|  |
| --- |
| Data Science Fundamentals (CSE-519) , Intermediate Project Report |
| Baltimore City Heat Map |
| Michael Liuzzi (109524485), Sharang Bhat (110920852), Leena Shekhar (110944612) |

|  |
| --- |
|  |

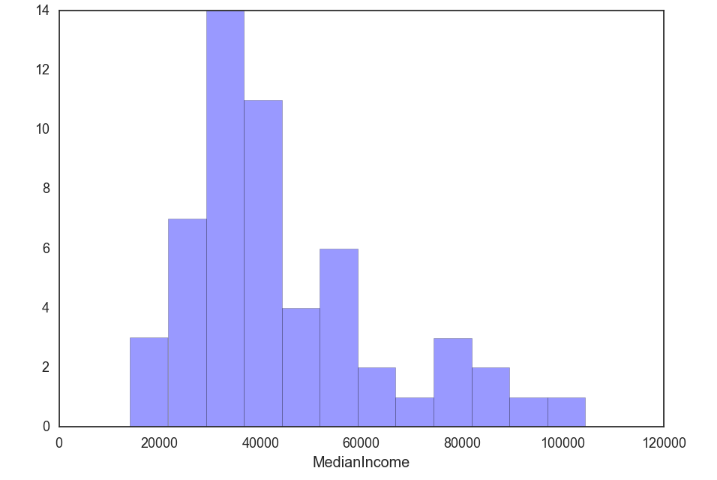
**Introduction**

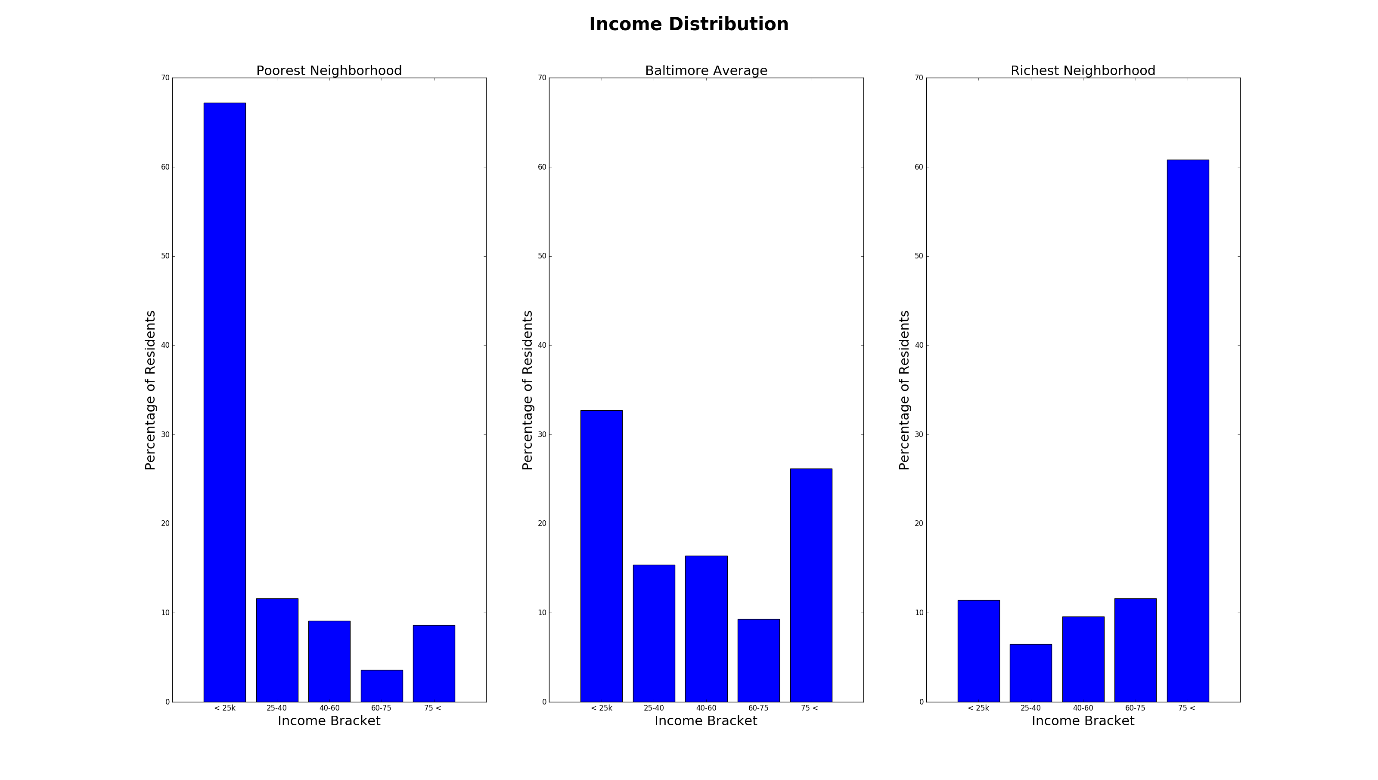
Baltimore is the largest city in Maryland, and the 29th most populous city in the country. It is also the largest independent city in the United States of America. This city has always been a city of neighbourhoods, everyone a universe unto itself. Each section has its own history, culture, and personality-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

