CSE 519 Data Science Project

Baltimore Neighborhood Heat Map

Introduction

Baltimore is the largest city in Maryland, and the 29th most populous city in the country. It is also the largest independent city in the United States of America. This city has always been a city of neighbourhoods. Each section has its own history, culture, and personality- an identity that can only be understood after years of experiences.

We all know that moving to a new city can be a daunting task, even with the help of a professional moving company. However, things start to look easy and comfortable once you know more about the area. Decision making can be made easier with a little bit of extra information up your sleeve. Is Baltimore on your list of potential place to move? Are you planning to rent or buy a house there? If so, then keep reading because this will guide you to pick your favourite neighbourhood because everybody loves a nice neighbourhood!

Motivation

In our last report we used the data provided by the BNIA organization to construct a baseline model for ranking neighborhoods. We primarily focused on two parameters: time on market -- a measure of neighborhood "hotness" -- and median housing cost -- a proxy for neighborhood quality. We also provided a thorough analysis and visualization of selected insights that we obtained from the dataset.

In this report we sought to improve upon the last by focusing on three prongs: we wanted to add more granularity to our ranking model, to test the predictive power of our data, and finally to study and implement a rigorous evaluation of our ranking. To those ends, this final report is structured to revisit the notion of ranking neighborhoods and then proceed towards enriching our baseline model and, our linear regression, and eventually to evaluation.

Rankings

The goal of this project is to output rankings based on the wealth of data provided through the Baltimore Neighbourhood Indicators Alliance. But what sort of rankings? Well, one way to do it would be to output ordered lists that reflect key neighbourhood vitals. A perfectly valid ranking under this single variable ordering would be the neighbourhoods ordered by median income. You'd end up with something like this:

Neighbourhood	Median Income
Greater Roland Park/Poplar Hill	104481.89
Canton	91735.65
Inner Harbor/Federal Hill	88854.21
South Baltimore	88487.05
North Baltimore/Guilford/Homeland	81450.63
Fells Point	77433.38
Mount Washington/Coldspring	76262.58
Highlandtown	71660.18
Lauraville	66195.71
Hamilton	63986.00

This is interesting -- we know where the rich prefer to live. But it doesn't get us much else. There are a multitude of interesting ways to rank

Community	2014 Data
Downtown/Seton Hill	243.3
Harbor East/Little Italy	159.4
Washington Village/Pigtown	139.4
Greater Mondawmin	102.9
Highlandtown	88.8

Interesting -- If we valued our safety perhaps we'd best avoid the Downtown/Seton Hill area. Looks like trouble.

But these rankings don't tell the whole story. They don't even tell half the story. You might say they're rather...one dimensional. We should them out. But how? If you're rich, well, perhaps you can simply trust the taste of other rich folks and choose the area with the highest median income. If you did that, you'd be using median income as a proxy for quality. But do people of the same economic cohort really want the same things as you? I'd venture a guess that they don't.

And what if you're not rich? Maybe you should find the safest neighbourhood that you can afford and stick to it. That might be a sensible approach. But what if you're a tough young guy and would tolerate a little less safety for better proximity to the music venues you like. Then you'd need some sort of intersection of the rankings for safety, entertainment, and price.

Baseline Model

We needed a metric that could first measure for quality, and our assumption was, that using housing prices as a proxy for quality could be a good start, further upon which, we could penalize or credit a neighbourhood depending on the aforementioned metrics.

How should we parametrize this model? What sorts of data can and should it include? These are important questions whose correct answer will guarantee we get the correct answer. Our first attempt was to rank neighbourhoods depending on the median Time on Market, that is a measure of how fast a house in a neighbourhood is bought. We assumed, that a neighbourhood, which has a lower time on market, is a favourable place to stay, as opposed to another where people don't buy real estate as fast. But what is the answer we want? Well, if we were to create a scoring function that output a ranking of neighbourhoods based on which was "best" we could do that. But defining what is best is, as we noted above, tricky. Should best be safest? Artist? The most connected to public transit? Some arbitrary combination of everything?

We looked at the deviations of the Time on Market for each of the neighbourhoods, and decided to rank the neighbourhoods by median prices divided by these deviations. Following are the results.

Neighbourhood	MedianHousePrice	MedianTOM	TOM Deviations	Dev Score
South Baltimore	289900.0	23	19.745455	12.604348
Inner Harbor/Federal Hill	320000.0	28	14.745455	11.428571
North Baltimore/Guilford/Homeland	325000.0	33	9.745455	9.848485
Canton	275000.0	30	12.745455	9.166667
Greater Roland Park/Poplar Hill	310000.0	36	6.745455	8.611111
Mount Washington/Coldspring	287000.0	34	8.745455	8.441176
Fells Point	249000.0	33	9.745455	7.545455
Highlandtown	265000.0	41	1.745455	6.463415
Patterson Park North & East	172000.0	28	14.745455	6.142857
Midtown	200000.0	33	9.745455	6.060606

This model, we felt, over-penalized a neighbourhood with lower prices than the highest, and we also realised, that the Price Per Day of Time on Market doesn't have any substantial meaning to our goal. We needed a better way to penalize a neighbourhood with high Time on Market.

For a baseline notion of "best" we eventually settled on the following formula:

