Study of Gentrification in New York Metropolitan Area

Citadel European Regional Datathon

Dmitry Silantyev
Padmanaba Srinivasan
Stefan Tionanda
Bahdan Zviazhynski

October 25, 2020

Contents

1 Executive Summary

In this report, we set out to develop a robust methodology to identify areas with ongoing gentrification. Despite ample anecdotal evidence, there is still no unifying framework that allows to pin down specific census tracts as subject to gentrification, let alone make conclusion about cause and effect of the phenomenon. Perhaps the starkest illustration of this is a recent study by Preis et al. [4] that applied four different methodologies to the Boston area and found only seven tracts identified as gentrifying or at risk by all four methods. Having reproduced the often-cited algorithm of Prof. Freeman from Columbia University [2] (henceforth referred to as the Columbia or the Governing Magazine method) and compared it to the results of a group of researchers from UC Berkeley [1], we also found the number of overlapping tracts to be in low double digits. This result has often been attributed by researchers to the difference in city-specific assumptions about what gentrification is. Here, we propose a simple and elegant clustering algorithm that is theory-agnostic yet powerful in identifying areas that have meaningfully different statistical properties comparing to their population and that can serve as an aid in policy making.

While the picture painted in the BuzzFeed article and the Governing report is that of newcomers (mostly white and better educated) displacing inhabitants of historically non-white neighbourhoods by pricing them out, averages taken across the metropolitan area don't yet lend support to this story. Over the course of 2009-18, the share of population with at least a bachelor's degree grew by only 2 p.p. and the share of Hispanic population (+3 p.p.) increased at the expense of white (-5 p.p.) and African American (-1 p.p.) groups (see Table 1). Correlations (see Figure 1), while being in line with the story, are modest in magnitude and don't indicate existence of strong linear relationships between variables (except in the case of race-based variables that are tied by linear relationship by definition and the link between higher educational attainment and income, which is not related to gentrification per se). Even more surprisingly, tracts identified as gentrifying based on the methodology designed by Governing [3] (relying on trends in education, income and home prices), do not reveal almost any pattern in the corresponding racial dynamic or crime data that we collected separately (see Figure 4).

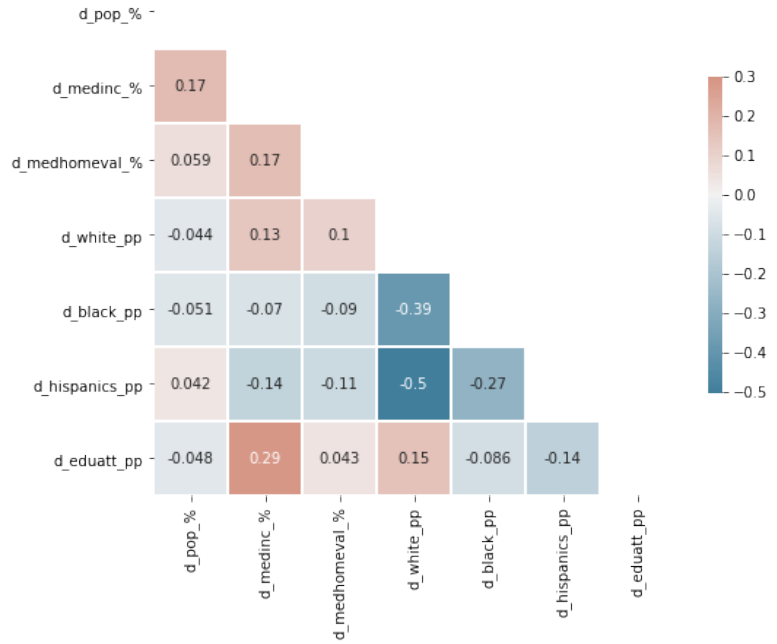


Figure 1: Correlation Matrix for Changes over 2009-18

Since there is no consensus about which properties a gentrifying tract should exhibit, this problem cannot be cast into a supervised classification problem for a machine learning algorithm. Therefore, without imposing any specific set of assumptions, we ran a series of unsupervised clustering algorithms to see if they can pick up some structure in the data. The results are very promising, as one can see in Figure 2. Out of ≈4,000 census tracts in New York metropolitan area, our best iteration of K-Means Clustering algorithm found a cluster of 130 tracts that all exhibit the pattern described by BuzzFeed, i.e. an ongoing displacement of African American and Hispanic populations by educated white newcomers, associated with a rapid increase in median income and median house price. Statistical tests conducted to see if the distribution of changes in the analysed variables across all tracts is different from that for the cluster, all confirm that the distributions are distinct in statistical sense. In that group, which we will further refer to as gentrified tracts, average (across tracts) median (within tract) income increased by 86% in 2009-18, the share of population with at least a bachelor's degree increased by 10 p.p. and an increase in the share of white population (+10 p.p.) was associated with a material decrease in non-white population (-6 p.p for African American and - 7 p.p. for Hispanic).