We all know that moving to a new city can be a daunted 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 lace 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!

**Exploratory Analysis**

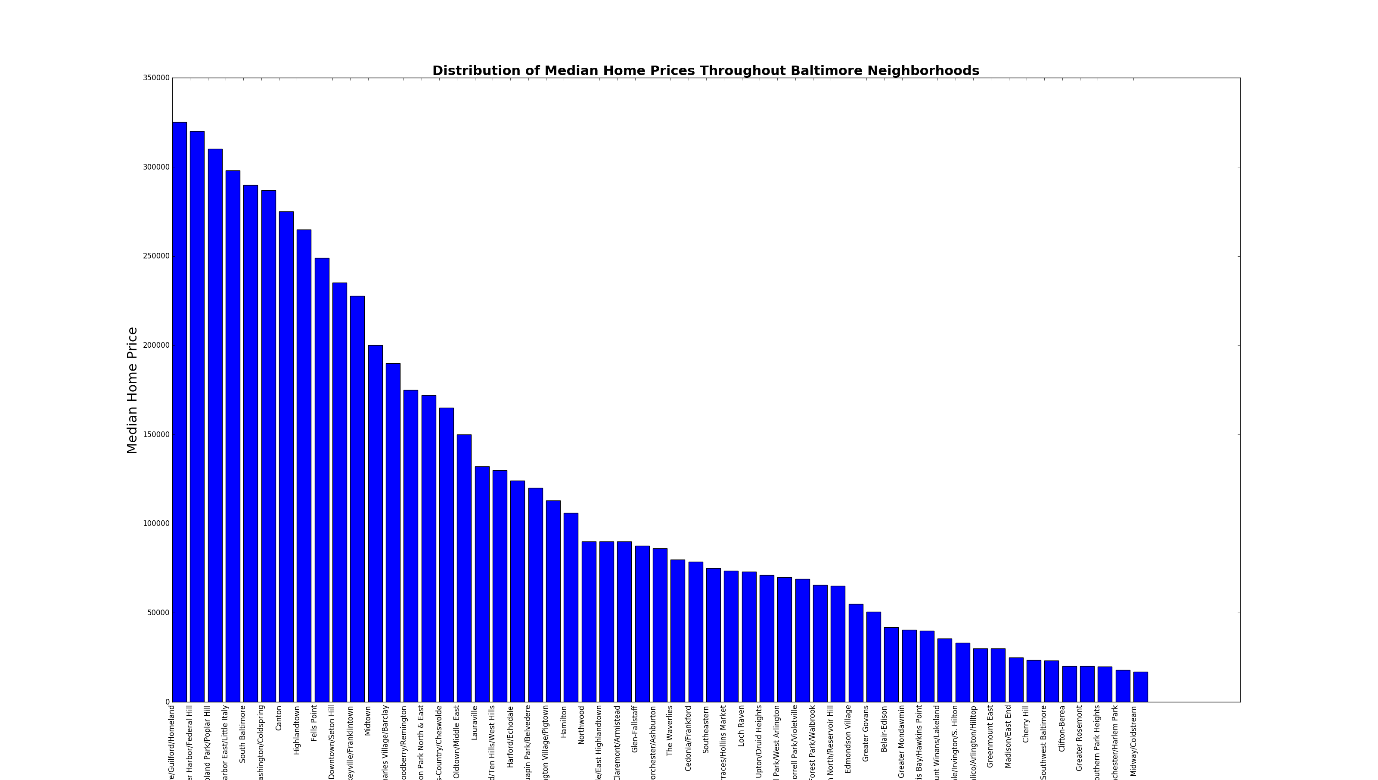
The first bit of analysis was inclined towards analysing distributions of metrics such as income, housing prices, and using the results, to look at the distributions in the richest and the poorest neighbourhoods in Baltimore.