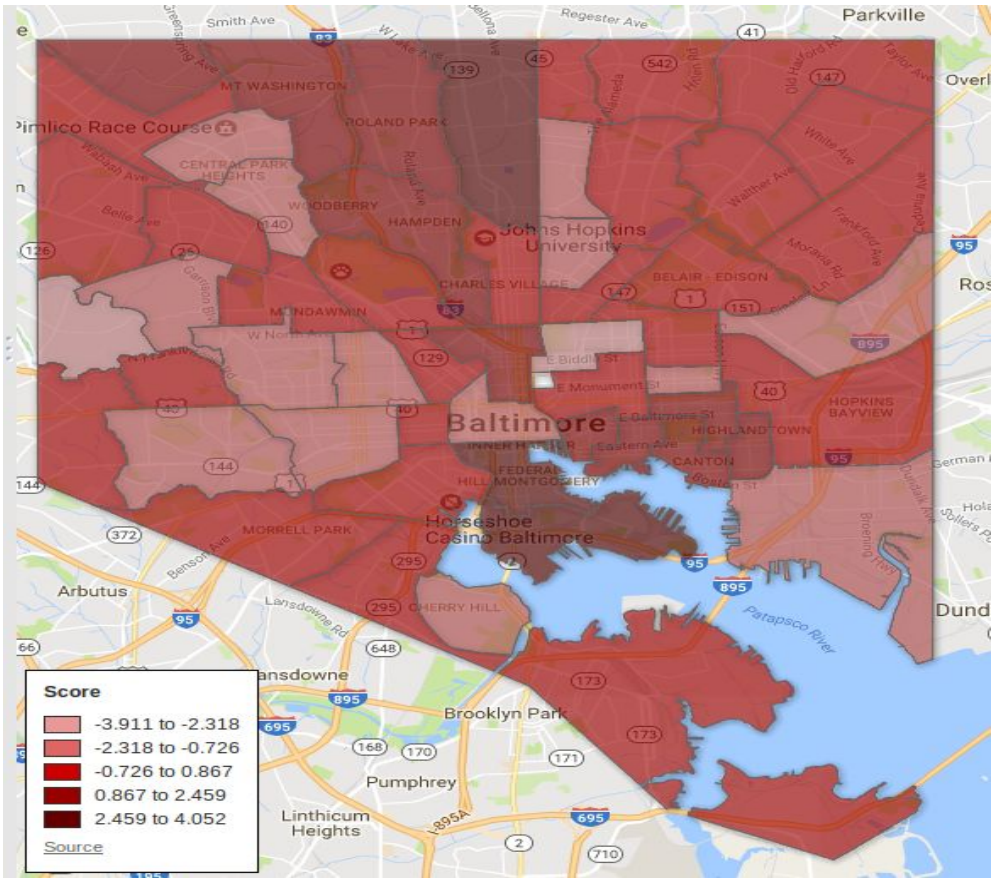
$$Zscore(MedianHousePrice) - Zscore(MedianTimeOnMarket)$$

What does this formula capture? Well, we think it captures a few things very compactly. One way to compute the baseline model would be to have a lot of different parameters that we could weight to our liking. But there is no ground truth. Plus, as you might imagine, people don't choose the neighbourhood they live in based upon the Baltimore Neighbourhood Indicators Alliance Vital Signs report. The theory that underlies this model is the one of efficient markets. The price of the home in each neighborhood should encode a lot of information about the area that the home is in. Succinct Perhaps we should stop here then. That's it. The market has told us everything we need to know.

Well, another bit of information that the market provides is the time on the market, or TOM. The time on the market is a useful indicator because it provides us with a sense of popularity. Areas with a higher time on the market are less popular for new homebuyers. Therefore, we use the negative sign. Combined, this simple model should give us some notion of the most desirable neighbourhoods. And it does. We've done some validation with lists available to us (aggregated neighborhood reviews and the like) that appear to confirm this initial baseline.

CSA2010	Median Number of Days on the Market (2014)	Median Price of Homes Sold (2014)	Median Prices Z Score	Median Days Z Score	Score
Inner Harbor/Federal Hill	28	320000	2.153482	-0.898578	3.052060
South Baltimore	23	289900	1.830947	-1.203274	3.034221
North Baltimore/Guilford/Homeland	33	325000	2.207059	-0.593881	2.800940
Greater Roland Park/Poplar Hill	36	310000	2.046328	-0.411063	2.457391
Canton	30	275000	1.671287	-0.776699	2.447986
Mount Washington/Coldspring	34	287000	1.799872	-0.532942	2.332814
Fells Point	33	249000	1.392685	-0.593881	1.986566
Highlandtown	41	265000	1.564132	-0.106367	1.670499
Patterson Park North & East	28	172000	0.567596	-0.898578	1.466174
Midtown	33	200000	0.867628	-0.593881	1.461510

And as plotted geo-spatially -



Expanded Ranking Model

Creative Ranking

We also wanted to look at neighborhoods from a different perspective- rather than only by direct measures like prices, education standards, and transportation we wanted to evaluate each neighborhood based on how it ranks in art. People have become more sensitive and aware about these things when they choose a neighborhood to live in. In fact this awareness has been encoded in law. Neighborhoods like Station North in Baltimore have actually been designated as an “Arts and Entertainment District” -- a legal and economic distinction that incentivizes creative revitalization in the hope of promoting urban renewal. Places with high creativity index are usually preferred by people, and we would like to test his hypothesis by comparing the ranking obtained by art index with other models.

We have chosen the following parameters to calculate art index for each neighborhood:

1. Businesses involved in creative economy- this takes into account all the profit and nonprofit businesses involved directly or indirectly in arts- industries that support artistic and cultural skillsets. 2. Employment in creative businesses- total number of people employed in the businesses we took into account. This does not include people who identity themselves as artists or are involved in part time art related activities.

3. Total public art- museums, murals etc

4. Events permits requested- art related events organized to support such activities

Our creativity index is calculated by summing the above mentioned parameters after they have been normalized.

