A Web-Based Mapping Interface for Site Selection

YI XU

Abstract

The procedures of using GIS for site selection can be sophisticated. The work was usually completed by an entire data analyst team. With the help of scripting language, API and online database, an individual has a chance to access massive open data and built applications for site selection. This paper gave a try to build a web-based mapping interface for site selection for Dunkin' Donuts. The core of the application is an OLS regression model. The platform for storing data and building the map is CartoDB.

Introduction

GIS is a powerful software technology for site selection. It allows companies to consider many possibilities, understand potential, review the impact of different investments, store and produce configurations, and analyze changing trends in the retail landscape. Companies like Starbucks use Esri's Business Analyst Online for site selection. Esri's Business Analyst Online is one of the most representative solutions for retail business processes including market planning, site selection, and customer segmentation. It utilizes GIS-based site selection analysis. Although with the help of accurate data, Esri's Business Analyst Online is quite powerful, the pricing is high and the reports are sophisticated. The basic plan starts from $1,500 and the standard plan stars from $7,000. It may not be expensive for big companies but is absolutely not cheap for individuals or small teams.

As API becomes popular, more and more organizations make their data public through API. Independent developers are able to have access to more data. With the help of scripting language, API and online database, individuals can perform a full process from data mining to data visualization. In the past, such work required a complete team to finish.

This paper gave a try to build a web-based mapping interface for site selection. The service object is Dunkin' Donuts but not limited to Dunkin' Donuts. Anyone with interests in site selection or data science is welcome to learn the procedures and use the interface.

Dunkin' Donuts is an American global coffeehouse chain and has become one of the largest coffee chains in the world. The company primarily competes with Starbucks, which gained market competitiveness through finding the best real estate locations. The web-based mapping interface built in this paper is not intended to compete with Esri's Business Analyst Online but to show the possibility for individual developers to get involved in retail site selection and build a scalable application.

Like Esri's Business Analyst Online, the application utilizes GIS-based site selection analysis and demographic data. Unlike Esri's Business Analyst Online, the application provides suggestions on site selection in a simpler way rather than generate various sophisticated reports. The financial cost of developing the application is almost zero because only open-source, free software would be used. The core of the application is a predictive model. An OLS regression model was used to predict the retail-site sales. A previous research found that a straightforward linear regression can be developed rapidly and cost-efficiently using open-source, free software. (Real Estate Site Selection With Predictive Modeling in the Open-Source R Language).

The interactive interface allows users to know the predicted sales of any site on the map through background calculation. At the same time, a report will be generated explaining how the predicted sale is calculated. This is a more straightforward way to make suggestions on site selection. With more accurate data added into the regression model, the model can be improved so that it will perform better and better.

Overview of the procedure

The techniques used in this paper include Python, R, JavaScript and ArcMap (optional).
The procedure is as followed.

1. An OLS regression model is built to find factors related to coffee sales.
2. API is used to get market value of properties and data about significant predictors.
3. The site suitability is derived from the difference between sales and market value.
4. Data collected is processed and uploaded to CartoDB.
5. JavaScript is used to build a web-based mapping service based on CartoDB.

Model

It started from a regression model.

In a GIS course I took at University of Pennsylvania, I learned how to use OLS regression model to predict the sales of volume of coffee shops given the spatial and aspatial qualities of coffee shops in PA. OLS is the abbreviation of ordinary least squares. It is also called linear least squares. OLS is a method for estimating the unknown parameters in a linear regression model. In the class, we used several possible predictors to build a model to predict the sales of volume. The dataset is not limited to Dunkin' Donuts but is about all coffee shops in Pennsylvania. It is a good training set.

Regression Model

Possible Predictors:

Variable	Explanation
dist_Hwy	Distance to the nearest highway
CoffeeDist	Distance to the nearest coffee stores other than itself
DistShop	Distance to the nearest supermarkets
popDens	Population density
POP	Population
HHs	Households
Families	Families
Homes	Homes
Med_Inc	Median income
Med_Rent	Median rent
Med_Value	Median house value
Pct_White	Percentage of White population
Pct_le_5yr	Percentage of population under 5 years old
Avg_HHSze	Average household size
Pct_Col2	percentage of population with Bachelor's degree or higher
Pct_BlPov	percentage of population in poverty
distEmpC	Distance to the nearest employment center
SALES_VOL	Sales of volume (the total amount of dollar sales in thousands)

Table 1: possible predictors

Multicollinearity

First, the correlations between each potential predictors need to be examined to see whether there is multicollinearity in the data set. Figure 1 is the correlation matrix of all potential predictors.

