David Ospina, Micah Achmoody, Nicholas Tan, Riley Krisch Sr.

Professor Kevin Gold

DS110 Introduction to Data Science with Python

May 3, 2023

<u>The Search for the GentriValue</u>

Index

GentriFactor: The Factors that contribute to quantification of gentrification

GentriValue: the probability of a neighborhood being gentrified

GentriValue Equation: The combination of the The Freeman Model, The Ellen and O'Regan Model, and the McKinnish Model.

****Note: The code contains all of the work that we have done over the course of the project, since most of it was data cleaning. The parts relevant to the latest iteration of the project, however, are limited to just the section titled, "New Idea (Mixture of Old Ideas)".***

Abstract

Gentrification has become a growing issue in the 21st century, as policymakers, urban planners and other stakeholders grapple with data to try and understand the trends influencing the phenomenon, in order to inform their decisions made and actions taken to address it. In this paper, our team analyzes the rate and scale of gentrification occurring in New York City at the

turn of the millennium, by applying three established models by leading figures in the field and then utilizing machine learning to evaluate if such results do indeed hold true.

Introduction

Upon discovering that the term “Gentrification” has been loosely utilized to support hot topics and decision making in urban planning, one very glaring yet simple question arises. *What is Gentrification?*

According to National Geographic, it describes gentrification as [1] “a process where wealthy, college-educated individuals begin to move into poor or working-class communities, often originally occupied by communities of color.” Gentrification has been used to support arguments in favor of unveiling racial disparities and consequences to urban planning and long term real estate expansion yet most definitions often lack enough substance to support with factual or numerical evidence.

The goal of this project is to take as many quantitative definitions of gentrification and combine them into a Machine Learning model that can predict the likelihood of gentrification in a metro area. The metro area we studied is New York City which consists of a little over 250 neighborhoods.

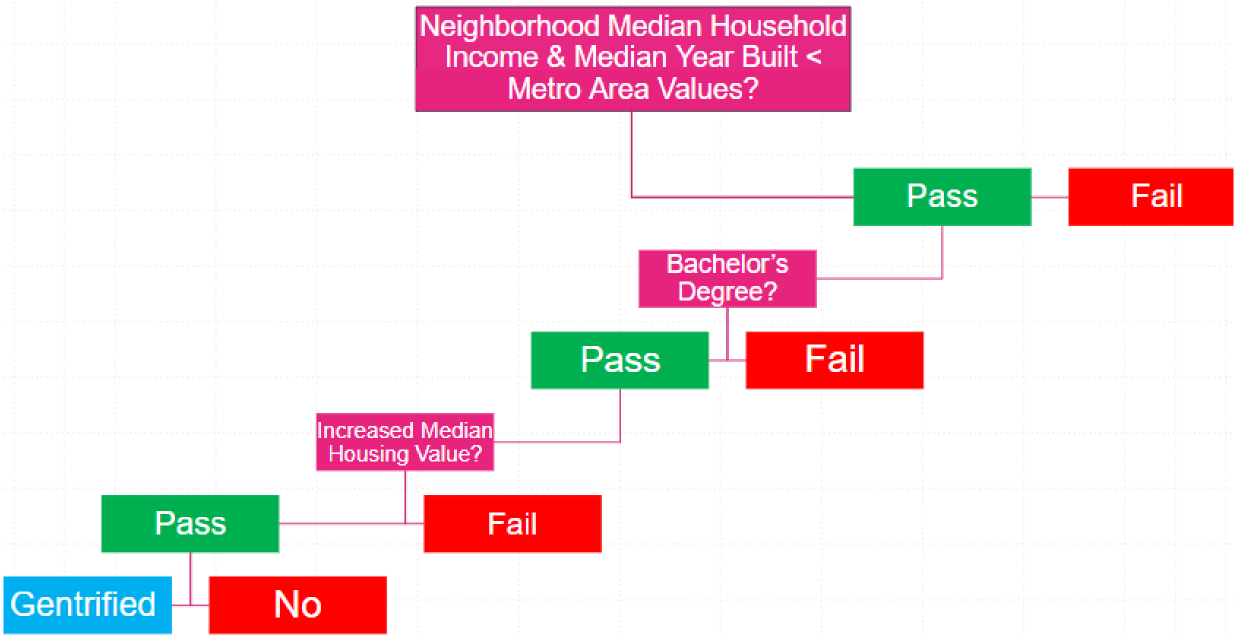
[1] “Gentrification.” Education. Accessed May 3, 2023. https://education.nationalgeographic.org/resource/gentrification/.