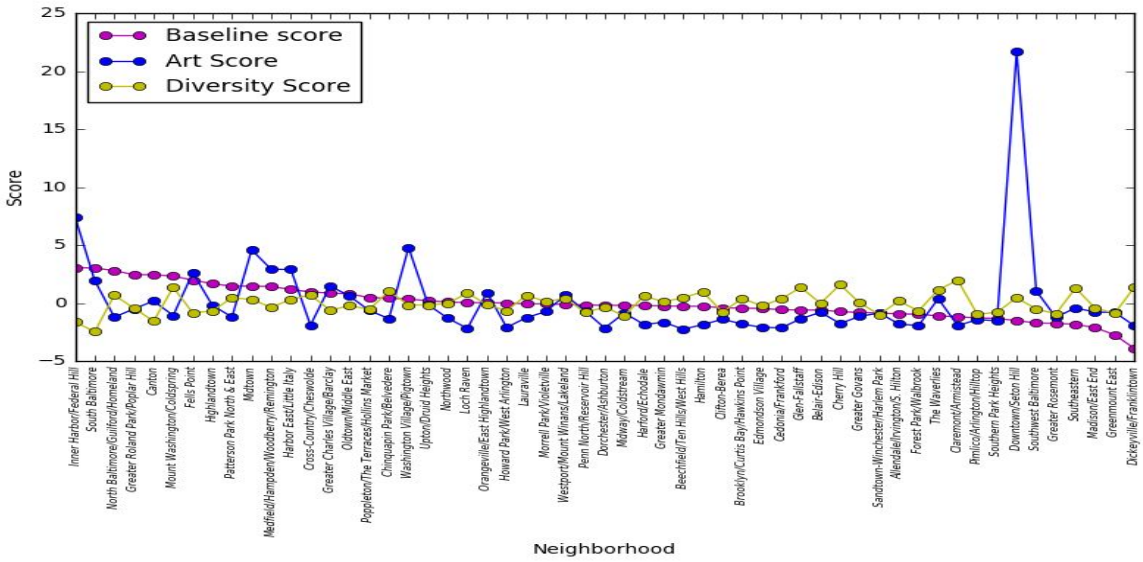
Diversity

Diversity is one thing which has different meaning to it. For our purpose we planned to consider the following parameters:

1. Racial diversity- this takes into account the fact that when two people are picked randomly what are the chances that they belong to different races- White, Asian, Black, Hispanic and non Hispanic

2. Female population- this shows percentage of females in the area.

Insights

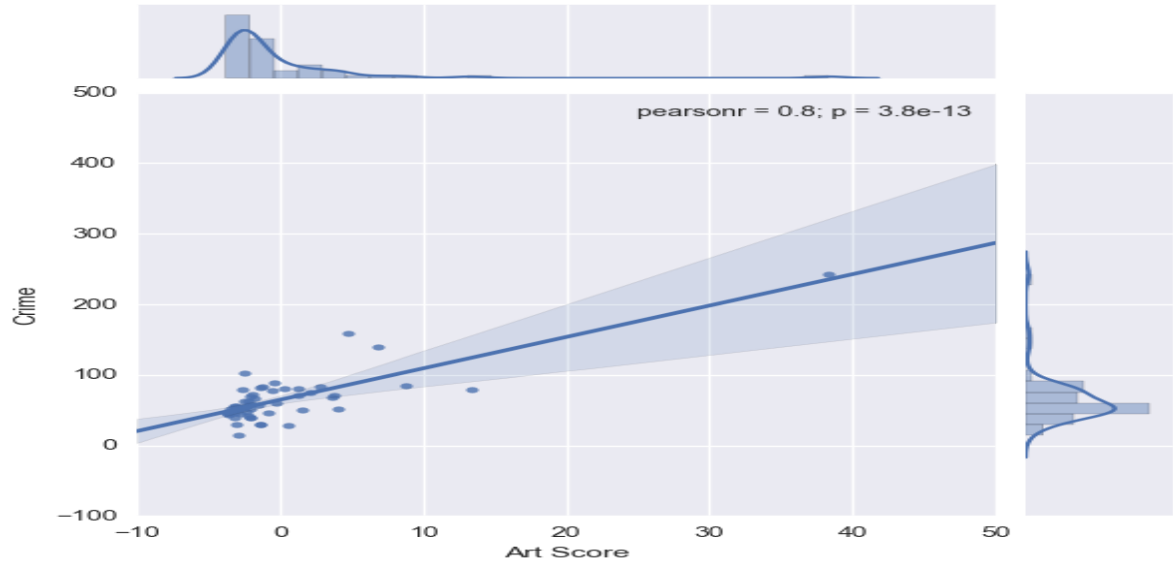


We can see that Federal Hill ranks the highest according to our baseline model, but it does not score very well in terms of art or diversity index. Further scouring the web for its demographics reveals that majority of people in this neighborhood are white (87.3%) even though it has almost equal representation from both the genders.

From the graph one striking thing we observe is that Downtown/Seton hill surpasses all other neighborhoods in terms of its art score. This place is in close proximity with Mount Vernon which is famous for its nightlife, cultural excitement, and LGBT hangout places. It is also famous for shows put up by galleries and emerging artists. A recent poll on Nextdoor shows that people love Seton hill for its architectural charm and historic appeal.

In our baseline model we have taken house prices as a proxy for quality and places which rank high according to that model have a bad diversity score. From the graph we can see that the first 10 places have below average diversity score. Toward the middle of the graph we can see that all the three graphs come close in their rankings and as we move further right places which are ranked lower as per our baseline model have either average or higher diversity score. Even though this is something we expect because we know that white community in Baltimore is richer than its counterparts, but when the data reveals such statistics it gets interesting!

We further narrowed down in to using only the number of public murals, the population involved in the creative economy in an area, number of art-houses and population earning their income from the arts industry per thousand population as a ranking function., and we saw a rather interesting correlation. The art score of an area, is highly correlated with the crime rate!

