Tufts University | DATA-200

Assessing the Status of Gentrification in Metropolitan Areas in the United States

Vishwesh Srinivasan, Vanessa Venkataraman, and Hanzhen Wang

AGENDA

01 Introduction

02 Literature Review

03 Data Description

04 Our Hypothesis

05 Models and their Results

06 Conclusions and Future Work

Introduction

“Gentrification is the process by which central urban neighborhoods that have undergone disinvestments and economic decline experience a reversal, reinvestment, and the in-migration of a relatively well-off middle- and upper middle-class population.” - Neil Smith

Positives of Gentrification: increased investment, commercial development, increased property values, improvement in economic opportunity

Negatives of Gentrification: high rate of displacement in communities, higher housing costs, higher property taxes, forced relocation, community cultural changes, changes in businesses, dislocation

Gentrification disproportionately impacts low-income families and people of color.

In order to counteract the negative impacts of gentrification, it is important to recognize which neighborhoods are susceptible to gentrification and which communities are most vulnerable to displacement.

Literature Review

Kolko (2007) studied the determinants of gentrification using the Neighborhood Change Database. His research supported the fact that low-income city neighborhoods closer to the city center and with older housing stock witness higher income growth.

Kolko (2009), in another study, evaluated the effect of employment location on gentrification. His findings suggested that an increase in the mean neighborhood job pay influences the tract's income growth, by which he inferred that people follow jobs.

Boustan et al. (2019) studied if condominium structure led to gentrification in central cities. They did not see a statistically significant association between the condo density and the typical characteristics of gentrified regions – income levels and education.

Rucks-Ahidiana (2021) studied the effect of racial composition on the way gentrification happens in a community. He found that the gentrifying tracts with residents of color, compared to tracts with white people, display more characteristics of poverty.

Literature Review

Rigolon & Németh (2019) took a broad approach and considered the effect of three different categories of variables – people, place, and policy on gentrification.

They used the criteria given by Chapple et al. (2017) to identify the areas susceptible to gentrification in 2000. Then they again used the definition Chapple et al. (2017) gave to identify the areas that had actually been gentrified in 2015.

Finally, they tested how all the variables they had selected in the three categories for their model are beneficial in predicting whether a gentrification susceptible region will get gentrified or not.

They tested their model in five Combined Statistical Areas (CSAs). These CSAs are Chicago, Los Angeles, New York, San Francisco, and Washington, DC.

Their research showed that all three categories assist in predicting the probability of gentrification of the tracts. Some of the significant factors were: race, distance from downtown, units older than 30 years and population density.

Data Description - Neighborhood Change Database (NCDB)

Census tracts are locally determined geographic units, typically including between 2,500 and 8,000 persons. Tracts are meant to approximate "neighborhoods" by capturing a group of residents with similar social characteristics, economic status, and housing conditions.

A metropolitan area or metro is a region that consists of a densely populated urban agglomeration and its surrounding territories sharing industries, commercial areas, transport network, infrastructures and housing.

A metro area usually comprises multiple principal cities, jurisdictions and municipalities: neighborhoods, townships, boroughs, cities, towns, exurbs, suburbs, counties, districts, as well as even states and nations like the euro districts. As social, economic and political institutions have changed, metropolitan areas have become key economic and political regions.

A powerful feature of the NCDB is its ability to match tracts across all the five census years, enabling users to examine changing tract characteristics between 1970 and 2010.

Since the boundaries of many tracts change between the decennial censuses (either splitting into several tracts, merging with other tracts, or appearing for the first time), a methodology has been developed to link such tracts and their associated data to standard geographic boundaries.

Data Description - Neighborhood Change Database (NCDB)

Classification of Variables

In accordance with Census Bureau conventions, NCDB variables are separated into two general groups: population and housing. Population variables are further classified into nine categories:

General Population Characteristics
Family Structure/Marriage
Mobility/Transportation
Education
Employment/Labor Market
Poverty/Public Assistance
Income and Earnings
Age Distribution
Language Ability

While housing variables are grouped into four categories:

Housing Tenure/Occupancy
Housing Characteristics/Utilities
Housing Costs/Affordability - Owners
Housing Costs/Affordability - Renters

Additionally, there is a category for geographic identifier variables.

Our Hypothesis

The three factors that contribute the most to determining if a neighborhood is susceptible to gentrification are...

- PEOPLE
- PLACE
- POLICY

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_{people} \cdot features_{people} + \beta_{place} \cdot features_{place} + \beta_{policy} \cdot features_{policy} \quad (1)$$

β_i – represents the set of coefficients for each feature under the i category

$features_i$ – represents the set of features under the i category

Our Hypothesis - Model design and Output Variable (Y)

We selected the following two regions for our analysis:

- New York-Northern New Jersey-Long Island, NY-NJ-PA Metropolitan Statistical Area (MSA)
- Los Angeles-Long Beach-Santa Ana, CA Metropolitan Statistical Area (MSA)

We choose to study the process of gentrification between the years 2000 and 2010. To do so, we use different variables from the year 2000 as predictors (X) and create the labels (Y) as 'Gentrified' (1) or 'Not Gentrified' (0).