Literature Review

Existing literature provides no concrete method to define gentrification. Generally, definitions hinge upon the type of research being done, and the scope of research which researchers intend to pursue. However, there are multiple major methods that can attempt to quantify gentrification. [2]According to the SaportaReports Report, *How Do Researchers Define Gentrification?,* three models are at the forefront of this definition.

The Freeman Model classifies gentrification as occurring in an area where the median household income and share of housing built in the prior 20 years are both less than the metro-area values. For an area to be gentrified, the share of residents with college degrees also has to be greater than the metro value, and there needs to be an increase in house prices.

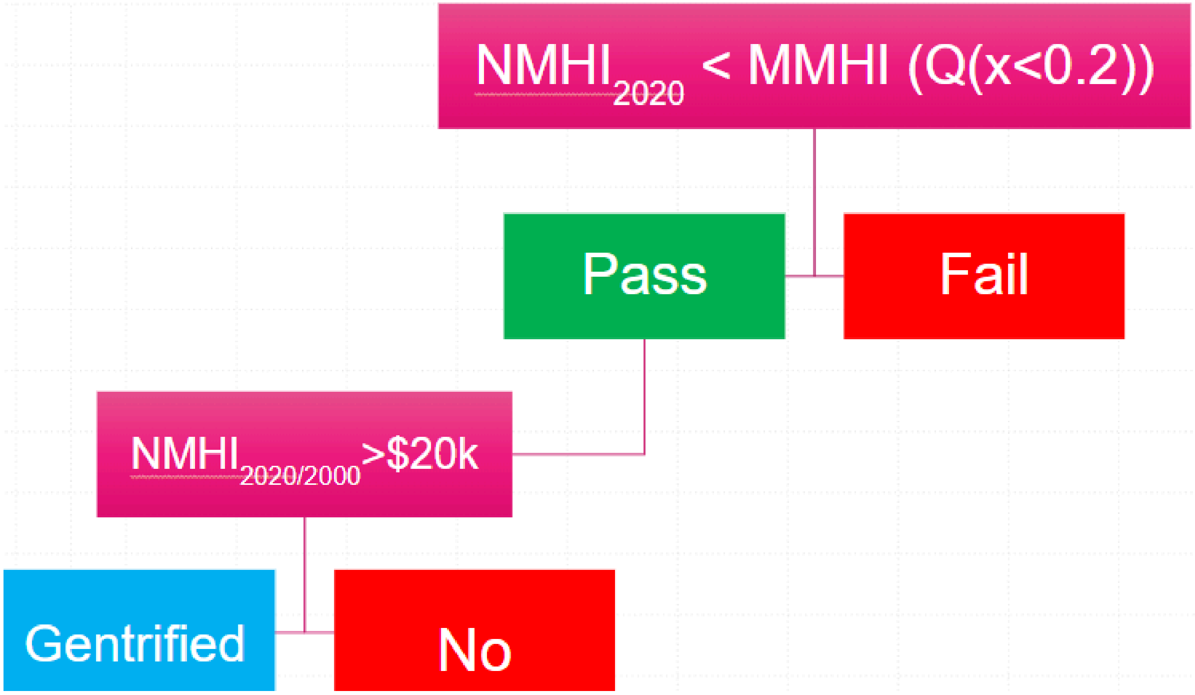
Fig 1.1 Schematic Diagram of the Freeman Model


Neighborhood Median Household Income & Median Year Built < Metro Area Values?
Pass
Fail
Bachelor's Degree?
Pass
Fail
Increased Median Housing Value?
Pass
Fail
Gentrified
No

[2] Vashi, Sonam. "How Do Researchers Measure Gentrification?" SaportaReport, May 20, 2019. https://saportareport.com/how-do-researchers-measure-gentrification/main-slider/sonam-vashi/#:~:text=The%20Freeman%20model%20classifies%20gentrification,than%20the%20metro%2Darea%20values.