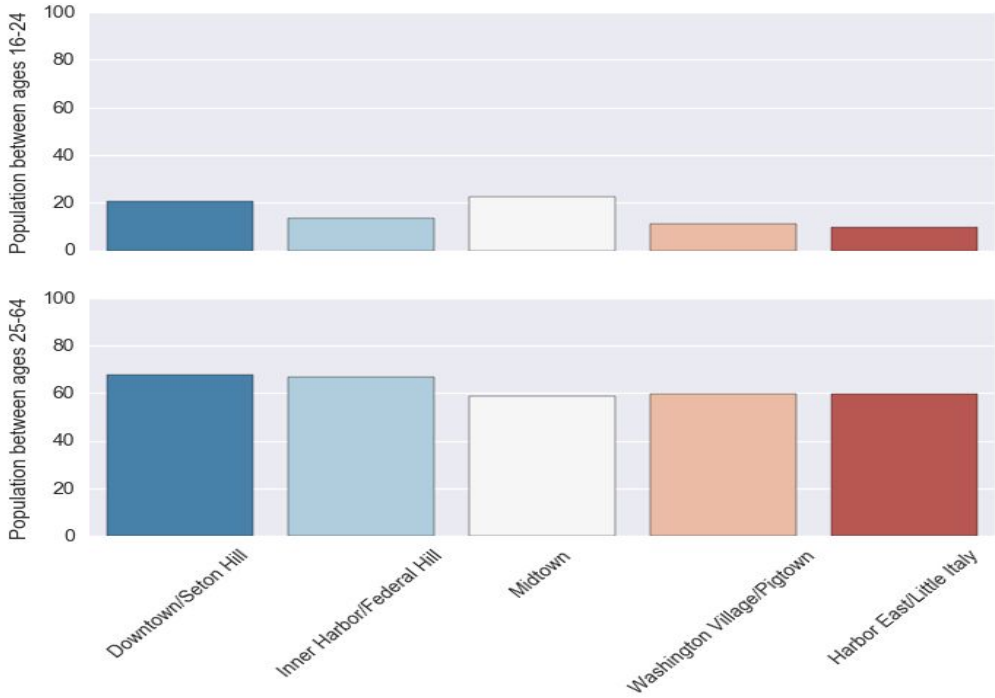


Community	Total Score
Downtown/Seton Hill	10.753992
Inner Harbor/Federal Hill	4.817525
Midtown	3.819838
South Baltimore	2.724911
Medfield/Hampden/Woodberry/Remington	1.982407
Washington Village/Pigtown	1.960907
Fells Point	1.903663
Harbor East/Little Italy	1.670173
Greater Charles Village/Barclay	1.039396
Oldtown/Middle East	0.291340

Though we would not like to delve too much into explanation two possible rationales do come to mind. Areas which are poor are more likely to have cheap housing which is conducive to the creative economy. Further, areas which are poor may also be the object of beautification efforts by the government -- mural painting comes to mind. Thus the art score of the area would be inflated.

Let's now look at the population distribution across these areas

In the figure below, we can see, that there exists a greater population belonging to the working class in areas such as Seton and Federal Hill. We can infer from this that a vast population of youth does prefer these two areas as opposed to the others. Federal Hill ranks highly in most rankings and seems to be a good combination of diversity, art and culture and quality of living.



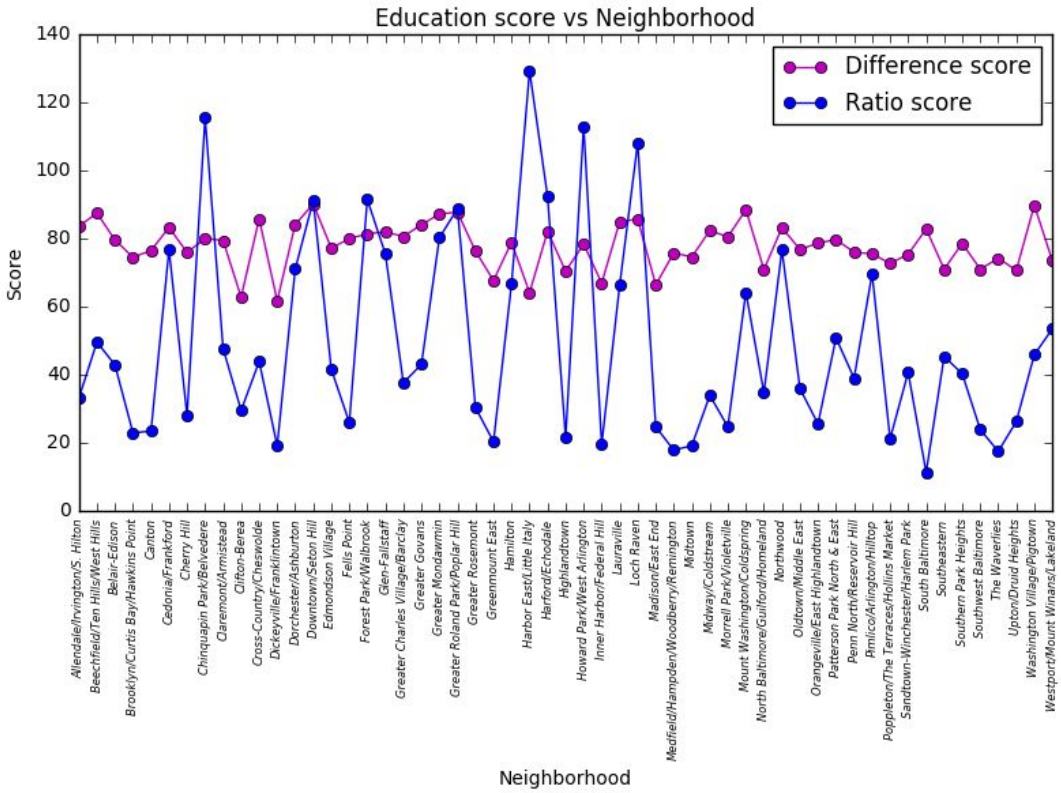
Education

Because the American school system divides neighborhoods into school districts a family with young children must consider the 'education index' of the neighborhood to make sure their kids are brought up with the appropriate resources. Therefore we wanted to take this factor into consideration while ranking the neighborhoods. We rewarded each neighborhood for having a higher percentage of children who successfully completed high school, and penalized for those who either withdrew or dropped out of school. Thus input to our model are:

1. Children who successfully complete high school

2. Children who withdrew or dropped out

We think this is a better parameter than taking the count of school per neighborhood as kids do not necessarily go to the school in their neighborhood because of private schooling. Parents may prefer to put their kids in the best school irrespective of their proximity. This model is also better than taking the ratio of the above two parameters as in that case we penalize a lot for withdrawing from school.