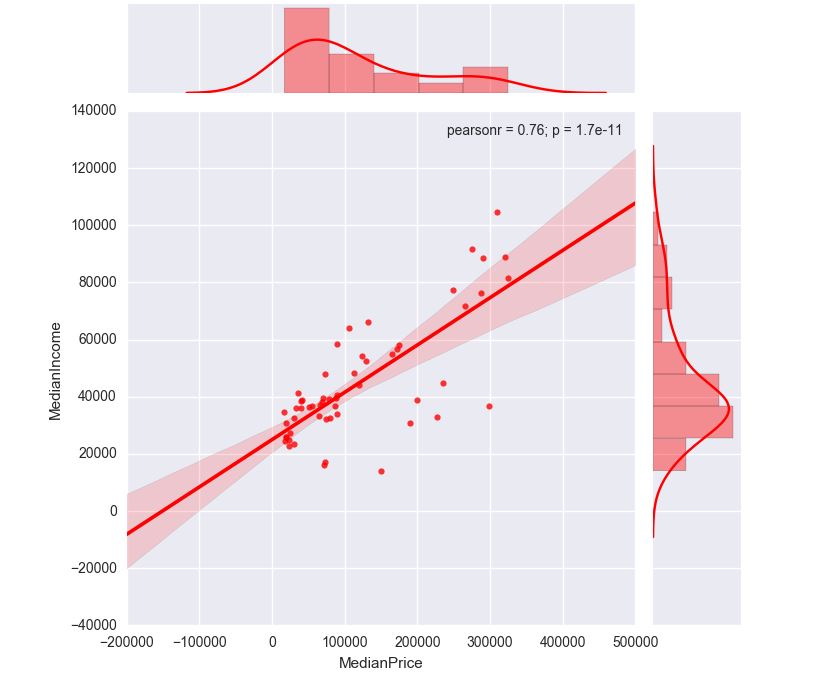


This chart confirms something that is something of a truism when we think about cities -- that neighbourhoods are extremely segregated and well defined on income boundaries. As we can see from these distributions the richest and poorest areas do not look like the Baltimore city average at all. Instead, they are, respectively, very rich, and very poor. This stark inequality represents one of the most troubling facets of city living. The inequality in wealth distribution is interesting in and of itself but it may also imply something important about the geography of the city. If wealth were more evenly spread would we see more "viable" neighbourhoods? It is hard to think that is not the case.

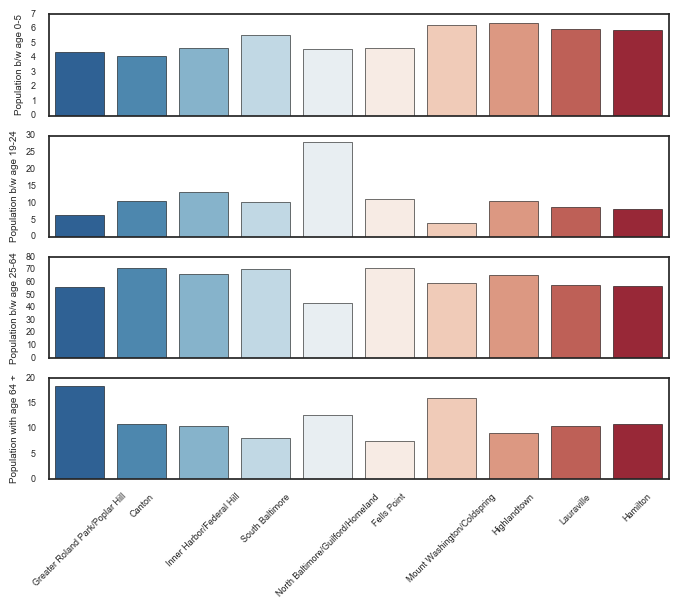
Our next move, was to analyse the distribution of housing prices across the city.