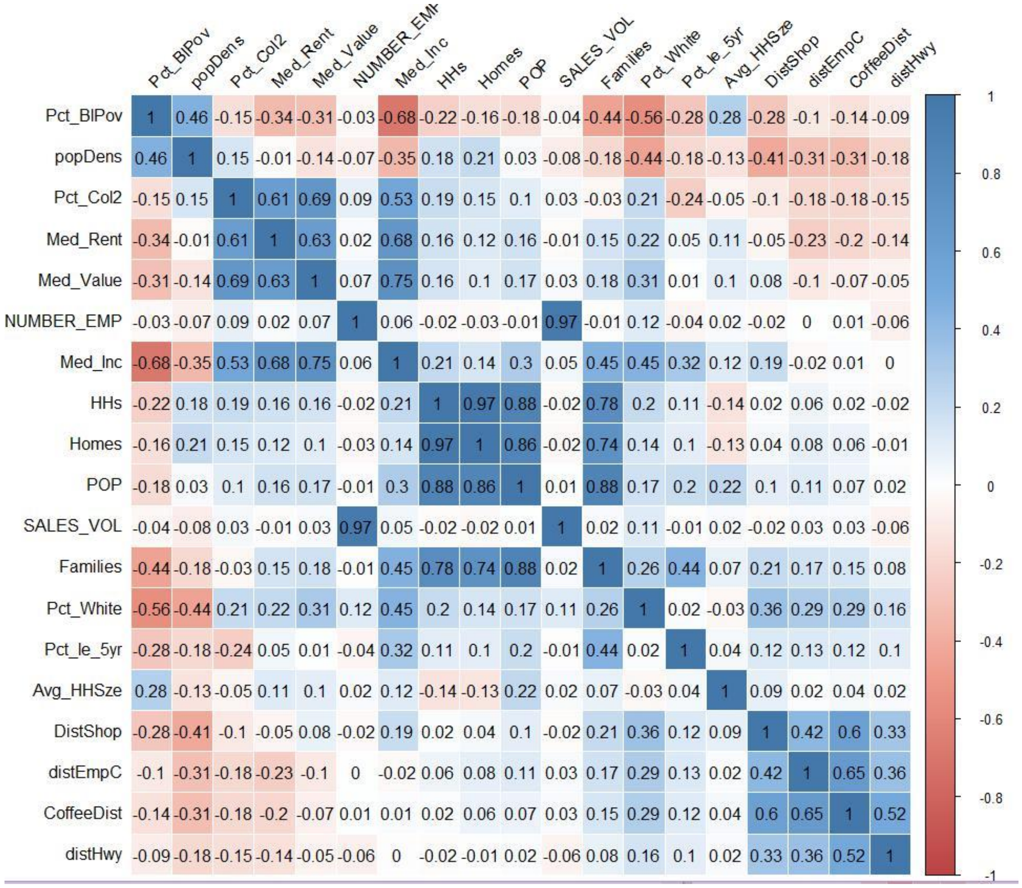

Figure 1: the correlation matrix

We can see that homes, households, population and families are highly correlated (correlation > 0.8). When applying OLS regression, we should only keep one of these predictors.

OLS regression:

The dataset was split into an 80% training set and a 20% hold-out validation set. In the class, we used these predictors to build a regression model to predict the sales of volume and found that the number of employees was a significant predictor. However, I altered the model here. I used the ratio of sales of volume and number of employees as the dependent variable. The reason is that the model is used for site selection. That is to say, the number of employees is unknown when building the regression model. Therefore, I use the ratio of sales of volume and number of employees as the dependent variable. The result of OLS regression is as followed.

```
Call:
lm(formula = SALES_VOL ~ distHwy + CoffeeDist + DistShop + popDens +
    HHS + Med_Inc + Med_Rent + Med_Value + Pct_White + Pct_le_5yr +
    Avg_HHSze + Pct_Col2 + Pct_BlPov + distEmpC + NUMBER_EMP +
    isDunkin, data = training2)

Residuals:
    Min      1Q  Median      3Q     Max
-813.76  -42.19  -10.28   29.82  726.89

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.000e+01  1.743e+01  -1.721  0.08546 .
distHwy     -7.019e-04  8.645e-04  -0.812  0.41700
CoffeeDist   5.388e-04  5.434e-04   0.992  0.32158
DistShop    -1.785e-03  1.346e-03  -1.326  0.18501
popDens      1.256e+03  8.963e+02   1.401  0.16149
HHS          5.591e-04  2.992e-03   0.187  0.85181
Med_Inc      4.660e-04  3.518e-04   1.324  0.18558
Med_Rent    -7.722e-02  1.755e-02  -4.401 1.17e-05 ***
Med_Value   -1.547e-04  6.732e-05  -2.297  0.02175 *
Pct_White    3.528e-01  1.293e-01   2.729  0.00643 **
Pct_le_5yr   1.631e-01  1.360e+00   0.120  0.90456
Avg_HHSze   -9.708e-01  4.469e+00  -0.217  0.82806
Pct_Col2    -3.022e-01  2.214e-01  -1.365  0.17253
Pct_BlPov    4.047e-02  3.467e-01   0.117  0.90711
distEmpC     1.436e-04  1.221e-04   1.176  0.23966
NUMBER_EMP   4.722e+01  2.639e-01 178.905  < 2e-16 ***
isDunkin     1.359e+02  4.778e+00  28.442  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 80.45 on 1316 degrees of freedom
Multiple R-squared:  0.9626,    Adjusted R-squared:  0.9622
F-statistic:  2120 on 16 and 1316 DF,  p-value: < 2.2e-16
```

Figure 2: the result of OLS regression

As is shown in Figure 2, significant predictors are median rent, median value, percentage of White population, number of employees, and a dummy variable “isDunkin” indicating whether a coffee shop is Dunkin’ Donuts. Note that the number of employees is actually a sign of the size of coffee shop. The more employees, the larger the coffee shop is.

The R-squared value is 0.9626 which means the model can explain 96% of the variance in the dependent variable. The low p-value associated with the F-ratio shows that we can reject the null hypothesis that all coefficients in the model are 0.

Final Model:

```
Call:
lm(formula = SALES_VOL ~ Med_Rent + Med_Value + Pct_White + NUMBER_EMP +
    isDunkin, data = training2)

Residuals:
    Min      1Q  Median      3Q     Max
-811.97  -42.54   -9.80   30.59  723.15

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.237e+01  9.942e+00  -2.251  0.02458 *
Med_Rent    -7.209e-02  1.429e-02  -5.046 5.15e-07 ***
Med_Value   -1.461e-04  4.624e-05  -3.160  0.00161 **
Pct_White    3.847e-01  9.802e-02   3.924 9.14e-05 ***
NUMBER_EMP   4.720e+01  2.611e-01 180.781  < 2e-16 ***
isDunkin     1.383e+02  4.640e+00  29.816  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 80.46 on 1327 degrees of freedom
Multiple R-squared:  0.9623,    Adjusted R-squared:  0.9622
F-statistic:  6779 on 5 and 1327 DF,  p-value: < 2.2e-16
```

Figure 3: final model

Accuracy:

