# The Gentrification Process in the New York metropolitan area

Citadel Open Datathon

Team 23
Yan Le, Yuhao Sun, Ricky Choi, Stanley Cheung

26 October, 2020

## 1 Executive Summary

Gentrification is defined as the transformation of a working-class or vacant area of the central city to a middle class residential and/or commercial use. Gentrification is a contemporary issue plaguing communities across the country, a phenomenon identified by the reconstruction and renovation of older, deteriorating neighborhoods into wealthy one. Gentrification process is controversial topic because it brings many benefits and drawbacks to the community at the same time. Gentrified tracts can benefit from investments in more affluent residents and corporations, in terms of increased property value. But gentrification also drives up the rent of properties and reduces affordable housing to original residents, resulting in displacement of original residents.

The aim of this study is to define the factors affect the gentrification process. First, we identified over 500 tracts gentrified from 2011 to 2018 in the NY-NJ-PA MSA compared to the based year of 2010. Based on the hypothesis, we identified certain features correlated with gentrification which includes census dataset in the New York county, 311 complaints and education attainment. It was found that most of these tracts are located around Hudson River between New Jersey state and New York state. Gentrification were much rarer in other counties. We also showed that increase in median home value is a better indicator for the criteria than increase in median household income.

In this research, We quantify the change of each identified feature by looking at the average percentage change of the feature over time. Specifically, we calculated the average percentage change of each identified feature relative to the last year, and averaging overall of the years before gentrification. A percentage change allows us to see the general trend of each feature. This percentage change is simplest to implement, but may prose some problems as the baseline for the percentage change calculation changes throughout the years

After developing random forest model to predict gentrification using analyzed dataset, it was found that the predictions of the years of gentrified tracts with excellent accuracy.

The optimized model performance reached MSE score of 0.28. The top 2 features are expected, the median household income and home value, as they are somehow inherit in the definition of gentrification. Additionally, the number of Caucasians, number of African American, education attainment and variety category of calls were highly correlated with the predictions. The results reinforced our notion that the census dataset and 311 calls have strong relationship for predicting the year of gentrification.

# 2 Background and Research Question

## 2.1 Motivation

Gentrification is a process is a process that creates both quantitative and qualitative changes in neighborhood character. there are measurable changes in a neighborhood such as increases in average household income, the increased presence of college-educated residents, and increases in housing values and rents that census data is capable of capturing. Gentrification also causes changes in such things as the local retail mix that census data does not capture and for which good quantitative measures either do not exist or require some form of fieldwork that makes comparisons across cities and, more importantly, at different points in time extremely difficult.

The census tracts of New York-Newark-Jersey City, NY-NJ-PA Metropolitan Statistical Area (MSA) have undergone significant change throughout the past decade. With a large number of tracts being gentrified on a yearly basis, this topic is a common discussion in social studies. Ranging from the positive, negative externalities and ethical justification of gentrification, to the socioeconomic impact it has on surrounding areas. From the investors' point of view, it is hard to predict how attractive the gentrification projects might be, due to the complexity coming with identifying and quantifying the factors affecting the success of these projects.

## 2.2 Research Question

In this caes, we propose the following question in analyzing the datasets:

- Topic Question: How can we identify and quantify the change that tracts undergo during the gentrification process?

In this study, we will answer the topic question by attempting to quantify tracts' socioeconomic environment through analysis of the 311 complaints and government census datasets. The gentrification from a quantitative direction was analyzed in both cross-sectional and historical approach.

# 3 Methodologies on Gentrification

For this analysis, we analyzed several given datasets and external datasets downed from census data.

1. Census (given) We used census data for year from 2010 to 2018 for each tract in the New York-Newark-Jersey City, NY-NJ-PA Metropolitan Statistical Area

2. 311 calls (given) We used 311 complaints" from New York City's open data portal from 2010 - 2018.

3. Education attainment (external)

4. Geographic data of census tracts from US Census data (external)

Different quantitative approaches were applied to identify to gentrified tracts. There is no universal definition on gentrification. Researchers use different methodologies to determine if a tract was gentrified over a time period. The criteria set in this research is a widely acceptable methodology from Govening Report [1]. Firstly, Year 2010 was determined as based year and the tracts in 2010 eligible for gentrification were identified by following criteria:

1. Have a population of at least 500 residents throughout the time period.

2. Have their median household income and median home value below the 40th percentile of all tracts within the NY-NJ-PA MSA at the beginning of the decade.

The rationale of this process is to identify tracts that were less affluent than other tracts to ensure that the rise in home value is caused by gentrification, not by other economic activities. Those gentrification-eligible tracts were then determined to be gentrified over a time period if they met the following requirements:

1. The tract's median home value increased when adjusted for inflation,

2. the percentage increase in the median home value of the tract was in the top third percentile of all tracts within the NY-NJ-PA MSA.

