Assessing the status of gentrification in metropolitan areas in the United States

Vishwesh Srinivasan, Vanessa Venkataraman, and Hanzhen Wang

22 December 2022

Abstract

Gentrification is a widely studied topic to help policymakers design inclusive policies to minimize the adverse effects of gentrification, like displacement of poor communities and people of color. We hypothesize that the factors leading to the gentrification of a region can be broadly classified into three categories – people, place, and policy. We test this in two metropolitan areas – New York and Los Angeles. We use classification machine learning algorithms to predict the likelihood of each census tract in the two mentioned metropolitan regions getting gentrified in the period of ten years. Under the people category, we found the following predictors helpful – race, age demographics, and median family income. Under the place category, we found the following predictors to be beneficial – the proportion of people living in the same metro area they are working, the proportion of people under the poverty level only for Los Angeles, and the proportion of nonfamily, rented, and vacant households only for New York. We found none of the policy variables to be statistically significant.

1 Introduction

Gentrification is defined by Neil Smith, a professor of geography and anthropology, as "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" (Smith, 1998). The word "gentrification" is often associated with positive change due to its meaning being related to the transformation of neighborhoods from low economic value to high economic value (Centers for Disease Control and Prevention, n.d.). Although gentrification has many positive aspects, such as increased investment, commercial development, increased property values, and improvements in economic opportunity, many negative impacts go along with the process that historical definitions leave out (*What are gentrification and displacement*, n.d.).

The gentrification process causes a high displacement rate in communities due to higher housing costs and property taxes (Centers for Disease Control and Prevention, n.d.). Low-income residents who can no longer afford to remain in their homes are forced to relocate as wealthier residents replace them. This process alters many aspects of the community, specifically cultural aspects and neighborhood character (*Understanding Gentrification and Displacement*, n.d.). Gentrification-induced displacement has historically had the most significant adverse impact on low-income communities and people of color. The process of gentrification has traditionally been shown to increase the displacement of low-income families and people of color as wealthier, predominantly white individuals move into the neighborhood (*What are gentrification and displacement*, n.d.). This process also alters businesses in the area, forcing them to shift their strategies to appeal to new residents. This directly impacts the culture in the neighborhood and often causes remaining residents to feel dislocated even though they have not left their homes (*Understanding Gentrification and Displacement*, n.d.).

Gentrification and gentrification-induced displacement are processes rooted in the unequal treatment of people of color and low-income communities (*What are gentrification and displacement*, n.d.). It is essential to bring attention to the negative impacts of these processes on minority groups and find solutions so that residents of gentrified areas can remain in their homes and enjoy the positive developments gentrification brings to neighborhoods (*Understanding Gentrification and Displacement*, n.d.). To counteract the negative impacts of gentrification, it is crucial to recognize which areas are susceptible to gentrification and which communities are most vulnerable to displacement. The goal is to be proactive and create effective strategies to prevent displacement during gentrification. This can be most effective if there is a way to predict which neighborhoods are susceptible to gentrification ahead of time.

We hypothesize that the three factors that contribute the most to determining if a neighborhood is susceptible to gentrification are people, place, and policy. More specifically, the neighborhood's demographics, the neighborhood's location, and the current policies in the area are all main determining variables of gentrification. We use classification machine learning techniques to determine whether a neighborhood will get gentrified between 2000 and 2010. More precisely, we use the conditions of an area in the year 2000 to predict if the place would get gentrified in the year 2010.

2 Literature Review

Gentrification is a common phenomenon studied by policymakers to better design policies that will help to foster inclusive growth of people living in gentrification-susceptible regions. Numerous factors lead to gentrification, as discussed in the previous section. Various factors have been studied in the past, and few studies focused on the impact of one or more aspects in isolation. Others have taken a holistic approach, included diverse factors, and studied their effects on gentrification. A few of the studies have been briefly explained in this section.

Various researchers have defined factors that are commonly seen in gentrification-susceptible regions. We have included the most recent two sources. According to Bates (2013), "vulnerable" areas include those with a large proportion of renters, low-income residents, persons of color, and residents who have not attended college. Chapple et al. (2017) defined a very similar rule-based system, in which he states that a particular region is "vulnerable" to gentrification if that area satisfies three out of the following four criteria: low-income households, a large proportion of residents have not attended college, high non-white population and high ratio of residents living in rented spaces.

Rigolon & Németh (2019) have argued that the definition given by Chapple et al. (2017) is better in many ways compared to the others in the past because of the following reasons:

(i) It comprises a comprehensive set of parameters about racial/ethnic identity, residence, and socioeconomic conditions
(ii) Parameters like income and the proportion of people living in rented spaces show gentrification susceptible as linked to past cycles of disinvestment
(iii) The definition is one of the recent definitions, and hence it serves as a generalized version of all the past attempts to define gentrification for different regions
(iv) It contains variables that can all be obtained through the U.S. Census Bureau, enabling the application of gentrification susceptibility in a cross-national study.
(v) This definition seeks to capture similarities across different regions, even though gentrification susceptibility and gentrification differ by geographical context.

Similarly, there are various methods to ascertain whether a particular area has actually been gentrified, and the most holistic one is defined by Chapple et al. (2017). He and his collaborators argued that a specific region is gentrified when it witnesses growth economically - an increase in housing prices both in terms of rent and cost of buying or constructing a house, an increase in the proportion of white people, an increase in the number of people receiving a college education and subsequently the median income of the residents. This definition is holistic since it includes variables from an individual's socioeconomic status, race, and external environmental conditions.

Kolko (2007) studied the determinants of gentrification using the Neighborhood Change Database. His study supported the fact established by previous studies that low-income city neighborhoods closer to the city center and with older housing stock witness higher income growth. He identified differences in the applicability of these models to different regions. His results showed that the factors of closeness to the city center and having older housing stock accumulated more wealth for the south and midwest regions.