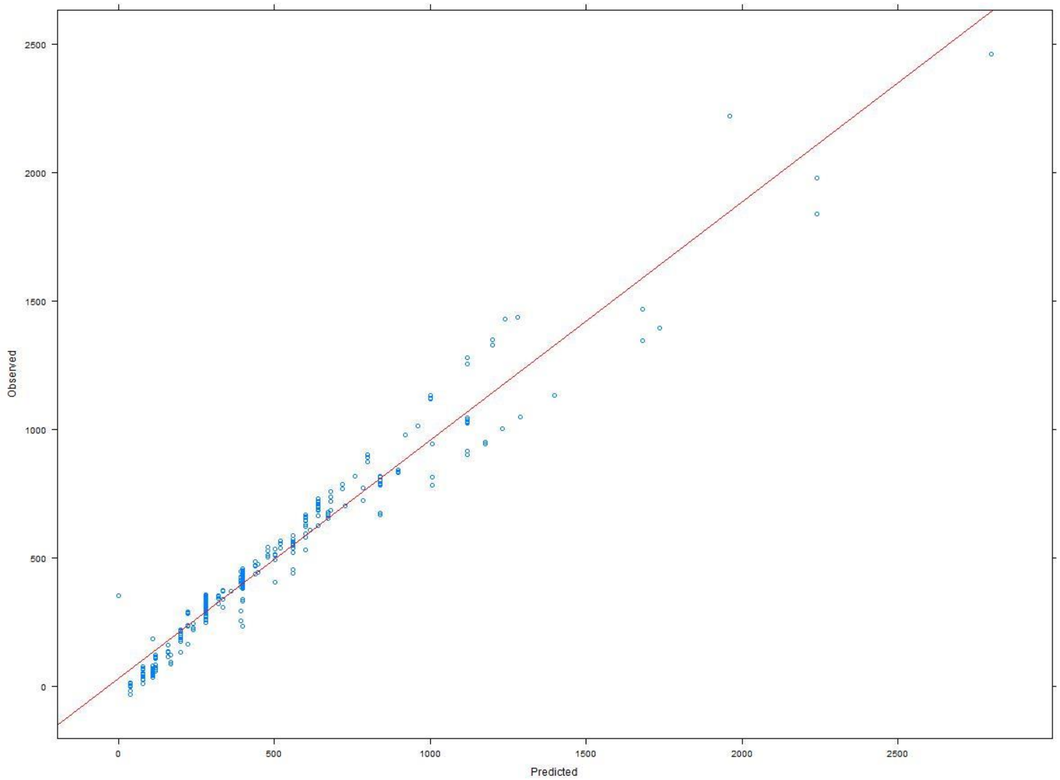

Figure 4: accuracy

Figure 4 is the plot of predicted value -observed value. The slope of the red line is 1. The plot of out of sample prediction suggest the model is quite robust.

Data Collection

Real Estate Data from Zillow

Zillow is an online real estate database company that was founded in 2006. It now provides Zillow API to public so everyone can get Zillow data through the API. The API covers data on Zestimate, rent Zestimate, home valuations, and other property details. More information is on the website: http://www.zillow.com/howto/api/APIOverview.htm.

The main reason why I would like to get data from Zillow is its signature product Zestimate. The Zestimate home value is Zillow's estimated market value, computed using a proprietary formula. The Zestimate is calculated from public and user-submitted data, taking into account special features, location, and market conditions. Zillow also produces a Zestimate forecast, which is Zillow's prediction of a home's Zestimate one year from now, based on current home and market information. Nationally, it has a median error rate of 7.9%. In Philadelphia, the Zestimate has a median error rate of 7.8% (Zillow, What is a Zestimate?).

Data Coverage and Zestimate Accuracy Table

Choose a location type below to change data:

Top Metro Areas
States/Counties*
National

	Zestimate Accuracy	Homes on Zillow	Homes With Zestimates	Within 5% of Sale Price	Within 10% of Sale Price	Within 20% of Sale Price	Median Error
Miami-Fort Lauderdale, FL	★★	2.5M	2.4M	31.1%	54.3%	79.4%	8.9%
Minneapolis-St Paul, MN	★★★	1.2M	1.1M	35.8%	60.5%	84.6%	7.6%
New York, NY	★★★	5.3M	4.9M	32.4%	55.3%	77.0%	8.6%
Orlando, FL	★★★	875.5K	803.0K	37.1%	61.9%	83.0%	7.2%
Philadelphia, PA	★★★	2.1M	2.0M	35.5%	58.1%	79.1%	7.8%
Phoenix, AZ	★★★★	1.7M	1.5M	43.0%	68.0%	87.8%	6.2%
Pittsburgh, PA	★★	969.4K	901.1K	28.1%	47.9%	70.3%	10.7%
Portland, OR	★★★★	809.2K	742.2K	43.2%	69.2%	88.8%	6.1%
Riverside, CA	★★★★	1.6M	1.3M	43.3%	67.2%	85.8%	6.1%
Sacramento, CA	★★★★	793.9K	690.8K	39.4%	64.5%	84.6%	6.8%

Last updated: February 09, 2016

Figure 5: Zestimate Accuracy Table

An alternative data source is property characteristic and assessment information from the Office of Property Assessment. It is available on opendataphilly.org and is updated twice per week. There is no information about the accuracy on the website but when taking a look at the dataset, there are quite a few errors. As is shown in Figure 6, we care about the sale price or the market value but there are some records with $1.00 sale price or $0.00 market value. These abnormal records account for 33.59% of the whole dataset which is a rather high ratio. Therefore, I choose Zillow as the data source.

Sale Price	Unfinished	Assessmer	Market Valu	Market Value
$367,426.00		########	#########	$50,000.00
$1.00		########	#########	$375,800.00
$1.00		12/02/000	12/02/0002	$0.00
$1.00		########	#########	$0.00
$80,000.00		########	#########	$80,000.00
$1.00		########	#########	$298,000.00
$280,000.00		########	#########	$298,000.00
$179,620.00		########	#########	$298,000.00
$232,878.00		########	#########	$298,000.00
$1,100,000.00		########	#########	$125,000.00
$4,306,000.00		########	#########	$636,800.00
$515,000.00		########	#########	$374,000.00
$116,750.00		########	#########	$273,400.00
$55,000.00		########	#########	$150,600.00
$45,000.00		########	#########	$133,700.00
$1.00		########	#########	$223,200.00
$237,500.00		########	#########	$181,200.00
$3.00		########	#########	$220,900.00
$335,000.00		########	#########	$216,000.00
$356,000.00		########	#########	$214,200.00
$179,500.00		########	#########	$228,100.00
$60,000.00		########	#########	$167,400.00
$335,000.00		########	#########	$300,100.00
$39,900.00		########	#########	$181,200.00
$360,000.00		########	#########	$220,900.00

Figure 6: abnormal records

The API used for this project from Zillow is called GetDeepSearch-Results API. It finds a property for a specified address. The result set returned contains the full address, Zestimate data and property data like lot size, year built, last sale detail etc. The required parameters are zws-id, address, and citystatezip. The details about these parameters are shown as followed.

The parameters of the API are:

PARAMETER	DESCRIPTION	REQUIRED
zws-id	The Zillow Web Service Identifier. Each subscriber to Zillow Web Services is uniquely identified by an ID sequence and every request to Web services requires this ID. Click here to get yours.	Yes
address	The address of the property to search. This string should be URL encoded.	Yes
citystatezip	The city+state combination and/or ZIP code for which to search. This string should be URL encoded. Note that giving both city and state is required. Using just one will not work.	Yes

Figure 7: API parameters

Zws-id is an ID provided by Zillow once you have registered on the website. The city is Philadelphia and the state is PA (Pennsylvania). Therefore the real problem is that how would I provide address. The other data source mentioned before is a solution.