As we can see from the home price chart home prices are high in a relatively small number of neighbourhoods before declining rather drastically. This seems to reinforce the above: a few neighbourhoods contain much of the wealth of Baltimore. This can be further substantiated by the following graph, which demonstrates the relationship between median price and income. There exists a reasonably high correlation coefficient of 0.76 at a p-value less than 0.05.



Further, we looked at age distributions amongst the neighbourhoods the neighbourhoods with highest incomes.

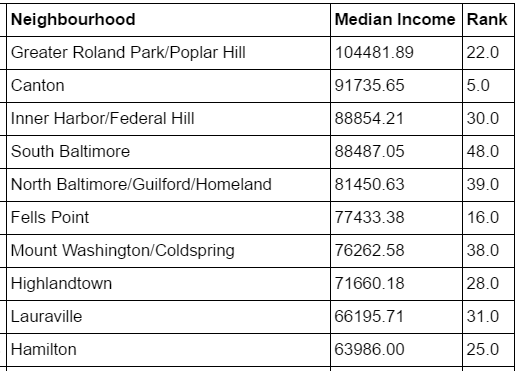
s

Comparing this chart to the income distributions across the neighbourhoods, we observed that the neighbourhoods with generally high incomes, are ones with a high population of people between the ages 19-64, that is, a working-class population. (North and South Baltimore). This gives us a fair idea of where the working population on the city is clustered. Also, we observed, that areas like Cold Spring have high populations of elderly people, and a relatively large population of children under the age of 5, and between the age of 25-64. This gives us an insight into the where families, with people involved in profession can be found.

We also considered racial diversity indices, percentage of people living below poverty line and the gender ratio for each neighbourhood. The above plots, were, in our opinion, the most relevant to our eventual mode.

**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:



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

Adult Arrests Per Thousand



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.

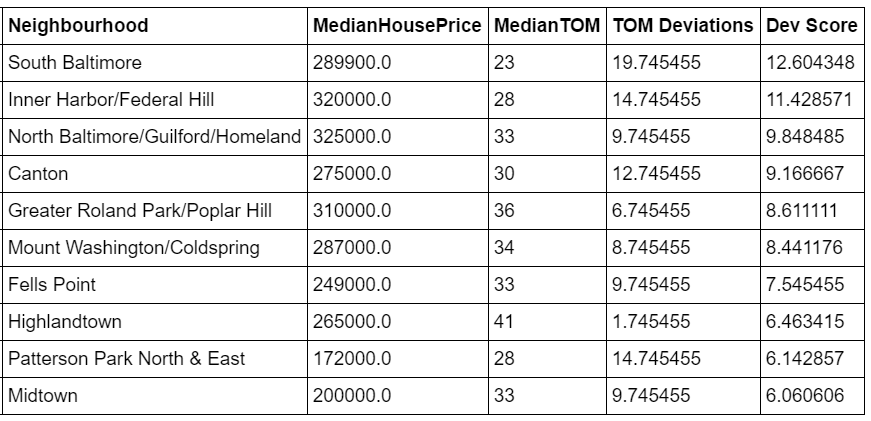
And what if you're not rich? Maybe you should find the safest neighbourhood that you can afford and stick to it. That might be a sensible approach. But what if you're a tough young guy and would tolerate a little less safety for better proximity to the music venues you like. Then you'd need some sort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