However, we did not require the educational attainment to increase for gentrified tracts. Instead, we will look at their correlation. We note that some researcher used the criteria that the tract is gentrified when its increase in median home value is larger than the increase in the median of median home value of all tracts in a MSA [2] by 50%. However, such threshold has limited applicability in our dataset. Since the average median value of all tracts between 2011 and 2017 was lower than that in 2010, this definition became insensible. The definition given by Governing Magazine, however, works in any time period, even if the median of median home value of all tracts decreases.

Some researchers also proposed to measure the increase in the median household income of the tracts instead of median home values. However, [1] also argued that many of the first residents to "gentrify" a neighborhood are often young professionals whose income might not be much higher than their new neighbors'. Such changes to a community are not reflected in income levels. We showed in the visualisation that median household income is less correlated to gentrification.

# 4 ET Processing of datasets

## 4.1 Data Extraction

This report uses the census data and 311 complaints as the input to identify their relationship (needs clarifying). A majority of these datasets were provided in the Datathon Materials. However, we also imported census data on educational attainment from the Census Department [3]. Additonal dataset of the geographic locations of the tracts were also used (provided by the Census department as well).

## 4.2 Data cleaning

Before conducting data transformation on the features of gentrified tracts, we needed to assess data quality and completeness and remove out-of-range and blank data fields. They can impede the accuracy of statistics on household income and home value of tracts, which are the threshold on the eligibility of tracts being gentrified.

For example, the median home value and household income value of some tracts are either missing or negative. These values are clearly out of range so these tracts were removed in advance of our data analysis. More importantly, the tracts in 2009 are totally different from those between 2010 and 2019 because the US Government updated its census tracts every decade [4]. Hence we discarded the census data in 2009 to maintain consistency of tracts, and we chose to use 2010 as our baseline year (since it is the earliest year with clean and operable data) and 2018 as our final year (since it was the most recent year with data).

For the 311 complaints dataset, we first removed any calls that had no record of address or coordinates, as these calls have no geographically identifiable features, thus making it impossible to identify which tract it belongs to.

## 4.3 Data Transformation

The locations of 311 complaints were given in different forms, such as X and Y Coordinates, latitude and longitude. To combine the number of complaints with the census data on each tract, we need to transform the locations of 311 complaints into the form of Geographical Identifiers.

This is done using the `geopandas` library in Python. Each call had either a coordinate pair in longitude and latitude, or an address associated with it. If a call had either, the coordinate pair is taken as the identifier as coordinate lookup is a lot faster than address lookup. For the coordinate lookup, we used an addition dataset from the US Census with the geographic boundaries for each tract in both the NYC and Newark regions [3]. This dataset contains the information of the specific shapes of each tract, and this was loaded up in `geopandas` and turned into polygons. Using the spatial join feature of `geopandas`, we were able to combine the call dataset with the tract shape dataset and associate each call with a tract. Using the in built spatial join heavily decreased the runtime for the data transformation. `geopandas` For the address lookup, we used the `censusgeocode` package to lookup census tracts from the address. We used a batch address lookup to speed up

the process. We were then able to group the calls by department per tract by using the labels in the 311 dataset.

## 4.4 Identifying Gentrified Tracts

These criteria can be implemented easily in Python. We first imported these files as `pandas Dataframes`. The data was then sorted by its geographical identifier to ensure that the indices of each row matched each other across the time period. The gentrified tracts were obtained simply as the intersection between the set of eligible tracts and the set of tracts with percentage growth in its median home value in the top third percentile.

To assess how gentrified tracts has changed since 2010 to 2018, the number of gentrified tracts, the number of did not gentrify, the number of not Eligible to Gentrify and etc were analysed. Table 1 summarizes the extent of the gentrified tracts in New York–Newark–Jersey City from 2011 to 2018. The lower-income Census tracts experienced a slightly growth from 2011 to 2014 and then followed by a significantly decreases in 2023. The reason may be because non-gentrified tracts have a higher chance of receiving investments than gentrified tracts. the home-valued of non-gentrify tract increased quicker than the gentrified tract which leads to a drop the number of gentrified tracts. We should also note that gentrification is irreversible. As gentrification involves in improvements in housing, its reverse are unlikely to happen in real life. Consider that housing estate businesses are keen on replacing old buildings with new ones, but not the other way around. Therefore, tracts gentrified in the early stage may appear de-gentrified recently, simply because their growth in median home values slows down and are no longer in the top third of all tracts.