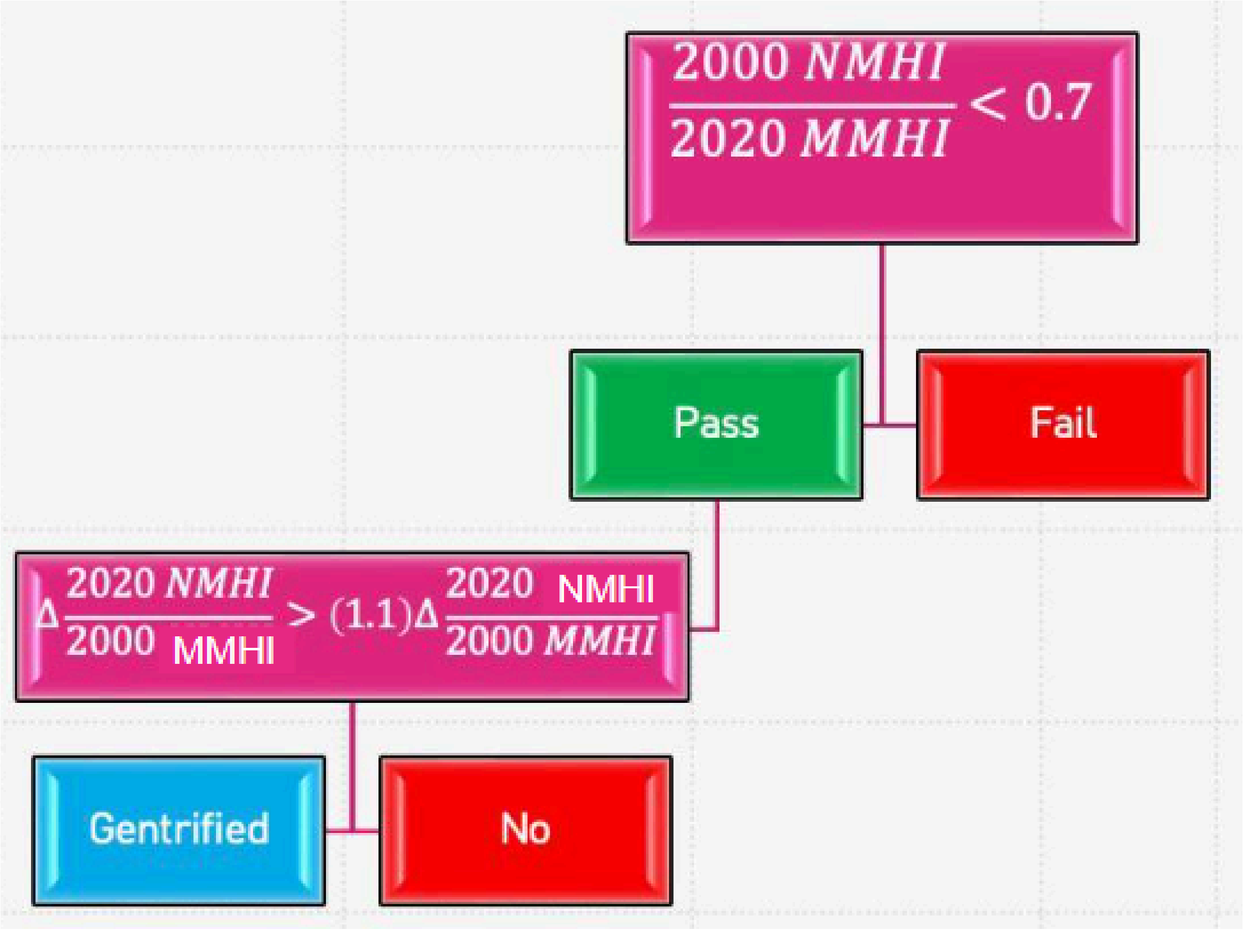
The McKinnish Model says an area is gentrified if the neighborhood average family income is in the bottom 20 percent of all urban neighborhoods nationwide and if there's been a real increase of at least $10,000 in the neighborhood's average family income within the last decade.

Fig 1.2 Schematic Diagram of the McKinnish Model



The Ellen & O'Regan model calls an area "gentrified" if the ratio of the neighborhood's household income at the start of the decade, compared to the metro average household income, is less than 0.7—and there needs to be at least a 10 percentage-point increase in the ratio of neighborhood to metro average household income over the past decade.

