Is Gentrification a Race Phenomenon or a Wealth Phenomenon?
Citadel Datathon 2020

Team 9

October 2020

Executive Summary

There is no universally-agreed definition of gentrification. Papers such as Maciag (2015) [1] suggest that under these conflicting definitions many tracts may be identified or misidentified as being gentrified. This suggests that applying modelling to "classify" tracts as gentrified is arbitrary nor provides actionable insights for policymakers.

Most definitions propose hypotheses behind gentrification, identifying race and income as explanatory variables – for example an influx of wealthy (Caucasian) individuals and investments may lead to a rise in home prices, displacing original inhabitants [2]. This is a complex definition that includes multiple causal relationships. To this end, we instead modelled the *impacts* of gentrification, identifying the changes in median home price (by quantiles) in a tract as our measure. We selected this measure, as qualitatively it presents the "worse" aspects of gentrification - unaffordable housing - and has wider implications across disciplines geography, politics, real estate.

With this in mind, we aimed to understand the (causal) relationships which might motivate changes in home values. We identified *race* and *wealth* as our key variables, based on qualitative sources [3] and literature, and sought to **investigate whether race and wealth were more important drivers behind changes in median home prices** for a given tract.

By utilising both linear and nonlinear model based approaches on longitudinal data (US census data), and making use of modern explainability techniques such as Shapley values, **we determine that the race and wealth are jointly significant**; the likelihood that a tract is gentrified is highly correlated to the change in median Caucasian income, rather than just median income.

1 Introduction: What is Gentrification?

There is a great deal of disagreement among social scientists regarding how gentrification should be defined and measured [4]. Ruth Glass [5], who coined the term in 1964, describes the process as:

> “One by one, many of the working class quarters have been invaded by the middle class – upper and lower ... Once this process of ‘gentrification’ starts in a district it goes on rapidly until all or most of the working class occupiers are displaced and the whole social social character of the district is changed”.

Modern definitions, such as that of Governing magazine [1], instead tend to focus on measurable variables such as household income and the median home value. This is supported by the following intuitive line of reasoning: an influx of affluent individuals results in an increase in home prices, causing previous residents to be displaced.

These types of definitions, however, are somewhat problematic in that they encompass several nodes in a causal chain, each of which may have several exogenous causes (as seen in Figure 1). Accordingly, a subtle but important point in this investigation is distinguishing between *metrics* for gentrification and *causes* of gentrification. It is illogical, for instance, to define gentrification as a rise in house prices then to use the rise in house prices as an input to a model – we do *not* want the model to simply recover our measurement function.

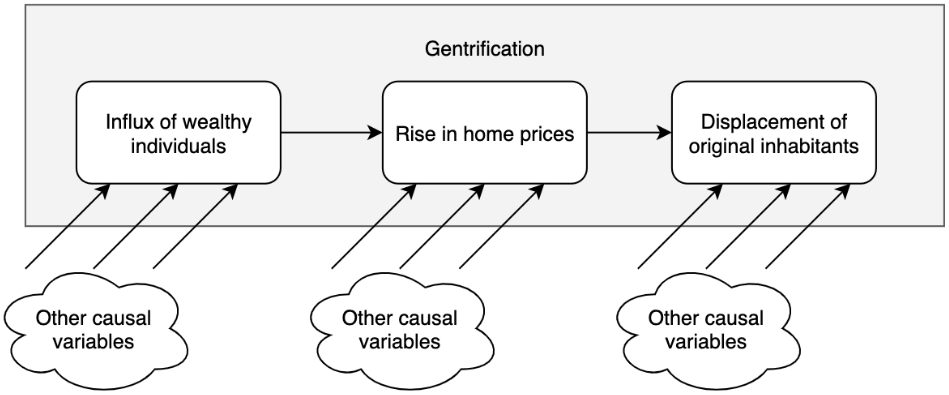

Figure 1: Causal diagram for gentrification

To address these issues, we have adopted a simple approach: gentrification will be measured solely by a rise in the median home price (technically, a rise in the median home price *quantile*). This approach is further discussed in Section 2. A corollary of this definition is that gentrification exists on a spectrum, rather than being a binary phenomenon (unlike in [3]).

This report will be structured as follows. Section 2 discusses the exploratory data analysis, while Section 3 discusses feature engineering and data modelling. We present our conclusions in Section 4.

2 Exploratory Data Analysis

2.1 Data overview

We were given a dataset of US Census data for New York, for the years 2009-2018, and a dataset of 311 (lifestyle complaints). To augment these, we downloaded further data from the Census API, including: median income by race, population data by education level.

We noticed that even for the original census dataset, only a small fraction of the tracts (c. 10%) contained data for the full 10 years, which limits our ability to build predictive long-term models.

2.2 Data Cleaning

First, we note that many of the columns are **not the actual population statistics, but estimates** (since it is survey data). We observe some values of "-66666.0" in the **Median Household Income** and **Median Home Values**, and interpreted them as missing values. In addition, we observed some "0" values, which could again be missing data.

We consider the dimensions of gentrification discussed in the Governing article: incomes, home value, and demographics - most of all, the proportion (and change) of Caucasian population for a tract. These distributions are shown in Figure 2.