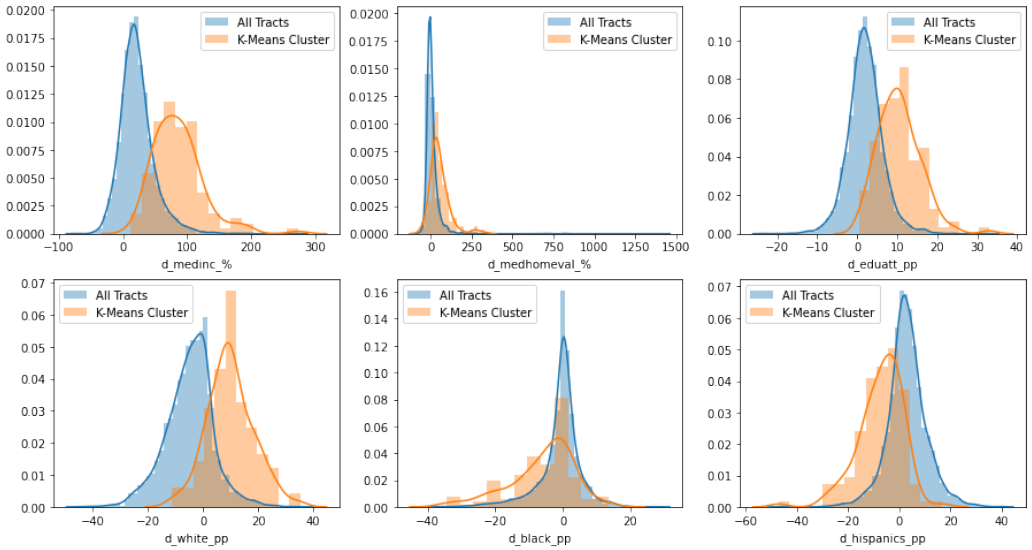


Figure 2: Distribution of Changes in All Tracts vs. Gentrified Tracts

		Mean		St. Dev.	
	KS p-val	All	Gent.	All	Gent.
Δ Median Income, %	0.0 ***	21.904	85.560	26.781	39.267
Δ Educ. Attainment, pp	0.0 ***	2.018	10.520	4.391	5.301
Δ Median Home Value, %	0.0 ***	8.046	54.250	47.301	60.731
Δ White Population, pp	0.0 ***	-4.923	9.920	8.711	8.413
Δ Black Population, pp	0.0 ***	-0.792	-5.983	6.344	9.513
Δ Hispanic Population, pp	0.0 ***	2.996	-7.294	7.656	8.833

Table 1: Kolmogorov-Smirnov Test Results and Distribution Statistics

Out of 130 tracts identified as distinct by K-Means Clustering and interpreted as gentrifying by us, only 14 overlap with tracts picked by our implementation of the Governing methodology. However, 106 tracts in the cluster are classified by the group of researchers from UC Berkeley as undergoing some form of exclusion or gentrification, ranging from "Ongoing Exclusion" to "Advanced Gentrification".

In order to ensure the validity of our findings out-of-sample, we reserve several variables as external, following a similar approach taken by Freeman [2]. One of reserved variables is the percentage change in the number of violent crimes committed on the territory of a tract and sourced from Crime Open Data Base (CODE) for over 1,000 census tracts in New York City. Here the

difference between gentrified and non-gentrified tracts is not so stark but still supports our thesis: while the number of violent crimes declined by 4.6% in NYC over 2009-18, it fell by 8.5% in gentrified tracts. Figure 3 illustrates this point as more tracts identified by our algorithm as gentrifying saw a material decrease in violent crime rates, comparing to the all tracts in New York City for which crime data was available.

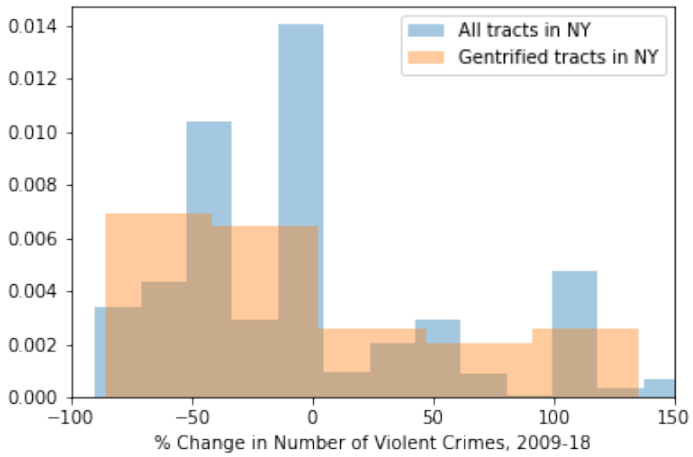


Figure 3: Distribution of % Changes in Number of Violent Crimes in All Tracts vs. Gentrified Tracts