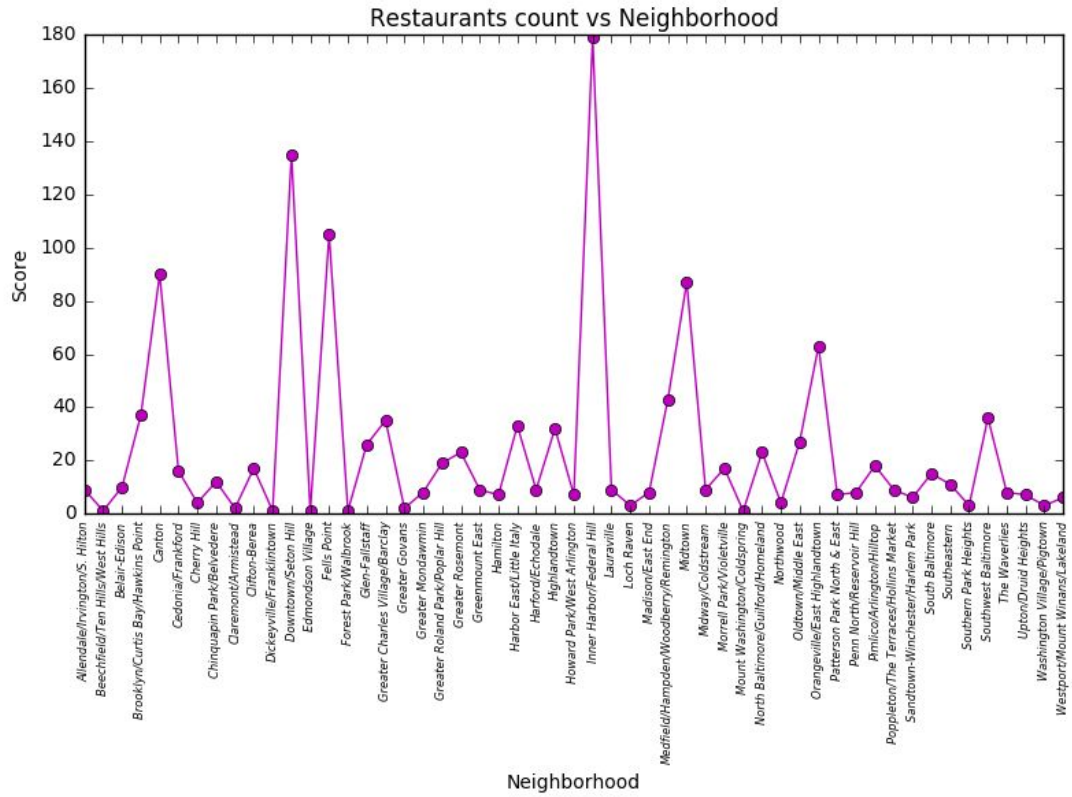


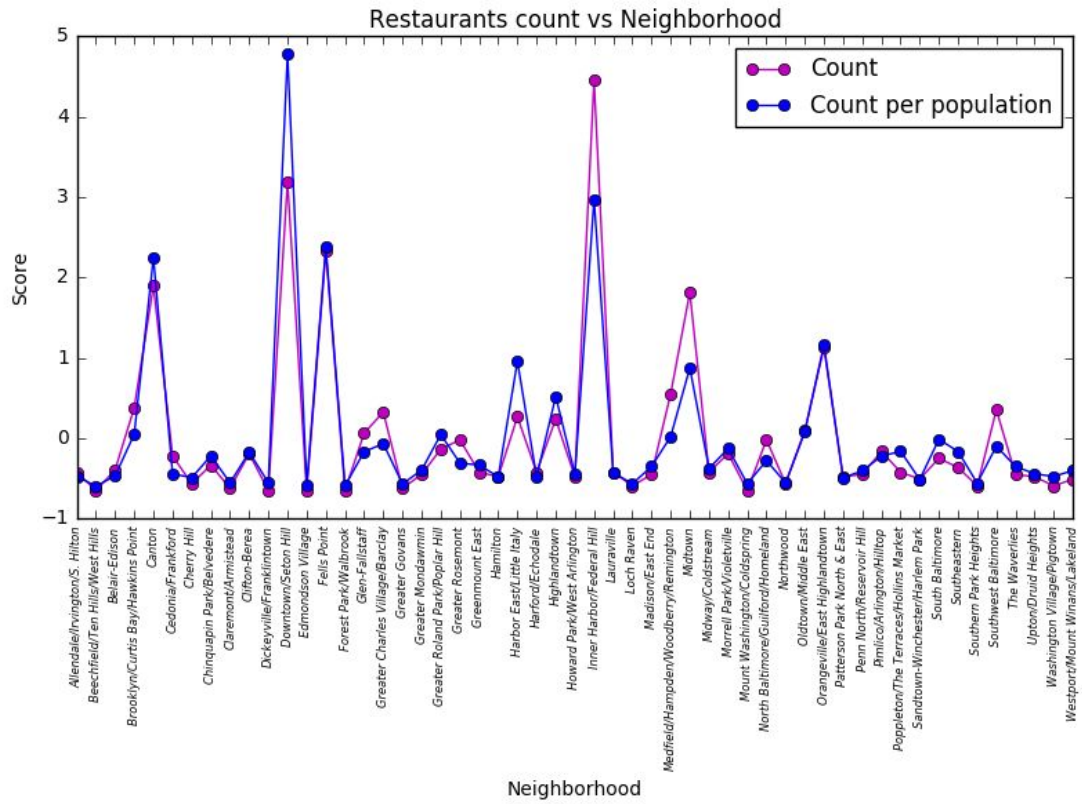
Nightlife

For ranking neighborhoods based on nightlife we used the restaurant dataset provided by Open Baltimore. We aggregated restaurants to neighborhoods in order to get the count to get the count and used that as our basis for assigning score to all the places. We think it is a good parameter to rank the places as famous places have more number of restaurants and of good quality due to high competition than places with fewer number of restaurants.

If we look at the graph which shows count of restaurants we can easily pick top neighborhoods, which stand out among all the neighborhoods. Federal Hill tops the list by a good margin.

We also plotted count per population to check that our count score does not vary much due to area of the neighborhood. Count per population line matches pretty well with the count line and maintains the relative position except for Seton Hill, Federal Hill, and Southwest Baltimore.