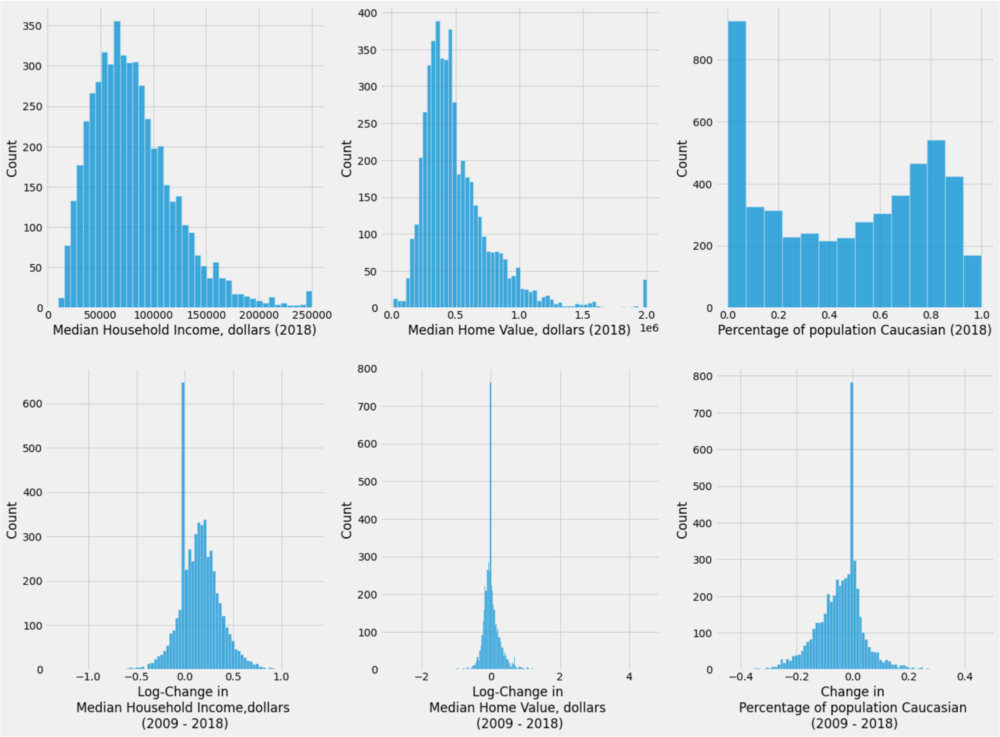

Figure 2: Distributions of median income, median home value, percentage caucasian and changes

We observe that incomes and home values (logically) follow a positively skewed distribution, though for the percentage of Caucasian population, there are many tracts with 0% Caucasian population. A closer look at the keys for the census data indicates that all of the provided data (given it is a census) are *estimates*. Thus whether the percentage is really 0%, or there is missing data, is unclear.

We take the log change in median household income and median home value (given the skewedness of the distributions) and observe that they exhibit more kurtosis; for both there are a lot of values of 0; and many extreme values. For the change in percentage Caucasian, we take the difference given the original values are bounded by [0,1] (giving a value from $[-1, 1]$). Again we observe that many values are exactly 0.

2.3 Naive Method – Quantiles

A very simple method to test for gentrification could be to inspect the changes in **home value, percentage Caucasian, and household income** across 2009 - 2018. We could say, for example, if the z-scores of our chosen features exceed thresholds then the tract is gentrified (per discussion of [1]) :

$$\text{gentrified if: } \frac{x_i - \mu_i}{\sigma_i} \geq \Phi^{-1}(a_1), i = 1..p$$

.

This could be a simple but robust method for policymakers to decide which districts are gentrified and where to allocate resources. This is very similar to the methodology of the Buzzfeed gentrification article [3]) which is in term inspired by the methodology of Michael Maciag [1].

To implement this, we calculate the changes in these features:

$$\log(x_{2018}/x_{2009}) \text{ for median income, median home value}$$

$$\frac{\text{Caucasian}_{2018}}{\text{total}_{2018}} - \frac{\text{Caucasian}_{2009}}{\text{total}_{2009}} \text{ for percentage Caucasian}$$

This also serves to stationarise all the features and considerably reduces the dimensionality. We run a Jarque-Bera test to test the normality assumption, and find that no feature follows a normal distribution (**Table 1**, see Appendix). Thus instead, we look to the empirical quantiles and select, for example: the tracts that exceed the 95th percentile for all of home value, percentage Caucasian, and household income (**Table 2**, see Appendix). Below we provide an example of tracts identified as gentrified under this rule, at the 75% threshold for all variables.

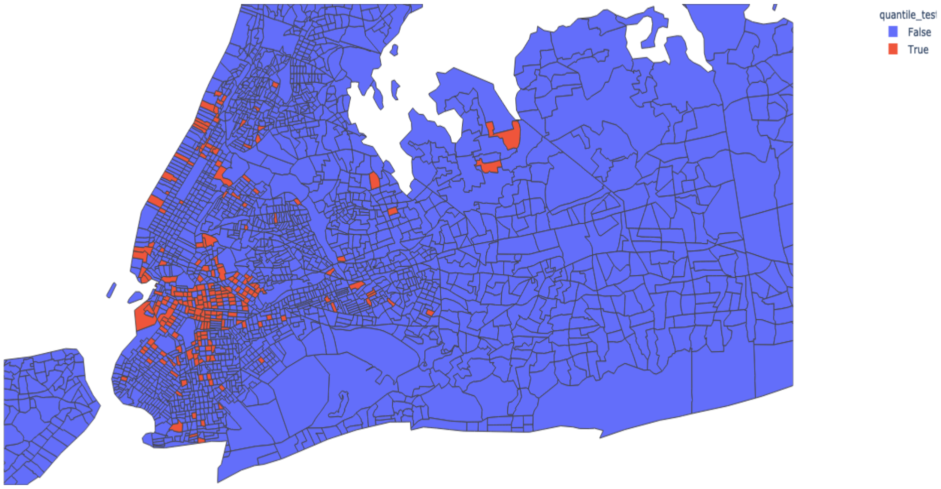

Figure 3: Tracts above the 75th quantile for changes in household income, Caucasian percentage, home values (denoted in red)

2.4 PCA

We look to an unsupervised approach to data exploration. Given our measurement of gentrification, we look to apply Principal Components Analysis on our data to explore the underlying factor structure.