The other way that we check if our methodology offers new insights into the effects of gentrification is by studying Manhattan's real estate market in depth. We dedicate Section 3 of the report to this case study. In summary, our K-means clustering algorithm is better able to find tracts that exhibit higher rates of property price growth than in surrounding tracts that haven't been identified as undergoing/undergone gentrification. Our method is able to identify these areas sooner and with key insight that alludes to one key indicator of gentrification: an increased desirability and demand for properties. However, the Columbia method was better able to find gentrified tracts with higher rates of eviction suggesting an increased rate of displacement of residential tenants in these tracts - another key indicator of gentrification. Figures 13 and 14 demonstrate these effects. We conclude that neither method wholly captures what gentrification is and suggest future work in extending the K-means method or other clustering methods forward with the inclusion of rental data to better identify gentrification.

2 Technical Exposition

2.1 Governing Magazine Methodology to Identify Gentrified Tracts

2.1.1 Gentrification Criteria and Data Availability

According to the methodology developed by Governing and implemented by BuzzFeed [6], the criteria for a tract to be eligible to gentrify are:

- tract is in the bottom 40% for median house value and income
- tract has the population of at least 500 people

If the above criteria are met, then to establish gentrification, the following criteria are used:

- inflation-adjusted house prices have gone up
- increase in inflation-adjusted home price is in the top third of all tracts
- increase in educational attainment is in the top third of all tracts

The corresponding criterion for gentrification is known to be one of the strictest compared to other studies [5], hence only 1.6% of the tracts are classed as gentrified by the end of 2018. One of the possible reasons is that according to the methodology, if data on any of the above criteria is not available, the tract is not eligible to gentrify, which is highly relevant for the census data as many values for house prices were replaced with a large negative placeholder value. Having those in the data shifts the actual value of the 40^{th} percentile for median house price and affects the results of the algorithm.

We first reconstructed the method as is and ran it on the tracts data in 2009 and 2018. It was found that 79 tracts were classified as gentrified, however 13 of them have missing median house price data, which makes these tracts not eligible to gentrify in the first place. We replaced all missing or invalid values with NaN, such that the percentile calculation ignores them but the data point itself is preserved as to not distort the distribution of other variables. This alternative implementation of the BuzzFeed algorithm identified 66 tracts as having gentrified. Taking a closer look, it can be noticed that all the tracts identified by our algorithm are identified by the baseline algorithm as well. That brings us to conclusion that our algorithm is more accurate than the baseline BuzzFeed algorithm, since it satisfies all the necessary criteria without dropping any tracts from the analysis completely and losing information.

2.1.2 Comparing Eligible and Gentrified Tracts

Below we compare the distributions of changes in variables for tracts that gentrified and tracts that were eligible but did not. It can be seen that by imposing a somewhat arbitrary cutoff point on two variables (median house prices and educational attainment), the method also affects the distributions of other variables (including income and racial variables) but not in a particularly meaningful way. Those tracts that the algorithm identified as gentrifying actually saw negative dynamic (on average) for all three racial groups we are considering in our analysis, which doesn't fit the very definition of gentrification that Governing sets. Moreover, some of the tracts that saw an increase in educated population are completely excluded from the sample. At the same time, the distributions of changes in variables for tracts that were eligible to gentrify but didn't, don't show material differences in statistical properties from the distributions for all tracts in the area. In our view, this calls for a different approach, not based on theory-driven and somewhat artificial constraints imposed on the data.