It is true that in property characteristic and assessment information from the Office of Property Assessment, there are quite a few abnormal records. However, the addresses in the dataset are accurate. They can be used to query data from Zillow. Another advantage of using addresses from the property data from the Office of Property Assessment is that it contains the category of properties. The category can help filter out places where are not suitable for retail stores. For siting coffee shops, I selected vacant and commercial properties.

Even though I kept only vacant and commercial properties, the number of instances is still striking. There were over 70 thousands addresses. A zws-id is limited to query one thousand times to call the Properties Details API per day (Zillow, Term of Use). I had to spend over 70 days query data from Zillow. I could create more accounts indeed but it is tedious and inefficient. As a result, I decided to limit my study area to west Philly. By taking a look at the population distribution map of Philadelphia, there are some places which are meaningless for siting coffee shops. The reason why I chose West Philly is that there is University City and many residential areas but few Dunkin' Donuts stores. If Dunkin' Donuts wants to expand its business, West Philly is quite potential for opening a new Dunkin' Donuts store.

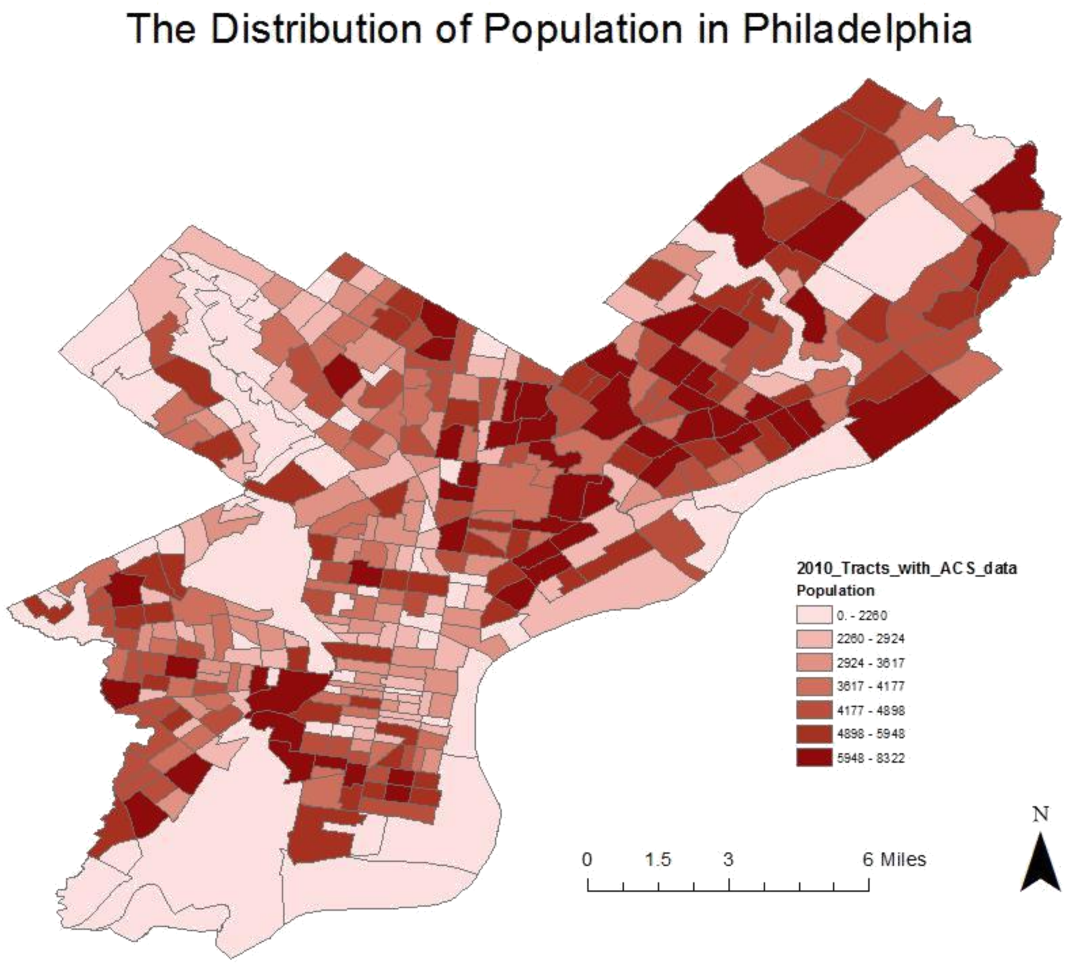

Figure 8: the distribution of population in Philadelphia

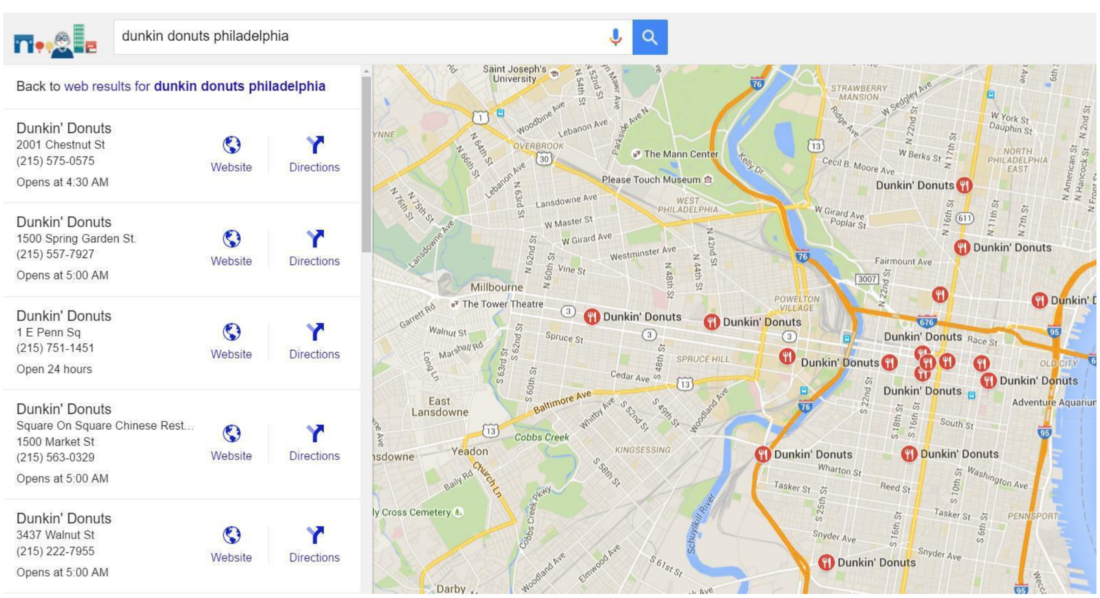

Figure 9: Dunkin' Donuts in Philadelphia

The final filtered dataset has about 7000 instances. I wrote a Python script to query data through Zillow API. The responses were in the format of XML. XML is a textual data format with strong support via Unicode for different human languages. There is a library in Python called xml helping parse XML files. Here is an example of the parsed data in Excel.