We use the following definition to identify if a tract is gentrified or not in the year 2010:

- Change in median household income > Change in MSA median
- Change in % college educated > Change in MSA percentage
- Change in median gross rent > Change in MSA median

(OR)

- % Increase of home value > % increase in MSA median

(Change between the year 2000 and 2010. Given by Rigolon & Németh (2019), adapted by them from Chapple et al. (2017))

Our Hypothesis - Predictors (People)

SHRWHT0	Proportion of White population
SHRBLK0	Proportion of Black/African American population
SHRAMI0 = SHRAMI0N/SHR0D	Proportion of American Indian/Alaska Native population
SHRASN0 = SHRASN0N/SHR0D	Proportion of Asian population
SHRHIP0 = SHRHIP0N/SHR0D	Proportion of Native Hawaiian and other Pacific Islander population
SHRHSP0	Proportion of Hispanic/Latino population
ADULT0	Proportion of persons who are adults 18+ years old
CHILD0	Proportion of persons who are children under 18 years old
OLD0	Proportion of persons who are 65+ years old
MDFAMY0	Median family income last year ($)

Our Hypothesis - Predictors (Place)

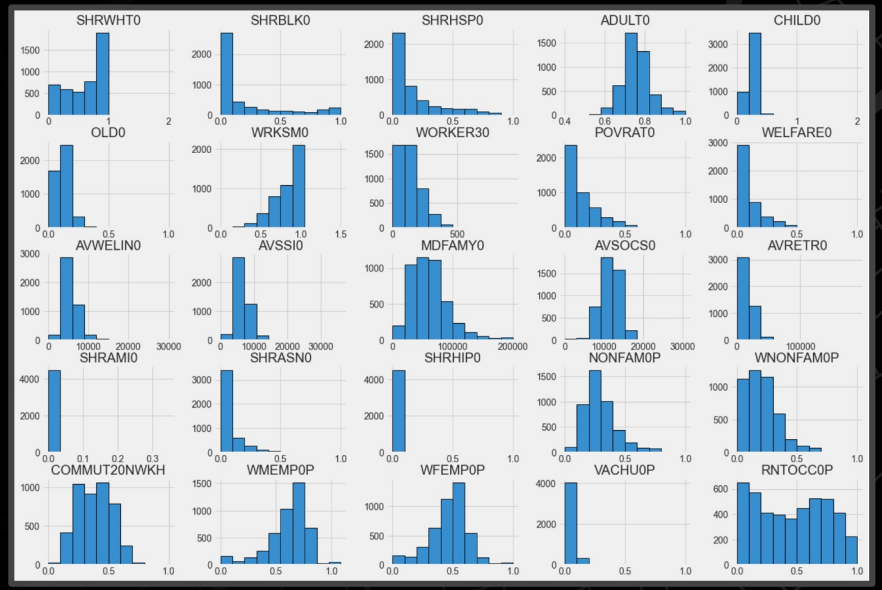
POVRAT0	Proportion of total persons below the poverty level last year
WFEMP0P = WFEMP0/WF16P0	Proportion of Employed White females 16+ years old
WMEMP0P = WMEMP0/WM16P0	Proportion of Employed White males 16+ years old
WORKER30	Families with 3+ workers last year
WRKSM0	Proportion of workers who live in a metro area and work within the same metro area
NONFAM0P = NONFAM0/NUMHHS0	Proportion of Nonfamily households
WNONFAM0P = WNONFAM0/NUMHHS0	Proportion of White nonfamily households
COMMUT20NWKH = (COMMUT20 - WKHOME0)/WRCNTY0D	Proportion of Workers 16+ years old not working from home and with travel time to work less than 25 minutes
VACHU0P = VACHU0/TOTHSUN0	Proportion of vacant housing units
RNTOCC0P = RNTOCC0/TOTHSUN0	Proportion of renter-occupied housing units

Our Hypothesis - Predictors (Policy)

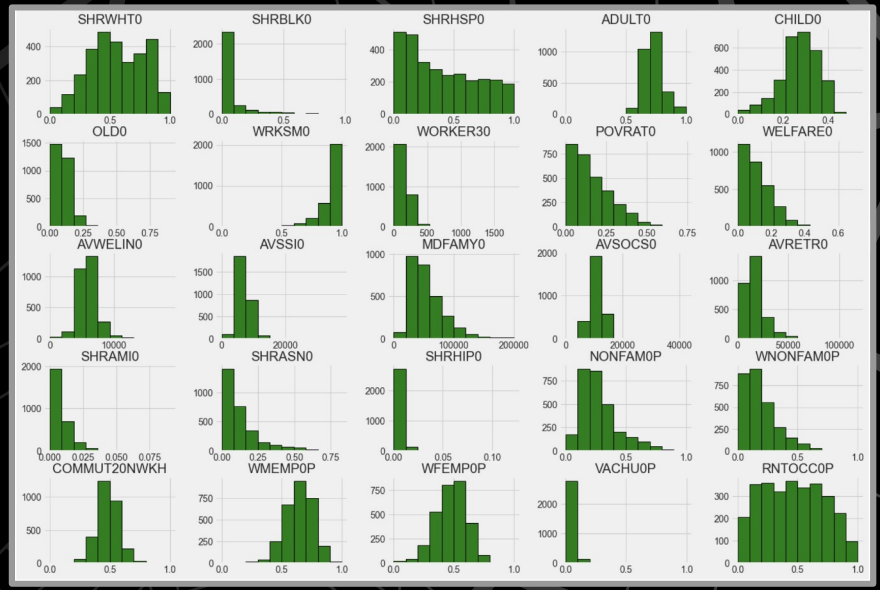
WELFARE0	Proportion of households with public assistance income (incl. SSI) last year
AVWELIN0	Average public assistance income for households on public assistance last year ($)
AVSSI0	Average Supplemental Security Income for households receiving SSI last year ($)
AVSOCS0	Average household social security income last year ($)
AVRETR0	Average household retirement income last year ($)