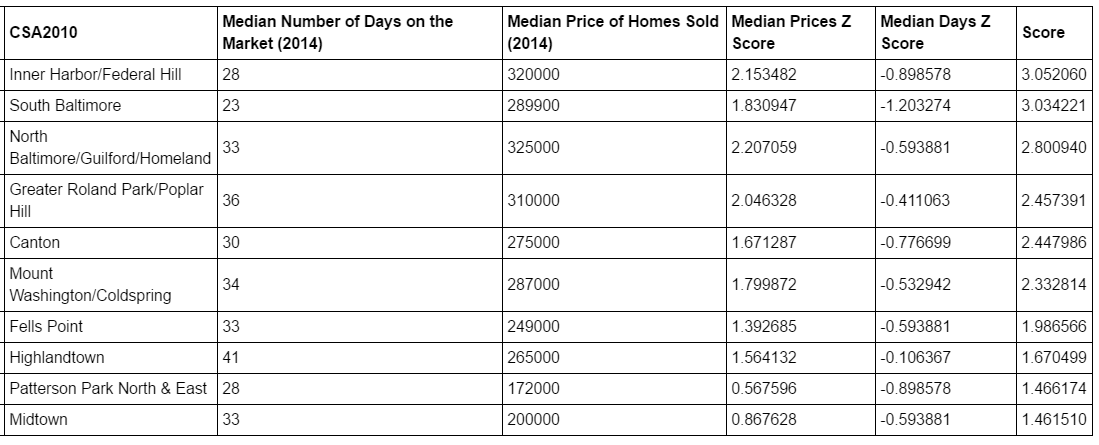


This model, we felt, over-penalized a neighbourhood with lower prices than the highest, and we also realised, that the Price Per Day of Time on Market doesn’t have any substantial meaning to our goal. We needed a better way to penalize a neighbourhood with high Time on Market.

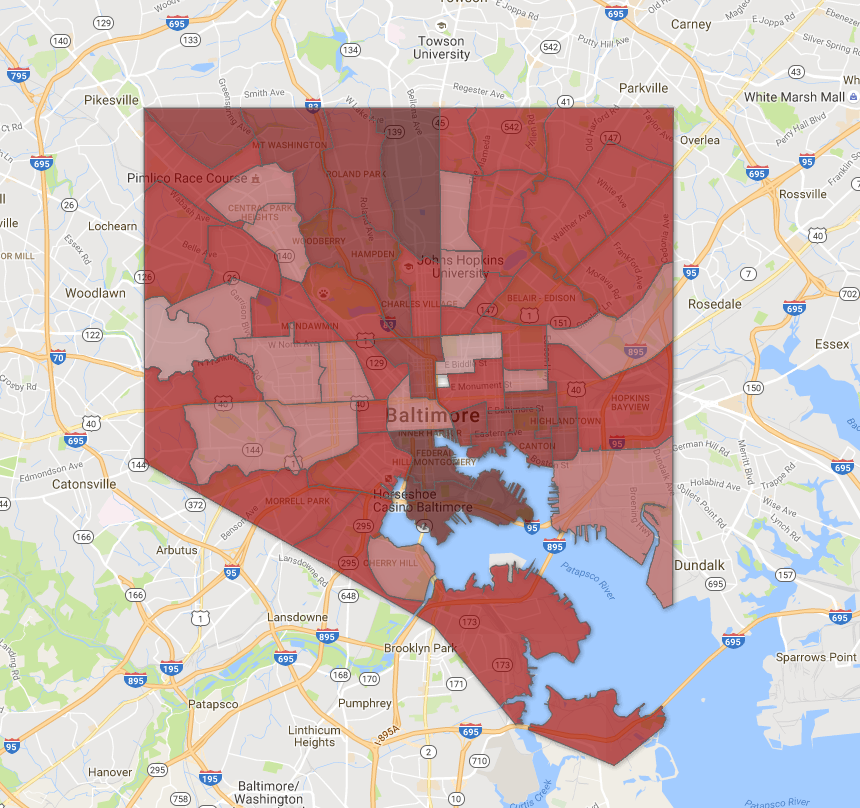
For a baseline notion of "best" we eventually settled on the following formula-

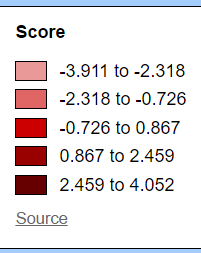
What does this formula capture? Well, we think it captures a few things very compactly. One way to compute the baseline model would be to have a lot of different parameters that we could weight to our liking. But there is no ground truth. Plus, as you might imagine, people don't choose the neighbourhood 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.