zpid	street	zipcode	lat	lon	lotSqFt	value	link
10321055	2235 Grays Ferry Ave	19146	39.94489	-75.1799	763	498815	http://www.zillow.com/homes/10321055_zpid/
80950573	2237 Grays Ferry Ave	19146	39.94484	-75.1799	763	461764	http://www.zillow.com/homes/80950573_zpid/
80944765	2243 Grays Ferry Ave	19146	39.94476	-75.1801	808	416854	http://www.zillow.com/homes/80944765_zpid/
10320945	625 S 23rd St	19146	39.94467	-75.1802	1359	843031	http://www.zillow.com/homes/10320945_zpid/
80953109	2314 South St	19146	39.9453	-75.1803	1295	782171	http://www.zillow.com/homes/80953109_zpid/
118355687	2620 Brown St	19130	39.97025	-75.1804	871	581171	http://www.zillow.com/homes/118355687_zpid/
10320137	2318 South St	19146	39.94529	-75.1804	1447	822377	http://www.zillow.com/homes/10320137_zpid/
122059903	500 S 24th St	19146	39.94648	-75.1807	720	721988	http://www.zillow.com/homes/122059903_zpid/
10244435	2701 Brown St	19130	39.9708	-75.1813	1667	677144	http://www.zillow.com/homes/10244435_zpid/
10210293	2500 Pine St	19103	39.94746	-75.1816	888	1150986	http://www.zillow.com/homes/10210293_zpid/
10320142	2522 South St	19146	39.94561	-75.1825	928	668227	http://www.zillow.com/homes/10320142_zpid/
10244389	2738 Brown St	19130	39.97052	-75.1826	1000	432428	http://www.zillow.com/homes/10244389_zpid/
80952042	2615 South St	19146	39.9461	-75.1838	1520	683600	http://www.zillow.com/homes/80952042_zpid/
118364183	3118 Spring Garden St	19104	39.96317	-75.1879	709	230597	http://www.zillow.com/homes/118364183_zpid/
10291952	3231 Powelton Ave	19104	39.96066	-75.1891	1829	310449	http://www.zillow.com/homes/10291952_zpid/
10292065	3221 Hamilton St	19104	39.96263	-75.1891	2178	278366	http://www.zillow.com/homes/10292065_zpid/
10292510	3221 Mount Vernon St	19104	39.96504	-75.1896	1306	226491	http://www.zillow.com/homes/10292510_zpid/
10292598	3221 Wallace St	19104	39.96565	-75.1897	1393	193414	http://www.zillow.com/homes/10292598_zpid/
118337380	3233 Spring Garden St	19104	39.96333	-75.1899	2205	809370	http://www.zillow.com/homes/118337380_zpid/
10292845	601 N 34th St	19104	39.96419	-75.1914	1693	236143	http://www.zillow.com/homes/10292845_zpid/

Figure 10: the parsed data in excel

I kept zpid for the purpose of possible further queries. I can use it to get more detailed information about the property.

Census tract with ACS 5-year data

In the regression model, the significant predictors are median income, median rent, average household size, percentage of population with Bachelor's degree or higher, and percentage of population in poverty. To fill the equation and calculate the result of the prediction model, data on these features in each census tract is required.

I got the census tract data from opendataphilly.org (OPENDATAPHILLY). The newest one I can get is Census Tracts (2010).

The data of predictors was gathered from American Community Survey (ACS) 5-Year-Data (2010-2014) (Bureau). The ACS covers a broad range of topics about social, economic, demographic, and housing characteristics of the U.S. population. The API call url is http://api.census.gov/data/2014/acs5?. To get the information you want, you need to provide the acs5 code for different characteristics and the census tract id including the id for state and city. To search the right acs5 code for the required characteristics is painful because the dataset is so detailed that there may be thousands of instances describing similar things.

The responses are different from what I got from querying Zillow API. They are strings rather than XML files. In terms of parsing, strings are easier to parse. However, if you got any errors during gathering data, it is easier to recover data in XML files.

Join

The next step is joining the census tracts, ACS data and Zillow properties data together in ArcMap. ArcMap is not a free software. If you want to use other free open source software, I recommend CartoDB which I will discuss it further in the next section. CartoDB offers a solution to join spatial dataset either by attributes or by spatial location. It is similar to the join function in ArcMap.

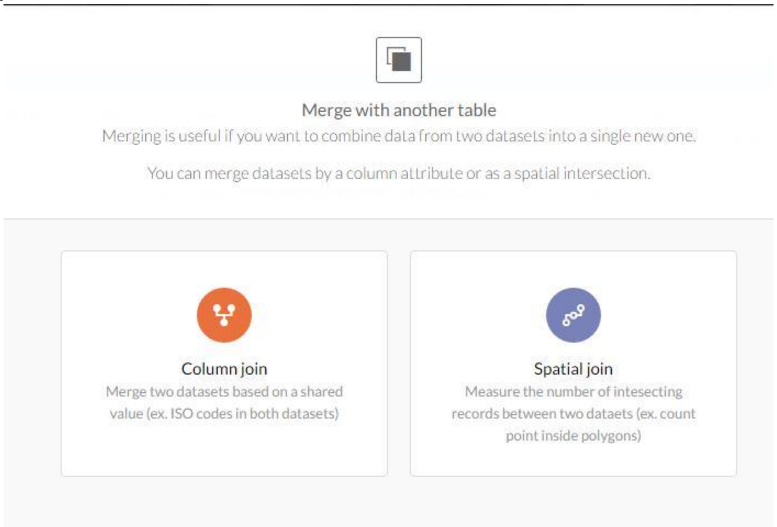

Figure 11: the procedure on CartoDB for join

The census tracts and ACS data were joined by tract id. The census tracts with ACS data was then joined to Zillow properties data by spatial location. Here are the results.