By performing PCA on data that has been separated by whether the tracts have been gentrified or not (using our naive ”classification”), we can find the factors that explain most of the variance in the data. By comparing the first principal component of both datasets, we are able to contrast the differences in factors that explain the variances. First we examine the amount of variance explained by the components for both the gentrified and non-gentrified tracts' data. These are shown in Figures 4 and 5 respectively.

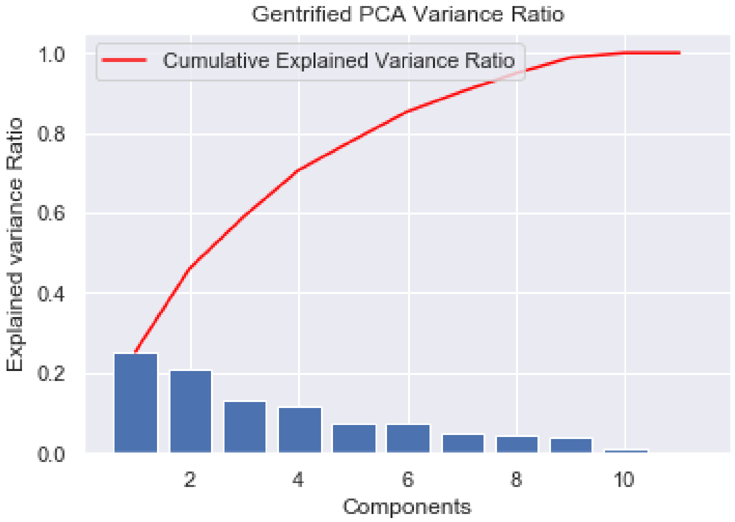

Figure 4: Gentrified PCA Variance Ratio

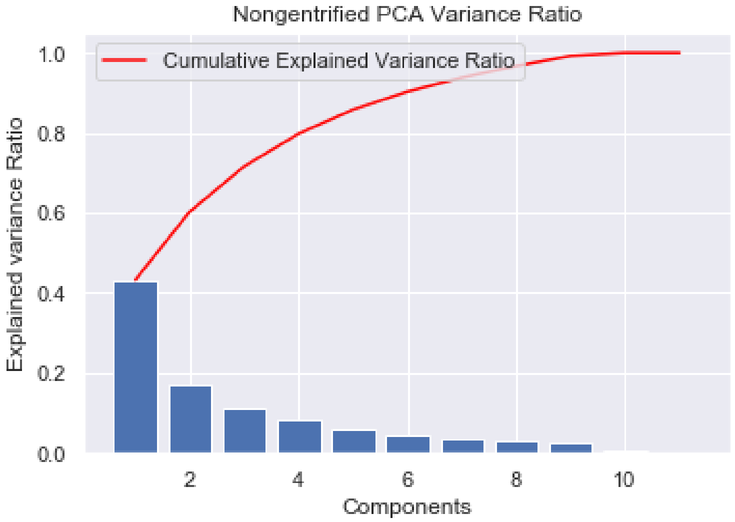

Figure 5: Nongentrified PCA Variance Ratio

PCA consists of an eigendecomposition of the covariance matrix – intuitively, rotating feature space to find the axes that maximise variance. A biplot (Figure 6) shows a 2D projection of these new axes. We can infer a few things from the biplot. Firstly, we notice that all the vectors are in the same quarter plane. This gives us a sign that the factors, albeit having different weights, have the same type of impact in explaining the variance of the data. However, when considering the angle of the vector, we notice that median household income is a much bigger explainer of variance for non-gentrified tracts than for gentrified areas. Also, we notice that the number of Caucasians has a very similar impact in explaining variance for both gentrified and non-gentrified tracts.

We can understand the relative importance a bit more if we consider the relative magnitudes of the principal components. In the case of the number of Caucasians v.s. the median household income, we notice that the ratio of the two is significantly higher for gentrified tracts than it is for non-gentrified tracts. This means that for gentrified tracts, median household income is less of an explainer of variance compared to number of caucasians, and for non gentrified tracts, this is flipped.

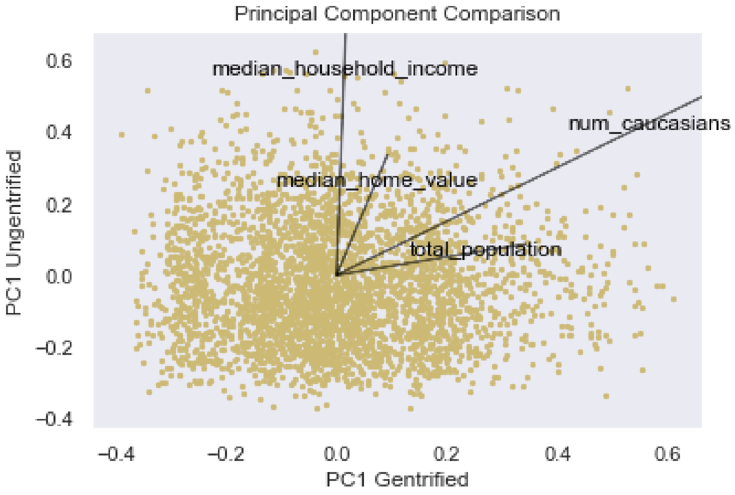

Figure 6: Biplot of Principal Components of Gentrified vs Nongentrified

3 Modelling

In the exploratory data analysis, our measure of gentrification (based on [1] and [3]) used a combination of several variables. As discussed in the introduction, however, this confuses cause and effect and greatly limits our ability to model the data. To that end, moving forward, we explore gentrification in terms of the changes in median home value. We will later define the **gentrification score of a tract to be the average annual change in the quantile of the tract's median home value across the 2009-2018 time period**. This definition conveniently removes the overall trends in house prices and focuses on the relative expense between tracts.

