The Data Open

Impact of Gentrification on the socioeconomic environment in the New York metropolitan area

Europe Regional Datathon 2020

Felix Teufel, Marouen Chafrouda, Max Johannes de Bock, Qinmeng Luan

October 25, 2020

1 Topic Question

In New York Metropolitan area, and more generally in the US there is growing fear of social injustice and the need to solve this issue is of utmost importance to policy makers, which is why we wanted to study the gentrification phenomena and provide insights on whether it has anything to do with social injustice and more generally with the changes in the socioeconomic environment in the New York metro area. which brings us to our main question :

- **What is the socioeconomic impact of gentrification in the New York metropolitan area?**

We hope this study can help governments have a better understanding of the impact of gentrification and give early indicators on where it will happen so that targeted policies can be put in place to ensure equality.

2 Executive Summary

We investigated the socioeconomic effects of gentrification on a census tract level. Through our analysis of the data, we discovered the following key insights.

- Gentrification is dependent on location : The spatial distribution of gentrified tracts (Figures 1,2), reveal that they cluster in certain urban areas.
- The increase in median household income caused by gentrification does not affect all ethnic groups equally: Caucasians see the largest increase in income while other ethnic groups do not undergo the same trend (Figure 5).
- Gentrifying tracts defy general demographic trends : the ratio of Caucasians steadily increases in gentrifying areas, and the African American population declines (Figure 6). Gentrifying tracts see no increase in Latino population, whereas other tracts in general do.
- In **New York city**, Gentrifying tracts have a consistently lower number of Emergency Medical Service incidents compared to non-gentrifying tracts.

From these insights, we constructed a set of features that are predictive of gentrification. Using these features we were able to design a regression model that can detect ongoing gentrification at early stages of the process from publicly available data.

3 Technical Exposition

3.1 Determining the gentrified and non-gentrified tracts

The methodology applied to this work to capture the gentrification process is based on the method proposed by *Governing Magazine* [1]. We start our analysis by finding tracts that are **eligible** for gentrification among all tracts in the NY-NJ-PA Metropolitan Statistical Area (MSA) based on these criteria:

1. The tract had more than 500 residents in 2009 and 2018.
2. The tract's median household income was in the bottom 40% in 2009.
3. The tract's median home value was in the bottom 40% in 2009.

Since the tracts eligible for gentrification are what we're interested in, **in all the rest of our analysis we will focus on eligible tracts.**

In order to test whether an eligible tract has **gentrified** in the time period of interest we use the following criteria:

1. The increase in a tract's educational attainment is in the top third percentile of all tracts within a metro area.
2. The tract's inflation-adjusted median home value increased and the increase was in the top third percentile of all tracts. We used Consumer Price Index (CPI-U) data provided by the U.S. Department of Labor Bureau of Labor Statistics [2] for the adjustment.

Based on the above criteria, of the 5023 tracts in the NY metropolitan area, 1523 were eligible for gentrification and 171 gentrified over the 2009-2018 period, yielding a **gentrification ratio** of **11.2%**.

In order to be able to rigorously test our findings, we chose to exclude 20 % of the data before starting our exploratory analysis as this analysis will be the main driver of the features engineering process and we want to keep a testing dataset that is not subject to any bias from our side.

3.2 Impact of geography on gentrification

In order to have a better view of how all the tracts are situated, and get a better sense on whether the location could play a role in the gentrification process we start by plotting the different tracts on a map in Figure 1.