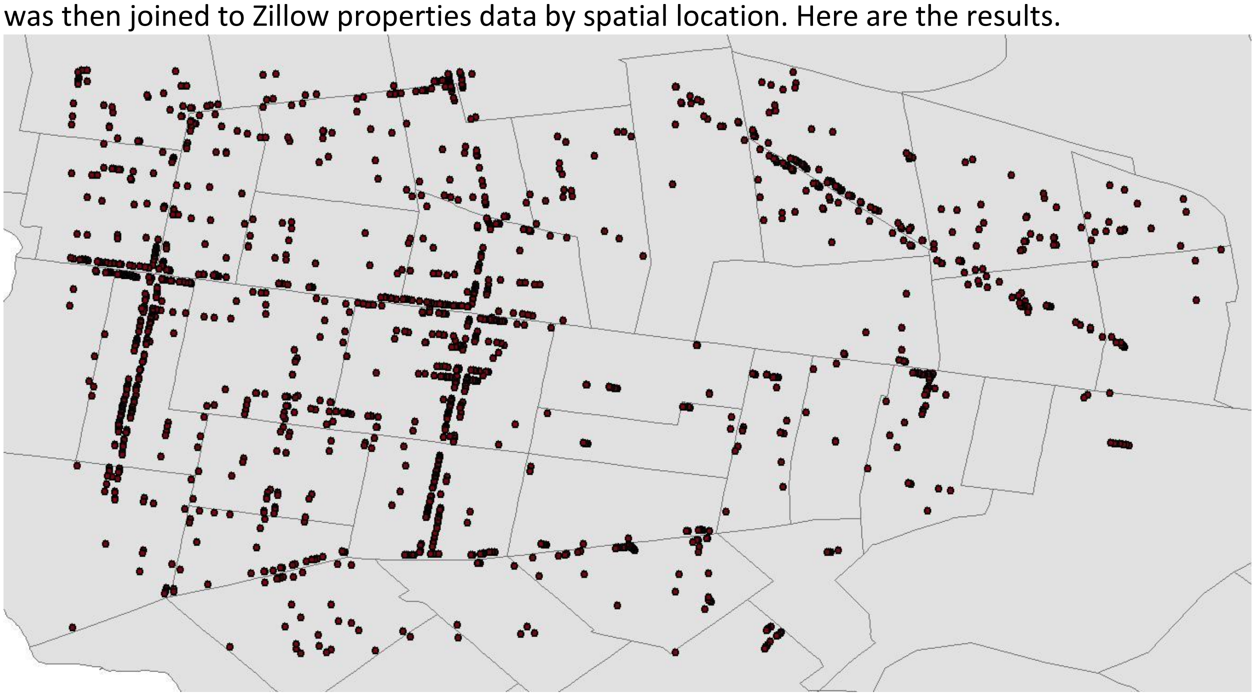

Figure 12: points in ArcMap

FID	Shape	FID_1	zpid	street	zipcode	lat	lon	lotSqFt	value	link	VS_ratio	TRACTCE10	GEOID10	POP	Med_Rent	Med_Val	White
0	Point	833	10189338	5401 Race St	19139	39.963716	-75.228126	1263	124650	http://www.zillow.com/homes/10189338	98.693587	009400	42101009400	4106	387	10661	41
1	Point	835	10189891	164 N 54th St	19139	39.963474	-75.228225	912	48420	http://www.zillow.com/homes/10189891	53.092105	009400	42101009400	4106	387	10661	41
2	Point	849	80948950	100 N 54th St	19139	39.962155	-75.228505	912	36671	http://www.zillow.com/homes/80948950	40.20943	009400	42101009400	4106	387	10661	41
3	Point	852	10189963	63-65 N Yewdall St	19139	39.961974	-75.228767	1674	109325	http://www.zillow.com/homes/10189963	65.307646	009400	42101009400	4106	387	10661	41
4	Point	855	80951286	5403 Market St	19139	39.960701	-75.228807	1048	128777	http://www.zillow.com/homes/80951286	122.878817	009400	42101009400	4106	387	10661	41
5	Point	861	80949802	5411 Market St	19139	39.960729	-75.229031	1048	112772	http://www.zillow.com/homes/80949802	107.60687	009400	42101009400	4106	387	10661	41
6	Point	862	80949802	5411 Market St	19139	39.960729	-75.229031	1048	112772	http://www.zillow.com/homes/80949802	107.60687	009400	42101009400	4106	387	10661	41
7	Point	864	80941752	5413 Market St	19139	39.960737	-75.229092	1179	104334	http://www.zillow.com/homes/80941752	88.493639	009400	42101009400	4106	387	10661	41
8	Point	865	80941752	5413 Market St	19139	39.960737	-75.229092	1179	104334	http://www.zillow.com/homes/80941752	88.493639	009400	42101009400	4106	387	10661	41
9	Point	867	80954272	5415 Market St	19139	39.960763	-75.229292	1048	80548	http://www.zillow.com/homes/80954272	76.858779	009400	42101009400	4106	387	10661	41
10	Point	868	80942878	5417 Market St	19139	39.960769	-75.229347	999	67533	http://www.zillow.com/homes/80942878	67.600601	009400	42101009400	4106	387	10661	41
11	Point	869	80949883	5419 Market St	19139	39.960776	-75.229399	993	72922	http://www.zillow.com/homes/80949883	73.436052	009400	42101009400	4106	387	10661	41
12	Point	871	80946944	5425 Market St	19139	39.960795	-75.229559	999	71036	http://www.zillow.com/homes/80946944	71.107107	009400	42101009400	4106	387	10661	41
13	Point	885	10190236	100 N Sickels St	19139	39.962358	-75.230127	901	47895	http://www.zillow.com/homes/10190236	53.157603	009400	42101009400	4106	387	10661	41
14	Point	888	10189372	5501 Race St	19139	39.963987	-75.230333	736	59926	http://www.zillow.com/homes/10189372	81.421196	009400	42101009400	4106	387	10661	41
15	Point	943	10192043	60 N 56th St	19139	39.962474	-75.232804	1306	70551	http://www.zillow.com/homes/10192043	54.020674	009400	42101009400	4106	387	10661	41
16	Point	946	80951851	5601 Market St	19139	39.961258	-75.232943	1824	194559	http://www.zillow.com/homes/80951851	106.666118	009400	42101009400	4106	387	10661	41
17	Point	949	80942712	5603 Market St	19139	39.961267	-75.233005	1536	50772	http://www.zillow.com/homes/80942712	33.054688	009400	42101009400	4106	387	10661	41
18	Point	951	80944671	5605 Market St	19139	39.961273	-75.233061	1536	52635	http://www.zillow.com/homes/80944671	34.267578	009400	42101009400	4106	387	10661	41
19	Point	952	10191649	5607 Market St	19139	39.96128	-75.233119	1536	57383	http://www.zillow.com/homes/10191649	37.358724	009400	42101009400	4106	387	10661	41