The causal diagram in Figure 1, based on typical definitions of gentrification, makes no reference to race, although anecdotal accounts like that of the Buzzfeed article [3] strongly hint at a racial component. With our new definition of gentrification in mind, we can now model the data to understand the relative importance of race and wealth in determining whether a tract becomes gentrified.

3.1 OLS Regression

The naive quantiles method from Section 2.3 is simple (and possibly robust), but does not account for the *joint distribution* in median home value, median income, and the percentage Caucasian. Additionally, the choice of thresholds is arbitrary; we may find more or less gentrified tracts depending on our choices of a_1, a_2, a_3.

To improve on this, we instead run a Linear Regression, the most simple being of the following form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i,2} + \epsilon_i$$

where
y_i is the log-change in median home value (2009 - 2018)
$x_{i,1}$ is the net change in Caucasian percentage of the population (2009 - 2018)
$x_{i,2}$ is the log-change in median the change in (2009 - 2018)
ϵ_i is the error term

The motivation for this regression specification, is that we wish to test the following hypothesis:
an increase in Caucasian population is associated with higher home values (our measure of gentrification), and identify a possible *causal relationship*

Dep. Variable:	median_home_value_change	**R-squared:**	0.035
Model:	OLS	**Adj. R-squared:**	0.034
Method:	Least Squares	**F-statistic:**	90.56
Date:	Wed, 21 Oct 2020	**Prob (F-statistic):**	2.31e-39
Time:	14:19:18	**Log-Likelihood:**	-981.41
No. Observations:	5023	**AIC:**	1969.
Df Residuals:	5020	**BIC:**	1988.
Df Model:	2		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	0.0052	0.006	0.925	0.355	-0.006	0.016
percentage_caucasian_change	0.3267	0.050	6.551	0.000	0.229	0.424
median_household_income_change	0.2261	0.020	11.418	0.000	0.187	0.265

Omnibus:	2856.368	**Durbin-Watson:**	1.873
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	173373.342
Skew:	1.946	**Prob(JB):**	0.00
Kurtosis:	31.517	**Cond. No.**	12.2

The results of the regression indicate several things:

1. The coefficient for `percentage_caucasian_change` is positive but small: a 1% increase in Caucasian population is associated with a $e^{0.3267/100} - 1 = 0.3\%$ change in home values. A possible explanation for this minuscule effect, is that there are many zeros – which may be missing values – in the data; our initial EDA in 2 indicates that the distribution of median home values is *skewed*. This coefficient is also a measure of global, or average effect across all tracts.

2. The residual plots (and result of the Jarque-Bera test) suggests heteroskedasticity. Combined with the former, there is very likely *Omitted Variable Bias* in this simple regression - explanatory variables and heterogeneity in the tracts - clusters - that we have yet to identify.

3. R^2 is low at 0.035, indicating that the model have significant predictive power nor explain most of the variance.

An interesting suggestion is that we can convert the output of any model as a ”classification rule” for gentrification. In specific, the residuals of the regression. - large residuals indicate unexplained variation – possible outliers – which may be used to identify gentrified districts. For example, we can say:

$$\text{Tract}_i \text{ is gentrified if } \epsilon_i > \mathbb{E}[\epsilon_i] + 2Var(\epsilon_i)^2$$

Gentrified tracts under this rule are shown in Figure 7.

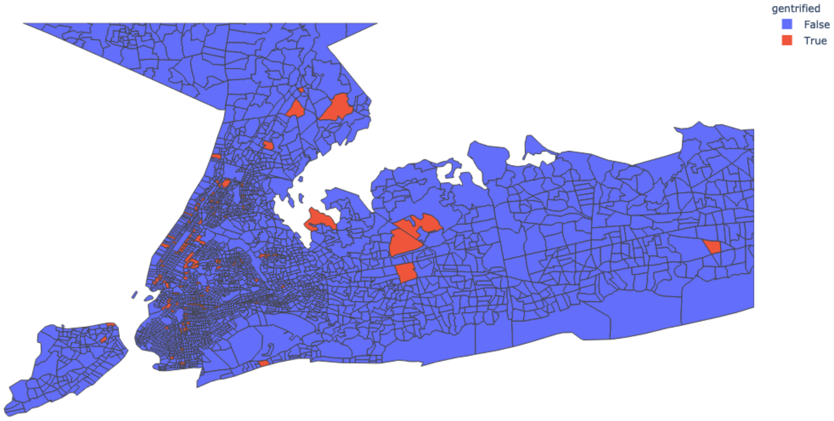

Figure 7: Tracts Identified by the OLS Residuals

Given the weaknesses of this initial regression, we proceed with two particular improvements:

- Instead of using the raw log changes, we use the changes in quantiles over the the 2009-2018 year period. This will address the skew in the distribution of home values and the distributional drift. This leads us to specifying our out come of interest y as the **gentrification score: the average annual change in the quantile of the tract's median home value across the 2009-2018 time period**

- As the initial regression did not explain much variance, we add more explanatory variables, for example education, and income by race. In addition, we will encode some information about the initial conditions of each tract instead of using a regression of changes on changes.