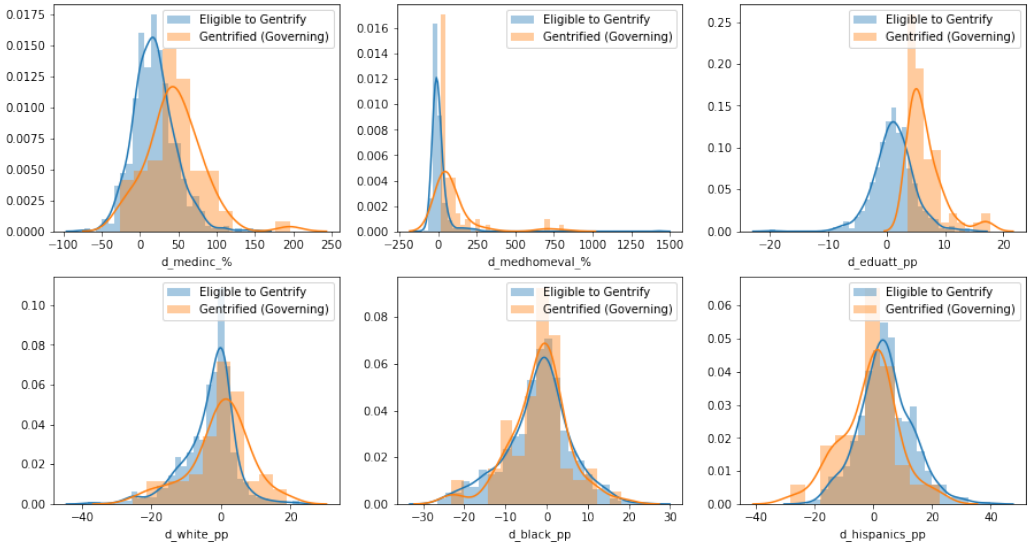


Figure 4: Distribution of Changes in Eligible vs. Gentrified Tracts

		Mean		St. Dev.	
	KS p-val	Eligible	Gent.	Eligible	Gent.
Δ Median Income, %	0.000 ***	21.030	45.249	29.504	37.027
Δ Median Home Value, %	0.000 ***	15.552	95.990	94.121	154.871
Δ White Population, pp	0.001 ***	-3.058	-0.243	7.709	8.872
Δ Black Population, pp	0.883	-2.079	-1.762	7.977	7.030
Δ Hispanic Population, pp	0.004 ***	3.596	-1.214	9.311	9.813
Δ Educ. Attainment, pp	0.000 ***	1.715	6.600	3.787	3.070

Table 2: Kolmogorov-Smirnov Test Results and Distribution Statistics

2.2 Clustering Algorithm to Identify Gentrified Tracts

Since we operated under the assumption that the distribution of changes in selected variables over 2009-18 period should differ between the population of all tracts and the sample of tracts labelled as gentrified, our first step was to apply Isolation Forest as an outlier detection algorithm and see if the points that it selects as outliers can be interpreted as gentrifying tracts. In Isolation Forest, the process of constructing a decision tree with random splits, that divide up an n-dimensional space using orthogonal cuts, is repeated to create an ensemble. For each data point, the average number of splits required to isolate the point is a metric for the regularity or normality of the point. The algorithm has the added benefit of random splits being computationally cheap and averaging across all trees considers the multiple manners to isolate the data. The downside is that "contamination", i.e. the share of points to be selected as outliers, is a hyperparameter of the algorithm and we would like to avoid specifying that ex-ante. However, having run the algorithm for a grid containing different possible values of contamination, we did not identify a set of points that would possess characteristics allowing us to interpret them as gentrifying tracts.

One of the complications that we were facing in working with available data was that only two variables (change in median household income and change in median home value), explained over 94% of variance in the data set. As can be seen in Figure 5, the first two principal components of standardized data basically coincide with the above mentioned variables. This property of the data makes separating it along all dimensions more challenging than if the features were distributed more similarly.

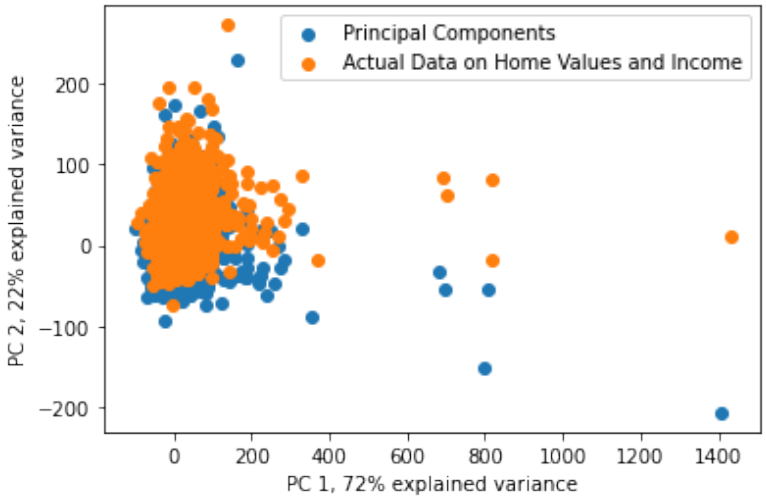


Figure 5: First Two Principal Components vs. Original Data

Naturally, we turned our attention to clustering algorithms that separate unlabelled data in feature space. The first such algorithm that we tried was Gaussian Mixture, that produces probabilistic cluster definitions. However, having conducted both D'Agostino-Pearson and Shapiro-Wilk normality tests for the distributions of changes in independent variables (that can be see in Figure 2), we did not identify a single distribution that would be statistically indistinguishable from normal. In addition, despite our expectation that the clusters have elongated form, which Gaussian Mixture is good at identifying, it did not pick up any cluster that would possess the properties consistent with gentrification phenomenon.