On the other hand, the closeness to the city center remained a significant factor for northeast regions, but the presence of older housing stock did not matter. In the west, proximity to the city center was not statistically substantial, and the black population percentage had a positive and statistically significant effect. The income of residents of a tract had the same impact across all the regions. He stated that including this factor does not reduce the relationship between gentrification and either proximity or aging stock.

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. Also, he saw a strong relationship between household income and job salary in tracts closer to the CBD (central business district) and large metros. He argued that this phenomenon is because of the high commuting cost per mile. Furthermore, he stated that within 2 miles of CBD (central business district), there is a strong relationship between the job salary and the tract's income. Thus, he concluded that the relationship between job location and gentrification is significant for areas nearer to downtown and for regions in larger metropolitan areas.

Rucks-Ahidiana (2021) studied the difference in the impact of gentrification on neighborhoods. He mainly focused on 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. On the other hand, gentrifying regions with white people experience an increase in income. He concluded that areas majorly occupied with white residents experience more inflow of white and high-income people. In contrast, regions mainly with people of color see an influx of not very high-income and white people. Thus, these regions become more racially diverse.

In the 1960s, the condominium structure was introduced. This system allowed people to own units in multi-family buildings. Boustan et al. (2019) studied if this system led to gentrification in central cities. They concluded that condo development did not attract high-income individuals, whereas builders chose regions with high-income residents to build condos. To prove this, they selected cities with restrictive condo ordinances and compared them with cities that did not have such restrictions relative to their neighboring areas. They did not see a statistically significant association between the condo density and the typical characteristics of gentrified regions – income levels and education.

The above all mentioned literature studied the relationship between gentrification and different factors. Some considered a few elements together, whereas others considered a single population characteristic or a new policy. Rigolon & Németh (2019) took a broad approach and studied the effect of various factors on gentrification.

Rigolon & Németh (2019) considered three different categories of variables – people, place, and policy. They studied their interactions and identified the variables individually or in conjunction with other variables, which would help predict whether a particular region would gentrify from 2000 to 2015. They used the definition 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. They used the U.S. Census Bureau and the Longitudinal Tract Data Base (LTDB) databases to get demographic and housing data. In addition to these data sources, they use a few other datasets like the National Historic Geographic Information System to get CSA-level data, the U.S. Department of Homeland Security and the U.S. Department of Housing and Urban Development to get data about the transit stations and housing as part of subsidized schemes. Their statistical analysis was done using IBM SPSS version 23.0.

For their set of predictors, they chose variables like household income, people with bachelor's degree, proportion of people who are a racial and ethnic minority as part of the 'people' category. Variables such as the percentage of rented households, median gross rent, the median value of owner-occupied households, distance from the downtown area, nearest rail station, and population density under the category of 'place.' Finally, under the category of 'policy,' they used the proportion of housing units subsidized by HUD programs as a predictor. They also considered the households under the HUD programs receiving the HCV, public housing, HUD-supported multi-family housing, and Low-Income Housing Tax Credit units. In total, they considered 13 variables across the three categories. They found that none of the variables correlated at a statistically significant level, so they did not eliminate any of them from their analysis. Then they used logistic regression to build a binary classification model with inputs as all these 13 variables to predict whether a tract would gentrify or not. They followed a hierarchical approach - adding variables of each category sequentially and analyzing the impact of each category on the model's ability to predict the process of gentrification.

Their analysis showed that across all five regions, all three categories assist in predicting the probability of gentrification of the tracts. Out of the people's class, they found race to be a strong predictor. They mainly found the percentage of Latino residents to be a significant predictor across all five regions. The distance from downtown was a strong predictor from the 'place' category. Also, variables like the percentage of multi-family households, proportion of units older than 30 years, and population density were other significant predictors in predicting the likelihood of gentrification. In addition, their results suggested that policy variables are less strong predictors than people or place variables. They could not find a consistent pattern of policy variables across the five regions. In New York and San Francisco CSAs, they observed that an increase in the percentage of HUD-subsidized housing units decreased the likelihood of gentrification. In contrast, in Washington and Chicago CSAs, the opposite was observed.

In conclusion, the scholarly work by Rigolon & Németh (2019) found the following:

(i) A region's racial and ethnic composition determines the odds of it becoming gentrified to a large extent.
(ii) Neighborhoods closer to downtown areas and with old housing units are likely to get gentrified.
(iii) Lastly, they found evidence of HUD's multi-family housing program in New York and HCVs in San Francisco limiting gentrification. In other CSAs, they did not find enough evidence of these programs affecting the process of gentrification.

3 Data Description and Visualization

The Neighborhood Change Database (NCDB) contains social, demographic, and housing data on census tracts in the United States for 1970, 1980, 1990, 2000, and 2010 (GeoLytics, n.d.). The source for this database is the decennial censuses performed by the U.S. Bureau of the Census.

Census tracts are locally determined geographic units, typically including between 2,500 and 8,000 persons (GeoLytics, n.d.). Tracts capture a group of residents with similar social characteristics, economic status, and housing conditions.

The main idea behind developing the NCDB is to allow users to compare tracts' characteristics across all the five years 1970 to 2010, even though the tracts' definitions change over time. There are three main types of tract changes that occur between censuses. Few tracts are combined over the years. Sometimes, a tract is divided into two or more tracts. In the third type, two or more tracts are changed into different sets of tracts. An analysis of 1990 and 2000 census tract boundaries showed that about 49 percent of all tracts were redefined between these two census years (GeoLytics, n.d.). Hence this feature of NCDB corrects this major issue and enables comparison of tracts' characteristics over the years which would not have been possible otherwise.

Over the years, information like detailed population, economic, and housing data were collected as part of the decennial census. To keep the data regularly updated, this system was changed in the year 2010 when all these data were started to be collected through the American Community Survey (ACS). While the advantage is that data updates are more frequent than was the case with the decennial census, the ACS's sample size is much smaller than that of the long census form.

It is also essential to recognize that the NCDB aggregates all the data to the census tract level instead of reporting it at an individual level. This is done to preserve individuals' confidentiality guaranteed by federal law. Since the NCDB depends on the decennial census, it can be updated only every ten years. Although data from these periods can be used in many ways, it fails to capture the exciting changes that happened between them.

