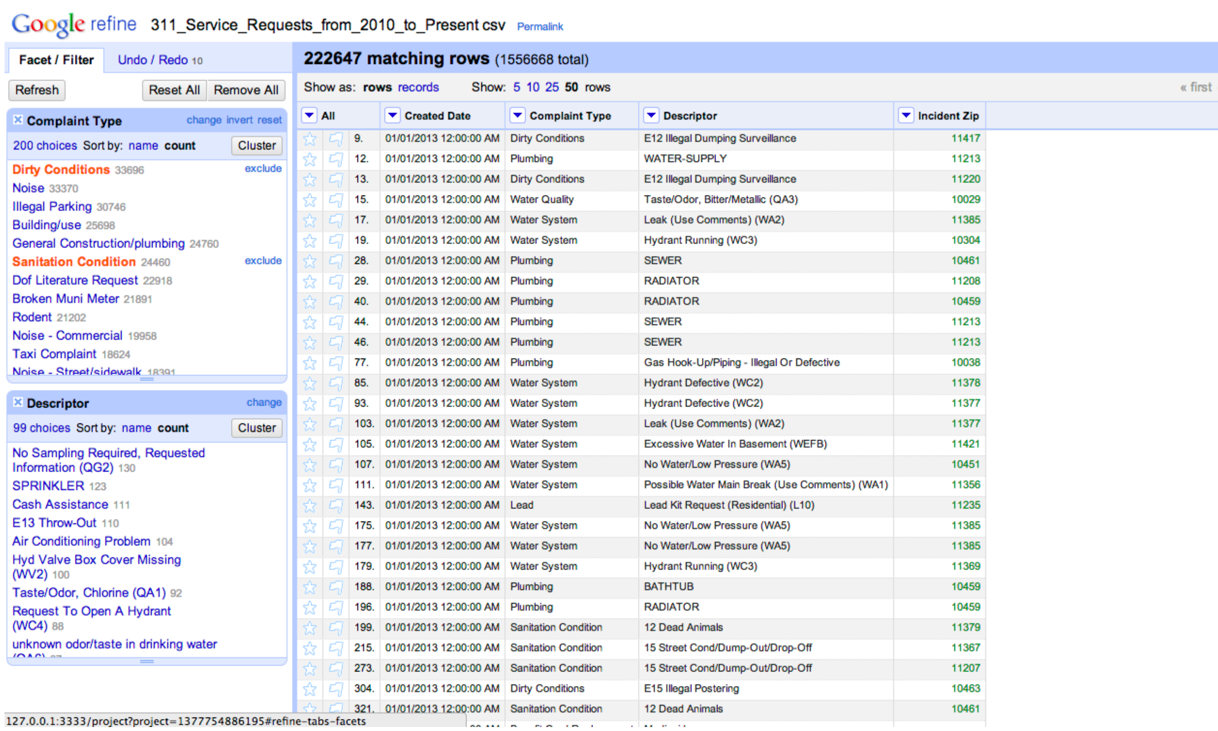
The second part is a storytelling sketch. It depicts buttons which would set the visualization to a certain view, as well as creating annotations, with the purpose of creating a storytelling layout.

II. Data

In order to carry out our expansion plan, we had to both reconsider the existing 311 data that we had as well as gather new data. First, while we had assembled an interesting and novel dataset for Project 2, with a variety of different 311 complaints from NYC's Open Data initiative, we understood that not all of it was relevant or important for our storytelling. Ice cream truck complaints were amusing and neat, but sadly added nothing to our attempt to reveal class and income disparities in 311 complaints.

As such, we went back to the original NYC 311 data to extract different complaint categories. After once again importing our data into Google Refine, we identified complaints that we thought might have correlation with income or other demographic data (age, race). Some notable categories included Plumbing, Lead, Asbestos, Vermin, Medicare Card Replacement, Food Stamp Card Replacement.

We, again, faced the issue that Complaint Type contained a more specific Descriptor. As can be seen in the screenshot below, "Water System" contains, among many others, "Hydrant Defective (WC2)", "Leak", "Excessive Water in Basement", "No Water/Low Pressure", "Possible Water Main Break", etc.