| Year | Gentrified Tract | Did Not Gentrify | Not Eligible to Gentrify | Total Census Tracts | Share of Gentrified Tracts (%) | Share of overall Gentrified Tracts (%) |
|---|---|---|---|---|---|---|
| 2011 | 300 | 558 | 4200 | 4500 | 53.76 | 6.67 |
| 2012 | 299 | 559 | 4204 | 4503 | 53.49 | 6.64 |
| 2013 | 318 | 540 | 4160 | 4478 | 58.89 | 7.10 |
| 2014 | 314 | 544 | 4167 | 4481 | 57.72 | 7.01 |
| 2015 | 223 | 635 | 4184 | 4407 | 35.12 | 5.06 |
| 2016 | 220 | 638 | 4195 | 4415 | 34.48 | 4.98 |
| 2017 | 221 | 637 | 4217 | 4438 | 34.69 | 4.98 |
| 2018 | 211 | 647 | 4199 | 4410 | 32.61 | 4.78 |

Figure 1: A summary of the extent of the gentrified tracts in New York–Newark–Jersey City from 2011 to 2018

### 4.5 Machine Learning Model

To implement the machine learning model to learn and quantify changes in features, the correlation matrices among the features of original census data, educational attainment and New Yorkers' 311 complaints were identified. From the observable pattern shown in the Figure 2., the total population is highly correlated with total population of people age over 25 years old which is 0.95. Therefore, the population in different ages are redundant features which were removed. It is found that the household income and home-valued are negatively correlated with number of non-Hispanic of Latinos. The increased number of residents of a historically Latino neighbourhood in NYC in the area may have correlation with the rate of gentrification and more in-depth analysis can be carried on. In terms of right figure, there are very little correlation between the different categories of the calls in NYC so these litter correlation data can be used for machine learning model.

In order to predict the year of gentrified tracts, random forest regression model was used to make predictions. After developing random forest model to predict gentrification using analyzed dataset, it was found that the predictions of the years of gentrified tracts with excellent accuracy.The optimized model performance reached MSE score of 0.28.

## 5 Results

### 5.1 Definition of 'change'

Our topic question aims to identify and quantify the change of the tracts when they are gentrified. We can quantify the change of each identified feature by looking at the average percentage change of the feature over time. Specifically, we calculated the average percentage change of each identified feature relative to the last year, and averaging over all of the years before gentrification. A percentage change allows us to see the general trend of each feature. This percentage change is simplest to implement, but may prose some problems as the baseline for the percentage change calculation changes throughout the years.

### 5.2 Visualisation

We use Tableau to visualise the distribution of gentrified tracts across the NJ-NW-PA MSA. Tableau does not have built-in maps for census tracts. Hence we first imported the spatial files of NJ-NY-PA MSA from the Census Bureau [5]. Tracts gentrified were then displayed on the map. As shown in Fig. 5, the gentrified tracts mainly centered around the Hudson River. This means wealth and investment have been concentrated around New York City, while other census tracts of the United States have remained impoverished.