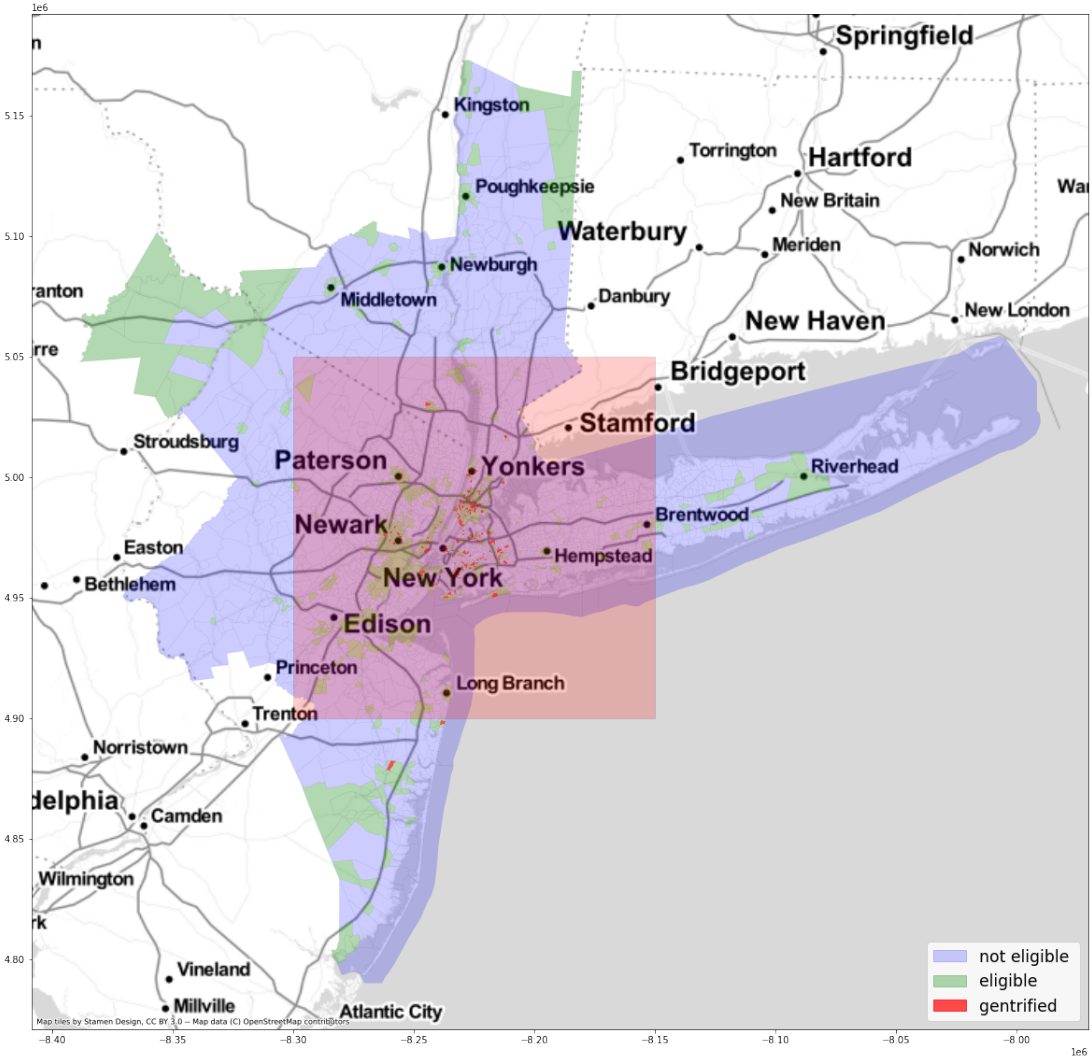

Figure 1: The NY-NJ-PA Metropolitan Statistical Area (MSA) map showing eligible and gentrified tracts in the year 2018. The area shaded in red is visualized in more detail in Figure 2

We observe in Figure 1 that the majority of gentrified areas are clustered in the New York City area, which is visualized in more detail in Figure 2. We conclude that the geographic location of a tract affects its tendency to be gentrified.

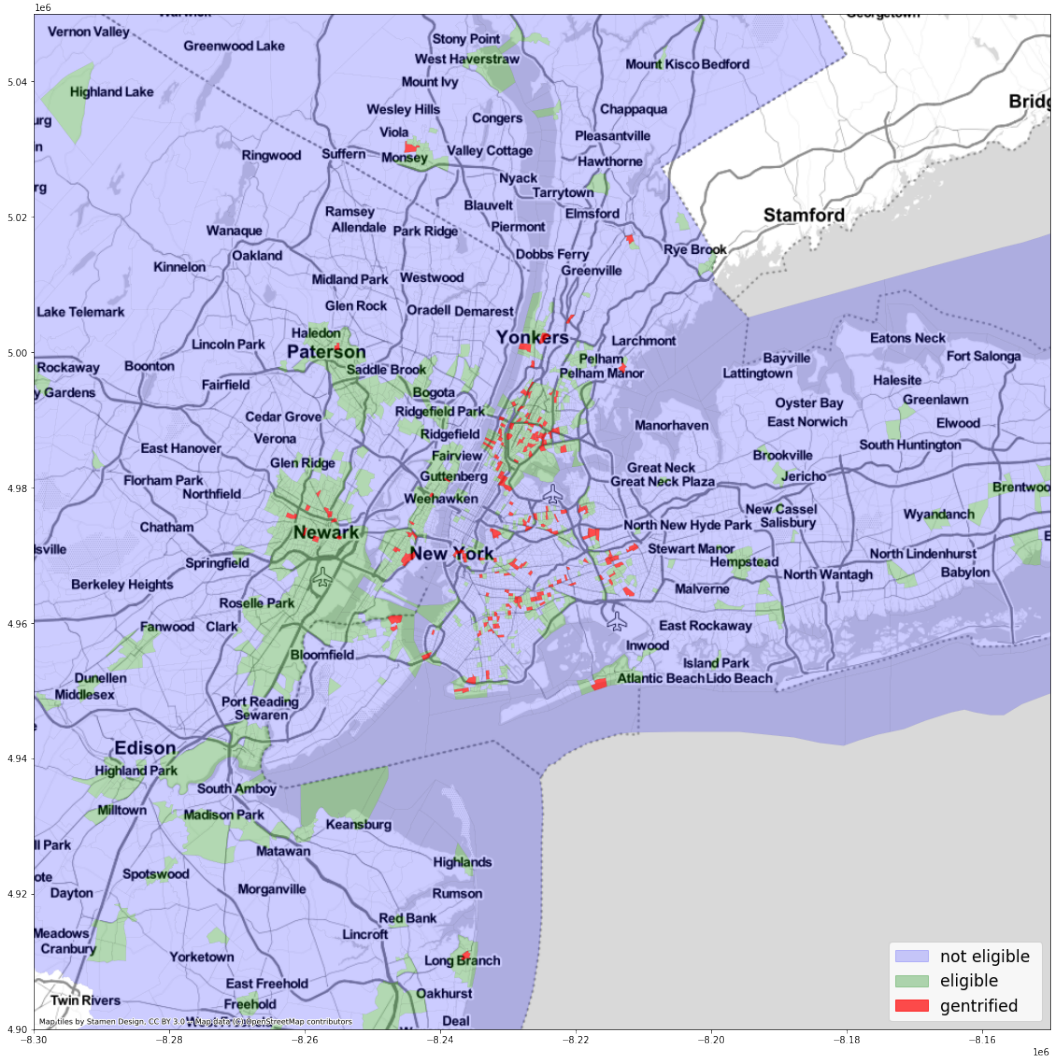

Figure 2: The NYC area map showing eligible and gentrified tracts in the year 2018.

In order to verify this, we computed the ratio of gentrified tracts among eligible tracts in each county. The results were as expected from the geographic visualization: Gentrification is present in some counties way more than in others (Figure 3).