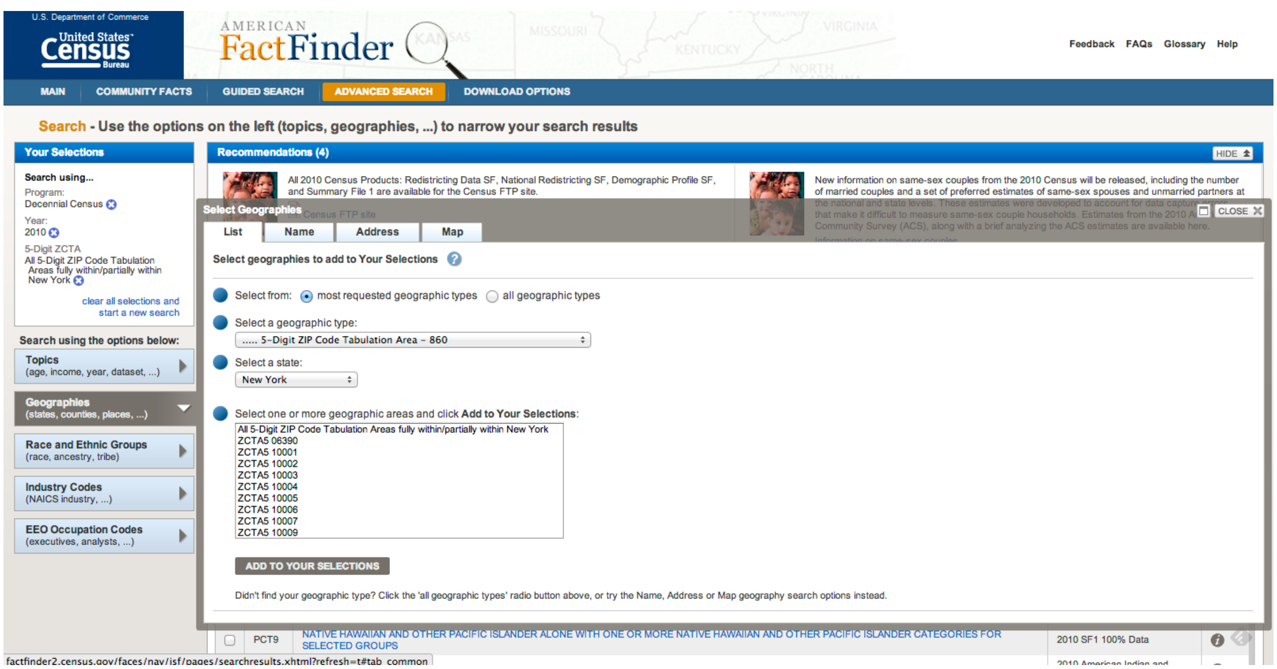
Even worse, neither category was at the exactly appropriate level of detail (Plumbing is specific enough—knowing whether a bathtub or shower is at fault is not important [see screenshot]—but in Road Conditions, we were interested in a subset called Potholes). As such, some of the Complaint Descriptors were the categories we were interested in while in other cases, the Complaint Type was. To overcome this issue, we combed through each Complaint Type and Description to find categories we wanted. Once we identified them, we had to merge the two columns manually, choosing the unified name as appropriate.

The final complaints we chose were: "Fire Hydrant Emergency (FHE)", "Rat Sighting", "Loud Music/Party", "Pothole", "School Maintenance", "Boilers", "Dirty Conditions","Broken Elevator", "Heating", "Plumbing", "Double Parked Blocking Vehicle", "Taxi Driver Complaint", "Vermin", "Dead Animals", "Graffiti", "Street Light Out", "Derelict Vehicle", "Sewer Backup."

Census Data

Next, we had to obtain the demographic data that would allow us to observe income/race trends in the 311 complaint data. We first turned to the US Census website. There, we learned that data collection in the United States is quite complicated. The decennial census that we are all familiar with was not particularly useful as it was not conducted in ZIP codes, but rather in Census plots. Recoding all the Census data into ZIP codes was simply not feasible, so instead we turned to the American Community Survey.

The ACS is a statistically based survey run by the Census Bureau which collects information similar to the decennial census but on a more frequent basis. The ACS is also presented in “Zip Code Tabulation Areas,” which we discovered are essentially equivalent to ZIP codes. ZIP codes as postal codes are not formally defined and change frequently according to postal needs. Thus, the Census Bureau defined “Zip Code Tabulation Areas” as more constant and reliable definitions of area. For our purposes, they were essentially interchangeable as discrepancies were block-sized and would not affect our visualization in any meaningful way.



At first glance, the ACS had too much data, including labor force participation by demographic status, commuting statistics, and much, much more.

View Geography Notes | View Table Notes

Although the American Community Survey (ACS) produces population, demographic and housing unit estimates, it is the Census Bureau's Population Estimates Program that produces and disseminates the official estimates of the population for the nation, states, counties, cities and towns and estimates of housing units for states and counties.