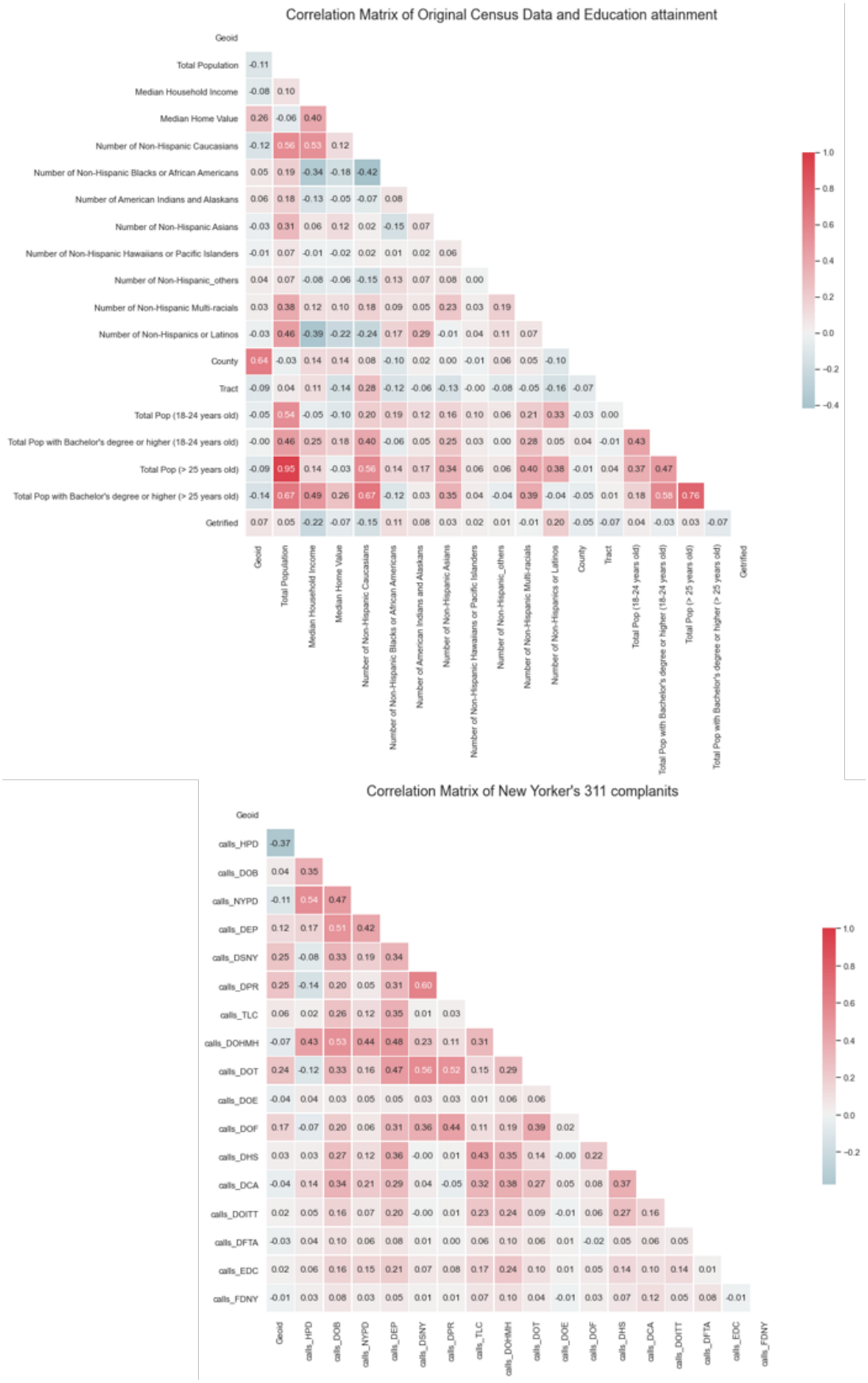


Figure 2: Correlation matrix among the features of original census data and extra dataset of the educational attainment; correlation matrix of New Yorkers' 311 complaints in different categories.

## 5.3 Key features different between gentrified tracts and other tracts

We compared the census data on three groups of tracts: eligible tracts (containing gentrified tracts and not-gentrified tracts) and ineligible tracts (tracts ineligible to gentrify). We found a few important differences in the median home value, median household income, population by ethnicity, and educational attainment.

Fig. 6. shows the distribution of median home value and household income of tracts from 2011 to 2018. The median home values of ineligible tracts have been higher than the other two groups, which is due to our requirement that gentrification-eligible tracts be the bottom 40th percentile of all tracts.

We can also see that the median home value of gentrified tracts slowly increased over the time, and consistently higher than those of tracts that did not gentrify. On the other hand, there are no observable trends in the median home values of other two groups of tracts. Hence using the median home value of gentrified tracts as a requirement allows us to separates gentrified tracts from eligible tracts.

In terms of median household income, the median household income of ineligible tracts has been higher than the other two groups as well, indicating that median household income value and median value has strong correlation.

As mentioned above, some researchers used increase in median household income of tracts as a requirement for gentrification. However, we can see from Fig. 6. that the median household income did not separate gentrified tracts from eligible tracts. This shows that median household income value performs poorer in separating gentrified tracts.

In terms of educational attainment and population by ethnicity, Fig. 7. shows the ratios of people with bachelor's degree or higher to the population in all tracts and the ratios of Latinos to the population in all tracts. The ratios of people with bachelor's degree or higher in tracts eligible to gentrify doubles that in tracts ineligible to gentrify. This shows that people living in tracts eligible to gentrify receive poorer education than people in other tracts. Also the ratios of Latinos in the total population of tracts eligible to gentrify doubles that in tracts ineligible to gentrify.

## 5.4 Insights on the gentrified tracts' trends.

Only the 10 most important features are shown in the Fig. 3. Note that the features are in terms of average change, as defined above. The top 2 features are expected, the median household income and home value, as they are somehow inherit in the definition of gentrification. Additionally, the number of non-Hispanic Caucasians, number of African American, education attainment and variety category of calls were highly correlated with the predictions. The results reinforced our notion that the census dataset and 311 calls have strong relationship for predicting the years taken to gentrify the tracts.

We have visualized the mean values for 10 most important factors. The results, seen in Fig. 4 have aided us into understanding the overall trends by analyzing mean values of each feature over the gentrification process. The main takeaways are as follows:

- The number of African Americans increase during the gentrification period, while other demographic minorities, such as Hispanic, Latinos and Asians inhibit a down-

ward trend. Therefore, it can be deduced that the gentrified tracts become less diverse and are mainly dominated by a few demographic groups.