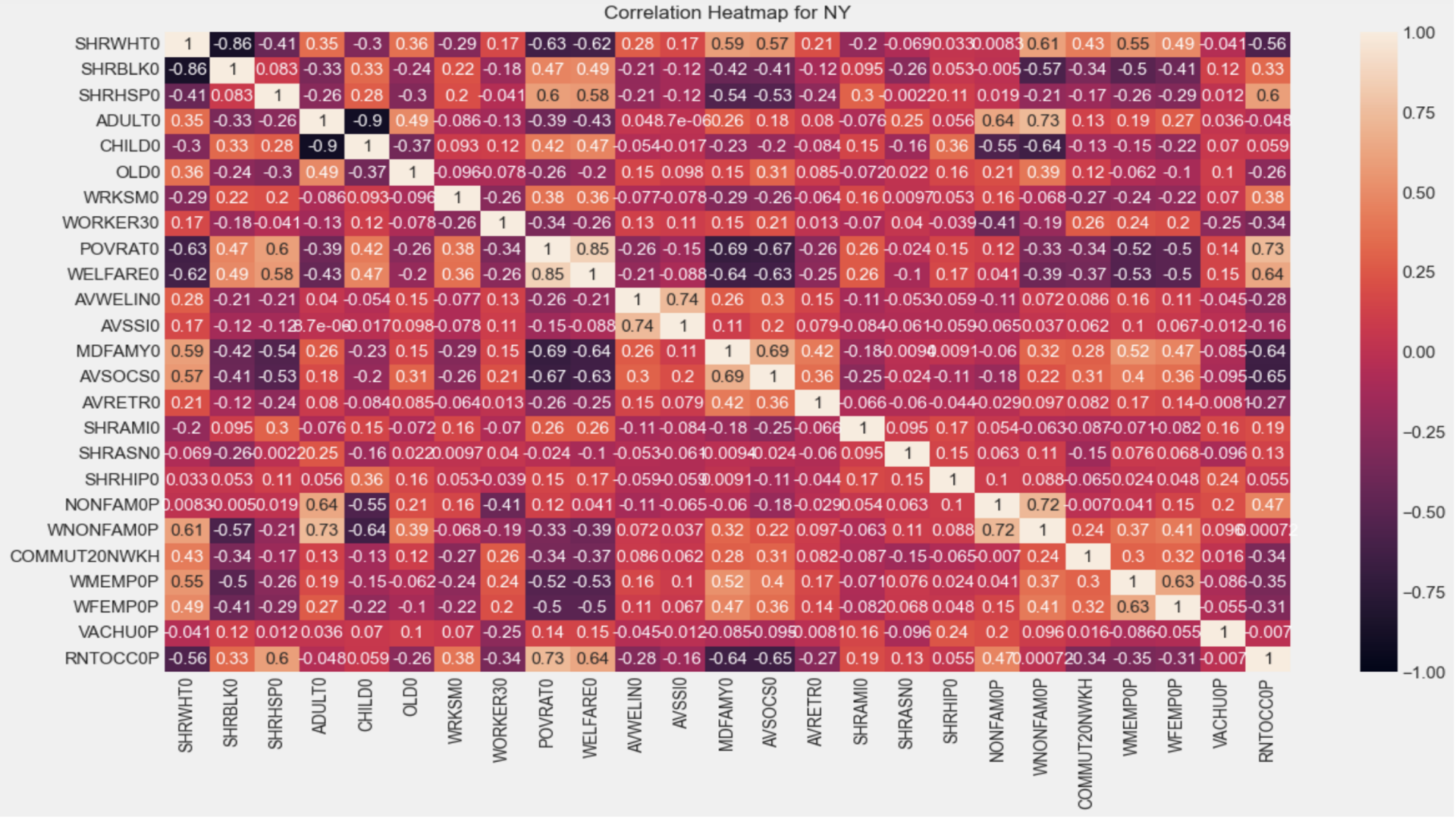
Data Visualizations

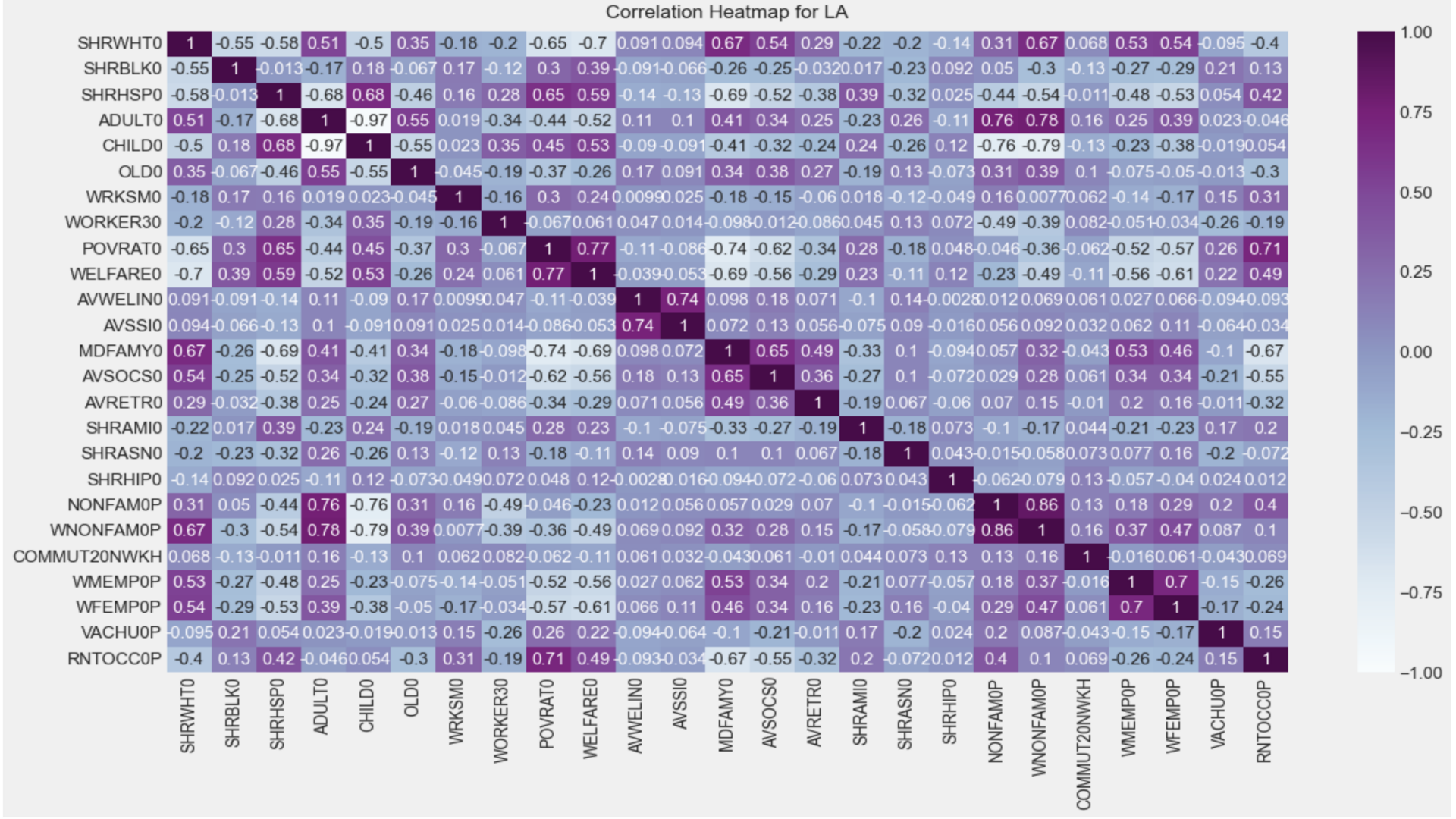
New York Metropolitan Area



Los Angeles Metropolitan Area




Correlation Heatmap for NY
SHRWHT0 1 -0.86 -0.41 0.35 -0.3 0.36 -0.29 0.17 -0.63 -0.62 0.28 0.17 0.59 0.57 0.21 -0.2 -0.069 0.033 0.0083 0.61 0.43 0.55 0.49 -0.041 -0.56
SHRBLK0 -0.86 1 0.083 -0.33 0.33 -0.24 0.22 -0.18 0.47 0.49 -0.21 -0.12 -0.42 -0.41 -0.12 0.095 -0.26 0.053 -0.005 -0.57 -0.34 -0.5 -0.41 0.12 0.33
SHRHSP0 -0.41 0.083 1 -0.26 0.28 -0.3 0.2 -0.041 0.6 0.58 -0.21 -0.12 -0.54 -0.53 -0.24 0.3 -0.0022 0.11 0.019 -0.21 -0.17 -0.26 -0.29 0.012 0.6
ADULT0 0.35 -0.33 -0.26 1 -0.9 0.49 -0.086 -0.13 -0.39 -0.43 0.048 7e-06 0.26 0.18 0.08 -0.076 0.25 0.056 0.64 0.73 0.13 0.19 0.27 0.036 -0.048
CHILD0 -0.3 0.33 0.28 -0.9 1 -0.37 0.093 0.12 0.42 0.47 -0.054 -0.017 -0.23 -0.2 -0.084 0.15 -0.16 0.36 -0.55 -0.64 -0.13 -0.15 -0.22 0.07 0.059
OLD0 0.36 -0.24 -0.3 0.49 -0.37 1 -0.096 -0.078 -0.26 -0.2 0.15 0.098 0.15 0.31 0.085 -0.072 0.022 0.16 0.21 0.39 0.12 -0.062 -0.1 0.1 -0.26
WRKSM0 -0.29 0.22 0.2 -0.086 0.093 -0.096 1 -0.26 0.38 0.36 -0.077 -0.078 -0.29 -0.26 -0.064 0.16 0.0097 0.053 0.16 -0.068 -0.27 -0.24 -0.22 0.07 0.38
WORKER30 0.17 -0.18 -0.041 -0.13 0.12 -0.078 -0.26 1 -0.34 -0.26 0.13 0.11 0.15 0.21 0.013 -0.07 0.04 -0.039 -0.41 -0.19 0.26 0.24 0.2 -0.25 -0.34
POVRAT0 -0.63 0.47 0.6 -0.39 0.42 -0.26 0.38 -0.34 1 0.85 -0.26 -0.15 -0.69 -0.67 -0.26 0.26 -0.024 0.15 0.12 -0.33 -0.34 -0.52 -0.5 0.14 0.73
WELFARE0 -0.62 0.49 0.58 -0.43 0.47 -0.2 0.36 -0.26 0.85 1 -0.21 -0.088 -0.64 -0.63 -0.25 0.26 -0.1 0.17 0.041 -0.39 -0.37 -0.53 -0.5 0.15 0.64
AVWELIN0 0.28 -0.21 -0.21 0.04 -0.054 0.15 -0.077 0.13 -0.26 -0.21 1 0.74 0.26 0.3 0.15 -0.11 -0.053 -0.059 -0.11 0.072 0.086 0.16 0.11 -0.045 -0.28
AVSSI0 0.17 -0.12 -0.12 8.7e-06 0.017 0.098 -0.078 0.11 -0.15 -0.088 0.74 1 0.11 0.2 0.079 -0.084 -0.061 -0.059 -0.065 0.037 0.062 0.1 0.067 -0.012 -0.16
MDFAMY0 0.59 -0.42 -0.54 0.26 -0.23 0.15 -0.29 0.15 -0.69 -0.64 0.26 0.11 1 0.69 0.42 -0.18 -0.0094 -0.0091 -0.06 0.32 0.28 0.52 0.47 -0.085 -0.64