1 - 18 of 7,176

Subject	ZCTA5 06390				ZCTA5 10001				ZCTA5 10002				ZCTA5 10003				ZCTA5 10004	
	Estimate	Margin of Error	Percent	Percent Margin of Error	Estimate	Margin of Error	Percent	Percent Margin of Error	Estimate	Margin of Error	Percent	Percent Margin of Error	Estimate	Margin of Error	Percent	Percent Margin of Error	Estimate	Margin of Error
EMPLOYMENT STATUS																		
Population 16 years and over	247	+/-108	247	(X)	19,050	+/-1,029	19,050	(X)	70,231	+/-2,024	70,231	(X)	52,037	+/-1,741	52,037	(X)	2,326	+/-260
In labor force	156	+/-63	63.2%	+/-21.7	13,700	+/-918	71.9%	+/-2.2	39,827	+/-1,613	56.7%	+/-1.6	35,834	+/-1,276	68.9%	+/-2.1	2,092	+/-242
Civilian labor force	156	+/-63	63.2%	+/-21.7	13,700	+/-918	71.9%	+/-2.2	39,827	+/-1,613	56.7%	+/-1.6	35,834	+/-1,276	68.9%	+/-2.1	2,092	+/-242
Employed	156	+/-63	63.2%	+/-21.7	12,501	+/-873	65.6%	+/-2.6	35,665	+/-1,557	50.8%	+/-1.6	33,377	+/-1,191	64.1%	+/-1.9	1,904	+/-247
Unemployed	0	+/-89	0.0%	+/-12.3	1,199	+/-337	6.3%	+/-1.7	4,162	+/-555	5.9%	+/-0.8	2,457	+/-373	4.7%	+/-0.7	188	+/-94
Armed Forces	0	+/-89	0.0%	+/-12.3	0	+/-89	0.0%	+/-0.2	0	+/-89	0.0%	+/-0.1	0	+/-89	0.0%	+/-0.1	0	+/-89
Not in labor force	91	+/-82	36.8%	+/-21.7	5,350	+/-465	28.1%	+/-2.2	30,404	+/-1,403	43.3%	+/-1.6	16,203	+/-1,381	31.1%	+/-2.1	234	+/-96
Civilian labor force	156	+/-63	156	(X)	13,700	+/-918	13,700	(X)	39,827	+/-1,613	39,827	(X)	35,834	+/-1,276	35,834	(X)	2,092	+/-242
Percent Unemployed	(X)	(X)	0.0%	+/-18.7	(X)	(X)	8.8%	+/-2.3	(X)	(X)	10.5%	+/-1.3	(X)	(X)	6.9%	+/-1.0	(X)	(X)
Females 16 years and over	125	+/-61	125	(X)	10,034	+/-794	10,034	(X)	36,909	+/-1,369	36,909	(X)	27,421	+/-1,335	27,421	(X)	898	+/-192
In labor force	64	+/-30	51.2%	+/-21.0	6,392	+/-662	63.7%	+/-3.4	18,946	+/-1,007	51.3%	+/-2.0	17,526	+/-899	63.9%	+/-3.0	807	+/-193
Civilian labor force	64	+/-30	51.2%	+/-21.0	6,392	+/-662	63.7%	+/-3.4	18,946	+/-1,007	51.3%	+/-2.0	17,526	+/-899	63.9%	+/-3.0	807	+/-193
Employed	64	+/-30	51.2%	+/-21.0	5,755	+/-660	57.4%	+/-4.1	17,130	+/-996	46.4%	+/-2.1	16,470	+/-861	60.1%	+/-2.9	689	+/-183
Own children under 6 years	0	+/-89	0	(X)	810	+/-257	810	(X)	3,691	+/-592	3,691	(X)	1,549	+/-309	1,549	(X)	247	+/-161
All parents in family in labor force	0	+/-89	-	**	499	+/-251	61.6%	+/-18.4	2,225	+/-511	60.3%	+/-7.8	896	+/-238	57.8%	+/-10.2	220	+/-157
Own children 6 to 17 years	47	+/-36	47	(X)	1,329	+/-330	1,329	(X)	8,054	+/-817	8,054	(X)	1,840	+/-308	1,840	(X)	36	+/-36
All parents in family in labor force	23	+/-22	48.9%	+/-40.7	1,035	+/-363	77.9%	+/-13.6	4,960	+/-592	61.6%	+/-7.1	1,193	+/-255	64.8%	+/-9.1	36	+/-36
COMMUTING TO WORK																		
Workers 16 years and over	156	+/-63	156	(X)	12,321	+/-874	12,321	(X)	34,909	+/-1,558	34,909	(X)	32,790	+/-1,165	32,790	(X)	1,883	+/-248
Car, truck, or van -- drove alone	130	+/-52	83.3%	+/-10.6	589	+/-190	4.8%	+/-1.5	2,516	+/-432	7.2%	+/-1.1	1,320	+/-292	4.0%	+/-0.9	124	+/-69

The Census Bureau's software proved cryptic, and we were presented with the above table of every piece of data for every ZIP code at once, including statistical information (margin of error) Luckily, we soon understood that one unlabeled button allowed us to select particular categories to keep. In the screenshot below, we have selected Median and Mean Income.