3.2 Feature Engineering

To process the features, we used a similar transformation to the gentrification score: a quantile transformation (reflecting how a variable in a tract compared to the values in other tracts in the same year), followed by a differencing to find the average annual quantile change. We also used the initial 2009 quantile values as features, in order to capture "momentum" effects.

The raw features used in this model were as follows:

- Racial – the percentage of Caucasians, Blacks, and Hispanics.
- Education – the proportion of various education levels (e.g highschool, bachelor's degree, graduate degree).
- Income – median income, income by race, Gini coefficient.

3.3 Least-squares Regression

With this basic feature engineering out of the way, we proceeded to fit a simple OLS regression model to the dataset, giving the following summary:

Dep. Variable:	avg_qq_chg	**R-squared:**	0.140
Model:	OLS	**Adj. R-squared:**	0.123
Method:	Least Squares	**F-statistic:**	8.077
Date:	Fri, 23 Oct 2020	**Prob (F-statistic):**	4.43e-28
Time:	21:25:15	**Log-Likelihood:**	3333.0
No. Observations:	1316	**AIC:**	-6612.
Df Residuals:	1289	**BIC:**	-6472.
Df Model:	26		

The regression betas are shown in Figure 8. We can infer several important facts from this initial regression model:

- The overall regression only explains 14% of the variance in the change in home price quantiles. This is a significant increase over the initial regression.
- High home values are indeed associated with an out-migration of Blacks and Hispanics, as seen by the large negative coefficients. For every 10-percentile increase in a tract's median home price, the (quantile of the) Hispanic percentage of the tract decreases by 2.5 percentile points. This lends credence to the idea that increasing house prices result in displacement.
- Counter-intuitively, the change in the Caucasian percentage is also negatively related to the increase in home prices. This is further explored in Section 3.
- This out-migration disproportionately includes less educated strata of the population.
- On the other end of the spectrum, the factors most positively associated with an increase in home value are the percentage of graduate degree-holders, and an increase in *white* affluence (much more so than affluence overall).
- On the whole, the initial quantile values had less explanatory power compared to the average quantile changes. Accordingly, in the refined model, we shall remove the initial quantile features.

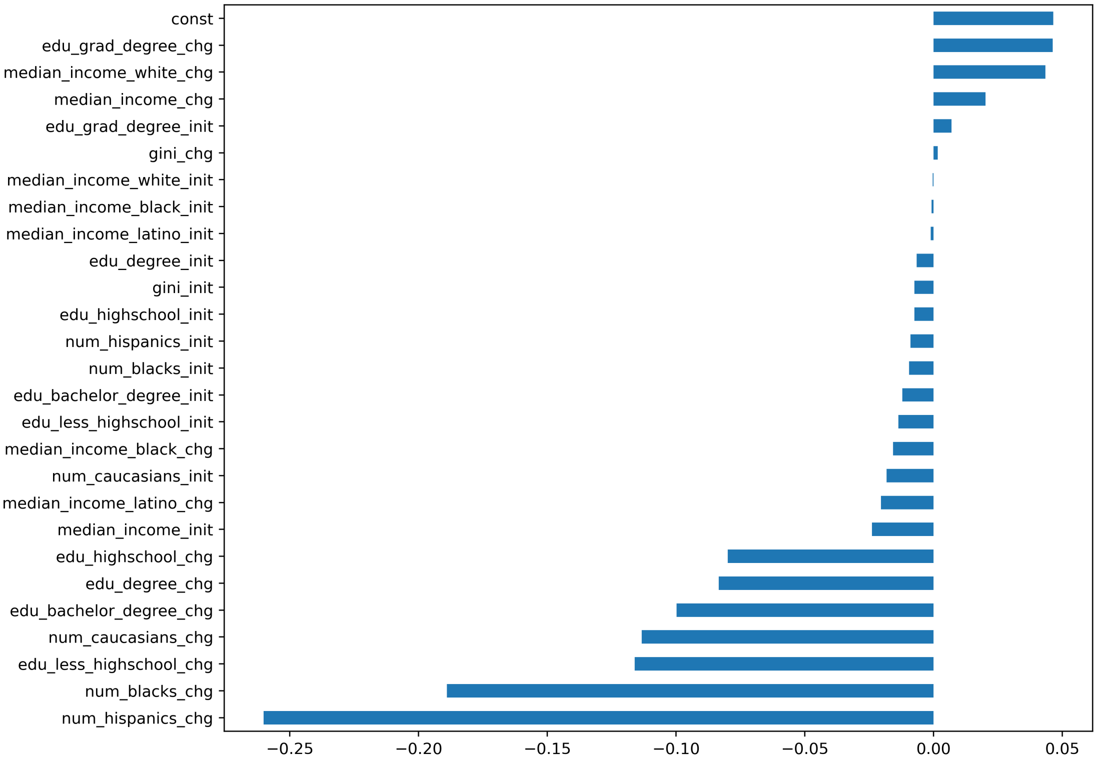

Figure 8: Regression betas

3.4 Panel Regression

The flaws of the previous regression approaches are that a lot of *information* is lost in the process. We lose information about *local structure*, idiosyncratic difference between each tract, which may be what we want to identify. In addition, by reducing to a single net change for 2009-2018, we lose the information about changes over *time*.