Fig 1.3 Schematic Diagram of the Ellen and O'Regan Model


$\frac{2000\ NMHI}{2020\ MMHI} < 0.7$
Pass
Fail
$\Delta\frac{2020\ NMHI}{2000\ MMHI} > (1.1)\Delta\frac{2020\ NMHI}{2000\ MMHI}$
Gentrified
No

Upon reviewing these Gentrification models, we found that there are lacking metrics which may be vital to quantifying gentrification. The research paper, [3]*A Quantitative Approach to Gentrification: Determinants of Gentrification in U.S. Cities, 1970-2010,* by University of Georgia professor, Richard W. Martin, cites:

(1) Median home values or median gross rents

(2) Median household incomes

(3) Central-City Manufacturing Employment Share

[3] Martin, Richard W. "A Quantitative Approach to Gentrification: Determinants of ..." Accessed May 3, 2023. https://www.terry.uga.edu/sites/default/files/inline-files/Determinants_of_Gentrification.pdf.

(4) Black population shares

(5) Percent of employed central-city residents in professional and executive occupations

(6) Percent of central-city residents aged 25 and over with at least a bachelor's degree

(7) Percent of housing units built before 1939

(8) Median housing age

(9) Central-city employment share

(10) Gentrification Potential

as eligible determinants of gentrification (with (10) being dependent on the rest). These factors, which we've shorted-handed as *GentriFactors*, have immense weight to this quantitative definition of gentrification but at least some of these *GentriFactors* have been neglected by the Freeman Model, The Ellen & O'Regan Model, and The McKinnish Model. More specifically, *GentriFactors*, 3, 4, 5, 8, 9, and 10.

Nonetheless, for simplicity of definition, the only *GentriFactors* being used in this project (according to the models) are Median Household Income in both the locally and the metro area (neighboring New York counties), median home value in both locally and in the metro area, people with bachelor's degrees above the age of 24 from 2000-2020, and median year built of the property.

Methodology:

Our Data was collected from Harvard Dataverse[4] and Policy Map[5], a highly specific census tool that consolidates United States Census Bureau data into visual data. Sifting through the vast sea of census data from Policymap and Harvard Dataverse, we utilized inner join and consolidated our data into Pandas, creating a DataFrame containing only the necessary *GentriFactors*, as well as the percent change over time from 2000-2021 for Martin's determining factors to observe the trends over a 10-20 year period.

80% of the raw collected data is then pushed through the three gentrification models. If two out of three of the models indicate that the neighborhood has been gentrified then it will be considered "gentrified". If not, it will be classified as "non-gentrified". The "Gentrified or not" data will be put into a new dataframe. This DataFrame will be used to train a decision tree. The 20% of the raw data not pushed through the three models will be the test data for the machine learning methods k-nearest neighbors and decision trees. These will yield their own set of independent results on whether or not the neighborhood in question is in fact gentrified or not. This will help evaluate the accuracy of the machine learning models.

[4] Johnson, Glen, 2020, "Gentrification Index Data for NYC", https://doi.org/10.7910/DVN/O56ZMB, Harvard Dataverse, V1, UNF:6:eYFpP1YXevmb/OXLh+t6+Q== [fileUNF]