Our second choice, the K-Means Clustering algorithm, actually performed much better and finally allowed us to separate the data in a way that was consistent with the definition of gentrification. The clusters in this algorithm are chosen to reduce the inertia, the within cluster sum of square distance, and therefore the objective function is

$$\min_{C_k} \sum_{k} \sum_{X_j \in C_k} \|X_j - \mu_k\|^2.$$

The centroid of a cluster μ_k is equal to

$$\mu_k = \frac{1}{|C_k|} \sum_{X_j \in C_k} X_j,$$

where $|C_k|$ is the number of points in cluster k. The components of the centroid are equal to the mean of each feature of all points assigned to the cluster. After choosing the starting locations of each cluster's centroid randomly, it:

1. assigns each point to a cluster based on which cluster centroid it's the closest to
2. calculates the centroid of resulting cluster using the points that have been assigned to the cluster
3. repeats the above steps until convergence is met

While the algorithm is guaranteed to converge, it will not converge to the global minimum. It is a greedy algorithm in a sense that it is relatively quick to run but we are at risk of obtaining a suboptimal solution and different starting positions will result in different solutions. To counteract reporting local minimum as our final result, we ran the algorithm thirty times and choose the best iteration. We also experimented with the number of clusters and ended up obtaining the best results when we set the number of clusters equal to eight, the number of categories the group of UC Berkeley researchers used to classify stages of gentrification. It is important to note that, while the number of tracts in each cluster almost exactly coincided with the distribution produced by that group, there was very little overlap in the correspondence of each category to our clusters. Final clustering distribution is depicted in Figure 6 below and the significance of the result is best illustrated by Figure 2.

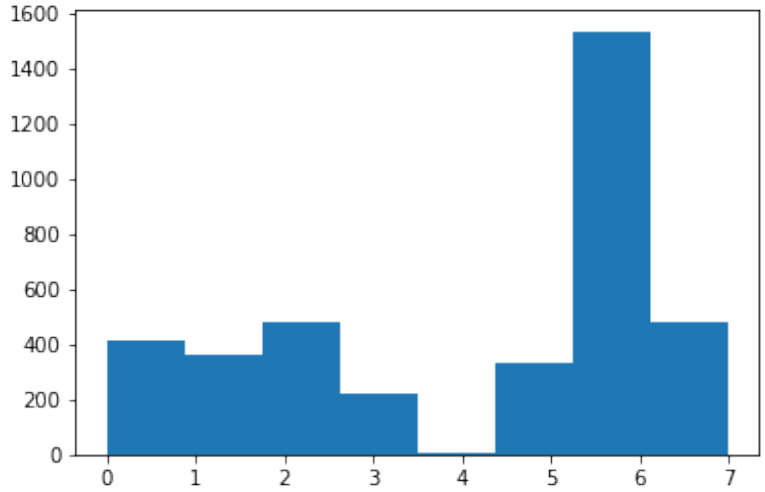


Figure 6: Cluster Size for K-Means Algorithm

3 Manhattan: A Case Study in Gentrification

In this section we explore how tracts identified by the Columbia methodology and the tracts identified by our methodology as having been gentrified compare in terms of the change in market value of condominiums along with the effect of the evolving property market, as well as the implications for renters. We consider Manhattan (County of New York), one of the five borough of New York City, as the focus of our study.
The Department of Finance (DOF) is required by New York state law to value condominiums or cooperatives as though they are residential apartment buildings and these valuations are made public through the DOF Condominium Comparable Rental Income dataset [7] (DOF Condo Valuation). The DOF uses a relative method of valuation combining rental information from similar properties to estimate a market value and more importantly, a market value per square foot which enables comparison between condominiums in different areas regardless of size. The DOF Condo Valuation dataset refers to each property as using a 'Borough, Block, Lot' [8] (BBL) parcel system used in New York City from which we find the GEOID of properties using a tool on the US Census Bureau website [9]. GEOIDs pertaining to Manhattan are filtered for yielding the dataset used in this case study.

3.1 Comparing condominium value change in gentrified tracts

We compare the tracts identified as having undergone gentrification by both the Columbia methodology as well as our own K-means method. The Columbia framework identifies nine tracts as gentrified of which six have sufficient market valuations to proceed. Our framework identifies sixteen gentrified GEOIDs of which two tracts (36061022900, 36061022301) were identified by the Columbia method as well. We compute the mean cumulative percentage change over all GEOIDs for tracts identified by each method in Figure 7.
However, analysing relative change in property values for gentrified tracts alone doesn't tell us about how prices in these tracts have changed compared to the other, non-gentrified tracts in Manhattan. We calculate the cumulative relative value average for all non-gentrified tracts identified by Columbia and K-means and calculate the excess relative change:

$$R_e = R_g - R_{ng} \tag{3.1}$$

R_e, Excess relative change

R_g, Relative change in gentrified tracts

R_{ng}, Relative change in non-gentrified tracts

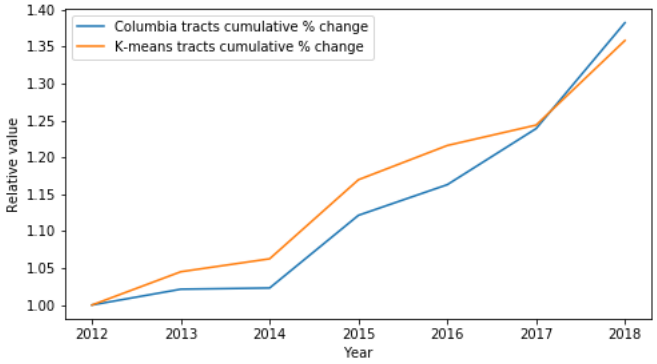


Figure 7: Comparing cumulative relative change in condominium values

We plot the excess relative change for gentrified tracts identified by each method in Figure 8. We see that the gentrified tracts identified by the Columbia method see very little excess increase over non-gentrified tracts up until 2016, after which a sharp increase in excess relative change occurs. The excess change for gentrified tracts identified by K-means, on the other hand, shows a far more gradual but consistent positive excess over non-gentrified tracts. But we are now faced with the question, how does this compare to the state of the general housing market in Manhattan? Specifically, how has the market driven this trend?

3.2 The property market in Manhattan

We are interested in the broad state of the housing market in New York City and more specifically, Manhattan. The magazine, *New York Curbed*, has performed an analysis of the general property marked in the five boroughs of New York City [10] and openly provides the datasets they used with total property sales per quarter from Q1 2010 to Q3 2019 and mean sale price for Manhattan for the same time frame. Analysing the state of the property market in Manhattan allows us to assess and explain excess relative change observed in the Section 3.1.

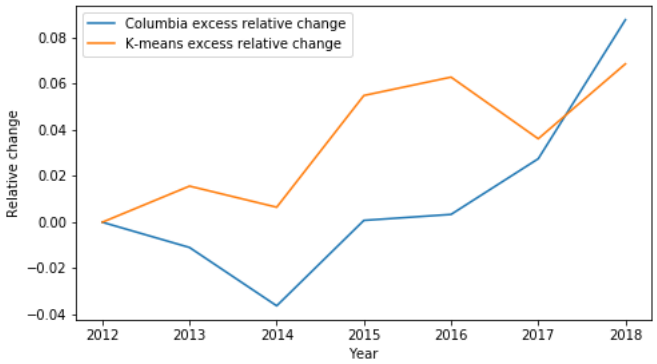


Figure 8: Excess relative change in condo prices

Using the aforementioned datasets, we plot the number of sales per quarter for all boroughs in Figure 9 where we see Manhattan dominating with most sales in the early half of the 2010's, ceding this title to Queens post approximately 2015. Plotting the relative change in sales for the period we are interested in, 2012-2018, (Figure 10) we see that sales peak in 2013 and 2014 before generally declining. However, during this period the average sale price of properties in Manhattan has increased monotonically (Figure 11) despite a reduction in overall sales leading to an overall increase in annual market value in Figure 12, except for 2018.