By Census Bureau conventions, NCDB variables are separated into two general groups: population and housing (GeoLytics, n.d.).

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

Metropolitan area definitions are determined by the Office of Management and Budget (OMB), which follows official standards created by the interagency Federal Executive Committee on Metropolitan Areas. The essential requirement for a region to be designated a metro area is the presence of a city with at least 50,000 residents or a census-defined "urbanized area" in which the constituent counties have at least 100,000 residents (GeoLytics, n.d.). The main idea behind defining a metro area is to capture centers and the surrounding areas that are economically or socially connected to it.

For our analysis, we consider two Metropolitan Statistical Areas – New York-Northern New Jersey-Long Island, NY-NJ-PA (referred to as New York MSA) and Los Angeles-Long Beach-Santa Ana, CA (referred to as Los Angeles MSA). Since our analysis is from the year 2000 to 2010 and census tracts are grouped by counties in NCDB, we use the U.S. Census Bureau's Historical Delineation file for Dec 2009 under the category 'Metropolitan statistical areas and components' (Bureau, 2009) to determine the counties under each of the two metropolitan areas. We specifically chose New York and Los Angeles MSAs since we wanted to see the differences in the process of gentrification on the east and west coast of the United States.

To define the target label, whether a census tract is gentrified or not in the period of ten years (2000 - 2010), we use the following definition given by Rigolon & Németh (2019), adapted by them from Chapple et al. (2017):

(i) Change in median household income > Change in MSA median
(ii) Change in % college educated > Change in MSA percentage
(iii) Change in median gross rent > Change in MSA median
(OR)
(iv) % Increase of home value > % increase in MSA median

Here, we calculate the change between the years 2010 and 2000.

If the criteria given in the above definition are satisfied for a tract, we label it as 'Gentrified.' If not, we define the tract as 'Not Gentrified. Table 1 illustrates the exact variables we chose from the NCDB to represent each required field in the above definition. Table 2 shows summary statistics of the variables separately for the New York and the Los Angeles MSAs for the years 2000 and 2010.

We make two crucial assumptions while choosing the variables for the above task listed above:

(i) The median gross rent value in the NCDB only considers the rent paid by cash. We assume that the rent paid by other means won't be substantially different from what is paid by cash, and thus, this value is very close to the actual median gross rent.

(ii) The median value of housing units only represents the home value associated with the housing units occupied by owners of the units. Here again, we assume the unit's value will not be substantially different just because renters occupy a unit. Hence, we believe this value is very close to the actual median home value.

Since in our analysis, we look at the conditions of a neighborhood in the year 2000 to determine the possibility of it becoming gentrified, we pick a set of variables covering the three categories – people, place, and policy from the NCDB. Our rationale behind choosing variables is inspired by the literature that we read about gentrification. Table 3 lists all the variables' names and their corresponding meaning that we consider for our initial analysis. Table 4 shows the summary statistics for all these variables separately for New York and Los Angeles MSAs.

To understand the distribution of the variables, we plotted histograms separately for the New York MSA and the Los Angeles MSA. Figures 1 and 2 illustrate the histograms obtained for New York MSA and Los Angeles MSA, respectively. Figure 1, representing the frequency of numerical data for each independent variable in the New York MSA dataset, reveals that a few variables are skewed. The variables with the most prominent skew are the proportion of total persons below the poverty level last year, the proportion of households with public assistance income last year, and the median family income last year. Figure 2, illustrating the frequency of numerical data for each independent variable in the Los Angeles MSA dataset, reveals that quite a few variables are skewed. Similar to the New York MSA dataset, the proportion of total persons below the poverty level last year, the proportion of households with public assistance income last year, and the median family income in the previous year are skewed. Additionally, the proportion of White nonfamily households, the proportion of workers who live in a metro area and work in the same metro area, and the proportion of the Asian population are also skewed.

To further understand the interactions and relationships between variables, we visualized the correlation between the variables using heatmaps separately for the New York MSA and the Los Angeles MSA. Figures 3 and 4 illustrate the heatmaps obtained for New York MSA and Los Angeles MSA, respectively. The heatmaps provide some interesting insights listed below:

- The proportion of the white population and the proportion of the black population are negatively correlated and have a value of -0.86 and -0.55 for New York and Los Angeles, respectively. This indicates the racial segregation of the population, which is more predominant in the York MSA.
- We observe a high positive correlation between the proportion of people below the poverty line and the proportion of households with public assistance. This indicates that the public assistance policy is being correctly implemented and that the correct group is receiving the benefit.
- We see a high positive correlation between the proportion of white males employed and the proportion of white females employed, which suggests that the job opportunities, to a large extent, do not discriminate against people based on gender.

After data cleaning, we had the following data composition:

- New York MSA: A total of 4512 census tracts, in which we determine 990 tracts as gentrified in the year 2010 and the rest 3522 tracts as not gentrified.

- Los Angeles MSA: A total of 2926 census tracts, in which we determine 583 tracts as gentrified in the year 2010 and the rest 2343 tracts as not gentrified.

Only ~ 25% of all the tracts are gentrified for both New York and Los Angeles MSAs. We deal with this data imbalance using an artificial data generation technique, explained in the next section.

4 Model

Our modeling process included the following steps:

(i) Normalizing the variables which did not have values between 0 and 1. This step is done since our algorithm for artificial data generation works better with normalized variables.
(ii) Splitting the dataset into training and testing sets (80% for training and 20% for testing)
(iii) Fitting and interpreting a logistic regression model using all the predictors we chose for the New York and Los Angeles MSAs separately.
(iv) Fitting a random forest classifier using only the statistically significant variables we determined from the logistic regression model.
(v) Using the ADASYN (Adaptive Synthetic Sampling) technique to generate more data points of the minority class artificially.
(vi) Using the balanced dataset to fit a random forest classifier using only the statistically significant variables, we determined from the logistic regression model.