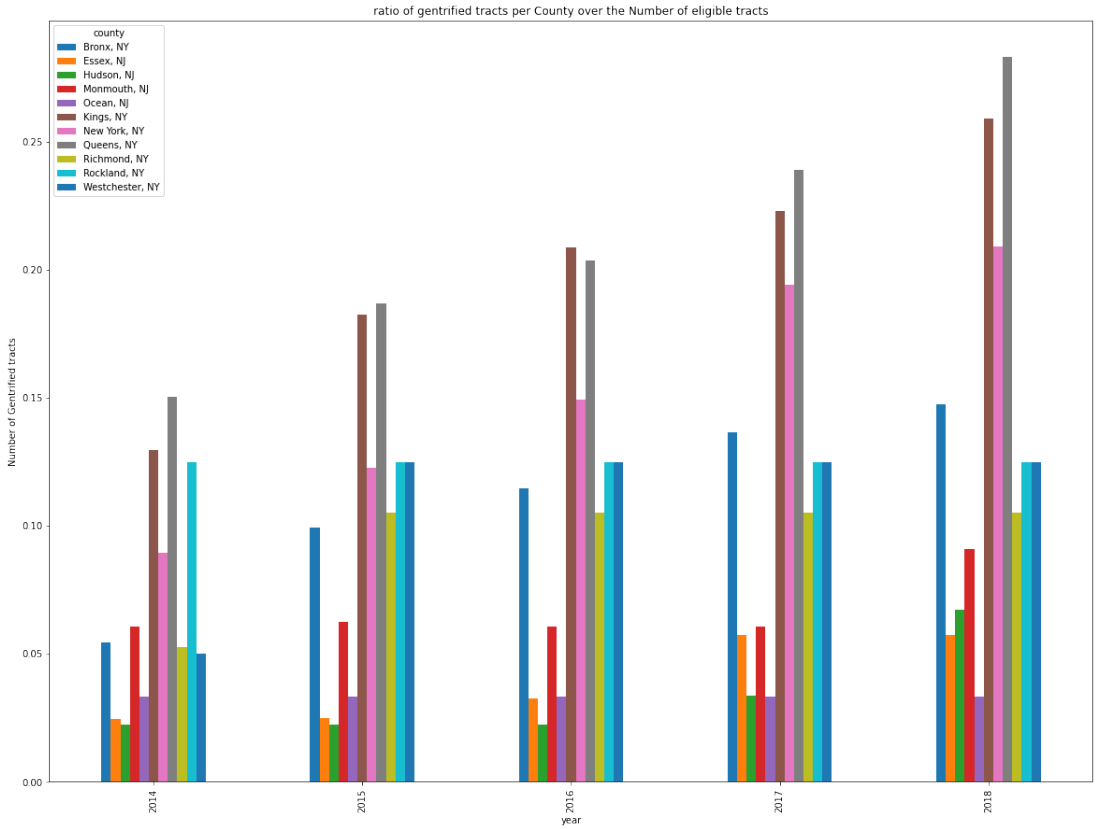

Figure 3: Ratio of gentrified tracts to eligible tracts in each county over the 2009-2018 period

3.3 Gentrification and income imbalance

One of the main issues we aimed at investigating was how gentrification is linked to social injustice, by asking what effect gentrification has on different ethnic groups. One aspect of this is its impact on household incomes. The available census data provides us with the median income of each tract. In Figure 4 we observe that on average, gentrified tracts experience a faster growth in median income than others. While an increase is expected - gentrification is the displacement of low income inhabitants - we emphasize that an increase in median income was not part of our gentrification endpoint criterion, so it can be used as an indicator on its own for predicting gentrification.

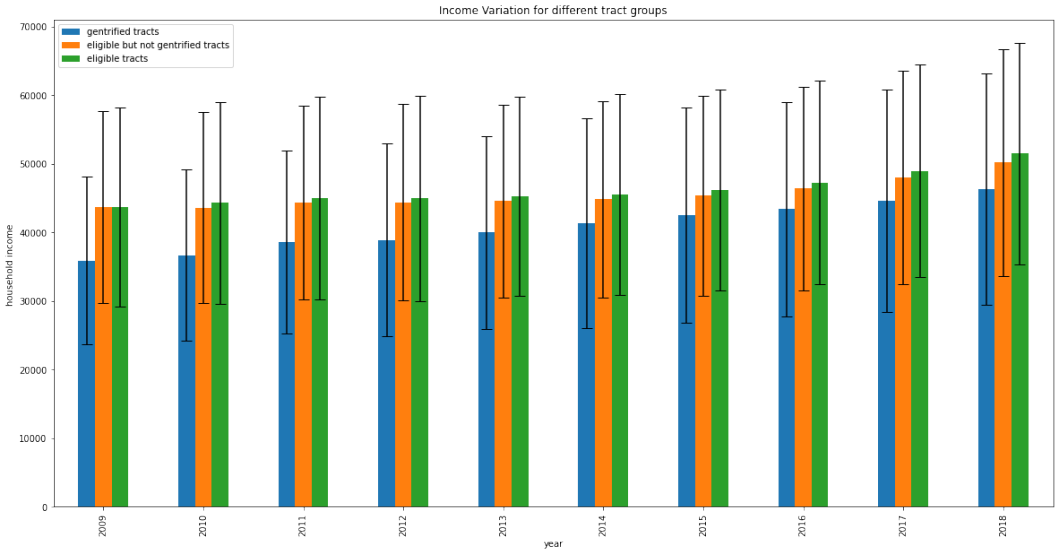

Figure 4: Mean household income over the 2009-2018 period for gentrified and not gentrified tracts.

To answer our question we were interested in how this increase breaks down to the different ethnic groups living in a tract. Are all affected the same? To verify this, we need the household incomes of the distinct groups and investigate whether some minorities might be disadvantaged and do not experience the general upwards trend. We do not have access to such data, but by the means of linear regression we can estimate the median income of each ethnic group in eligible and gentrified tracts. The procedure is described in 3.3.1. Our results in Figure 5 show that:

- **Caucasians** seem to be the group that experiences the largest increase in income
- **African Americans** and **Latinos** do not experience any income increase.
- **Asians** seem not to be affected by gentrification, as their median income is stable throughout the years.

The differences in income increase per ethnic group is very different in tracts that underwent gentrification from those that did not. The income imbalance in gentrified tracts greatly increased, driven by an increase in median income of Caucasians, while other ethnic groups experienced stagnation.