- The number of calls to the Police Department and Department of Transport inhibit a steep downward trend, pointing out at the decrease in crime and issues with the transportation systems. Consequently, it is safe to assume that gentrified tracts have seen an improvement in receiving basic public services.

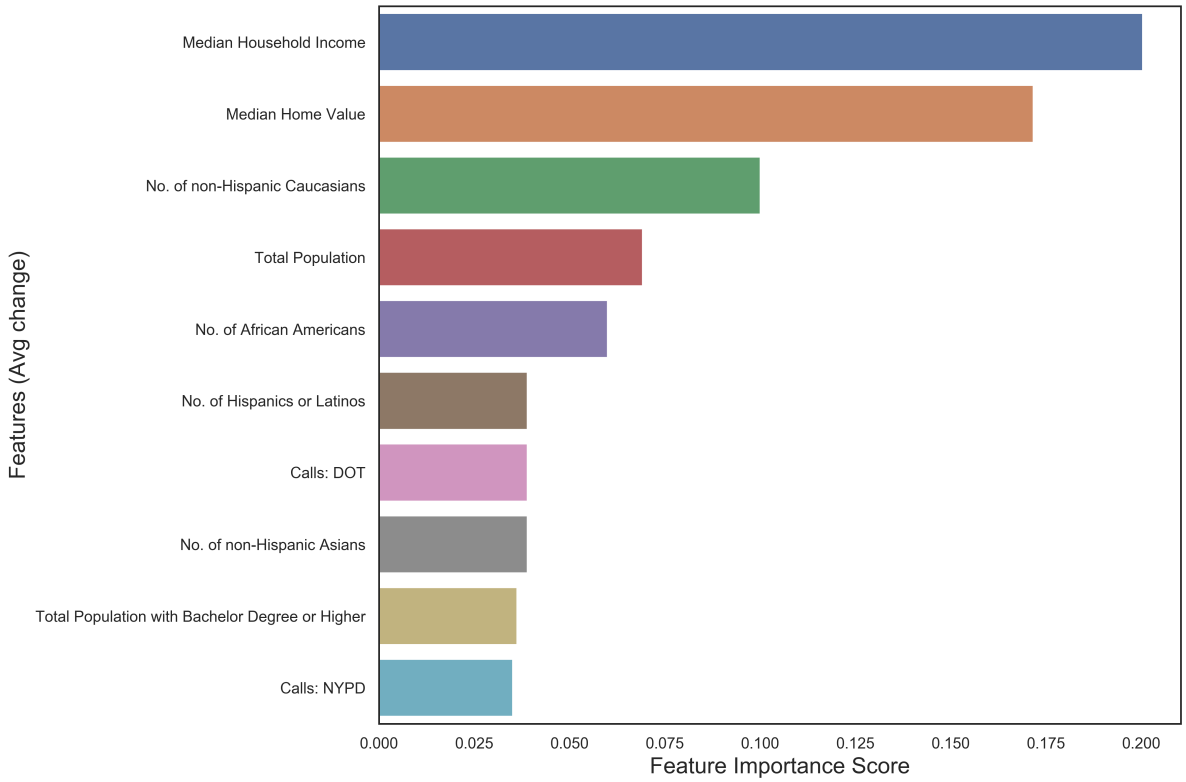


Figure 3: The feature importance score of the top 10 most important features determined by our machine learning model. Note all features are measured in terms of their average change, instead of their absolute value.