1 - 18 of 14,352

Subject	ZCTA5 06390								ZCTA5 10001								ZCTA5 10002	
	Households		Families		Married-couple families		Nonfamily households		Households		Families		Married-couple families		Nonfamily households		Households	
	Estimate	Margin of Error	Estimate	Margin of Error	Estimate	Margin of Error	Estimate	Margin of Error	Estimate	Margin of Error	Estimate	Margin of Error	Estimate	Margin of Error	Estimate	Margin of Error	Estimate	Margin of Error
Total	131	+/-50	86	+/-49	80	+/-48	45	+/-28	10,473	+/-447	3,029	+/-374	1,969	+/-286	7,444	+/-433	32,573	+/-617
Less than $10,000	0.0%	+/-21.7	0.0%	+/-30.5	0.0%	+/-32.1	0.0%	+/-44.2	8.6%	+/-1.9	6.8%	+/-3.5	0.8%	+/-1.0	9.3%	+/-2.3	17.5%	+/-1.5
$10,000 to $14,999	4.6%	+/-6.7	7.0%	+/-11.1	0.0%	+/-32.1	0.0%	+/-44.2	4.1%	+/-1.4	3.2%	+/-2.0	2.5%	+/-2.4	4.8%	+/-1.7	11.5%	+/-1.3
$15,000 to $24,999	12.2%	+/-10.6	0.0%	+/-30.5	0.0%	+/-32.1	35.6%	+/-27.6	10.6%	+/-2.4	8.5%	+/-4.4	3.3%	+/-2.1	11.5%	+/-2.6	13.7%	+/-1.4
$25,000 to $34,999	6.9%	+/-5.9	3.5%	+/-6.0	3.8%	+/-6.4	13.3%	+/-20.6	8.5%	+/-2.2	2.2%	+/-1.3	1.9%	+/-1.8	10.8%	+/-3.1	9.4%	+/-1.1
$35,000 to $49,999	0.0%	+/-21.7	0.0%	+/-30.5	0.0%	+/-32.1	0.0%	+/-44.2	9.5%	+/-1.8	13.1%	+/-3.7	11.0%	+/-4.5	8.6%	+/-2.2	9.6%	+/-1.3
$50,000 to $74,999	5.3%	+/-6.5	4.7%	+/-7.2	5.0%	+/-7.8	6.7%	+/-13.9	11.3%	+/-2.3	11.7%	+/-4.4	8.8%	+/-4.6	11.3%	+/-2.7	14.6%	+/-1.4
$75,000 to $99,999	10.7%	+/-10.4	5.8%	+/-10.7	6.3%	+/-11.7	20.0%	+/-22.7	9.8%	+/-1.7	11.7%	+/-4.3	9.6%	+/-5.0	9.0%	+/-2.0	8.3%	+/-1.3
$100,000 to $149,999	8.4%	+/-8.6	12.8%	+/-12.7	13.8%	+/-13.7	0.0%	+/-44.2	13.4%	+/-2.4	10.9%	+/-4.0	15.4%	+/-5.8	13.8%	+/-2.8	7.9%	+/-1.2
$150,000 to $199,999	34.4%	+/-22.7	52.3%	+/-23.7	56.3%	+/-23.4	0.0%	+/-44.2	8.0%	+/-2.0	8.9%	+/-3.9	13.1%	+/-5.7	7.6%	+/-2.1	3.7%	+/-0.8
$200,000 or more	17.6%	+/-16.9	14.0%	+/-13.9	15.0%	+/-15.2	24.4%	+/-33.0	16.3%	+/-2.9	23.1%	+/-5.0	33.6%	+/-7.3	13.3%	+/-3.0	4.0%	+/-0.6
Median income (dollars)	151,250	+/-89,088	165,400	+/-18,963	165,700	+/-9,552	65,417	+/-89,948	67,795	+/-6,489	81,319	+/-14,794	142,049	+/-20,945	58,083	+/-10,967	32,407	+/-2,484
Mean income (dollars)	123,712	+/-20,843	144,636	+/-23,015	N	N	83,724	+/-57,059	121,403	+/-11,145	157,583	+/-21,573	N	N	105,145	+/-12,234	57,063	+/-2,629
PERCENT IMPUTED																		
Household income in the past 12 months	45.0%	(X)	(X)	(X)	(X)	(X)	(X)	(X)	19.3%	(X)	(X)	(X)	(X)	(X)	(X)	(X)	31.1%	(X)
Family income in the past 12 months	(X)	(X)	38.4%	(X)	(X)	(X)	(X)	(X)	(X)	(X)	24.2%	(X)	(X)	(X)	(X)	(X)	(X)	(X)
Nonfamily income in the past 12 months	(X)	(X)	(X)	(X)	(X)	(X)	57.8%	(X)	(X)	(X)	(X)	(X)	(X)	(X)	16.9%	(X)	(X)	(X)

Source: U.S. Census Bureau, 2007-2011 American Community Survey

Similarly, we were able to remove extraneous statistical margin of error information:

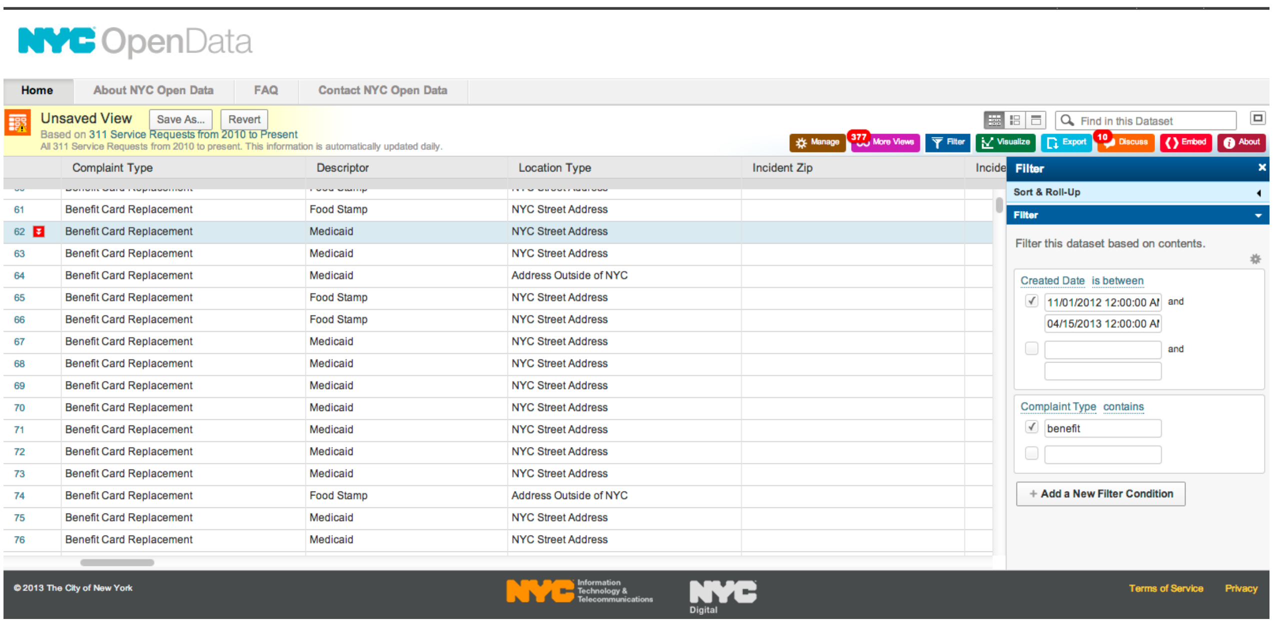
1 - 18 of 1,794

Subject	ZCTA5 06390	ZCTA5 10001	ZCTA5 10002	ZCTA5 10003	ZCTA5 10004	ZCTA5 10005	ZCTA5 10006	ZCTA5 10007	ZCTA5 10009	ZCTA5 10010	ZCTA5 10011	ZCTA5 10012	ZCTA5 10013	ZCTA5 10014	ZCTA5 10016	ZCTA5 10017	ZCTA5 10018
	Households	Households	Households	Households	Households	Households	Households	Households	Households	Households	Households	Households	Households	Households	Households	Households	Househol
	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate
Median income (dollars)	151,250	67,795	32,407	88,601	127,448	117,885	111,617	191,900	55,316	94,242	99,700	77,072	64,806	102,747	102,796	105,974	84,85
Mean income (dollars)	123,712	121,403	57,063	140,822	245,827	159,876	150,366	347,623	83,700	155,208	181,230	134,811	172,252	161,283	156,390	157,961	126,49

Source: U.S. Census Bureau, 2007-2011 American Community Survey

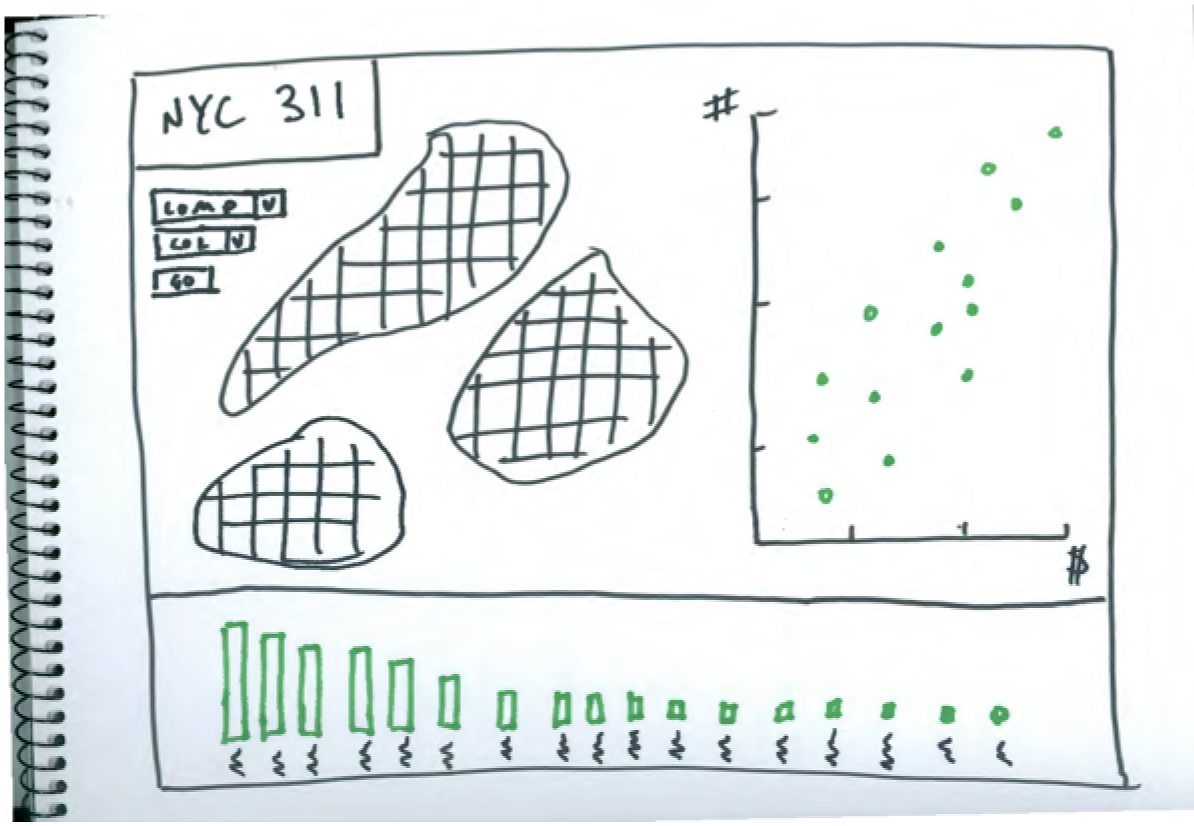
With our data in CSVs, we modified our Python script from Project 2 and used it to convert the data into JSON.

Unfortunately, some of our most promising data, Benefit Card Replacements (people calling to replace lost or damaged benefit cards), were not usable. We had expected that complaints about Medicare or Food Stamp cards would correlate very highly with income levels. However, for an unknown reason, these complaints did not have Incident Zip data (see below).



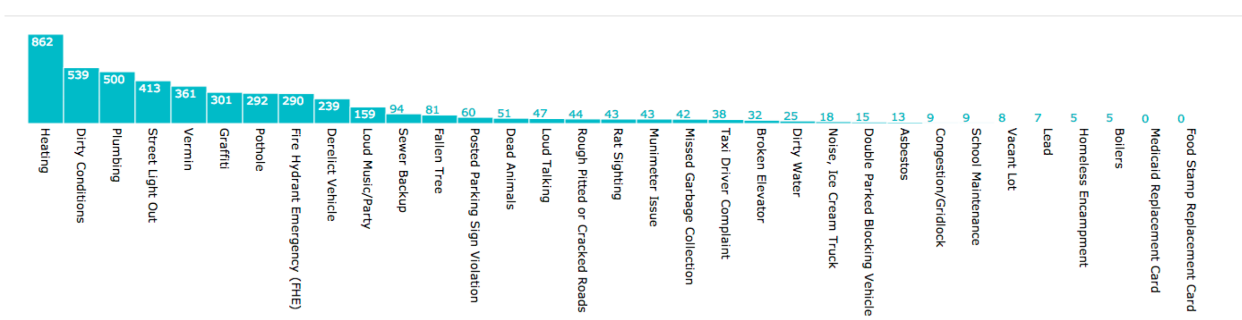
III. Implementation

We ultimately decided to implement a design quite similar to one of our original sketches:



Given the storytelling component, we wanted to emphasize the scatterplot over the bar graph. While the bar graph was an important component of Project 2, in Project 3 it had relatively little to add to our thesis about complaints and demographics. It only showed complaints per ZIP code and did not relate the data to income or race. However, we did decide to keep it given that it added context for the user in exploring how popular given complaints were on an absolute level (not just relative, as the choropleth shows).