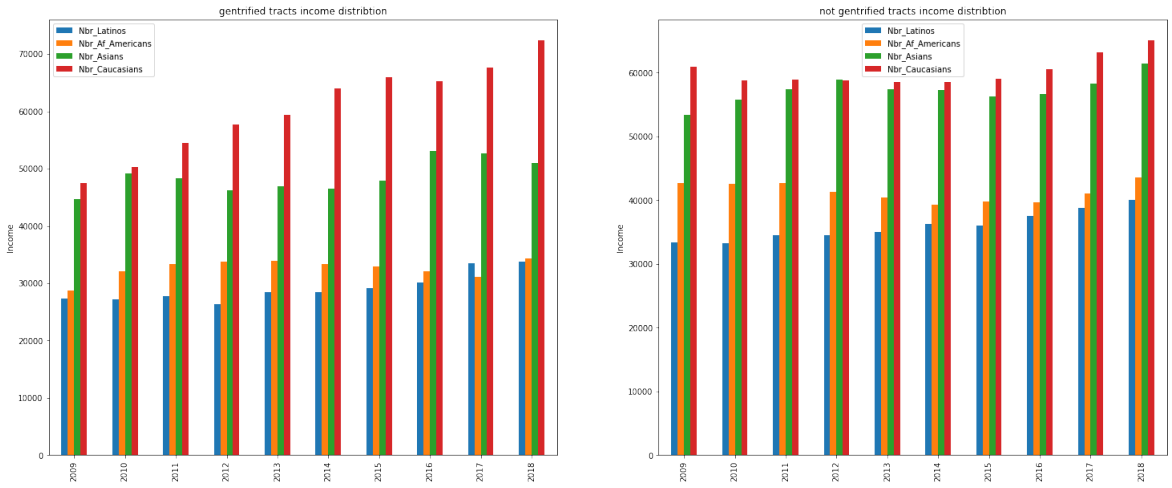

Figure 5: Relative variation of Income among ethnic groups over the 2009-2018 time period

Our results show that the median household income per ethnic group can be a good indicator of gentrification. We therefore decided to use the median household incomes of the major ethnic groups (Caucasians, African Americans, Latinos and Asians) as well as the trends of these incomes over the first 5 years (up to 2013) as features for prediction.

3.3.1 Non Negative Least Squares for income per ethnic group

In order to estimate the median household income per ethnic group we apply linear regression, with features X_i being the ratios of each ethnic group in a tract and the target being the median household income. Since some ethnic groups are always present at a very low rate, we group those together as "other" and we drop this feature as it is perfectly collinear with the sum of all the others (sum of frequencies = 1)

The resulting regression models the median household income as a function of the ratios of each ethnicity in a tract: $\mathbf{X}\beta = \mathbf{Income}$.

The model is fitted without an intercept, and this results in the coefficients of the regressor corresponding to the median household income of the respective ethnic group (to get the estimate, we set the ratio of the group in question to 1 and all the others to 0). However, in order for our coefficients to be meaningful, we constrain our coefficients to be positive in the linear regression optimisation, which yields the non negative least squares (nnls) optimization target:

$$\arg\min_{\beta} \|\mathbf{X}\beta - \mathbf{Income}\|_2 \text{ subject to } \beta \geq 0 \tag{1}$$

3.4 Change of population compositions

3.4.1 Impact on demography of tracts

The population of different ethnic groups and their evolution from 2009 to 2018 are visualized in Figure 6.

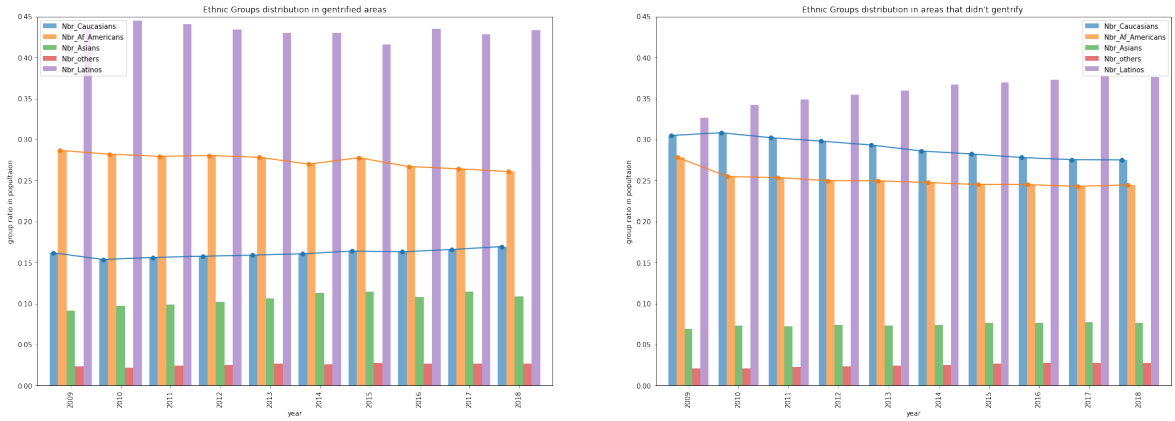

Figure 6: The change of population composition in gentrified tracts (left) and eligible-but-not-gentrified tracts (right) from 2009 to 2018 in the metropolitan statistical area. The ethnic groups shown in the the legend from top to bottom are Nbr_Caucasians (i.e. White), Nbr_Af_Americans, Nbr_Asians, Nbr_others, Nbr_multi_race and Nbr_Latinos respectively.

From Figure 6 we can observe the difference in demographic trends of gentrified and non-gentrified tracts:

- Tracts that gentrify on average have a lower ratio of **Caucasians** and a higher ratio of **Latinos** This indicates that the initial ethnic composition of a tract has an effect on possible gentrification.
- The ratio of **Caucasians** increases steadily over the years in gentrified tracts.
- In tracts that are eligible but do not gentrify, the ratio of **Caucasians** decreases steadily. It is not clear what causes this trend. One hypothesis would be that the displaced population of gentrified tracts moves to other eligible tracts, but this cannot be inferred from the available data.

Our analysis provides us with a more detailed understanding of how gentrification affects the demographics of tracts and indicates how both the ratios and the trends over time can be used as an early indicator of an ongoing gentrification process.

3.4.2 Entropy change of the ethnic groups

When comparing pie graphs of the ethnic composition in selected gentrified and eligible-but-not-gentrified tracts, we observed an increase in ethnic diversity in gentrified tracts. We sought to condense this qualitative observation into a quantitative measure. The entropy of the ethnic group distribution (Equation 2) is such a measure that is already established in literature [3].

$$Entropy = -\sum_{i=0}^{n} P(x_i) * log(P(X_i)) \tag{2}$$

$P(x_i)$ represents the population ratio of each ethnic group. The more ethnically diverse a tract is, the higher is its entropy.
The change in entropy over time could be used as a feature to predict gentrification. We investigated the evolution of entropy in gentrified, eligible-but-not-gentrified and not eligible tracts from 2009 to 2018 (Figure 7). We find that:

- The entropy of gentrified and eligible-but-not-gentrified tracts started from a similar value in 2009, but the gentrified tracts had a steeper increase of the entropy in the next following years.
- The gentrified tracts had a drop of the entropy in 2017 which could imply a suspended phase of gentrification.
- The not-eligible tracts started from a very low entropy and continuously had smaller value of entropy compared with eligible tracts.