And as plotted geo-spatially -





**Improvements**

"Best Fits"

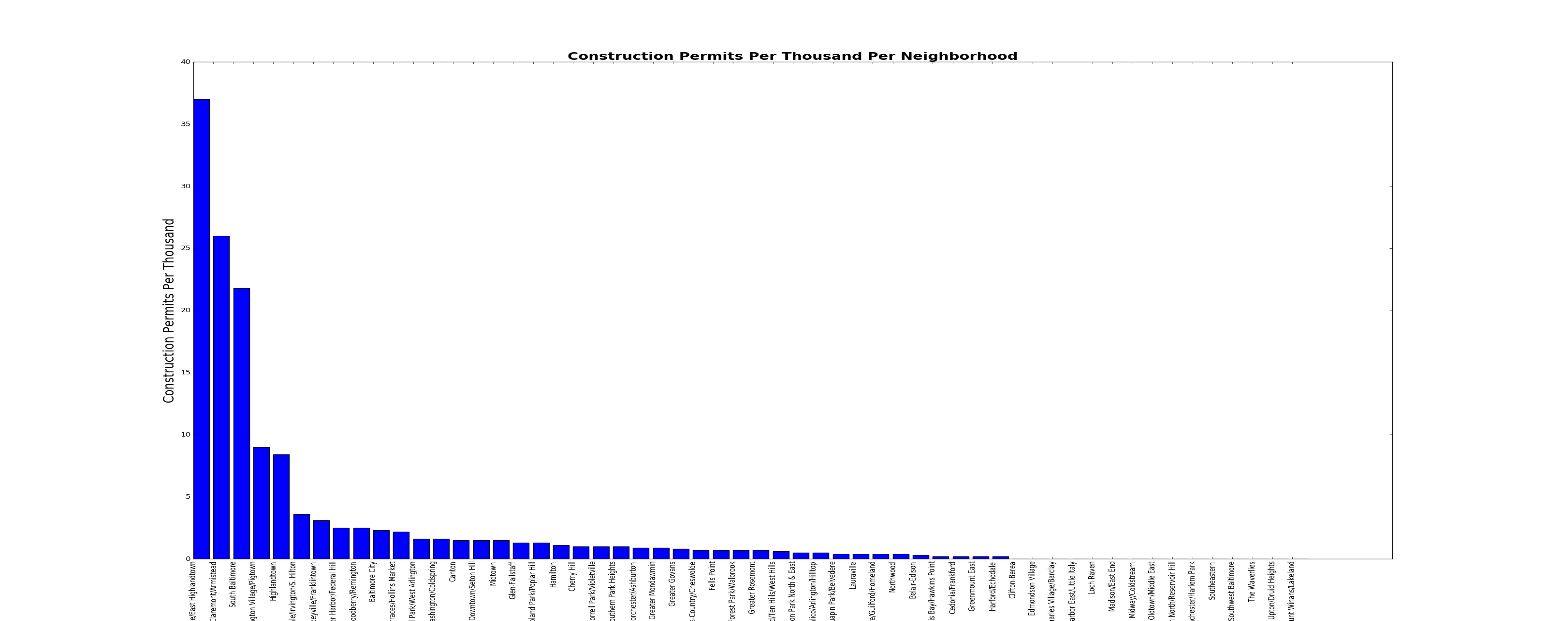
The model should also expose new and interesting areas that have not yet necessarily become hot spots. To this end, we will need to parametrize a new model that can tease out different dimensions of the data. And to that end we will provide profiles that are combinations of set points on the parameters we use.

As a matter of fact, Nate Silver actually did something to this effect for NYC five years ago.[1] We will likely use his categories as a starting point for our own modelling whilst taking into account data we don't have and including data that he didn't incorporate. These "categories" with their attached weights look something like this:

|  |  |
| --- | --- |
| Category | Weight |
| Housing Cost | 25 % |
| Transit: | 13% |
| Shopping and Services | 9% |
| Safety | 8% |
| Restaurants | 8% |
| Schools | 6% |
| Diversity | 6% |
| Creative Capital | 5% |
| Housing Quality | 5% |
| Green Space | 5% |
| Health and Environment | 5% |
| Nightlife | 4% |

**Additional Datasets/Metrics**

1) Construction- Areas with high construction might also be promising. While market prices give us a notion of current desirability for a given area construction permits give us some indication of FUTURE desirability as well. Construction is a long-term investment and gives us a high confidence indicator that the area will continue to be popular in the future. There is a significant and very lopsided concentration of residential construction per thousand residents. The top three areas for new construction are Orangeville/East Highlandtown at 36, Claremont/Armistead at 26.0 and South Baltimore at 21.8. The neighborhood with the fourth highest number of permits per thousand is Washington Village/Pigtown at a comparatively paltry 9. The construction quickly drops to near zero from there.