20	Point	975	80951409	157 N 57th St	19139	39.964188	-75.233997	1072	101306	http://www.zillow.com/homes/80951409	94.501866	009400	42101009400	4106	387	10661	41
21	Point	976	10592966	5647 Race St	19139	39.964529	-75.233998	4600	83170	http://www.zillow.com/homes/10592966	18.080435	009400	42101009400	4106	387	10661	41
22	Point	981	10192150	129 N 57th St	19139	39.963608	-75.234119	1072	35885	http://www.zillow.com/homes/10192150	33.474813	009400	42101009400	4106	387	10661	41
23	Point	984	80941635	5635 Market St	19139	39.96141	-75.234181	1728	209683	http://www.zillow.com/homes/80941635	121.344329	009400	42101009400	4106	387	10661	41
24	Point	989	80944288	101 N 57th St	19139	39.962871	-75.234265	1054	61063	http://www.zillow.com/homes/80944288	57.934535	009400	42101009400	4106	387	10661	41
25	Point	991	80949146	61 N 57th St	19139	39.962654	-75.234285	1296	55776	http://www.zillow.com/homes/80949146	43.037037	009400	42101009400	4106	387	10661	41
26	Point	999	10192810	200 N 57th St	19139	39.964461	-75.234451	997	65055	http://www.zillow.com/homes/10192810	65.250752	009400	42101009400	4106	387	10661	41
27	Point	1000	118353100	5645 Market St	19139	39.961445	-75.234467	1536	60470	http://www.zillow.com/homes/11835310	39.36849	009400	42101009400	4106	387	10661	41
28	Point	1005	10191656	5649 Market St	19139	39.961459	-75.23458	1536	54442	http://www.zillow.com/homes/10191656	35.44401	009400	42101009400	4106	387	10661	41
29	Point	1006	80946461	5651 Market St	19139	39.961466	-75.234642	1824	178036	http://www.zillow.com/homes/80946461	97.607456	009400	42101009400	4106	387	10661	41
30	Point	1015	118338770	54 N 57th St	19139	39.962741	-75.234848	1368	71727	http://www.zillow.com/homes/11833877	52.432018	009400	42101009400	4106	387	10661	41
31	Point	1045	10192869	201 N 58th St	19139	39.964644	-75.235905	1120	109660	http://www.zillow.com/homes/10192869	97.910714	009400	42101009400	4106	387	10661	41
32	Point	1049	10192213	131 N 58th St	19139	39.963854	-75.236072	1120	55278	http://www.zillow.com/homes/10192213	49.355357	009400	42101009400	4106	387	10661	41
33	Point	1050	80948639	111 N 58th St	19139	39.963332	-75.236166	1200	88961	http://www.zillow.com/homes/80948639	74.134167	009400	42101009400	4106	387	10661	41
34	Point	1051	80942318	109 N 58th St	19139	39.963291	-75.236173	1200	92171	http://www.zillow.com/homes/80942318	76.809167	009400	42101009400	4106	387	10661	41
35	Point	1040	10193329	5801 Haverford Ave	19131	39.968734	-75.235536	1613	122566	http://www.zillow.com/homes/10193329	75.986361	009500	42101009500	3205	452	6883	20
36	Point	1052	80943247	5821 Haverford Ave	19131	39.968909	-75.236187	1240	73319	http://www.zillow.com/homes/80943247	59.128226	009500	42101009500	3205	452	6883	20

Figure 13: the attribute table

Web-Based Mapping Interface

In the previous section, I successfully got the data. The next step is to build a web-based mapping interface to let others have access to the data. I used JavaScript, HTML and CSS to build the web-based mapping interface. These three language are three core technologies of World Wide Web content production.

CartoDB platform

There are different ways to store your data and query it. CartoDB is one of the solutions. It is free (for limited space) and easy to learn to use. I uploaded the CSV version of my data onto CartoDB database. It would geocode the data for me automatically.