AVSOCS0 0.57 -0.41 -0.53 0.18 -0.2 0.31 -0.26 0.21 -0.67 -0.63 0.3 0.2 0.69 1 0.36 -0.25 -0.024 -0.11 -0.18 0.22 0.31 0.4 0.36 -0.095 -0.65
AVRETR0 0.21 -0.12 -0.24 0.08 -0.084 0.085 -0.064 0.013 -0.26 -0.25 0.15 0.079 0.42 0.36 1 -0.066 -0.06 -0.044 -0.029 0.097 0.082 0.17 0.14 -0.0081 -0.27
SHRAMI0 -0.2 0.095 0.3 -0.076 0.15 -0.072 0.16 -0.07 0.26 0.26 -0.11 -0.084 -0.18 -0.25 -0.066 1 0.095 0.17 0.054 -0.063 -0.087 -0.071 -0.082 0.16 0.19
SHRASN0 -0.069 -0.26 -0.0022 0.25 -0.16 0.022 0.0097 0.04 -0.024 -0.1 -0.053 -0.061 0.0094 0.024 -0.06 0.095 1 0.15 0.063 0.11 -0.15 0.076 0.068 -0.096 0.13
SHRHIP0 0.033 0.053 0.11 0.056 0.36 0.16 0.053 -0.039 0.15 0.17 -0.059 -0.059 0.0091 -0.11 -0.044 0.17 0.15 1 0.1 0.088 -0.065 0.024 0.048 0.24 0.055
NONFAM0P 0.0083 -0.005 0.019 0.64 -0.55 0.21 0.16 -0.41 0.12 0.041 -0.11 -0.065 -0.06 -0.18 -0.029 0.054 0.063 0.1 1 0.72 -0.007 0.041 0.15 0.2 0.47
WNONFAM0P 0.61 -0.57 -0.21 0.73 -0.64 0.39 -0.068 -0.19 -0.33 -0.39 0.072 0.037 0.32 0.22 0.097 -0.063 0.11 0.088 0.72 1 0.24 0.37 0.41 0.096 0.0007
COMMUT20NWKH 0.43 -0.34 -0.17 0.13 -0.13 0.12 -0.27 0.26 -0.34 -0.37 0.086 0.062 0.28 0.31 0.082 -0.087 -0.15 -0.065 -0.007 0.24 1 0.3 0.32 0.016 -0.34
WMEMP0P 0.55 -0.5 -0.26 0.19 -0.15 -0.062 -0.24 0.24 -0.52 -0.53 0.16 0.1 0.52 0.4 0.17 -0.071 0.076 0.024 0.041 0.37 0.3 1 0.63 -0.086 -0.35
WFEMP0P 0.49 -0.41 -0.29 0.27 -0.22 -0.1 -0.22 0.2 -0.5 -0.5 0.11 0.067 0.47 0.36 0.14 -0.082 0.068 0.048 0.15 0.41 0.32 0.63 1 -0.055 -0.31
VACHU0P -0.041 0.12 0.012 0.036 0.07 0.1 0.07 -0.25 0.14 0.15 -0.045 -0.012 -0.085 -0.095 0.0081 0.16 -0.096 0.24 0.2 0.096 0.016 -0.086 -0.055 1 -0.007
RNTOCC0P -0.56 0.33 0.6 -0.048 0.059 -0.26 0.38 -0.34 0.73 0.64 -0.28 -0.16 -0.64 -0.65 -0.27 0.19 0.13 0.055 0.47 0.0007 0.34 -0.35 -0.31 -0.007 1
SHRWHT0 SHRBLK0 SHRHSP0 ADULT0 CHILD0 OLD0 WRKSM0 WORKER30 POVRAT0 WELFARE0 AVWELIN0 AVSSI0 MDFAMY0 AVSOCS0 AVRETR0 SHRAMI0 SHRASN0 SHRHIP0 NONFAM0P WNONFAM0P COMMUT20NWKH WMEMP0P WFEMP0P VACHU0P RNTOCC0P