We run a first differences panel regression of the following form:

$$\Delta y_{it} = \alpha_t + \beta_0 + \gamma_i \Delta x_{it} + \epsilon_{it}$$

Where:
y_{it} is the **log-change in median home value**
x_{it} is the **percentage change in Caucasians** for a given tract i and time t
γ_i is an **entity fixed effect** for a specific tract - a tract's sensitivity to changes in caucasian population
α_t is the **time fixed effect** - idiosyncratic differences across the years) for year t
The Panel Regression gives us a **coefficient for each tract**, γ_i, the **entity fixed effect**, accounting for idiosyncratic differences across the years - the **time fixed feffect**, α_t.

A large coefficient γ_i suggests - that **changes in the percentage of Caucasian population are associated with large changes in median home value**. In effect, it gives us a "propensity" for a tract which can be used to identify gentrified tracts.
(There are too many coefficients to plot (given there is one dummy for every tract), but in 9 weidentify the tracts with the top $N = 20$ coefficients.)

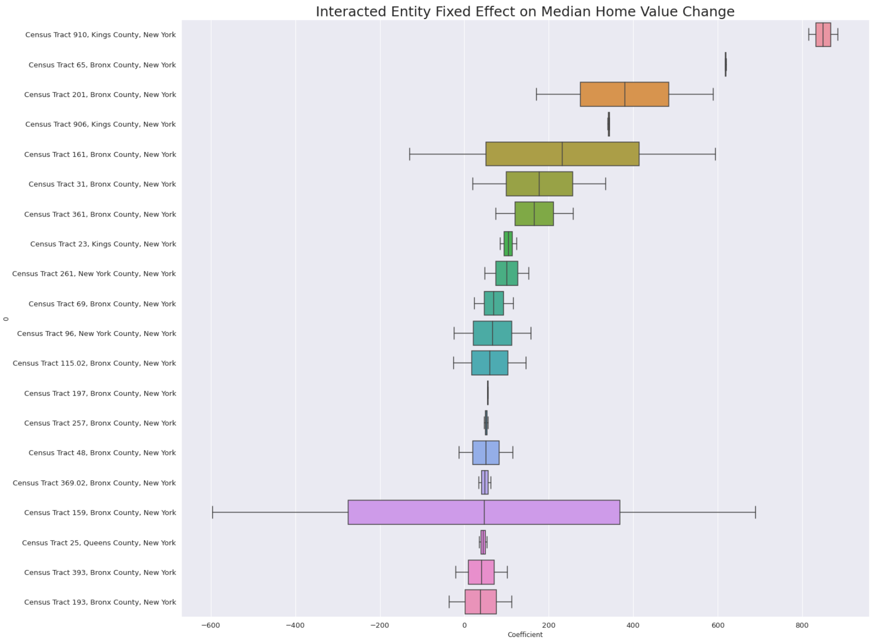

Figure 9: Entity Fixed Effects

3.5 Tree-Based Models

Machine Learning methods have been applied successfully to the analysis of gentrification [6]. Tree-based regression models, such as Random Forest, are able to capture nonlinear interaction terms between features. More importantly for our purposes, they have a high degree of explainability (as compared to black-box neural networks, for example). This is because the fundamental unit of a tree-based model – the decision tree – works in a highly intuitive manner, asking "yes/no" questions of the features to result in a prediction. Our motivation for using a tree-based learner to identify relationships is that they are **explainable**, should have more **predictive power** than OLS.

We fitted a Random Forest regression model to the data and computed Shapley values. Intuitively, Shapley values measure the contribution of a given feature to the final model. Figure 10 contains a wealth of information. Firstly, features are ranked in decreasing order of their contribution (sign-independent) to the model. This corroborates the results of the OLS in Section 3.3, with the most important features being the change in the quantiles of the Caucasian median income, the proportion of residents with a graduate degree, and the proportions of Blacks, Caucasians and Hispanics. Secondly, each point per row of the Shapley plot corresponds to a data point coloured by the magnitude of the feature value. The point's position on the horizontal axis shows whether the data point contributed positively or negatively to the resulting change in median home price.

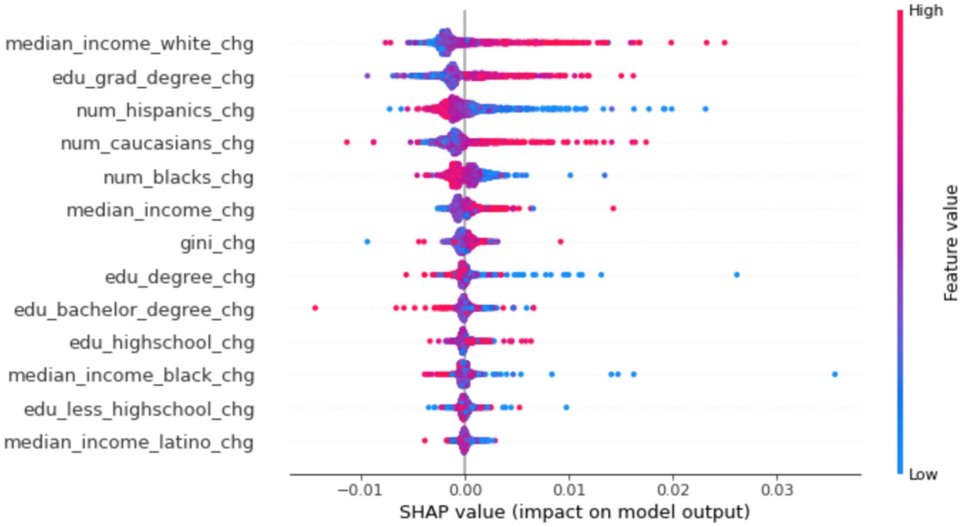

Figure 10: Shapley values for a Random Forest regressor

The relative strength of the `median_income_white_chg` feature compared to `median_income_chg` suggests that racial information is critical. This is supported by the high relative importance of `num_caucasians_chg` and `num_hispanics_chg`. We can gain a more nuanced understanding of the impact of a particular feature by plotting the relationship between Shapley values and individual features. For example, Figure 11 shows that the change in median Caucasian income has a nonlinear effect on the median home price. The pattern of colours suggests that this feature is highly correlated with `median_income_chg`, as expected.