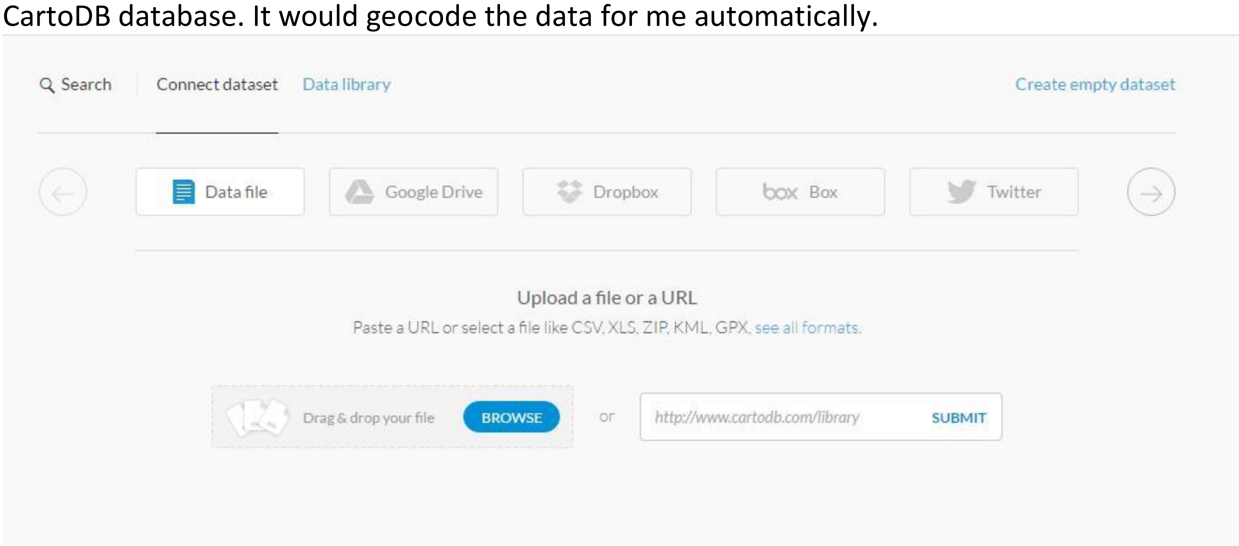

Figure 14: the procedure on CartoDB for uploading data

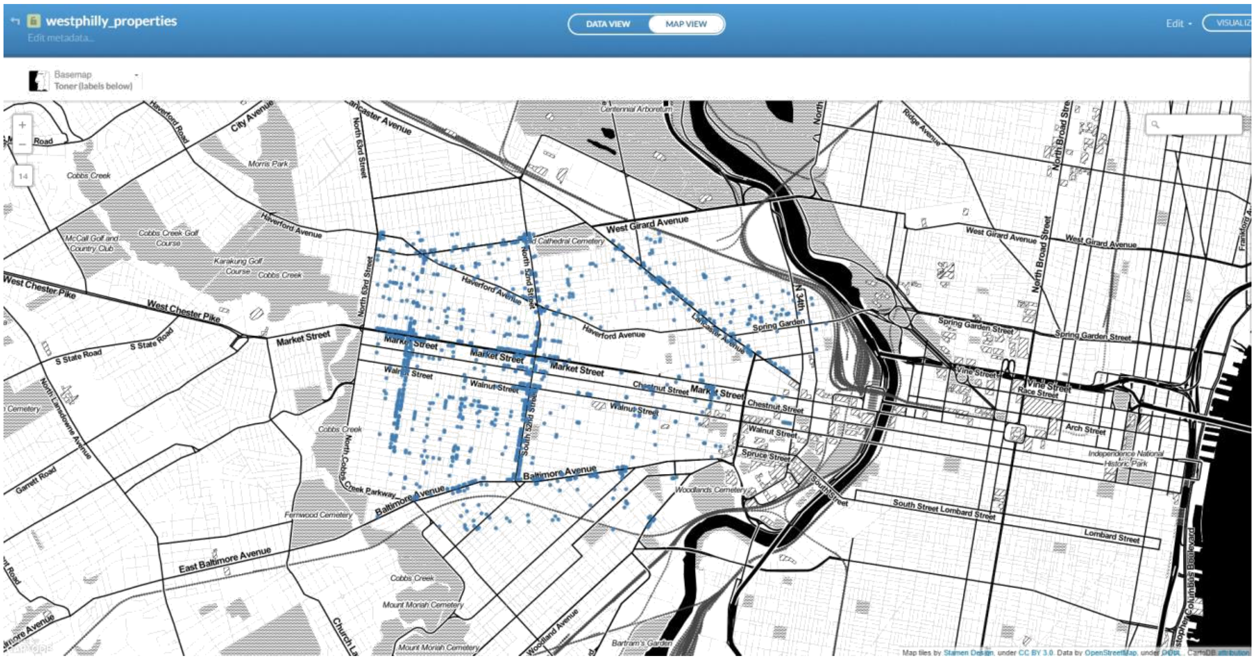

Figure 15: map view of the data

Interface

Figure 16: the web interface

As is shown in Figure 16, you can enter the range of lot size, choose a way to sort and define the number of records shown on the map. Click on the marker on the map, an info window will appear and show the details of the marker.

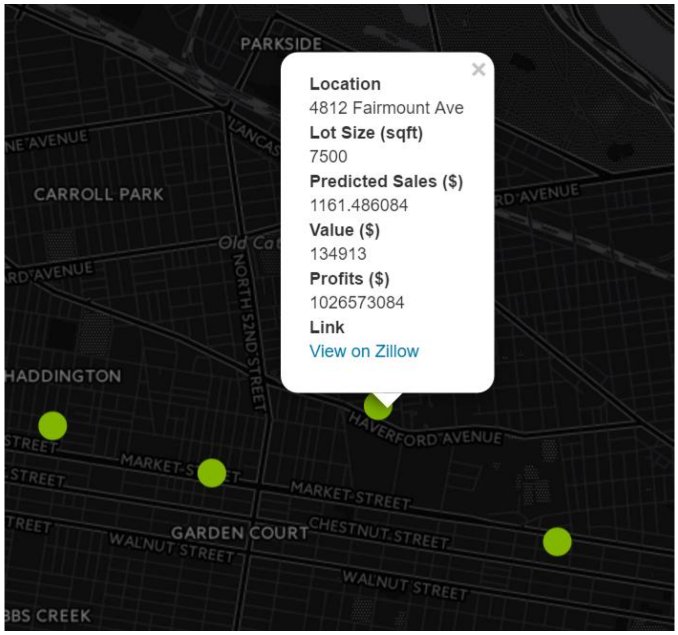

Figure 17: details of the marker

Conclusion

The project is hosted on GIithub (https://github.com/Roxyi/site_machine). The url of web-based interface is http://roxyi.github.io/site_machine/.

In general, the web-based mapping interface is built successfully. There is space to improve it. For example, more possible predictors should be examined. The predictors are all demographic or economic information. There must be some predictors that are related to sales spatially. More data can be collected to make the application not limited in West Philly.

In this project, I used Python to gather data through API, R to build an OLS regression model, JavaScript to build a web application and ArcGIS to process spatial data. These cover almost all what I learnt in the graduate school.

References

Bureau, U. S. (n.d.). *American Community Survey 5-Year-Data*. Retrieved from http://www.census.gov/data/developers/data-sets/acs-survey-5-year-data.html

OPENDATAPHILLY. (n.d.). *Census Tracts*. Retrieved from https://www.opendataphilly.org/dataset/census-tracts

Real Estate Site Selection With Predictive Modeling in the Open-Source R Language. (n.d.). Retrieved from Decision Analyst: http://www.decisionanalyst.com/CaseStudies/RealEstateSiteSelection.dai

Zillow. (n.d.). *Term of Use*. Retrieved from http://www.zillow.com/howto/api/APITerms.htm

Zillow. (n.d.). *What is a Zestimate?* Retrieved from Zestimate: http://www.zillow.com/zestimate/#what