Normalizing the variables:
Since most of the predictors we chose were in ratios, we did not normalize them. For the variables which were not, we applied the min-max scaler to make the value of the variables between 0 and 1.

Training-Testing split:
After we performed the training-testing split, we had the following data composition:

For New York MSA:
Training set – A total of 3609 census tracts, with 793 tracts as gentrified in the year 2010 and the rest 2816 tracts as not gentrified.
Testing set – A total of 903 census tracts, with 197 tracts as gentrified in the year 2010 and the rest 706 tracts as not gentrified.

For Los Angeles MSA:
Training set – A total of 2340 census tracts, with 462 tracts as gentrified in the year 2010 and the rest 1878 tracts as not gentrified.
Testing set – A total of 586 census tracts, with 121 tracts as gentrified in the year 2010 and the rest 465 tracts as not gentrified.

Logistic Regression:
For the logistic regression model, we use the below equation to quantify our model:

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

$$\beta_i - represents\ the\ set\ of\ coefficients\ for\ each\ feature\ under\ the\ i\ category$$

$$features_i - represents\ the\ set\ of\ features\ under\ the\ i\ category$$

$$\log\left(\frac{p}{1-p}\right) - it\ is\ the\ \log odds\ ratio, with\ p = the\ probability\ of\ a\ census\ tract\ to\ be\ gentrified$$

After fitting the logistic model, we determine the statistically significant predictors (p-value < 0.05) and use only them for further analysis. The significant predictors, why we think they are significant, and their coefficients are discussed separately for New York and Los Angeles MSAs in detail in the next section.

Tables 5 and 6 show the confusion matrix and other metrics obtained for the logistic regression model on the testing set of the New York MSA and Los Angeles MSA, respectively. Even though we got an ~78% accuracy in both metropolitan areas, we observed that the F-1 score was too low. 0.11 and 0.07 for New York MSA and Los Angeles MSA, respectively. This was also evident from the fact that only very few that were supposed to be labeled as gentrified were correctly labeled. 13 out of 197 in the case of New York MSA and 5 out of 121 in the case of Los Angeles MSA. We identified that this issue is because of the imbalanced dataset. Next, we tried a random forest classifier, which employs the bagging technique to improve classification models' performance.

Random Forest:
We consider only the statistically significant predictors obtained from the logistic regression model to fit the random forest classifier. Tables 7 and 8 show the confusion matrix and other metrics obtained for the random forest classifier on the testing set of the New York MSA and Los Angeles MSA, respectively. We observed a slight increase in the F-1 scores for both metropolitan areas. The F-1 scores were 0.16 and 0.14 for New York MSA and Los Angeles MSA, respectively. 20 out of 197 in the case of New York MSA and 12 out of 121 in the case of Los Angeles MSA were correctly labeled as gentrified. Since we did not observe any significant improvement, we employed the technique of artificially generating data points of the minority class to tackle the data imbalance issue.

Synthetic Data Generation:
SMOTE (Synthetic Minority Over-sampling Technique) is an over-sampling technique in which synthetic samples are created by randomly sampling the characteristics from occurrences in the

minority class. There are three primary techniques for over-sampling implemented in R – SMOTE, ADASYN (Adaptive Synthetic Sampling), and DB-SMOTE (Density-Based SMOTE). We tried all three methods and found the ADASYN technique most effective for our case. We obtained the highest increase in F-1 scores of the random forest classifier by using the ADASYN technique. In this technique, more samples are generated for the minority class instances, which are difficult to learn. We used this technique only on the training set and not on the testing set. After the re-sampling process, we obtained the following data composition:

For New York MSA:
Training set – A total of 5673 census tracts, with 2857 tracts as gentrified in the year 2010 and the rest 2816 tracts as not gentrified.

For Los Angeles MSA:
Training set – A total of 3842 census tracts, with 1964 tracts as gentrified in the year 2010 and the rest 1878 tracts as not gentrified.

Random Forest using the balanced dataset:
We again consider only the statistically significant predictors obtained from the logistic regression model to fit the random forest classifier using the new balanced dataset. Tables 9 and 10 show the confusion matrix and other metrics obtained for the random forest classifier on the testing set of the New York MSA and Los Angeles MSA, respectively. We observed a significant increase in the F-1 scores for both metropolitan areas. The F-1 scores were 0.38 and 0.40 for New York MSA and Los Angeles MSA, respectively. 78 out of 197 in the case of New York MSA and 55 out of 121 in the case of Los Angeles MSA were correctly labeled as gentrified.

Hence, after solving the data imbalance using the ADASYN (Adaptive Synthetic Sampling) technique, we observed that more percentage of the gentrified class is correctly labeled; therefore, the F-1 score increases by a significant factor.

5 Empirical Analysis

The logistic regression analysis results determine nine significant variables for the New York MSA and five significant variables for the Los Angeles MSA.

For the New York MSA, we found the significant variables are:

(i) the proportion of people who are 65+ years old
(ii) the proportion of workers who live and work within the same metro area
(iii) median family income in the last year
(iv) the proportion of Native Hawaiian and other Pacific Islander populations
(v) the proportion of nonfamily households
(vi) the proportion of workers that are 16+ years old not working from home and with travel time to work that is less than 25 minutes

(vii) the proportion of employed White females that are 16+ years old
(viii) the proportion of vacant housing units
(ix) the proportion of renter-occupied housing units

These variables are listed in bold, along with their coefficient values and p-values in Table 11. The variables for the proportion of workers who live in a metro area and work within the same metro area and the median family income in the last year overlap with the Los Angeles MSA results, so these variables will be discussed later in the section.