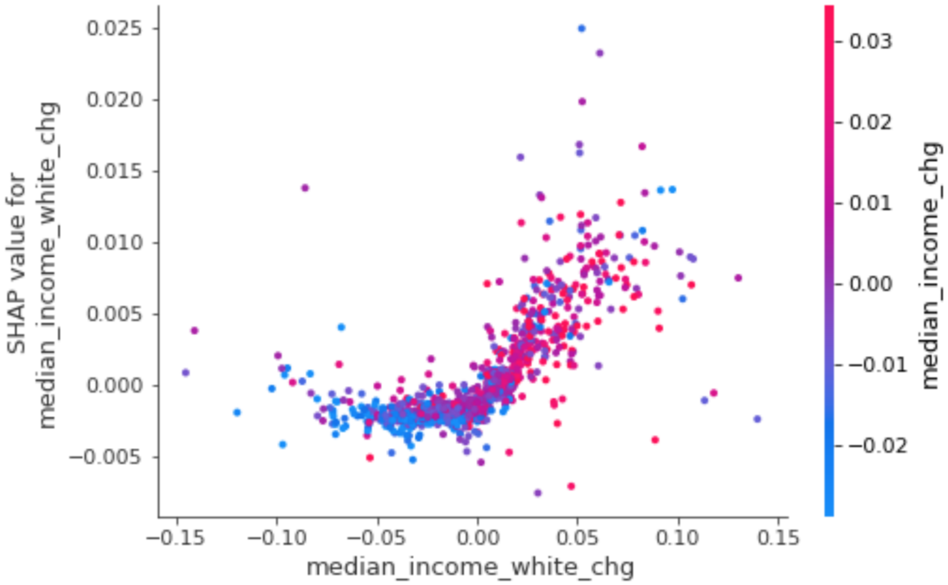

Figure 11: Plot of Shapley values against the change in median Caucasian income

4 Conclusion

In conclusion, the classic definition of gentrification, which considers the influx of wealthy individuals to be the ultimate source of displacement, neglects a key third variable: race. Our analysis shows that the median change in *Caucasian* income per tract is a much more important feature compared to the overall median change in income per tract when determining the change in the median house price in the tract.

However, it would be wrong for us to conclude that gentrification can be explained solely by adding race to the standard definition. Both the OLS coefficients and the Shapley values for the Random Forest regressor demonstrate that another important feature is the change in the percentage of people with graduate degrees. This suggests that current government initiatives, such as requiring private property developers to build affordable housing, will not necessarily prevent the displacement of a neighbourhood's original inhabitants [7].

References

[1] M. Maciag, “Gentrification report methodology,” 2015.

[2] R. Florida, “This is what happens after a neighborhood gets gentrified,” 2015.

[3] L. T. Vo, “They played dominoes outside their apartment for decades. then the white people moved in and police started showing up.,” 2018.

[4] M. Davidson, “Critical commentary. gentrification in crisis: Towards consensus or disagreement?,” *Urban Studies*, vol. 48, no. 10, pp. 1987–1996, 2011.

[5] R. Glass, “Aspects of change,” *The gentrification debates: A reader*, pp. 19–30, 1964.

[6] J. Reades, J. D. Souza, and P. Hubbard, “Understanding urban gentrification through machine learning,” *Urban Studies*, vol. 56, no. 5, pp. 922–942, 2019.

[7] M. Ponsford, “Developers must build low-cost homes or publish finances: London mayor,” 2017.

Appendix

Table 1 - Jarque-Bera Normality Test

feature	test_statistic	p_value
median_home_value_change	156300.401804	0.000...
median_household_income_change	1049.964079	0.000...
percentage_caucasian_change	721.586478	0.000...

Table 2 - Gentrified Tracts Identified by 95th Quantiles

NAME	Log-Change in Median Home Value	Log-Change in Median Income	Change in Percentage Caucasian
Census Tract 119, Kings County, New York	0.991585	0.533129	0.211894
Census Tract 121, Kings County, New York	0.687070	0.551961	0.093864
Census Tract 127, Kings County, New York	0.621981	0.688317	0.116148
Census Tract 15, Kings County, New York	0.520753	0.705523	0.103573
Census Tract 191, Kings County, New York	0.564497	0.695036	0.351430
Census Tract 217, Kings County, New York	1.457316	0.622713	0.140393
Census Tract 219, Kings County, New York	0.765115	0.746637	0.098002
Census Tract 229, Kings County, New York	0.514185	0.571043	0.211251
Census Tract 229, New York County, New York	1.264719	0.550725	0.109688
Census Tract 243, Kings County, New York	0.666215	0.867855	0.267937
Census Tract 261, Kings County, New York	0.545740	0.669966	0.119844
Census Tract 269, Kings County, New York	0.869701	1.316169	0.094384
Census Tract 275, Kings County, New York	0.520905	0.539971	0.112262
Census Tract 279, Kings County, New York	0.654916	0.677045	0.129122
Census Tract 289, Kings County, New York	0.850526	0.563277	0.092810
Census Tract 315, Kings County, New York	0.641010	0.723374	0.263230
Census Tract 317.01, Kings County, New York	0.872355	0.720066	0.271448
Census Tract 317.02, Kings County, New York	0.576349	0.646320	0.204126
Census Tract 450, Kings County, New York	0.664801	0.551111	0.234145