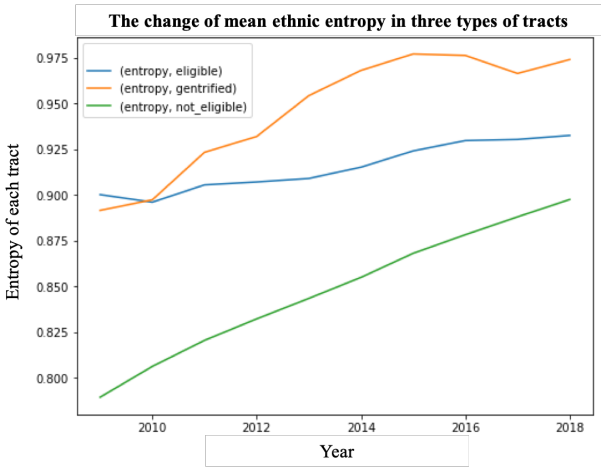

Figure 7: The change of mean ethnic group entropy in gentrified, eligible-but-not-gentrified and not-eligible tracts.

3.5 EMS Incident Dispatch Data

The 311 call data gives an excellent idea of the small and common events that happen all around New York. Unfortunately, it does not any shed light on the more serious crime that gets reported. Even worse, 911 data is not publicly available for New York. To get a better sense of this, it is a good idea to look into the emergency medical services incident dispatch data, which is public. NYC OpenData provides a complete collection of all EMS dispatches in the the City of New York starting in 2005 and being continuously updated. This data set can then be filtered for crime-related dispatch codes, such as STAB, DRUG, PD-13 (officer back-up), SHOT, etc. The original data set has a total of 22 million reports, of which 1.5 million (7%) are related to cases of severe violent crime.

As this data is aggregated and anonymised as much as possible for privacy reasons, it does not go any further than zipcode information. As one zipcode contain multiple census tracts, the data can best be analysed in the following two ways; by looking at the average gentrification rate of census tracts within the zipcode, and by analyzing individual census tracts based on their property of being in a zipcode with certain crime-related emergency data.

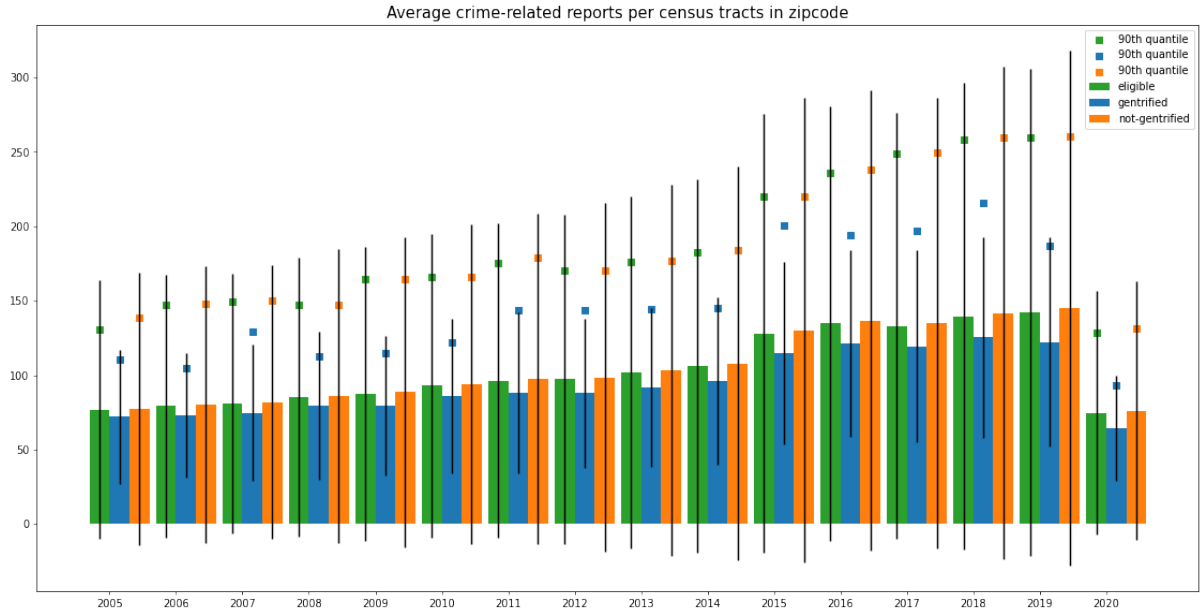

Figure 8: Crime related reports in gentrified and not-gentrified tracts: 2005-2020.

Starting with the latter, we take a closer look at figure 8. This data is formed by giving each tract the property of the average census tract with which it shares zip code. Then yearly average crime related dispatches are added up and plotted. Here, three things can be observed in these graphs.

Firstly, the averages seem to show little initial different pattern between those tracts that gentrification and those that did gentrify. It is highly likely that this is, among other things, due to the effect of the overall averaging that is caused by the fact that we can only see the average frequency of the zip code in which a certain tract exists. An interesting pattern that does appear is the rising difference in average crime-related reports between gentrified and non-gentrified tracts. This implies that gentrification does seem to have a reducing effect on severe crime in an area. An interesting question here would be whether this reduces overall crime, or just moves it. As shown earlier the low-income population tends

to be pushed out of gentrifying neighbourhoods and will possibly take these higher crime rates with them.

Secondly, there is a significant difference in the tail behaviour of the gentrified tracts compared to the non-gentrified. This can be seen in the quantiles plotted (squares) as well as the larger standard deviations (vertical line). To ensure that these quantiles are not too heavily influenced by the asymptotic bias of quantile estimation, it is important to not pick too high of a quantile. Therefore the 90 percent quantile is taken. Here it clear that not gentrified tracts have much longer tails than the gentrified one. It is possible that this is because of the small amount of zip codes where the tracts are largely unanimous. This is naturally bound to occur as our map shows the clustered property of gentrified area. Unfortunately, due to the high number of tracts combined with the large size of a zip code, these events are rare compared to the cases were zip codes are evenly represented, rendering this type of analysis not very helpful.