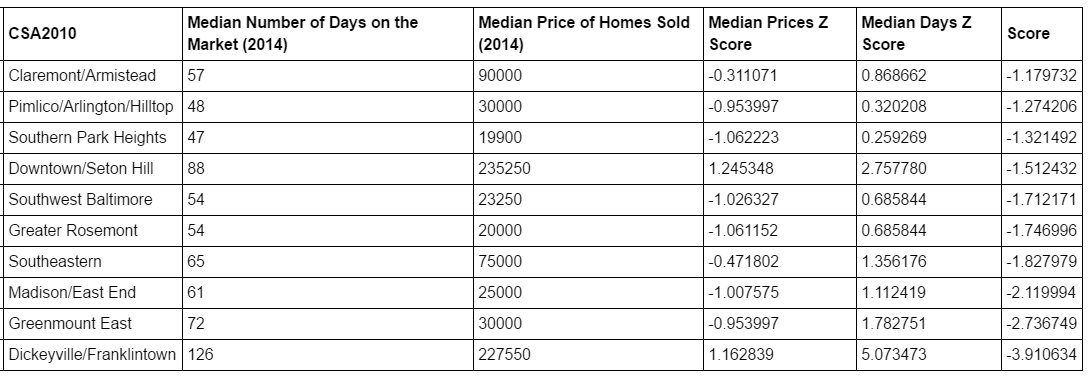


2) 311 Data- We have gathered data from the City of Baltimore open data portal[3], and categorized them by the location of the call and the complaint. We are planning to incorporate this data to incorporate citizen feedback into our model in near future.

3) Yelp data[2] to get information about restaurants, clubs, and pubs in each neighbourhood.

4) Creativity index using the number of murals, art centres, museums to see how neighbourhoods rank in terms of being artsy.

After focusing on these areas, it seems that there is good evidence to suggest that new construction reflects some measure of desirability. Take for example Orangeville/East Highland town. This neighbourhood abuts the currently very highly scoring Highland town which itself abuts the very highly scoring Canton neighbourhood. South Baltimore is currently the second highest performing neighbourhood per our scoring. The construction data strongly implies that it will retain, if not improve, upon that distinction. Washington Village/Pig town, while undergoing comparatively modest construction at 9 permits per thousand borders Inner Harbour/Federal Hill on its Eastern side and Downtown on it's Northern Border.



From the table above we can observe that for Green mount East(53) median prices is low compared to the average price, its availability in market is for a longer duration. After going through reviews and statistics about these neighbourhoods, we found that Green mount East is one of the worst places to live in Baltimore. We will further boost this hypothesis by finding support by incorporating more data into our model.

we also see that Inner Harbour/Federal Hill is quite in demand even though it is one of the most expensive neighbourhoods in Baltimore. Further web scouring indeed supports this finding.

We will also run clustering algorithms on these neighbourhoods to study their behaviour- ideally similar behaviours should get grouped together. This will also help in identifying any neighbourhood which exhibits any anomalous behaviour. We are positive that further exploration will reveal interesting facts about Baltimore and its people.

We see our final model developing as a recommender system which can guide a person looking for a specific neighbourhood by suggesting the best match.

**Similar work done**

1. Nearest Neighbourhoods in NYC?

Contiguity with popular areas? Neighbourhoods that border popular neighbourhoods but aren't particularly good themselves maybe should get a boost. Particularly if, as we see in the construction data, the area is "up and coming."

References

[1] <http://nymag.com/realestate/neighborhoods/2010/65374/>

[2] <https://www.yelp.com/baltimore>

[3] <https://data.baltimorecity.gov/City-Services/311-Customer-Service-Requests/9agw-sxsr>

[4] <http://bniajfi.org/>

[5] <https://fusiontables.google.com/data?docid=1g7xK0GWLHmc2Lo77Re3iZQ9XiMd-bgCDLzJz0LaZ#map:id=3>