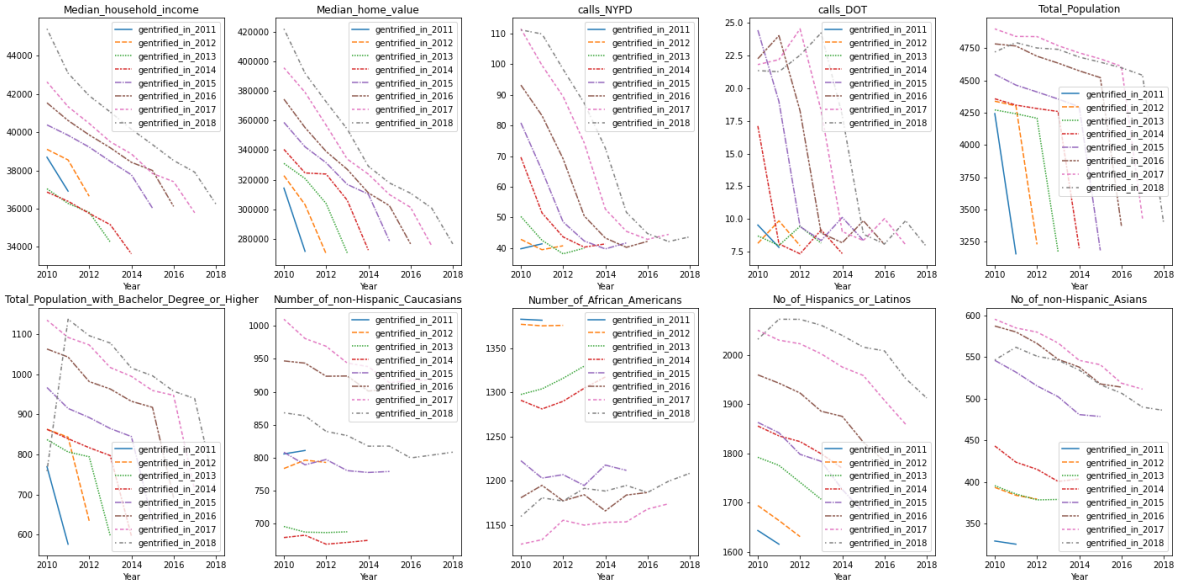


Figure 4: Mean values for each important factor for tracts that got gentrified over the 2011-2018 period. Please keep in mind that the values are averages, which are heavily affected by the outliers. Hence, the data is for trend illustrations only.

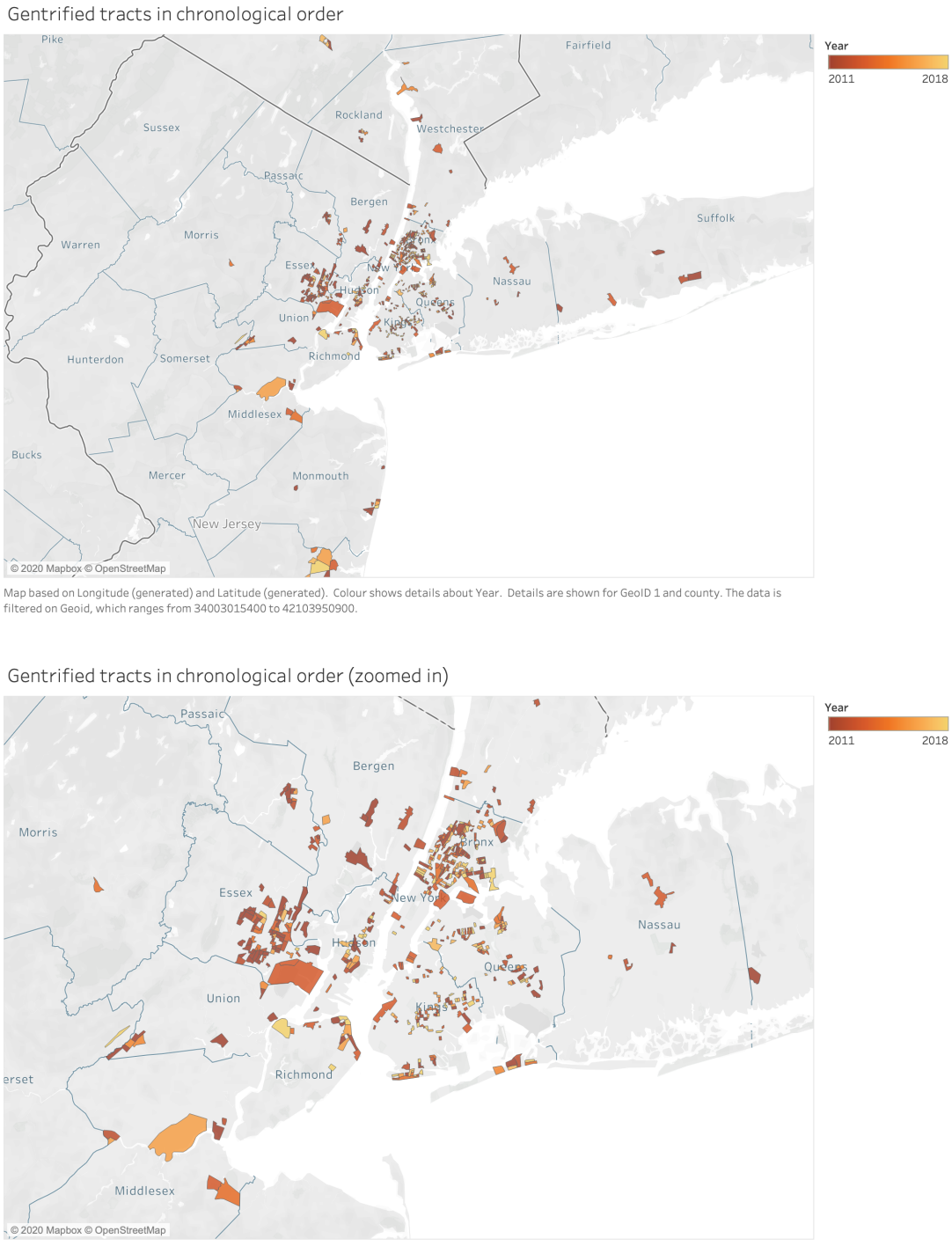


Figure 5: Gentrified tracts on the map, generated by Tableau. A few of the tracts were not included in the upper figure. These figures showed that the gentrified tracts mainly centered around Hudson River. 64% of gentrified tracts are located in 4 counties: Bronx, Essex, Queens and Kings.