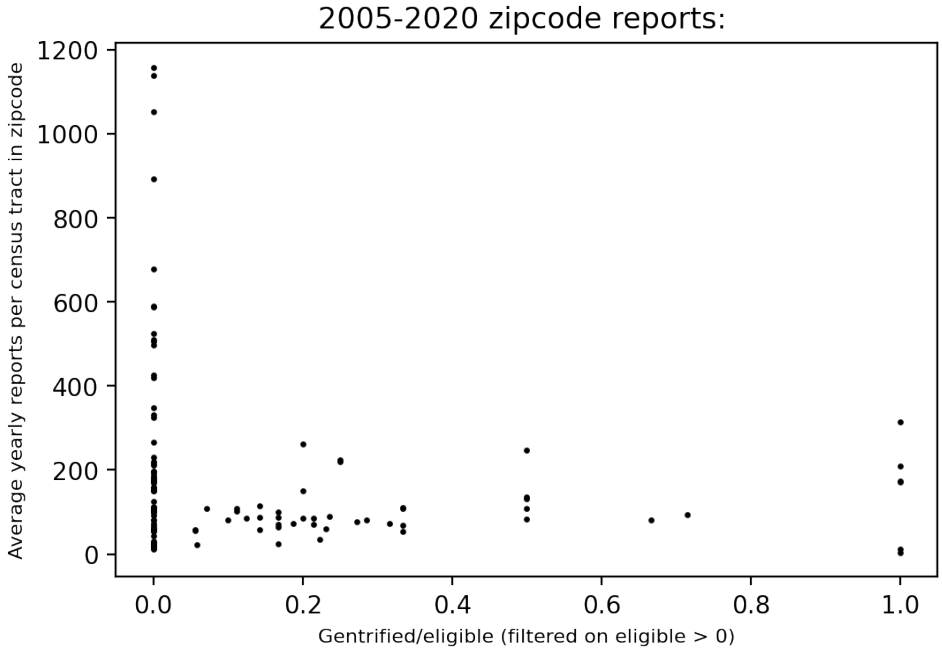

Figure 9: Average yearly reports per census tracts in zipcode over gentrificaion rate: 2005-2020

Another way of looking at the data is by looking at the average gentrification rate of census tracts within the zip code. In this case, the size of the zip codes is much less of a problem than before. As only the ratio of gentrified and eligible for gentrification is observed. This is plotted in figure 9. The report numbers were divided by the 16 years and the amount of tracts it contains. This plot seems to immediately confirm our previous suspicion about the rare occurrence of a zip code with a majority of tracts gentrifying. As the vast majority of zip codes with any gentrified tracts, still have relatively low gentrification rates overall. Although the graph seems to show little predictive power when report numbers are low, it is clear that it is very unlikely that there will be any tracts gentrifying when report numbers are high. To generate the graph, only zip codes with eligible tracts were analyzed. This seems to confirm the suspicion that the severe crime incidents, as detected with EMS dispatch data, has a negative impact on the likely-hood of an area gentrifying.

3.6 311 Data

As the task we are presented with is to analyze gentrification on the level of census tracts, we use the geographic coordinate information of each call to assign it to the tract in which it occurred. About 10% of all calls in a given year are lost in the procedure due to missing geographic information.

The 311 call data set consists of calls that are classified into 232 distinct categories. This categorization leads to a massive sparsity problem in the data: For most tracts, most call classes will have a count of 0 in most years, making comparison of temporal trends impossible. We therefore decided not to investigate timescales lower than a full year.

To further address the sparsity issue, we manually reclassified the calls, reducing the number of categories from 232 to 24. This resulted in a decrease of 0-counts in the data set from 82 % to 34 %. As these properties of the data still did not allow to investigate temporal trends - yearly counts on a per-tract level are stochastically 0 or above 0 over time, we examined whether the number of calls in a given year is related to the gentrification status that we inferred from the census data. To obtain a model-free estimate of this, we compute the mutual information between the gentrification label and the call category of interest. We do not discover any dependency of the investigated variables, the mutual information of all categories is close to 0 Figure 10.

We conclude from our analysis that the amount of available 311 call data on the single tract level is too low to allow the deduction of information on the tract from it.