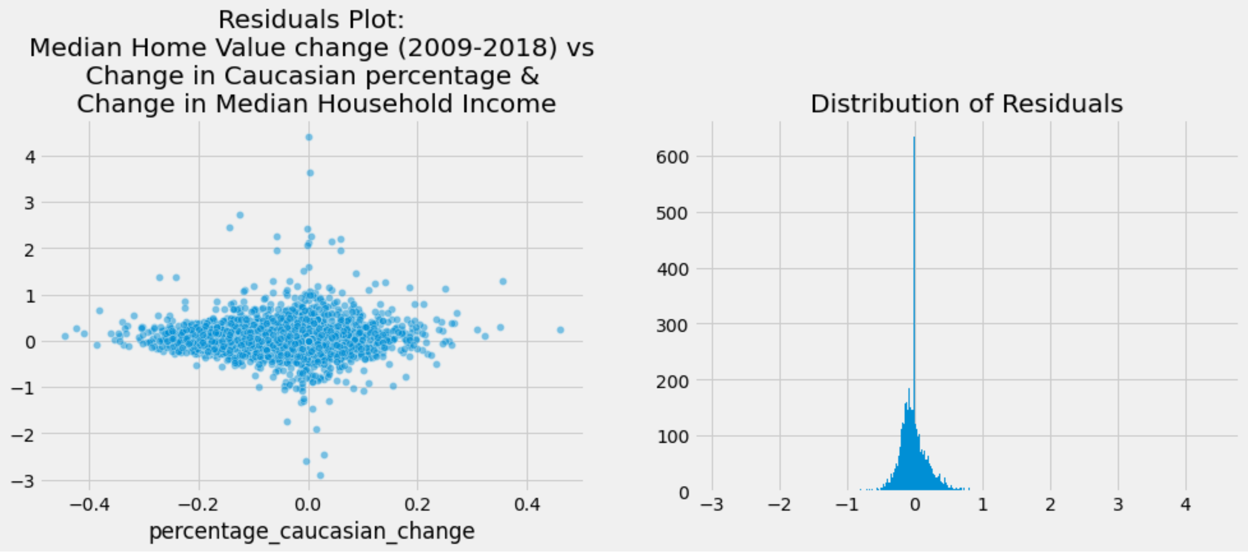

Table 3 - Linear regression results

Dep. Variable:	avg_qq_chg	**R-squared:**	0.140
Model:	OLS	**Adj. R-squared:**	0.123
Method:	Least Squares	**F-statistic:**	8.077
Date:	Fri, 23 Oct 2020	**Prob (F-statistic):**	4.43e-28
Time:	21:35:24	**Log-Likelihood:**	3333.0
No. Observations:	1316	**AIC:**	-6612.
Df Residuals:	1289	**BIC:**	-6472.
Df Model:	26		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	0.0464	0.012	3.969	0.000	0.023	0.069
num_caucasians_chg	-0.1133	0.082	-1.375	0.169	-0.275	0.048
num_blacks_chg	-0.1890	0.088	-2.142	0.032	-0.362	-0.016
num_hispanics_chg	-0.2601	0.081	-3.210	0.001	-0.419	-0.101
edu_less_highschool_chg	-0.1161	0.051	-2.259	0.024	-0.217	-0.015
edu_highschool_chg	-0.0799	0.038	-2.096	0.036	-0.155	-0.005
edu_degree_chg	-0.0834	0.030	-2.742	0.006	-0.143	-0.024
edu_bachelor_degree_chg	-0.0998	0.040	-2.482	0.013	-0.179	-0.021
edu_grad_degree_chg	0.0463	0.049	0.949	0.343	-0.049	0.142
gini_chg	0.0017	0.022	0.074	0.941	-0.042	0.046
median_income_chg	0.0202	0.044	0.454	0.650	-0.067	0.107
median_income_white_chg	0.0435	0.023	1.856	0.064	-0.002	0.089
median_income_black_chg	-0.0157	0.021	-0.767	0.443	-0.056	0.025
median_income_latino_chg	-0.0204	0.024	-0.853	0.394	-0.067	0.027
num_caucasians_init	-0.0183	0.006	-3.323	0.001	-0.029	-0.007
num_blacks_init	-0.0096	0.005	-1.831	0.067	-0.020	0.001
num_hispanics_init	-0.0089	0.006	-1.572	0.116	-0.020	0.002
edu_less_highschool_init	-0.0137	0.006	-2.143	0.032	-0.026	-0.001
edu_highschool_init	-0.0074	0.005	-1.563	0.118	-0.017	0.002
edu_degree_init	-0.0066	0.004	-1.877	0.061	-0.013	0.000
edu_bachelor_degree_init	-0.0121	0.005	-2.259	0.024	-0.023	-0.002
edu_grad_degree_init	0.0070	0.006	1.170	0.242	-0.005	0.019
gini_init	-0.0074	0.003	-2.201	0.028	-0.014	-0.001
median_income_init	-0.0239	0.006	-3.973	0.000	-0.036	-0.012
median_income_white_init	-0.0003	0.004	-0.092	0.926	-0.007	0.007
median_income_black_init	-0.0008	0.003	-0.217	0.828	-0.008	0.006
median_income_latino_init	-0.0011	0.004	-0.294	0.769	-0.009	0.007

Omnibus:	202.127	**Durbin-Watson:**	1.837
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1748.527
Skew:	0.425	**Prob(JB):**	0.00
Kurtosis:	8.583	**Cond. No.**	466.