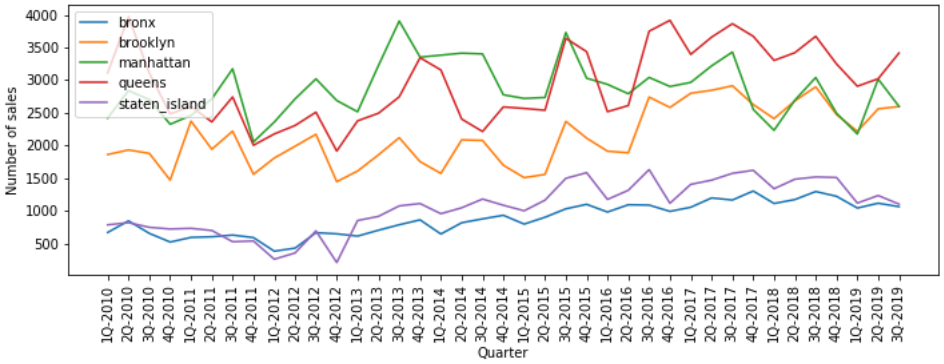


Figure 9: Total sales per quarter

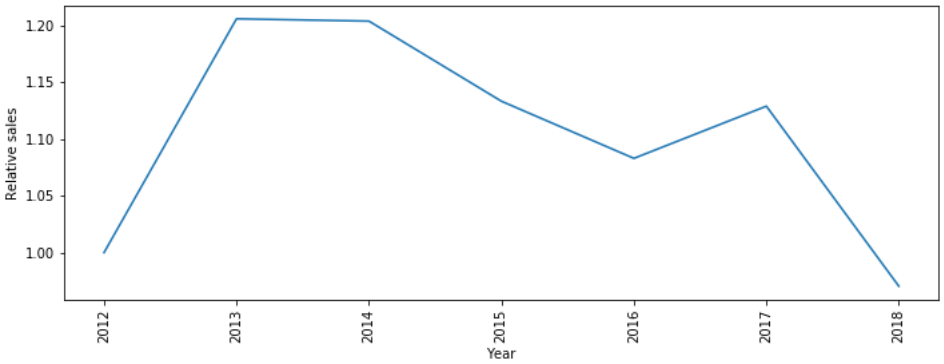


Figure 10: Manhattan relative property sales 2012-2018

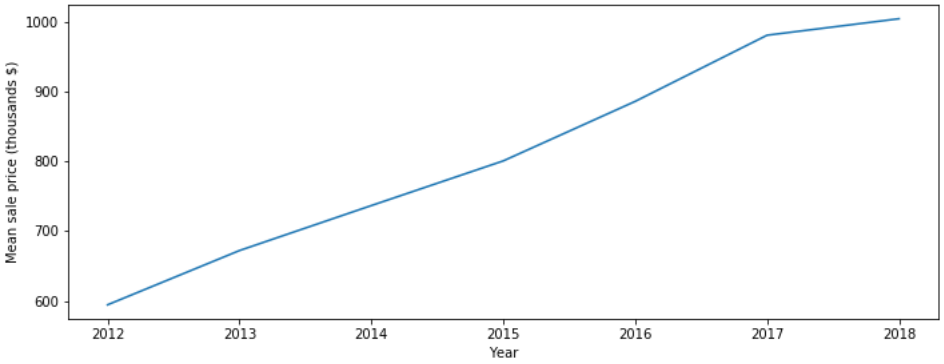


Figure 11: Manhattan mean sale price 2012-2018

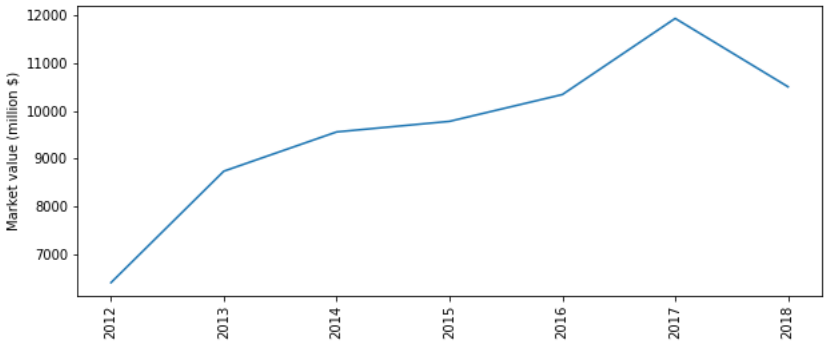
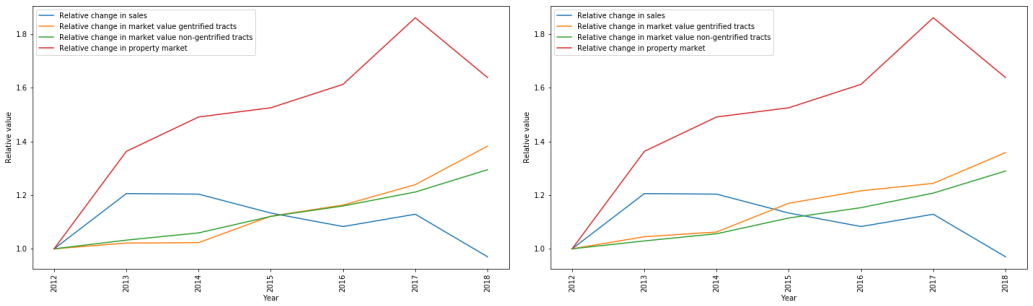


Figure 12: Manhattan property market value 2012-2018

2018 saw the worst performing Manhattan real estate market since 2009 as over the past decade [11], lack of available land for new construction led to

increasing land values which in turn encouraged any new properties to be built for the luxury market resulting in a lack of more affordable housing. Financial market performance in 2018 saw a souring of interest for Manhattan real estate from foreign investors leading to a declining property market.
Figure 13a shows that condo price growth in areas identified as gentrified according to the Columbia methodology has occurred almost exclusively after 2016, merely two years after the decline in Manhattan's property sales began. This suggests that gentrification in Manhattan is a very recent phenomenon that has occurred in response to dwindling interest in the high value properties. In comparison, condominium prices in areas identified as gentrified by our K-means model (Figure 13b) identifies more tracts that have seen excess increase in condo prices far earlier with weak signs of excess increase prior to 2014 and far more significant excess increase after 2014, which corresponds to gentrification as a response to a reduction in property sales. With our K-means method, we are able to identify tracts undergoing gentrification earlier than with the Columbia method, as well as relate gentrification more closely to the state of the property market in the area.