Restaurants count vs Neighborhood
Score
Neighborhood
0
20
40
60
80
100
120
140
160
180
Allendale/Irvington/S. Hilton
Beechfield/Ten Hills/West Hills
Belair-Edison
Brooklyn/Curtis Bay/Hawkins Point
Canton
Cedonia/Frankford
Cherry Hill
Chinquapin Park/Belvedere
Claremont/Armistead
Clifton-Berea
Dickeyville/Franklintown
Downtown/Seton Hill
Edmondson Village
Fells Point
Forest Park/Walbrook
Glen-Fallstaff
Greater Charles Village/Barclay
Greater Govans
Greater Mondawmin
Greater Roland Park/Poplar Hill
Greater Rosemont
Greenmount East
Hamilton
Harbor East/Little Italy
Harford/Echodale
Highlandtown
Howard Park/West Arlington
Inner Harbor/Federal Hill
Lauraville
Loch Raven
Madison/East End
Medfield/Hampden/Woodberry/Remington
Midtown
Midway/Coldstream
Morrell Park/Violetville
Mount Washington/Coldspring
North Baltimore/Guilford/Homeland
Northwood
Oldtown/Middle East
Orangeville/East Highlandtown
Patterson Park North & East
Penn North/Reservoir Hill
Pimlico/Arlington/Hilltop
Poppleton/The Terraces/Hollins Market
Sandtown-Winchester/Harlem Park
South Baltimore
Southeastern
Southern Park Heights
Southwest Baltimore
The Waverlies
Upton/Druid Heights
Washington Village/Pigtown
Westport/Mount Winans/Lakeland


Restaurants count vs Neighborhood
Count
Count per population
Score
Neighborhood
−1
0
1
2
3
4
5
Allendale/Irvington/S. Hilton
Beechfield/Ten Hills/West Hills
Belair-Edison
Brooklyn/Curtis Bay/Hawkins Point
Canton
Cedonia/Frankford
Cherry Hill
Chinquapin Park/Belvedere
Claremont/Armistead
Clifton-Berea
Dickeyville/Franklintown
Downtown/Seton Hill
Edmondson Village
Fells Point
Forest Park/Walbrook
Glen-Fallstaff
Greater Charles Village/Barclay
Greater Govans
Greater Mondawmin
Greater Roland Park/Poplar Hill
Greater Rosemont
Greenmount East
Hamilton
Harbor East/Little Italy
Harford/Echodale
Highlandtown
Howard Park/West Arlington
Inner Harbor/Federal Hill
Lauraville
Loch Raven
Madison/East End
Medfield/Hampden/Woodberry/Remington
Midtown
Midway/Coldstream
Morrell Park/Violetville
Mount Washington/Coldspring
North Baltimore/Guilford/Homeland
Northwood
Oldtown/Middle East
Orangeville/East Highlandtown
Patterson Park North & East
Penn North/Reservoir Hill
Pimlico/Arlington/Hilltop
Poppleton/The Terraces/Hollins Market
Sandtown-Winchester/Harlem Park
South Baltimore
Southeastern
Southern Park Heights
Southwest Baltimore
The Waverlies
Upton/Druid Heights
Washington Village/Pigtown
Westport/Mount Winans/Lakeland

Final Ranking

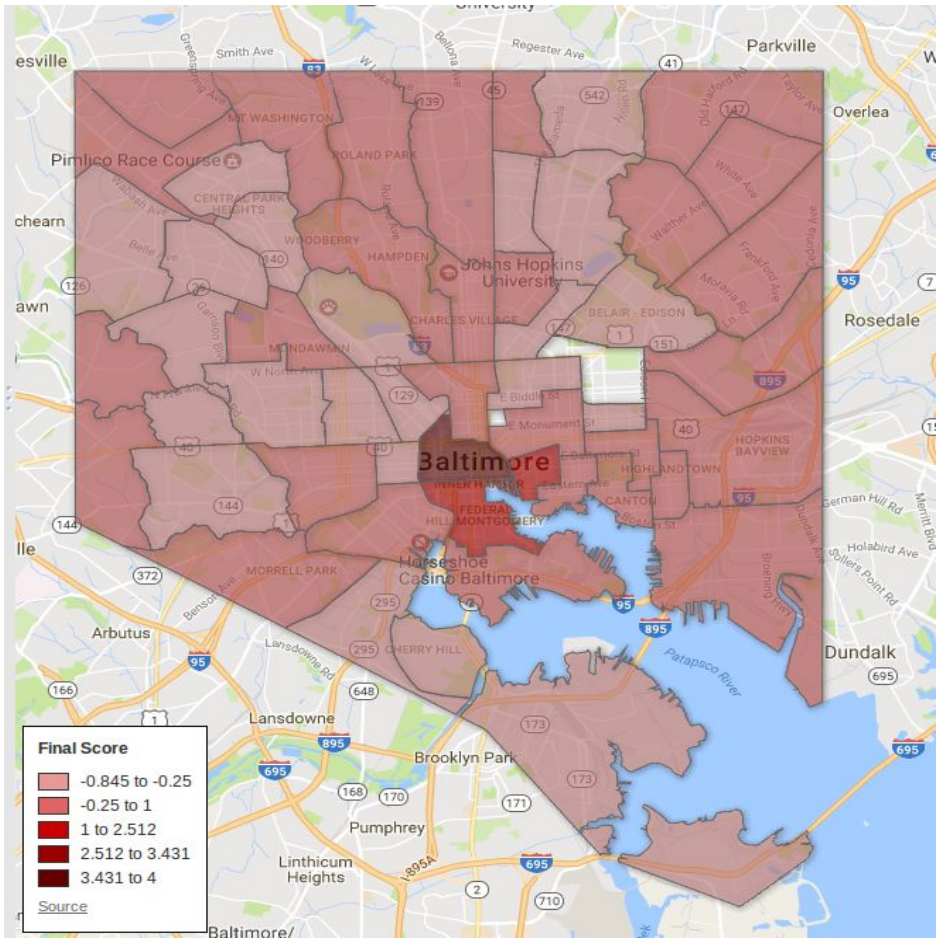
Below is the final ranking that we observed after incorporating the new parameters into our evaluation model.