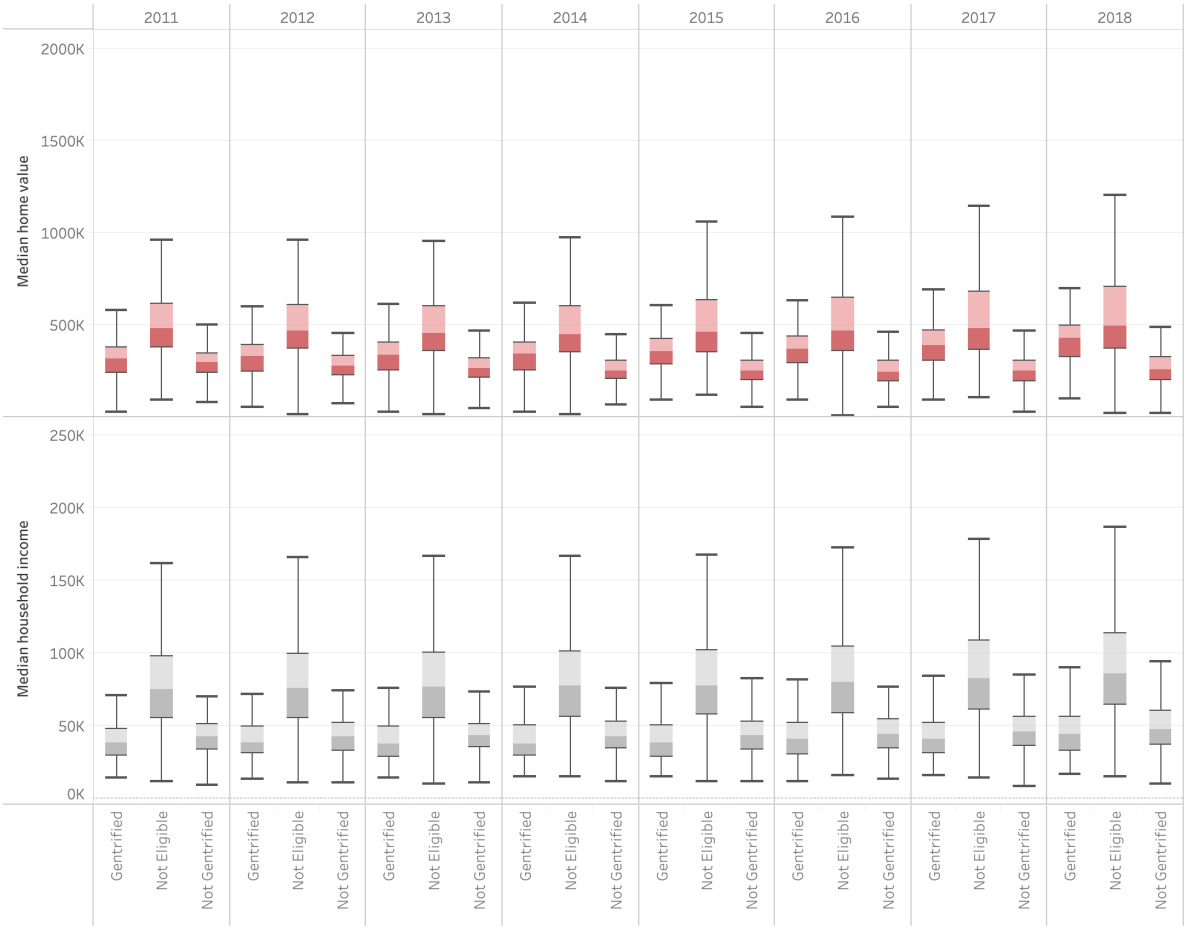


Figure 6: Distribution of median home value and median household value of tracts in NY-NJ-PA MSA. Not eligible - refers to tracts not eligible for gentrification in 2010.