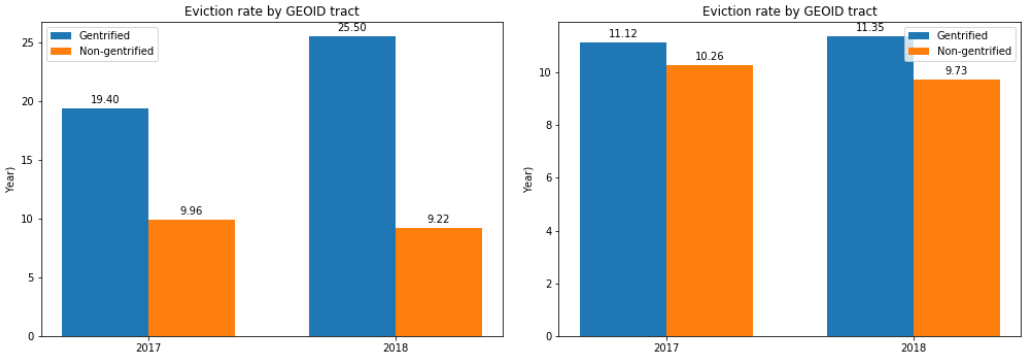
(a) Columbia method comparative plot (b) K-means method comparative plot

Figure 13: Comparative plots for Columbia and K-means method results

3.3 Considering rental properties

Thus far we have considered only the price of properties – price is a direct indicator of the desirability and demand for housing in an area – but we have failed to consider displacement. Renters are another key demographic in an area as they are ones that are likely to be 'priced out' of their homes during the process of gentrification. Yet we disregard this information with good reason; New York only provides data on evictions from the year 2017 [12]. This limits the extent to which we can use evictions data to identify

gentrified tracts, but in this study it provides a crucial way of evaluating the methodologies that find such tracts.
We consider residential evictions in Manhattan that were completed in 2017 and 2018, although we note that eviction proceedings take on average twelve weeks in New York, in addition to up to six months a judge can award a tenant to stay in a property before eviction.
To enable comparison between evictions in gentrified and non-gentrified areas, we normalise total evictions by the number of GEOID tracts they occurred in thereby obtaining an average evictions per tract metric for 2017 and 2018.
Figure 14a shows that gentrified tracts identified by the Columbia method show significantly higher eviction rates than non-gentrified tracts with the eviction rate increasing in 2018 for these gentrified tracts. By contrast, in Figure 14b, the K-means method gentrified tracts show a very similar eviction rate with both non-gentrified areas as well as across the two years. This potentially suggests that the Columbia method is more adept at recognising displacement due to gentrification, although more data of evictions throughout the 2012-2018 period would confirm this.



(a) Columbia method (b) Our K-means method

Figure 14: Evictions per tract for gentrified and non-gentrified tracts identified by Columbia and K-means methods

References

[1] Chapple, Karen, et al. "Developing a new methodology for analyzing potential displacement." (2017).

[2] Freeman, Lance. "Displacement or succession? Residential mobility in gentrifying neighborhoods." *Urban Affairs Review* 40.4 (2005): 463-491.

[3] Maciag, Michael. "Gentrification Report Methodology" *Governing* (2015) `https://www.governing.com/gov-data/gentrification-report-methodology.html`. Accessed 24 October 2020.

[4] Preis, Benjamin, et al. "Mapping gentrification and displacement pressure: An exploration of four distinct methodologies." *Urban Studies* (2020): 0042098020903011.

[5] Bhavsar, Kumar, Richman. "Defining gentrification for epidemiologic research: A systematic review." *PLOS ONE*. 15(5):e0233361.

[6] Vo, Lam Thuy. "They Played Dominoes Outside Their Apartment For Decades. Then The White People Moved In And Police Started Showing Up" *BuzzFeedNews* `https://www.buzzfeednews.com/article/lamvo/gentrification-complaints-311-new-york` Accessed 24 October 2020.

[7] NYC Open Data DOF Condominium Comparable Rental Income in NYC. `https://data.cityofnewyork.us/City-Government/DOF-Condominium-Comparable-Rental-Income-in-NYC/9ck6-2jew`. Accessed 20 October 2020.

[8] Borough, Block, Lot parcel number system. `https://en.wikipedia.org/wiki/Borough,_Block_and_Lot`. Accessed 20 October 2020.

[9] Finding the GEOID of an address. `https://geocoding.geo.census.gov/geocoder/geographies/onelineaddress?form`. Accessed 21 October 2020.

[10] NYC home prices nearly doubled in the 2010s. What do the 2020s hold? `https://ny.curbed.com/2019/12/13/21009872/nyc-home-value-2010s-manhattan-apartments`. Accessed 22 October 2020.

[11] Manhattan real estate closes 2018 as worst year since the financial crisis. https://www.cnbc.com/2019/01/02/manhattan-real-estate-closes-2018-as-worst-year-since-financial-crisis.html. Accessed 23 October 2020.

[12] NYC Open Data Evictions https://data.cityofnewyork.us/City-Government/Evictions/6z8x-wfk4. Accessed 22 October 2020.