	CSA2010	Final Score
13	Downtown/Seton Hill	3.749565
25	Harbor East/Little Italy	1.160344
29	Inner Harbor/Federal Hill	1.074473
34	Midtown	0.751496
15	Fells Point	0.730515
53	Washington Village/Pigtown	0.682169
11	Dickeyville/Franklintown	0.590958
37	Mount Washington/Coldspring	0.386896
47	South Baltimore	0.381654
21	Greater Roland Park/Poplar Hill	0.349679
27	Highlandtown	0.329329
18	Greater Charles Village/Barclay	0.279924
4	Canton	0.271830
48	Southeastern	0.204255
33	Medfield/Hampden/Woodberry/Remington	0.179143

The final ranking does in fact modify the ranking produced by the baseline model. This is clearly explainable by the weight vector which we applied to our parameter space. Our function, provided by the file final_ranking.py, takes the input data and a weight vector and multiplies the scores by their corresponding weight. The weights we used correspond roughly to those used by Nate Silver in his assessment of NYC neighborhoods. And the weights Nate Silver used were determined by a poll of American residents. The weight vector can be described by the following dictionary:

standard_weight_vector = { “price”: .25, “hotness”: .15, “crime”: .15, “nightlife”: .10, “education”: .10, “diversity”: .10, “creative”: .10, “art”: .05}

One distinct difference between the baseline and the final ranking is that price median housing price is a much smaller percentage of the overall total. In our new model housing price only accounts for 25 percent of the total whereas in the baseline model it accounted for 50 percent. This means that neighborhoods like Roland, which rank highly in price but relatively lowly on other metrics suffered, while more balanced neighborhoods like Downtown, which did not even rank in the top 10 in the first ranking, shot to the top!



Evaluation

In order to rigorously evaluate our ranking data we turned to several distinct datasets and approaches. I will begin with the most basic of our approaches and move towards the most sophisticated, detailing the relative merits and demerits, as well as the unique challenges posed by each, as I move along.

As part of their open data platform Baltimore provides anonymized 311 service data. A 311 service is an interface the city provides for non-emergency citizen reporting. Through 311 residents can flag problems for the city to rectify --in effect acting as a distributed citizen sensor network. Each row of the dataset captures this sensor data as an incident type, the neighborhood to which it applies, a start, status and due date, and an outcome.

Though the data is mostly well manicured the neighborhood schema which Baltimore employs differs from the community statistical area (CSA) model that the BNIA organization uses. Because the bulk of our statistics correspond to these CSA areas the sub-neighborhoods of the 311 dataset had to be aggregated into the CSA format. More subtly, the dataset actually uses two different timestamp formats within each row -- swapping the date and month. Close examination revealed several cases that ended before they started, or that had what appeared to be invalid month codes.

After cleaning the data of these irregularities we proceeded to our analysis. Our first approach was to use simple frequency counts to measure the dysfunction of each neighborhood. The assumption here would be that neighborhoods with more problems will have correspondingly higher 311 call totals. As it turns out, there is a negative correlation, as we expected.

Unfortunately, it is a rather weak one at best. This first pass yielded a correlation of -0.20023 and a relatively high p value of .14.

Perhaps then our first thought was too naive. 311 is a very general service that covers a huge array of incidents and not all of them are good proxies for quality. So next we sorted the categories by popularity and removed 311 data that we considered irrelevant to neighborhood quality and only kept measures related to health, housing, and pests. For instance, one such category is property sanitation. A 311 incident labeled "HCD-Sanitation Property" is used to denote an instance where a resident complains about the state of another residents property. If a neighbor's property were strewn with garbage or in disrepair then this HCD code would apply. Now operating under the assumption that certain classes of 311 call better mapped to neighborhood quality we again did the frequency tallying. The new dataset was only slightly less disappointing. This time we squeaked out a correlation of -.25 and a P value of .061.

So then, is using the 311 data to evaluate our model an idea dead in the water? Not yet. Our third approach utilized the time delta between when a case was opened and closed. The assumption here is that perhaps neighborhoods that are worse quality will also get lower quality service as the city might try to economize on labor, prioritize high value areas, and also satisfy it's most wealthy, and therefore influential, residents.

In order to get a handle on how the data looked we wanted to compute median and mean values for each incident type for each neighborhood. From there we could examine the variance between the median and mean values across all neighborhoods. The intuition was that we could focus on areas of high variance.

We began by binning all of the 311 calls on a neighborhood basis. This means that we mapped row numbers to incident type bins on a neighborhood by neighborhood basis. We then calculated the median and mean for each incident bucket.

This was not as simple as it seems. For one, we had to keep a raw count so we could validate that our statistics were meaningful. If a particular neighborhood or incident category had few meaningful samples across the board then it was discarded after a case by case examination. More troubling was the question of how we should represent our time deltas. While it made sense to store the data at its most granular resolution -- i.e. accurate to the second -- should we calculate variance there too? This was certainly a test of judgement. If we calculate variance at the level of seconds then the variance would be astronomical, but choose too large a timescale and you would see no resolution at all. All differences look small on the timeline to infinity. To this point we tested empirically what sorts of variance we would get at the second, minute, and hour timescale. Variance was astronomical if we used seconds, and very low at the hour end.

What does this mean? Well, perhaps that looking for variance to key us into categories to examine is the wrong approach. Instead, with that thinking in mind we moved on to calculating correlations between all valid incident categories (those that met a minimum count value) and the output of our scoring function.

Here to the interpretation was tricky. There were both positive and negative correlations (some reaching around .45 in either direction). What to make of that is unclear. Does the city prioritize certain kinds of service for some neighborhoods? Maybe, maybe not. It seems more likely that there is no real preference for neighborhood. Otherwise, would it make sense that 'TRS-Abandoned Vehicle' claims with an aggregate count of near 18000 incidents would be positively correlated with score --i.e. the higher the score the longer above the median time for the category it takes to close an abandoned vehicle case -- while 'TRS-Parking Complaint Commercial Veh. Residential' is negatively correlated? Or that 'HCD-Sanitation Property' is positively correlated (.35) while 'TRM-Illegal Sign Removal' would be negatively correlated (-.29)?

Both of these requests seem be beautification related and therefore high impact and probably high priority -- IF such a system of priorities did exist. Further, while it may be possible after seeing the data to come up with an ad-hoc rationalization we do not think that we can do so responsibly.