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Correlation Heatmap for LA
SHRWHT0 1 -0.55 -0.58 0.51 -0.5 0.35 -0.18 -0.2 -0.65 -0.7 0.091 0.094 0.67 0.54 0.29 -0.22 -0.2 -0.14 0.31 0.67 0.068 0.53 0.54 -0.095 -0.4
SHRBLK0 -0.55 1 -0.013 -0.17 0.18 -0.067 0.17 -0.12 0.3 0.39 -0.091 -0.066 -0.26 -0.25 -0.032 0.017 -0.23 0.092 0.05 -0.3 -0.13 -0.27 -0.29 0.21 0.13
SHRHSP0 -0.58 -0.013 1 -0.68 0.68 -0.46 0.16 0.28 0.65 0.59 -0.14 -0.13 -0.69 -0.52 -0.38 0.39 -0.32 0.025 -0.44 -0.54 -0.011 -0.48 -0.53 0.054 0.42
ADULT0 0.51 -0.17 -0.68 1 -0.97 0.55 0.019 -0.34 -0.44 -0.52 0.11 0.1 0.41 0.34 0.25 -0.23 0.26 -0.11 0.76 0.78 0.16 0.25 0.39 0.023 -0.046
CHILD0 -0.5 0.18 0.68 -0.97 1 -0.55 0.023 0.35 0.45 0.53 -0.09 -0.091 -0.41 -0.32 -0.24 0.24 -0.26 0.12 -0.76 -0.79 -0.13 -0.23 -0.38 -0.019 0.054
OLD0 0.35 -0.067 -0.46 0.55 -0.55 1 -0.045 -0.19 -0.37 -0.26 0.17 0.091 0.34 0.38 0.27 -0.19 0.13 -0.073 0.31 0.39 0.1 -0.075 -0.05 -0.013 -0.3
WRKSM0 -0.18 0.17 0.16 0.019 0.023 -0.045 1 -0.16 0.3 0.24 0.0099 0.025 -0.18 -0.15 -0.06 0.018 -0.12 -0.049 0.16 0.0077 0.062 -0.14 -0.17 0.15 0.31
WORKER30 -0.2 -0.12 0.28 -0.34 0.35 -0.19 -0.16 1 -0.067 0.061 0.047 0.014 -0.098 -0.012 -0.086 0.045 0.13 0.072 -0.49 -0.39 0.082 -0.051 -0.034 -0.26 -0.19
POVRAT0 -0.65 0.3 0.65 -0.44 0.45 -0.37 0.3 -0.067 1 0.77 -0.11 -0.086 -0.74 -0.62 -0.34 0.28 -0.18 0.048 -0.046 -0.36 -0.062 -0.52 -0.57 0.26 0.71
WELFARE0 -0.7 0.39 0.59 -0.52 0.53 -0.26 0.24 0.061 0.77 1 -0.039 -0.053 -0.69 -0.56 -0.29 0.23 -0.11 0.12 -0.23 -0.49 -0.11 -0.56 -0.61 0.22 0.49
AVWELIN0 0.091 -0.091 -0.14 0.11 -0.09 0.17 0.0099 0.047 -0.11 -0.039 1 0.74 0.098 0.18 0.071 -0.1 0.14 -0.0028 0.012 0.069 0.061 0.027 0.066 -0.094 -0.093
AVSSI0 0.094 -0.066 -0.13 0.1 -0.091 0.091 0.025 0.014 -0.086 -0.053 0.74 1 0.072 0.13 0.056 -0.075 0.09 -0.016 0.056 0.092 0.032 0.062 0.11 -0.064 -0.034
MDFAMY0 0.67 -0.26 -0.69 0.41 -0.41 0.34 -0.18 -0.098 -0.74 -0.69 0.098 0.072 1 0.65 0.49 -0.33 0.1 -0.094 0.057 0.32 -0.043 0.53 0.46 -0.1 -0.67