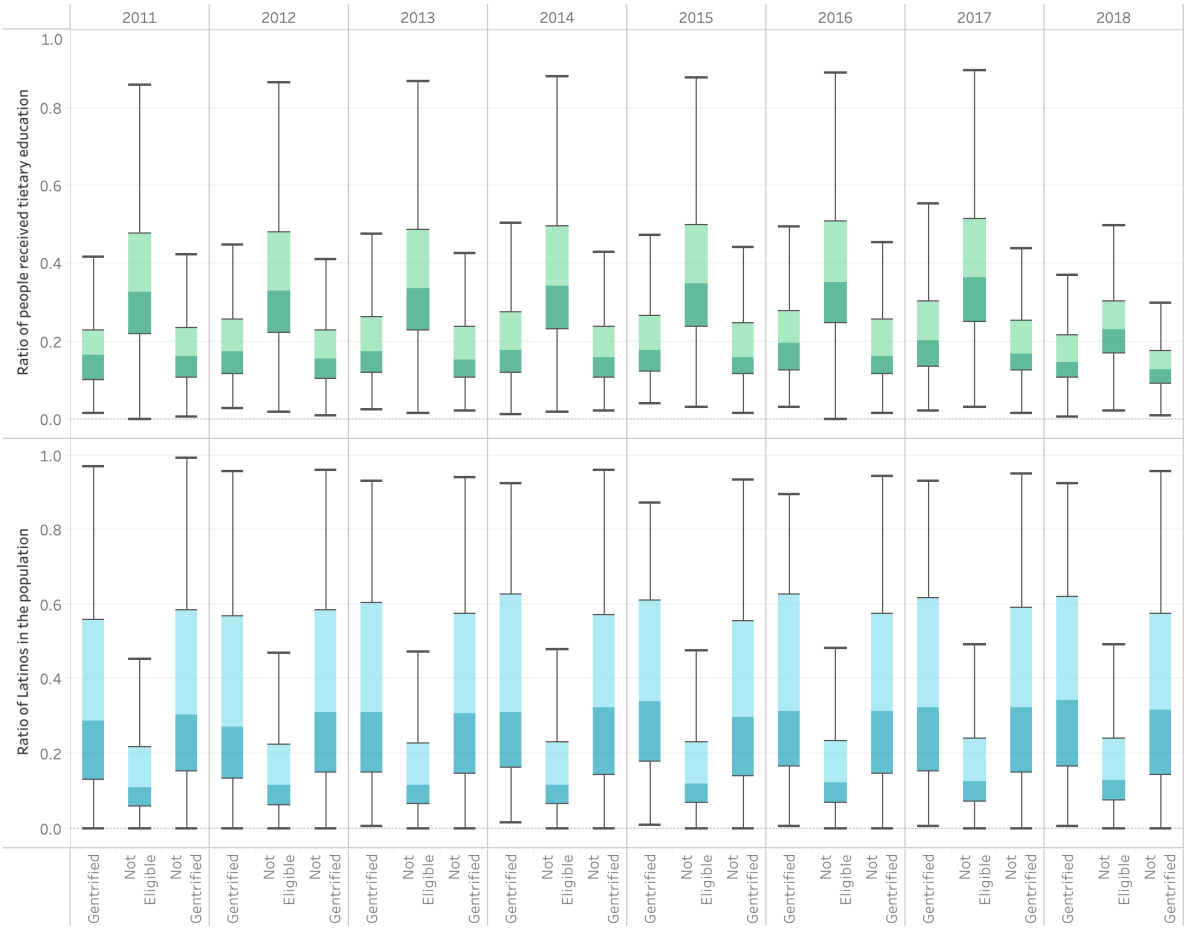


Figure 7: Ratio of people with bachelor's degree or higher to the population in all tracts, and ratio of the population Latinos to the whole population in all tracts.

# 6 Summary

The team have employed several data analysis techniques and frameworks in order to derive valuable insights from the given data. Ranging from data cleaning and feature engineering to Exploratory Data Analysis and Machine Learning, the resulting data and derived insights have been used in order to answer the topic question. As a result, we have quantified social factors, such as crime, and the quality of public services, like public transport, by locating queries from the 311 Calls data. Furthermore, we have included the changes in education levels and demographics to further gain insights into changes that tracts tend to go through during the gentrification process. We hope that findings presented in this report will further aid the process of evaluation of positive and negative externalities that gentrification process brings.

# References


[1] Maciag, M. (2015, January 31). Gentrification Report Methodology. Retrieved October 25, 2020, from https://www.governing.com/gov-data/gentrification-report-methodology.html

[2] Maciag, M. (2015, January 31). Gentrification Report Methodology. Retrieved October 25, 2020, from https://www.governing.com/gov-data/gentrification-report-methodology.html

[3] https://data.census.gov/cedsci

[4] https://www.census.gov/programs-surveys/geography/about/glossary.htmlpar$_t extimage_1$3

[5] https://www.census.gov/cgi-bin/geo/shapefiles/index.php