As such, we moved the bar chart down below the map and the scatterplot and expanded it to make better use of the newly available space. Still, we made sure to retain the essential elements of alignment and contrast so that readability was still high.

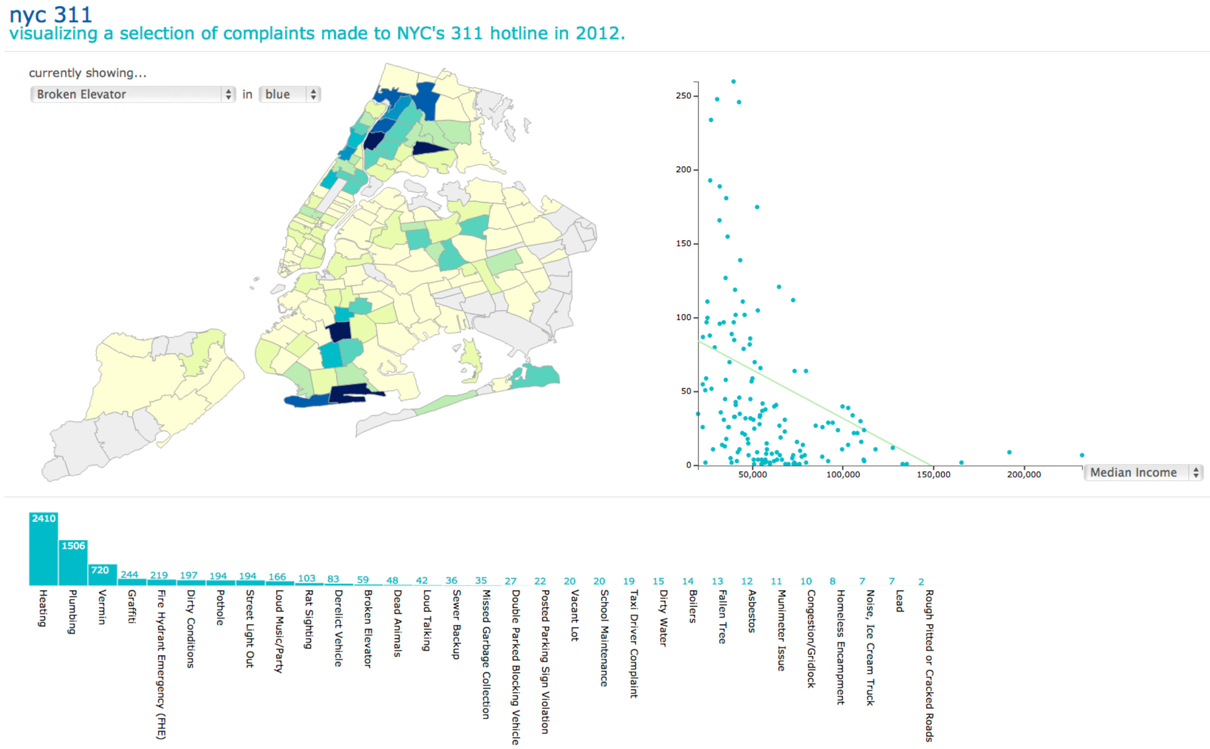


In putting the map and scatterplot side-by-side, we wanted to emphasize to users that the two views were interrelated and essential to fully understanding the story we wanted to tell. In order to emphasize this fact, we linked the two views so that brushing a particular ZIP code on the map or a particular datapoint on the plot would highlight its corresponding representation in the other view. This not only allowed users to see overarching trends, but also to explore where particular ZIP codes fell in the distribution (or vice versa).

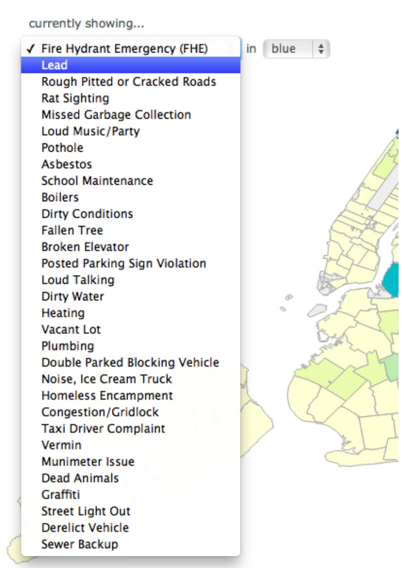
The screenshot below shows the early stages of our implementation of the scatterplot using D3:

At this point, we had also made the aesthetic decision to revert to a lighter color scheme. We felt that while the dark and black palette allowed the map to stand out previously, when we added more views (extending the bar chart and including the scatterplot), the dark colors felt cluttered and oppressive. The lighter, white-based appearance allowed for more breathability and white-space in the visualization. We also thought it might be less harsh on the eyes and seem more professional.