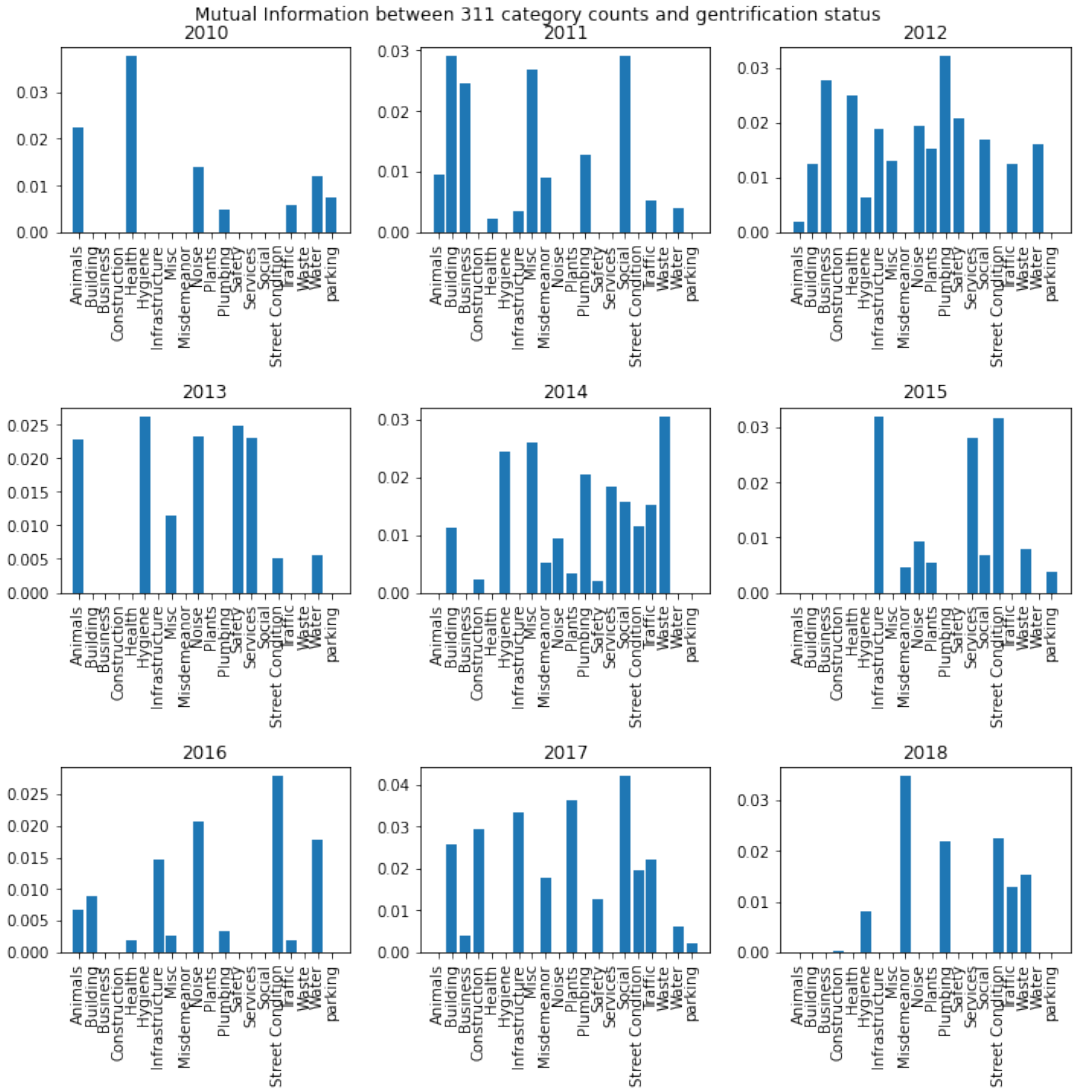

Figure 10: Mutual Information between 311 call counts per tract and gentrification status for all years. The counts were normalized by tract population size before analysis.

4 Predicting Gentrification

Our exploratory analysis revealed several differences between the tracts that undergo gentrification and tracts that are eligible but do not. These different changes can already occur in an early stage of the gentrification process and can serve as an indicator that gentrification will happen. We aimed at predicting gentrification from a five year horizon from a tract becoming officially considered as gentrified. The availability of an early prediction of whether a tract will gentrify in the coming years or not could help local governments in making targeted policies to help the low-income citizens that are at risk of becoming displaced and address the needs of the different ethnic groups that are affected by gentrification. Our objective thus was to **assess if we can predict gentrification over a five years horizon.**

4.1 Feature selection

We gathered all the features that were found to be meaningful in the exploratory analysis and computed their correlation to the gentrification status. Again, we only work with **eligible tracts** as the negative

set. The correlation matrix of the features and the labels are shown in Figure 11. Our features can be categorized in three types :

- Raw values : these are mid_Income per tract, mid_Income for each ethnic group, ethnic groups population ratios, and race entropy
- Trends : We use the trends of median incomes and ethnic groups ratios. Trends are computed in two ways :
  - Slope of a linear regression over the first 5 years
  - Trend=$\frac{\text{Average(2 Years)}}{\text{Average(5 Years)}} - 1$

    We kept the slope measure as it yielded better results, but both yield the same information.
- categorical : we kept counties (identified by the concatenation of county code and state code) as an indicator of the location, we hot encoded it since it's categorical .

from correlation with gentrification we can see that :

- Median income value has a significant negative correlation of -0.21 while its trend has a high positive correlation of 0.17.
- Asian ethnicity ratio trends has a positive correlation of 0.17.
- White group ratio and their income both have a negative correlation of -0.15 to the label, but trends on these two are positively correlated to gentrification.