AVSOCS0 0.54 -0.25 -0.52 0.34 -0.32 0.38 -0.15 -0.012 -0.62 -0.56 0.18 0.13 0.65 1 0.36 -0.27 0.1 -0.072 0.029 0.28 0.061 0.34 0.34 -0.21 -0.55
AVRETR0 0.29 -0.032 -0.38 0.25 -0.24 0.27 -0.06 -0.086 -0.34 -0.29 0.071 0.056 0.49 0.36 1 -0.19 0.067 -0.06 0.07 0.15 -0.01 0.2 0.16 -0.011 -0.32
SHRAMI0 -0.22 0.017 0.39 -0.23 0.24 -0.19 0.018 0.045 0.28 0.23 -0.1 -0.075 -0.33 -0.27 -0.19 1 -0.18 0.073 -0.1 -0.17 0.044 -0.21 -0.23 0.17 0.2
SHRASN0 -0.2 -0.23 -0.32 0.26 -0.26 0.13 -0.12 0.13 -0.18 -0.11 0.14 0.09 0.1 0.1 0.067 -0.18 1 0.043 -0.015 -0.058 0.073 0.077 0.16 -0.2 -0.072
SHRHIP0 -0.14 0.092 0.025 -0.11 0.12 -0.073 -0.049 0.072 0.048 0.12 -0.0028 0.016 -0.094 -0.072 -0.06 0.073 0.043 1 -0.062 -0.079 0.13 -0.057 -0.04 0.024 0.012
NONFAM0P 0.31 0.05 -0.44 0.76 -0.76 0.31 0.16 -0.49 -0.046 -0.23 0.012 0.056 0.057 0.029 0.07 -0.1 -0.015 -0.062 1 0.86 0.13 0.18 0.29 0.2 0.4
WNONFAM0P 0.67 -0.3 -0.54 0.78 -0.79 0.39 0.0077 -0.39 -0.36 -0.49 0.069 0.092 0.32 0.28 0.15 -0.17 -0.058 -0.079 0.86 1 0.16 0.37 0.47 0.087 0.1
COMMUT20NWKH 0.068 -0.13 -0.011 0.16 -0.13 0.1 0.062 0.082 -0.062 -0.11 0.061 0.032 -0.043 0.061 -0.01 0.044 0.073 0.13 0.13 0.16 1 -0.016 0.061 -0.043 0.069
WMEMP0P 0.53 -0.27 -0.48 0.25 -0.23 -0.075 -0.14 -0.051 -0.52 -0.56 0.027 0.062 0.53 0.34 0.2 -0.21 0.077 -0.057 0.18 0.37 -0.016 1 0.7 -0.15 -0.26
WFEMP0P 0.54 -0.29 -0.53 0.39 -0.38 -0.05 -0.17 -0.034 -0.57 -0.61 0.066 0.11 0.46 0.34 0.16 -0.23 0.16 -0.04 0.29 0.47 0.061 0.7 1 -0.17 -0.24
VACHU0P -0.095 0.21 0.054 0.023 -0.019 -0.013 0.15 -0.26 0.26 0.22 -0.094 -0.064 -0.1 -0.21 -0.011 0.17 -0.2 0.024 0.2 0.087 -0.043 -0.15 -0.17 1 0.15
RNTOCC0P -0.4 0.13 0.42 -0.046 0.054 -0.3 0.31 -0.19 0.71 0.49 -0.093 -0.034 -0.67 -0.55 -0.32 0.2 -0.072 0.012 0.4 0.1 0.069 -0.26 -0.24 0.15 1
SHRWHT0
SHRBLK0
SHRHSP0
ADULT0
CHILD0
OLD0
WRKSM0
WORKER30
POVRAT0
WELFARE0
AVWELIN0
AVSSI0
MDFAMY0
AVSOCS0
AVRETR0
SHRAMI0
SHRASN0
SHRHIP0
NONFAM0P
WNONFAM0P
COMMUT20NWKH
WMEMP0P
WFEMP0P
VACHU0P
RNTOCC0P
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Models and their Results