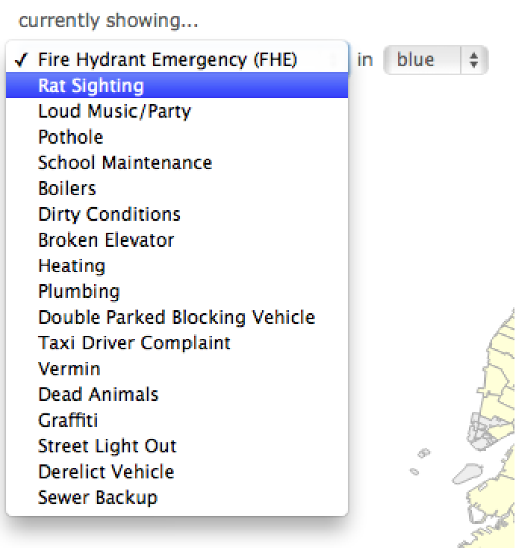
We were pleased with our initial rough implementation, but we felt that the scatterplot was not adequately conveying the trends in the data. As such, in order to make patterns more clear, we added a least squares linear trendline. As can be seen in the screenshot below, the line makes certain associations very visible:



At this point, we moved to refine our visualization. First, we decided to remove extraneous complaint and demographic data. Our initial list of 311 complaints was unnecessarily long and only proved overwhelming for users. At the same time, it made the bar chart below the map confusingly long (especially when the number of complaints in a category dropped off quickly—see above). We managed to prune this list:

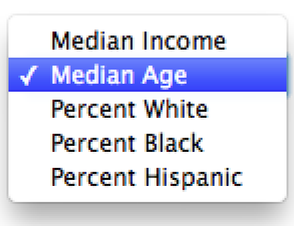


into a more manageable drop-down menu:

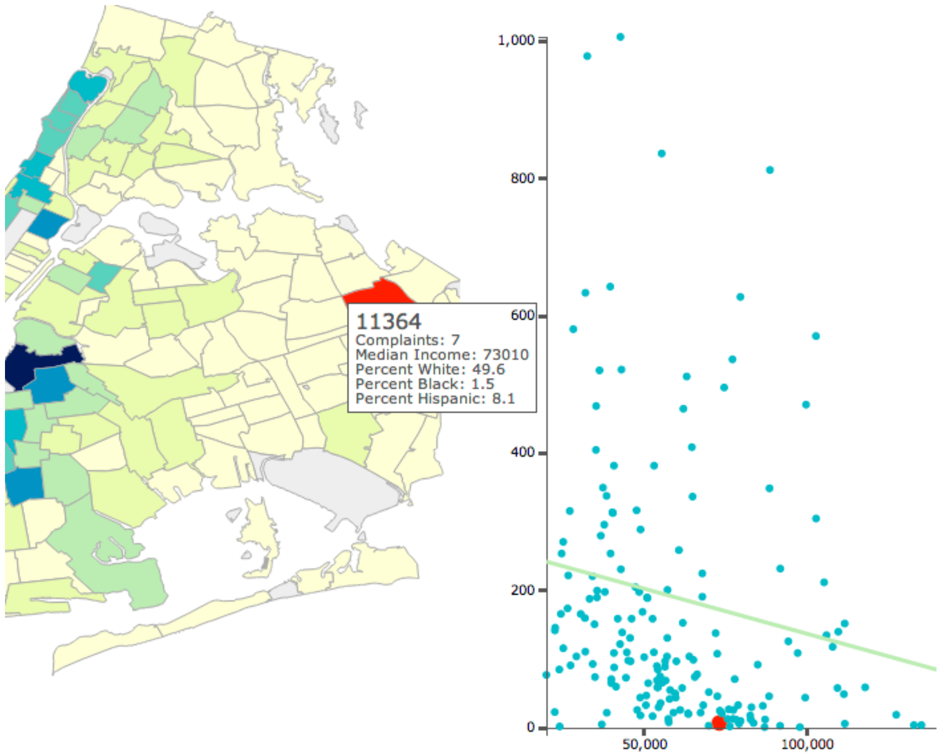


Similarly, with the demographic data, we had redundancies in keeping both median and mean income as well as extraneous information like total population. We decided to keep median income and discard the mean given large outliers, especially in New York. Total population showed meaningless correlations as the number of complaints did—surprise—correlate very highly with the population of a ZIP code.

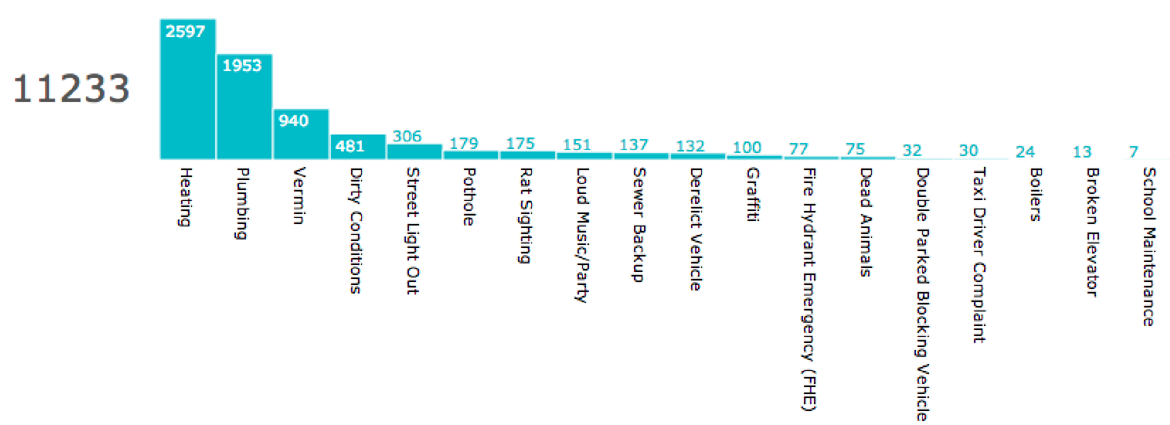
Median age proved to be a more complex decision. However, in examining potential trends across all the complaints, median age did not show any correlations. Furthermore, we had initially included age because we believed it would correlate with the number of Medicare complaints. However, without that data, age proved to be less relevant, and so we discarded it.