Aside from these variables, the proportion of people who are 65+ years old, the proportion of Native Hawaiian and other Pacific Islander populations, the proportion of workers that are 16+ years old not working from home and with travel time to work that is less than 25 minutes, and the proportion of renter-occupied housing units all produced negative coefficients. We believe that the proportion of people who are 65+ years old and the proportion of Native Hawaiian and other Pacific Islander populations are good indicators of gentrification because they are both minority populations. Gentrification causes low-income people and people of color to decrease in a metropolitan area, so it is reasonable that a decrease in these populations would indicate a higher probability of gentrification. We believe that the proportion of workers that are 16+ years old not working from home and with travel time to work that is less than 25 minutes, and the proportion of renter-occupied housing units are also good indicators of gentrification because as rent prices increase as a result of gentrification, people are forced to move farther away from their jobs and may even move outside of the metropolitan area where they can find affordable housing.

On the other hand, the proportion of nonfamily households, the proportion of employed White females that are 16+ years old, and the proportion of vacant housing units produced positive coefficients. We believe that the proportion of nonfamily households is a strong indicator of gentrification because gentrification causes lower-income families to move out of the area and for wealthier single people to move into the area. In this way, an increase in the proportion of nonfamily households would be correlated with a higher probability of gentrification. We also believe that the proportion of vacant housing units is a strong indicator of gentrification because gentrification causes many people to be displaced because they cannot afford the increasing rent prices. This would cause an increase in vacant housing units because people will leave the housing units to find a more affordable place to live. The proportion of employed White females 16+ years old was a significant variable we did not expect. This variable may need to be researched more to understand the positive correlation with gentrification.

For the Los Angeles MSA, the significant variables are:
(i) the proportion of the non-White population
(ii) the proportion of persons under 18 years old
(iii) the proportion of workers who live and work within the same metro area
(iv) the proportion of total persons below the poverty level in the last year
(v) median family income in the previous year

These variables are listed in bold, along with their coefficient values and p-values in Table 12. The variables for the proportion of workers who live in a metro area and work within the same metro

area and the median family income in the last year overlap with the results of the New York MSA, so these variables will be discussed later on in the section.

Aside from these variables, the proportion of persons under 18 years old and the proportion of total persons below the poverty level in the last year produced negative coefficients. We believe this is consistent with our hypothesis because gentrification causes low-income families to move out of metropolitan areas. As these proportions decrease, there is a higher probability of gentrification because the area is filled with wealthier single people. The proportion of the non-White population produced a positive coefficient. This is not consistent with our hypothesis because we believed that the proportion of the non-White population would be negatively correlated with gentrification. This variable will require more research to gain an understanding as to why this variable produced a positive coefficient.

If we compare the results of the New York MSA logistic regression and the Los Angeles MSA logistic regression results, we find that two significant variables appear in both. These significant variables are median family income in the last year and the proportion of workers who live in a metro area and work within the same metro area. In both metropolitan areas, the variable for median family income in the last year produced a negative coefficient, which can be seen in Tables 11 and 12. We believe this is the case because a low median family income creates a greater opportunity for people to move into the area, become educated, and open up businesses. This results in a higher risk of gentrification. In both metropolitan areas, the variable for the proportion of workers who live in a metro area and work within the same metro area produced a positive coefficient, which can be seen in Tables 11 and 12. We believe this is the case because many people are moving into the area for new job opportunities, so there is an influx of people that are moving into the metropolitan area specifically to work there. This results in a higher risk of gentrification.

We hypothesized that the three factors that contribute the most to determining if a neighborhood is susceptible to gentrification are people, place, and policy. The results of the logistic regression analysis support that people and place are major contributors to gentrification susceptibility but do not support the third factor of policy. This disparity is because the dataset did not contain much data on policy. Exploring this third factor may require additional research and data analysis.

6 Conclusion

From the 'people' category, we find that race is a significant variable in predicting whether a region would get gentrified. We did observe different groups of people having a different impact on the process of gentrification in the two metropolitan areas that we considered. However, our study corroborates previous research on race being an essential factor in the process of gentrification.

We also find that age demographics to be a significant factor, with the proportion of 65+ years to be of interest in New York MSA. In contrast, the proportion of children under 18+ is a deciding factor in Los Angeles MSA. This shows the differences in the process of gentrification between

the metropolitan areas on the east and west coast of the United States. The median family income shows significance in both the metropolitan regions considered.

From the 'place' category, people living in the same metro area where they work helped determine the likelihood of gentrification in both metropolitan regions. This is an exciting finding since this could help policymakers to intervene early in places where there is a sudden increase in job opportunities and make sure that job opportunities are inclusive in nature and do not force communities to move out.

The proportion of vacant and rented households in New York MSA is significant, whereas these variables are not significant in Los Angeles MSA. This is another factor that is different between the two metropolitan areas on the east and west coast of the United States.

From the 'policy' category, we did not find any variables that we considered significant. But we feel that the variables we chose were not accurate enough to capture the effect of policy changes on the process of gentrification. Policies on tenant protection measures such as rent control, anti-eviction ordinances, and condominium conversion regulations have impacted the gentrification process. Since NCDB does not have variables capturing this information, we feel that including them from other data sources would be a good step forward to study the impact of policy variables in conjunction with people and place variables on the process of gentrification.

We see two limitations in our analysis, the first being that we do not give the model any information on what happened between the years 2000 and 2010. Our study only considers the conditions in the year 2000 to predict the possibility of gentrification of a region in the year 2010. This can be changed to consider the dynamic changes happening between the periods. The second limitation is that we try to predict if a region is gentrified or not. But some areas can be in various stages of gentrification, and a gentrification score capturing these differences could act as a better target variable.

References

Bates, L. K., 2013. *Gentrification and Displacement Study: implementing an equitable inclusive development strategy in the context of gentrification,* Portland: City of Portland Bureau of Planning and Sustainability.

Boustan, L. P. et al., 2019. *Does Condominium Development Lead to Gentrification?,* National Bureau of Economic Research.

Bureau, U. C., 2009. U.S. Census Bureau. [Online] Available at: https://www2.census.gov/programs-surveys/metro-micro/geographies/reference files/2009/historical-delineation-files/list4.txt

Chapple, K. et al., 2017. *Developing a new methodology for analyzing potential displacement*, California: California Air Resources Board.