Data Composition:

- New York MSA: 990 tracts gentrified, 3522 tracts not gentrified
- Los Angeles MSA: 583 tracts gentrified, 2343 tracts not gentrified

Modeling Process:

1. Normalized the variables which were not in proportions (for step 5)
2. Built a logistic regression model
3. Identified the significant variables out of all the predictors
4. Used the identified variables to build a random forest classifier
5. Used ADASYN (Adaptive Synthetic Sampling) technique to artificially generate more data of the minority class
6. Used the balanced dataset to build a random forest classifier
7. Compared the results obtained

Models and their Results (New York MSA) - Logistic Regression Model

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-12.9445	9.2373	-1.401	0.16112	
SHRWHT0	-1.4586	1.2977	-1.124	0.26099	
SHRBLK0	-1.3948	1.3207	-1.056	0.29092	
SHRHSP0	-0.1774	0.7458	-0.238	0.81194	
ADULT0	14.6925	9.6997	1.515	0.12984	
CHILD0	10.8847	9.6264	1.131	0.25817	
OLD0	-4.3342	0.8668	-5.000	5.73e-07	***
WRKSM0	0.8315	0.2840	2.928	0.00342	**
WORKER30	0.3824	0.5380	0.711	0.47724	
POVRAT0	0.4840	0.8481	0.571	0.56822	
WELFARE0	-0.8786	0.9698	-0.906	0.36492	
AVWELIN0	0.5912	0.9110	0.649	0.51636	
AVSSI0	-0.5125	0.9145	-0.560	0.57518	
MDFAMY0	-2.8647	0.5195	-5.514	3.50e-08	***
AVSOCS0	0.8741	0.8578	1.019	0.30820	
AVRETR0	0.5164	0.7882	0.655	0.51233	
SHRAMI0	-5.1414	6.3257	-0.813	0.41634	
SHRASN0	-1.7052	1.3919	-1.225	0.22052	
SHRHIP0	-32.6760	16.0583	-2.035	0.04187	*
NONFAM0P	2.1535	0.9342	2.305	0.02116	*
WNONFAM0P	1.5820	0.9462	1.672	0.09455	.
COMMUT20NWKH	-1.7133	0.3605	-4.752	2.01e-06	***
WMEMP0P	0.1373	0.3554	0.386	0.69933	
WFEMP0P	0.8382	0.3965	2.114	0.03454	*
VACHU0P	1.1432	0.5732	1.994	0.04612	*
RNTOCC0P	-2.8206	0.3834	-7.357	1.88e-13	***

Confusion Matrix and Statistics

Prediction \ Reference	0	1
0	692	184
1	14	13

Accuracy : 0.7807
95% CI : (0.7523, 0.8073)
No Information Rate : 0.7818
P-Value [Acc > NIR] : 0.551

Kappa : 0.067

Mcnemar's Test P-Value : <2e-16

Sensitivity : 0.06599
Specificity : 0.98017
Pos Pred Value : 0.48148
Neg Pred Value : 0.78995
Precision : 0.48148
Recall : 0.06599
F1 : 0.11607
Prevalence : 0.21816
Detection Rate : 0.01440
Detection Prevalence : 0.02990
Balanced Accuracy : 0.52308

'Positive' Class : 1

Models and their Results (New York MSA) - Random Forest Classifier

```
[1] "Before Data Resampling"
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 681 177
         1  25  20

               Accuracy : 0.7763
                 95% CI : (0.7477, 0.8031)
    No Information Rate : 0.7818
    P-Value [Acc > NIR] : 0.6734

                  Kappa : 0.0916

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.10152
            Specificity : 0.96459
         Pos Pred Value : 0.44444
         Neg Pred Value : 0.79371
              Precision : 0.44444
                 Recall : 0.10152
                     F1 : 0.16529
             Prevalence : 0.21816
         Detection Rate : 0.02215
   Detection Prevalence : 0.04983
      Balanced Accuracy : 0.53306

       'Positive' Class : 1
```

```
[1] "After Data Resampling"
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 570 119
         1 136  78

               Accuracy : 0.7176
                 95% CI : (0.687, 0.7468)
    No Information Rate : 0.7818
    P-Value [Acc > NIR] : 1.0000

                  Kappa : 0.1972

 Mcnemar's Test P-Value : 0.3164

            Sensitivity : 0.39594
            Specificity : 0.80737
         Pos Pred Value : 0.36449
         Neg Pred Value : 0.82729
              Precision : 0.36449
                 Recall : 0.39594
                     F1 : 0.37956
             Prevalence : 0.21816
         Detection Rate : 0.08638
   Detection Prevalence : 0.23699
      Balanced Accuracy : 0.60165

       'Positive' Class : 1
```