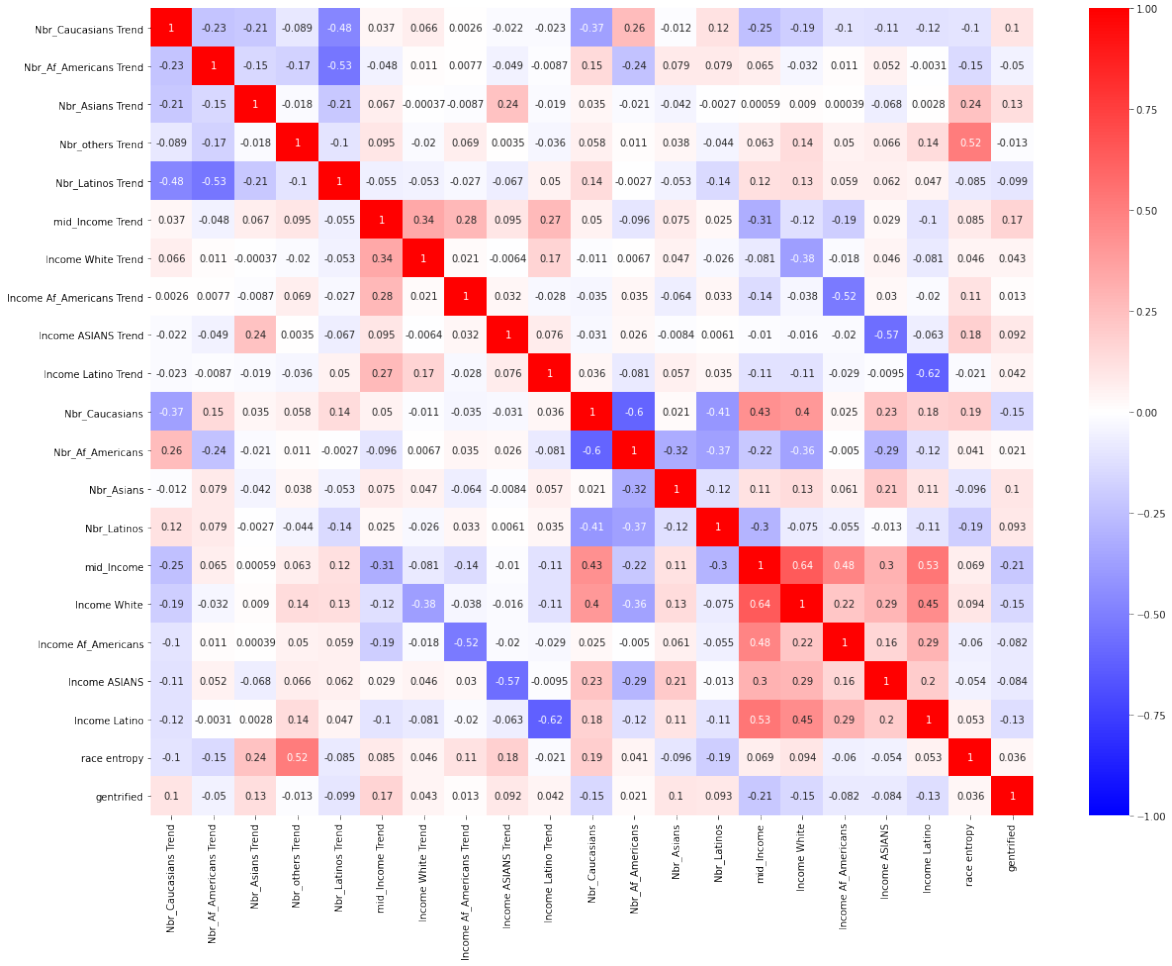

Figure 11: The correlation matrix of features derived from census data and gentrification. The census features include the population ratio of each ethnic group and the change of its ratio named as "trend".

4.2 Prediction model

Once we gathered all of our features, we tested 3 different models : Support Vector Machines, Logistic Regression and xgboost. We only report the logistic regression model, as it seems to perform well (compared to the other two) while keeping things simple.

The logistic regression is regularized with an **l2** penalty and 5-fold cross validation is used to determine the adequate regularization coefficient. The data is separated to validation and train sets and we also have an independent testing dataset (20% of initial data) that wasn't used in any of the analysis.

The logistic regression scores over all the datasets are reported in Table 1. The recall is relative high which means 85% of gentrified tracts are found by this model. The precision is relative low representing 31% of the predicted gentrified tracts are real gentrified tracts.

Table 1: Logistic regression model metrics for training, cross-validation and testing dataset.

Training dataset	precision	recall	f1-score	support
0	0.96	0.72	0.82	665
1	**0.35**	**0.85**	**0.49**	118
accuracy			0.74	783
macro avg	0.66	0.78	0.66	783
weighted avg	0.87	0.74	0.77	783
Validation dataset	precision	recall	f1-score	support
0	0.93	0.72	0.81	71
1	**0.38**	**0.75**	**0.50**	16
accuracy			0.72	87
macro avg	0.65	0.73	0.65	87
weighted avg	0.83	0.72	0.75	97
Testing dataset	precision	recall	f1-score	support
0	0.96	0.67	0.79	189
1	**0.31**	**0.85**	**0.46**	33
accuracy			0.70	222
macro avg	0.64	0.76	0.62	222
weighted avg	0.87	0.70	0.74	222

We also plotted the feature importance as learned by our model of the top 16 important features for

predicting gentrification (see Figure 12). We can see that location and income trends play a major role in gentrification. The model seems to have learned the same insights that we explained earlier.

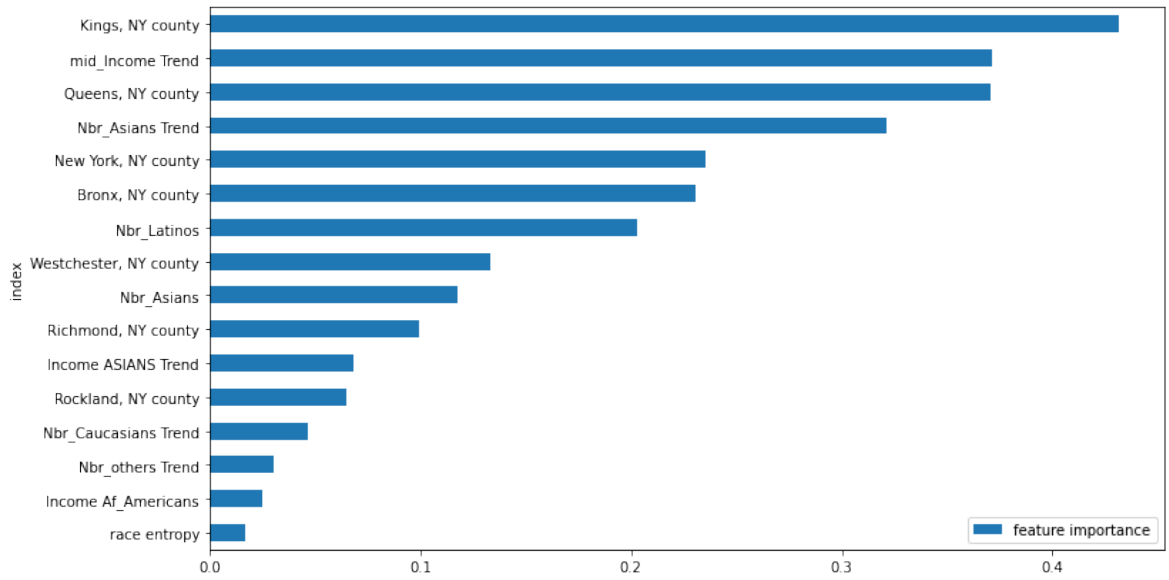

Figure 12: The top 16 important features obtained from the logistic regression model

Overall, the performance of the model is not as high as we hoped it would be, but it can still give us a decent idea on the gentrification process. Gentrification is somewhat rare to occur, and the study can be supplemented with other sources of data so that this phenomena can be predicted in a more precise manner. We tried to focus more on the impact of gentrification on the socioeconomic environment and give a view on what we can do to avoid its unwanted effects.

5 Conclusions

From our analysis the following conclusions can be drawn. Gentrification seems to be heavily influenced by its location. Both to the center as well as with respect to other gentrified areas. Due to the higher incomes of the Caucasians that are able to enter this now more expensive area, a larger discrepancy among population ethnicities arises. Another important differentiator of gentrifying areas, is the lower rates of violent crime that are reported. Using these discovered patterns, we were able to design a model that can predict gentrification at early stages from readily available data. The model can help local governments gain a better understanding of how gentrification happens. This allows them to address the issue appropriately by, for example, supporting those demographic groups most heavily affected.

References

[1] Michael Maciag. Gentrification report methodology. 2015.

[2] Consumer price index data from 1913 to 2020. `https://www.usinflationcalculator.com/inflation/consumer-price-index-and-annual-percent-changes-from-1913-to-2008/`. Accessed: 2020-10-25.

[3] Tom Carter. An introduction to information theory and entropy. 2014.