With our data finalized, we turned our attention to fine-tuning different parts of our implementation. One very important area was polishing the hovering and linking. Initially, we used a grey color to highlight the ZIP code in both views, but we found that it did not stand out enough. So, we switched it to a high-contrast color from other Color Brewer palettes (the contrast color changes in the different color schemes that we have: blue, red, purple, and green). We also decided to expand the tool-tip with more information, including the demographic data on income and ethnicity by ZIP code (see below).



Our final refinement with regards to the overall implementation was to finally fix a design issue that began in Project 2: clarifying for users that the bar chart was specific to an individual ZIP code. We solved the confusion by labeling the bar chart very clearly, which also had the added benefit of using balancing the white-space below the map/scatterplot:



IV. Storytelling

At this point, we knew to a large degree the stories that the combined data was telling us, and our focus now was to find a way to convey them to our viewers. We made two significant decisions about how to tell that story. First, we decided to align the title of our visualization with our discovery. It was no longer simply about exploring 311 data—we wanted to let users know that the 311 and Census data revealed some striking trends in who was facing the brunt of New York's quality of life issues.

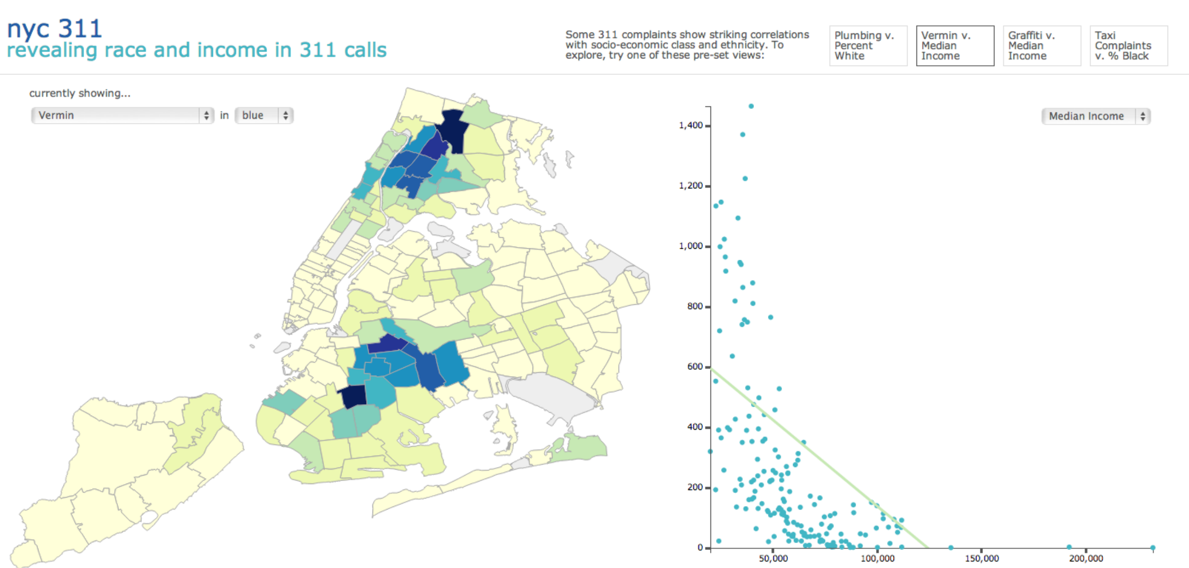
nyc 311
revealing race and income in 311 calls

Then, we devised a short sentence explaining our thesis to viewers. At the same time, we wanted them to draw their own conclusions without our being heavy-handed. As such, we decided to present the audience with 4 pre-set views that demonstrated the trends we had insinuated.

Some 311 complaints show striking correlations with socio-economic class and ethnicity. To explore, try one of these pre-set views:

Plumbing v. Percent White | Vermin v. Median Income | Graffiti v. Median Income | Taxi Complaints v. % Black

When a user selects one of the views, the box's border is darkened, and the visualization transitions to that view. Below, you can see Vermin v. Median Income, and a very strong negative trend can be seen, indicating that vermin complaints correlate very strongly with low incomes. The map further shows that the complaints are concentrated in traditionally low-income neighborhoods.



V. Conclusion

Project 3 has allowed us to improve our visualization significantly. While Project 2 enabled us to explore the fascinating NYC 311 data, we have now been able to create a more polished, cleaner, and more usable visualization that actually reveals deeper insight. To a degree we had not expected, certain NYC 311 complaints correlate very strongly with income and race. While this is not a stunning conclusion, it does demonstrate that standards of living and quality of life issues vary significantly by neighborhood and especially by income.

We really enjoyed working on Project 3, and we hope you find it as interesting as we did. Thank you.