Given all of this, we tried one last approach on the 311 data. We wanted to look at the top 3 neighborhoods as ranked by our baseline model and the bottom three and see whether or not things became more clear. The correlation value ceiling increased significantly but in most cases the corresponding p value naturally increased as well. One particularly tickling insight: the amount of time to close a citizen's complaint about a city employee was significantly correlated with score. Confirming, in my view, what we already intuitively know -- no one is more entitled than the rich.

Predictive Model

We also build a regression model to predict the most probable prices of the houses in a neighborhood. Based on our observations and analysis to this point, we believe following are the parameters that would be the best choice for evaluating the output. Parameters used to train the model are mentioned below:

1. Media TOM
2. Nightlife
3. Education quality
4. Diversity
5. Crime

We have performed sublinear feature and target scaling on all the parameters to make sure gradient descent converges more quickly. We decided to take logarithm rather than z score since z score is just linear transformation of data, and sublinear transformation is much more meaningful to build such models.

We also found correlation between each pair of features to make sure we do not have any highly correlated feature in our model. From the table we can see that Crime and Restaurant have a correlation coefficient of .45 and p value much less than .05. When we combined these two features and then trained the model, the model showed lower score 0.67. We decided to keep these two features separately and then train. We also checked whether there was any need to do singular value decomposition to figure out the most relevant features, but since the feature count is low we did not observe any particular need to do so.

	TOM	Crime	Restaurant	Education	Diversity
TOM	1.000000	0.221893	-0.023980	-0.258521	0.043518
Crime	0.221893	1.000000	0.451788	-0.012080	0.313616
Restaurant	-0.023980	0.451788	1.000000	-0.128404	0.306724
Education	-0.258521	-0.012080	-0.128404	1.000000	0.096767
Diversity	0.043518	0.313616	0.306724	0.096767	1.000000

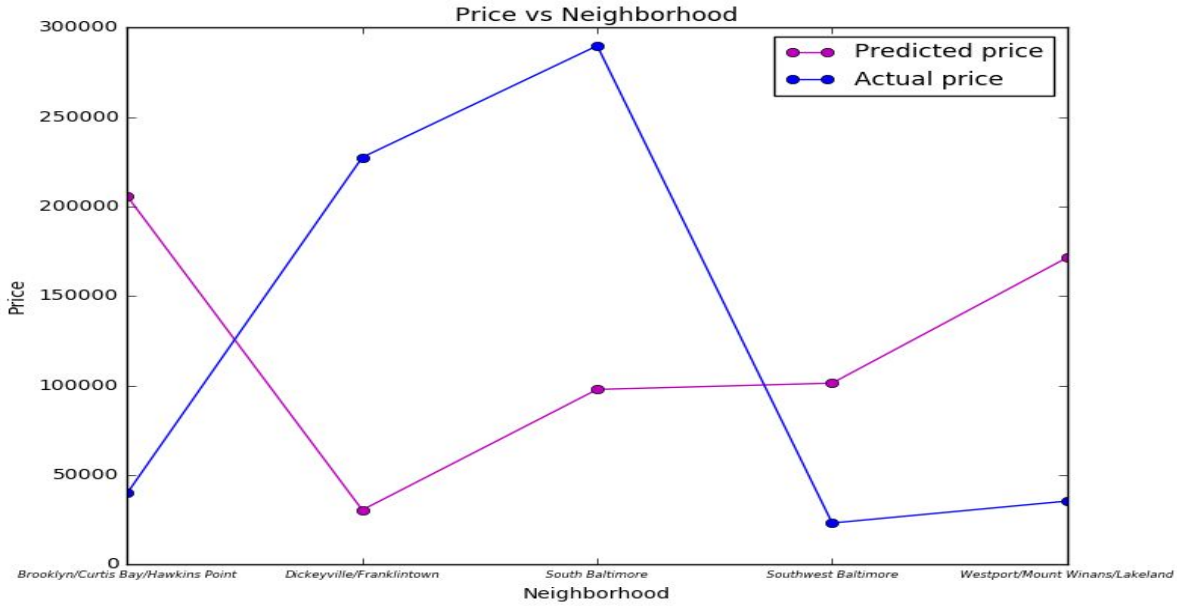
Model correctly penalised for having longer time on market for a house and having high crime rate and gives higher score for having a good stand in terms of number of restaurants, education, and diversity. Model score is 0.78 and residual sum of squares is 0.05.

From our initial analysis we can see that Brooklyn/Curtis Bay/Hawkins Point, Dickeyville/Franklintown, South Baltimore, Southwest Baltimore, and Westport/Mount Winans/Lakeland are outliers and they can really affect our linear regression fit.

Neighborhood	Actual Price	Predicted Price
Brooklyn/Curtis Bay/Hawkins Point	40000	205804
Dickeyville/Franklintown	227550	30524
South Baltimore	289900	97888
Southwest Baltimore	23250	101411
Westport/Mount Winans/Lakeland	35500	171505

From the graph we can see that the model has understated few and overstated few neighborhoods. Dickeyville/Franklintown has to suffer a lot due to its really high value of time in market. When we further look for the reason we find that the neighborhood is rural compared to its counterparts. We also see that contrary to our model South Baltimore is quite famous among its residents and is in high demand. Its population is quite diverse and the neighborhood is quite close to Federal Hill/Inner Harbor. These analysis further supports the fact that these places are

rather outliers.



Future Work

Model Enhancements: Below are the enhancements that we could do in future to check if the accuracy or robustness of the model ab be made better.

1. We can assign ratings to art places- museums, monuments, theatres etc to come up with a better and robust ranking system. This will require us to scour the web for such ratings as a comprehensive list of such places is difficult to get by per neighborhood basis.

2. We could assign ratings to the restaurants in addition to taking the count as good restaurants should be given more preference.

There could be a scope of improving our scoring function, evaluation and predictive models, but we believe that our analysis gives a complete and comprehensible picture of Baltimore's neighborhood.

Code Repository

We have made our analysis and source code available for any reference on GitHub, https://github.com/MichaelLiuzzi/BaltimoreRanking-519.

Team Members

1. Michael Liuzzi
2. Leena Shekhar
3. Sharang Bhat

This project is dedicated to our friend Michael, who is moving to Baltimore to start his career this January. All the best Michael.