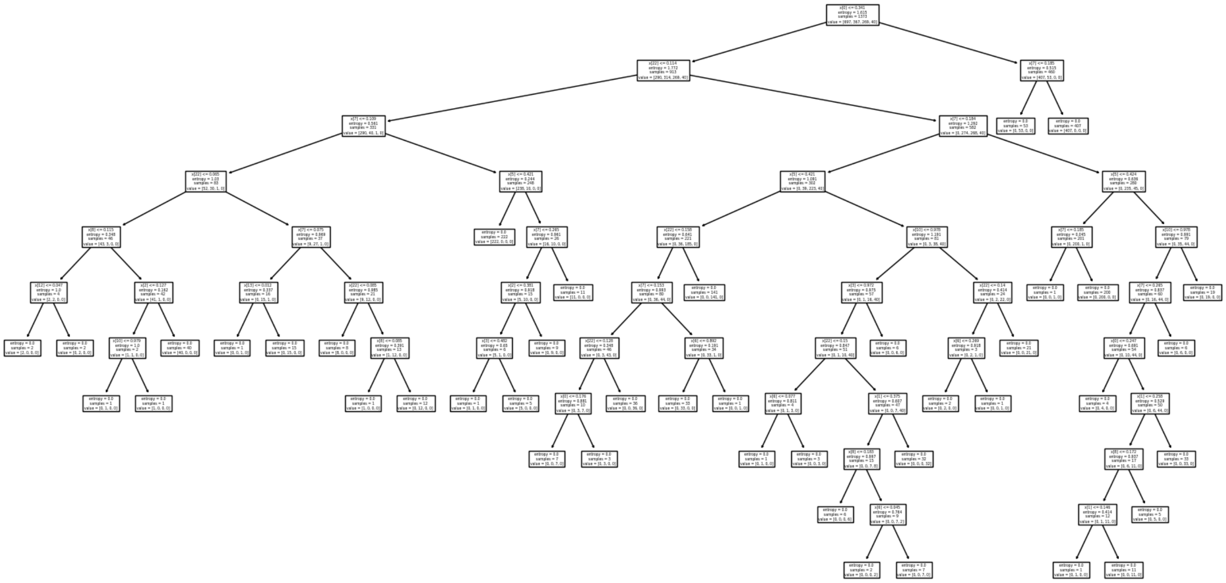
[5] "Home." PolicyMap, April 6, 2023. https://www.policymap.com/.

Results

Overall, the decision trees returned much more accurate results than k-nearest neighbors.

Decision trees: 96.5%

The Decision trees value represents an accuracy score for the train-test data sample. These are the results generated from the decision tree, which, when maximum depth was left untouched, returned a 96.22% accuracy value in terms of test results.



However, this was not the maximum accuracy score, which was achieved when maximum depth was set to 8. This yielded a 96.5% accuracy. As maximum depth was varied from 1 through 8, accuracy consistently showed signs of increment. However, once maximum depth was above 8, the accuracy remained constant at 96.22% no matter how the maximum depth was varied.

k-Nearest Neighbors: 67.44%

The k-Nearest Neighbors method produced some interesting results. Upon varying the number of neighbors, n, it was found that when n=1, the result was 59.3% accuracy, going against the theoretical answer of 100% when the test data is compared against one neighbor. This accuracy increases until n=13, to 67.44% accuracy, before it starts declining and stabilizes at n=331 at 56.1% accuracy, all the way until n=1373 which is the total size of the dataset. This could be explained by the fact that n=331 is a sizeable enough number of neighbors for the machine learning algorithm to satisfactorily produce the same result every time past that, even if there are more neighbors, which no longer alters the result.

Conclusion & Limitations

Ultimately, decision trees was the superior method in determining whether or not a neighborhood was gentrified, when compared against the three models used to define gentrification. It returned a much better accuracy rate as opposed to k-Nearest Neighbors. Decision trees could thus be a good predictor of whether or not neighborhoods may or may not become gentrified, given more data. We could have attempted to yield results that displayed test neighborhood names and their gentrification status based on the trained data. However, this was not possible due to the methods of machine learning used. Nonetheless, this is something that could be considered for future iterations of this research.

Citations

- “Gentrification.” Education. Accessed May 3, 2023. https://education.nationalgeographic.org/resource/gentrification/.
- “Home.” PolicyMap, April 6, 2023. https://www.policymap.com/.
- Johnson, Glen, 2020, "Gentrification Index Data for NYC", https://doi.org/10.7910/DVN/O56ZMB, Harvard Dataverse, V1, UNF:6:eYFpP1YXevmb/OXLh+t6+Q== [fileUNF]
- Martin, Richard W. “A Quantitative Approach to Gentrification: Determinants of ...” Accessed May 3, 2023. https://www.terry.uga.edu/sites/default/files/inline-files/Determinants_of_Gentrification.pdf.
- Vashi, Sonam. “How Do Researchers Measure Gentrification?” SaportaReport, May 20, 2019. https://saportareport.com/how-do-researchers-measure-gentrification/main-slider/sonam-vashi/#:~:text=The%20Freeman%20model%20classifies%20gentrification,than%20the%20metro%2Darea%20values.