Centers for Disease Control and Prevention. (n.d.). *Healthy places*. Centers for Disease Control and Prevention

GeoLytics, n.d. GeoLytics. [Online]. Available at: https://geolytics.com/user-guides

Kolko, J., 2007. The Determinants of Gentrification. *Available at SSRN: https://ssrn.com/abstract=985714 or http://dx.doi.org/10.2139/ssrn.985714.*

Kolko, J., 2009. Job Location, Neighborhood Change, and Gentrification. *Public Policy Institute of California.*

Rigolon, A. & Németh, J., 2019. Toward a socioecological model of gentrification: How people, place, and policy shape neighborhood change. *Journal of Urban Affairs,* 41:7, 887-909, DOI: 10.1080/07352166.2018.1562846.

Rucks-Ahidiana, Z., 2021. Racial composition and trajectories of gentrification in the United States. *Urban Studies,* 58(13), 2721–2741. https://doi.org/10.1177/0042098020963853.

Smith, N. (1998). The Encyclopedia of Housing.

Understanding Gentrification and Displacement. The Uprooted Project. (n.d.)

What are gentrification and displacement? Urban Displacement. (n.d.)

7 Appendix

Table 1: The list of variables and their corresponding variables names in the dataset for the years 2000 and 2010 used to define the Y labels.

Variable	Variable Name in 2000	Variable Name in 2010
Median Household Income	MDHHY0	MDHHY1A
% College Educated	(EDUC160/TRCTPOP0)*100	(EDUC161A/TRCTPOP1)*100
Median Gross Rent	MDGRENT0	MDGRENT1A
Median Home Value	MDVALHS0	MDVALHS1A

The definition given for the variables in the dataset data dictionary:
MDHHYy - Median household income last year ($)
EDUC16y - Persons 25+ years old who have a bachelor's or graduate/professional degree
TRCTPOPy - Total population
MDGRENTy - Median gross rent of specified renter-occupied housing units paying cash rent
MDVALHSy - Median value of specified owner-occupied housing units

(For 2000, y = 0 and for 2010, y = 1A)

Table 2: Summary statistics of the variables used to define the Y labels for New York and Los Angeles MSA in the years 2000 and 2010.

For New York MSA in the year 2000:

	TRCTPOP0	EDUC160	MDHHY0	MDVALHS0	MDGRENT0
Min.	0	0	0	0	0
1st Qu.	2686	274	34100	148650	692
Median	3843	572	49451	191700	799
Mean	4033	816	52923	219264	840
3rd Qu.	5169	1096	67152	254791	950
Max.	24834	9423	200001	1000001	2001

For New York MSA in the year 2010:

	TRCTPOP1	EDUC161A	MDHHY1A	MDVALHS1A	MDGRENT1A
Min.	0	0	0	0	0
1st Qu.	2782	377	42614	347000	984
Median	3968	749	63194	444900	1170
Mean	4160	989	68365	464374	1193
3rd Qu.	5336	1344	87994	580750	1418
Max.	26588	9616	250001	1000001	2001

For Los Angeles MSA in the year 2000:

	TRCTPOP0	EDUC160	MDHHY0	MDVALHS0	MDGRENT0
Min.	0	0	0	0	0
1st Qu.	3129	186	32164	158900	642
Median	4087	472	44519	195049	767
Mean	4222	690	49520	247129	851
3rd Qu.	5173	951	60685	279513	960
Max.	79512	40585	199754	1000001	2001

For Los Angeles MSA in the year 2010:

	TRCTPOP1	EDUC161A	MDHHY1A	MDVALHS1A	MDGRENT1A
Min.	0	0	0	0	0
1st Qu.	3295	267	41008	381100	982
Median	4226	656	57694	473400	1200
Mean	4380	857	63434	512820	1258
3rd Qu.	5383	1242	79594	628000	1502
Max.	21098	9384	242935	1000001	2001

Table 3: List of all the predictors that we chose to consider in the year 2000 for our classification algorithm.

Variable Name	Description
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 ($)
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
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 ($)
AVRETR0	Average household retirement income last year ($)

Table 4: Summary statistics of the predictors in the year 2000 used for the classification algorithm for New York and Los Angeles MSAs.

For New York MSA:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
SHRWHT0	0	0.34	0.72	0.62	0.9	2
SHRBLK0	0	0.014	0.055	0.21	0.293	1
SHRHSP0	0	0.047	0.097	0.19	0.256	1
ADULT0	0.4	0.71	0.75	0.76	0.79	1
CHILD0	0	0.21	0.25	0.24	0.29	2
OLD0	0	0.083	0.118	0.128	0.156	1
WRKSM0	0	0.72	0.89	0.82	0.95	1.5
WORKER30	0	68	119	136	187	926
POVRAT0	0	0.043	0.093	0.141	0.202	1
WELFARE0	0	0.033	0.064	0.107	0.144	1
AVWELIN0	0	4483	5308	5624	6448	30000
AVSSI0	0	5523	6457	6686	7588	35920
MDFAMY0	0	38213	57181	61499	77431	200001
AVSOCS0	0	9793	11651	11454	13134	30563
AVRETR0	0	11328	15436	17554	20860	189960
SHRAMI0	0	0	0.0013	0.0042	0.0048	0.3462
SHRASN0	0	0.015	0.042	0.081	0.099	1
SHRHIP0	0	0	0.00022	0.00133	0.00096	1
NONFAM0P	0	0.2	0.27	0.3	0.36	1
WNONFAM0P	0	0.1	0.19	0.21	0.28	1
COMMUT20NWKH	0	0.27	0.39	0.38	0.49	1
WMEMP0P	0	0.53	0.65	0.61	0.73	1.08
WFEMP0P	0	0.38	0.49	0.47	0.57	1
VACHU0P	0	0.022	0.036	0.055	0.059	1
RNTOCC0P	0	0.18	0.46	0.46	0.7	1