Models and their Results (Los Angeles MSA) - Logistic Regression Model

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.78803    4.22208   0.423   0.6719
SHRNWT0       3.85453    1.90171   2.027   0.0427 *
SHRWHT0       3.08880    1.82588   1.692   0.0907 .
SHRHSP0       0.95846    1.07062   0.895   0.3707
ADULT0       -4.81271    4.46384  -1.078   0.2810
CHILD0       -9.46631    4.53233  -2.089   0.0367 *
OLD0         -1.99893    1.37455  -1.454   0.1459
WRKSM0        1.46492    0.67769   2.162   0.0306 *
WORKER30     -0.34712    1.47142  -0.236   0.8135
POVRAT0      -6.69963    1.43379  -4.673 2.97e-06 ***
WELFARE0     -0.35601    1.51817  -0.235   0.8146
AVWELIN0     -1.06456    0.96546  -1.103   0.2702
AVSSI0       -0.11184    1.36278  -0.082   0.9346
MDFAMY0      -4.31326    0.95785  -4.503 6.70e-06 ***
AVSOCS0      -0.18423    1.43471  -0.128   0.8978
AVRETR0      -1.28245    0.84323  -1.521   0.1283
NONFAM0P     -0.96935    1.45259  -0.667   0.5046
WNONFAM0P     1.52376    1.55486   0.980   0.3271
COMMUT20NWKH  0.71423    0.64205   1.112   0.2660
WMEMP0P       1.22875    0.80505   1.526   0.1269
WFEMP0P       0.66118    0.80576   0.821   0.4119
VACHU0P      -1.00025    1.52945  -0.654   0.5131
RNTOCC0P     -0.05735    0.50697  -0.113   0.9099
---
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 454 116
         1  11   5

               Accuracy : 0.7833
                 95% CI : (0.7477, 0.816)
    No Information Rate : 0.7935
    P-Value [Acc > NIR] : 0.7482

                  Kappa : 0.026

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.041322
            Specificity : 0.976344
         Pos Pred Value : 0.312500
         Neg Pred Value : 0.796491
              Precision : 0.312500
                 Recall : 0.041322
                     F1 : 0.072993
             Prevalence : 0.206485
         Detection Rate : 0.008532
   Detection Prevalence : 0.027304
      Balanced Accuracy : 0.508833

       'Positive' Class : 1
```

Models and their Results (Los Angeles MSA) - Random Forest Classifier

```
[1] "Before Data Resampling"
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 435 109
         1  30  12

               Accuracy : 0.7628
                 95% CI : (0.7262, 0.7967)
    No Information Rate : 0.7935
    P-Value [Acc > NIR] : 0.9689

                  Kappa : 0.0457

 Mcnemar's Test P-Value : 3.694e-11

            Sensitivity : 0.09917
            Specificity : 0.93548
         Pos Pred Value : 0.28571
         Neg Pred Value : 0.79963
              Precision : 0.28571
                 Recall : 0.09917
                     F1 : 0.14724
             Prevalence : 0.20648
         Detection Rate : 0.02048
   Detection Prevalence : 0.07167
      Balanced Accuracy : 0.51733

       'Positive' Class : 1
```

```
[1] "After Data Resampling"
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 367  66
         1  98  55

               Accuracy : 0.7201
                 95% CI : (0.6819, 0.7562)
    No Information Rate : 0.7935
    P-Value [Acc > NIR] : 0.99999

                  Kappa : 0.2221

 Mcnemar's Test P-Value : 0.01549

            Sensitivity : 0.45455
            Specificity : 0.78925
         Pos Pred Value : 0.35948
         Neg Pred Value : 0.84758
              Precision : 0.35948
                 Recall : 0.45455
                     F1 : 0.40146
             Prevalence : 0.20648
         Detection Rate : 0.09386
   Detection Prevalence : 0.26109
      Balanced Accuracy : 0.62190

       'Positive' Class : 1
```

Conclusions

From the '**People**' category, we find the following variables statistically significant:

- We find '**race**' to be a significant variable, in NY MSA, just the proportion of Native Hawaiian and other Pacific Islander population, whereas in LA MSA, along with this proportion of population, the proportion of Black, American Indians and Asians are also significant.
- In NY MSA, we find the proportion of 65+ years significant whereas in LA MSA, we find the proportion of children under 18+ to be significant.
- We find the '**median family income**' to be significant for both the NY and LA MSAs.

From the '**Place**' category, we find the following variables statistically significant:

- We find the proportion of people living in the same metro area they are working in to be significant in both the NY and LA MSAs. In addition in NY MSA, we find that the proportion of people who travel less than 25 minutes daily to work also significant.

Conclusions

- In LA MSA, the proportion of people under the poverty level also to be significant.
- In NY MSA, variables like the proportion of Nonfamily households, proportion of employed white females, proportion of vacant and rented households are significant.

From the '**Policy**' category, we do not find any of the variables that we considered to be statistically significant.

Future Work

- To study the effect of other variables which we did not consider from the NCDB. These new variables could act as strong predictors and improve the classification model's performance.

- We have very limited variables pertaining to 'Policy' in the NCDB, hence we can explore other datasets which can give information about development programs and study their effect on the process of gentrification.

- Instead of having a binary output variable - whether a region will get gentrified or not, we could possibly change the output variable to a 'gentrification score,' which represents the degree of gentrification.


Q&A

Thank you!

Vishwesh Srinivasan, Vanessa Venkataraman, and Hanzhen Wang