For Los Angeles MSA:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
SHRWHT0	0	0.34	0.72	0.62	0.9	2
SHRBLK0	0	0.014	0.055	0.21	0.293	1
SHRHSP0	0	0.047	0.097	0.19	0.256	1
ADULT0	0.4	0.71	0.75	0.76	0.79	1
CHILD0	0	0.21	0.25	0.24	0.29	2
OLD0	0	0.083	0.118	0.128	0.156	1

WRKSM0	0	0.72	0.89	0.82	0.95	1.5
WORKER30	0	68	119	136	187	926
POVRAT0	0	0.043	0.093	0.141	0.202	1
WELFARE0	0	0.033	0.064	0.107	0.144	1
AVWELIN0	0	4483	5308	5624	6448	30000
AVSSI0	0	5523	6457	6686	7588	35920
MDFAMY0	0	38213	57181	61499	77431	200001
AVSOCS0	0	9793	11651	11454	13134	30563
AVRETR0	0	11328	15436	17554	20860	189960
SHRAMI0	0	0	0.0013	0.0042	0.0048	0.3462
SHRASN0	0	0.015	0.042	0.081	0.099	1
SHRHIP0	0	0	0.00022	0.00133	0.00096	1
NONFAM0P	0	0.2	0.27	0.3	0.36	1
WNONFAM0P	0	0.1	0.19	0.21	0.28	1
COMMUT20NWKH	0	0.27	0.39	0.38	0.49	1
WMEMP0P	0	0.53	0.65	0.61	0.73	1.08
WFEMP0P	0	0.38	0.49	0.47	0.57	1
VACHU0P	0	0.022	0.036	0.055	0.059	1
RNTOCC0P	0	0.18	0.46	0.46	0.7	1

Figure 1: Histograms showing the distribution of predictors chosen for the New York MSA.

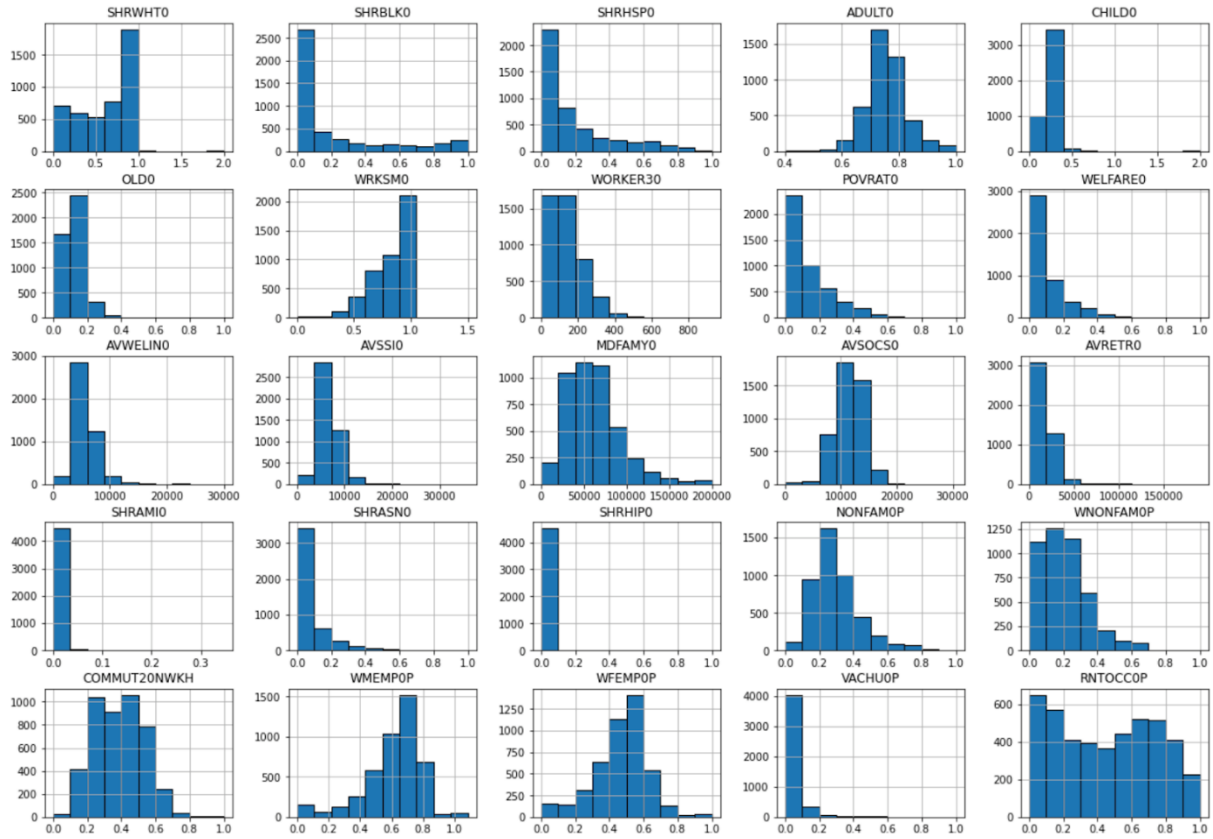


Figure 2: Histograms showing the distribution of predictors chosen for the Los Angeles MSA.

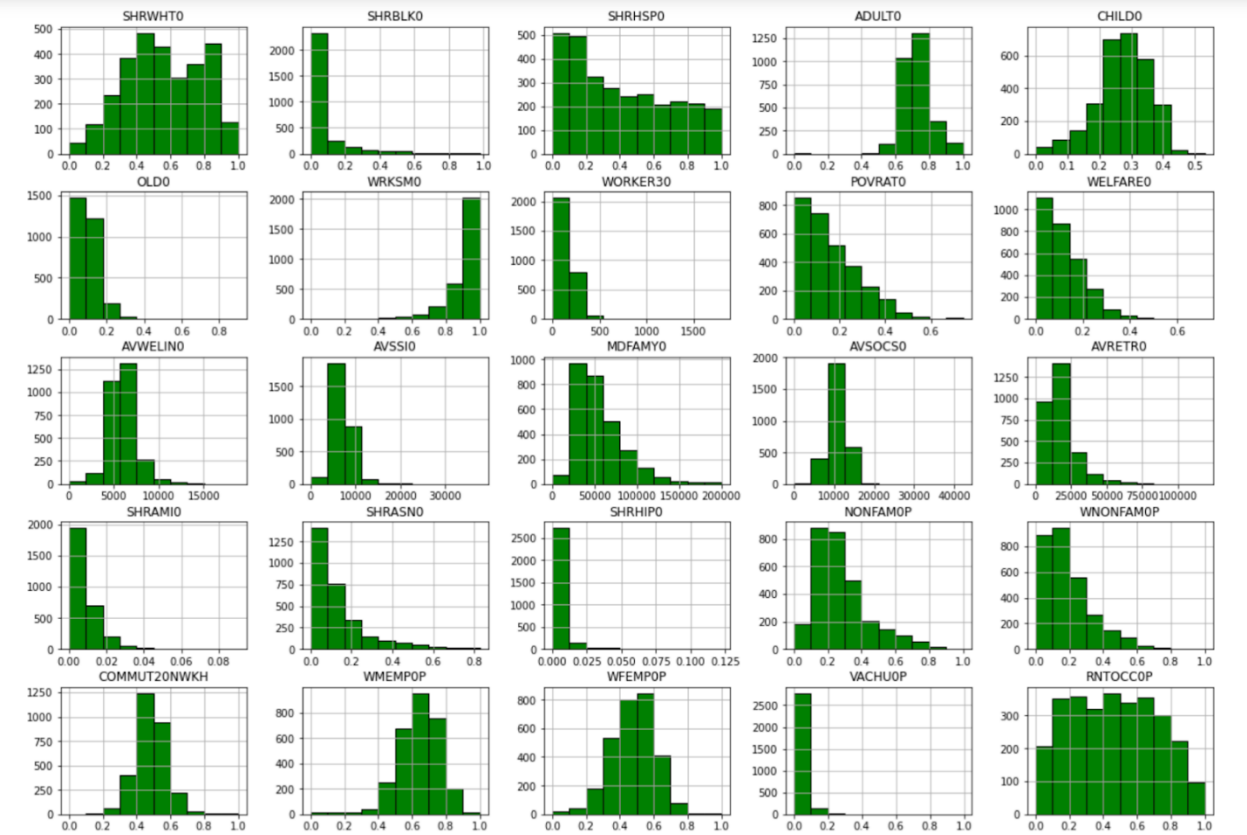


Figure 3: Correlation Heatmap for New York MSA variables

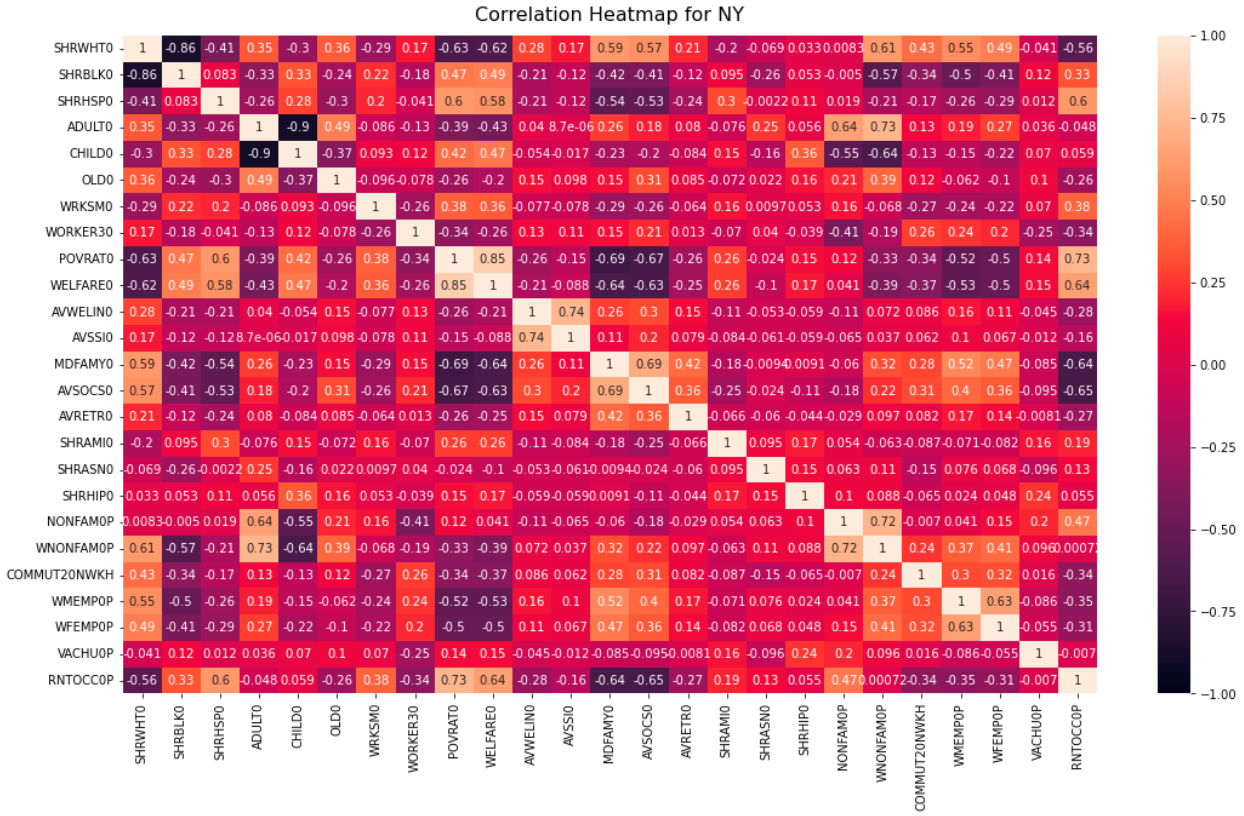


Figure 4: Correlation Heatmap for Los Angeles MSA variables

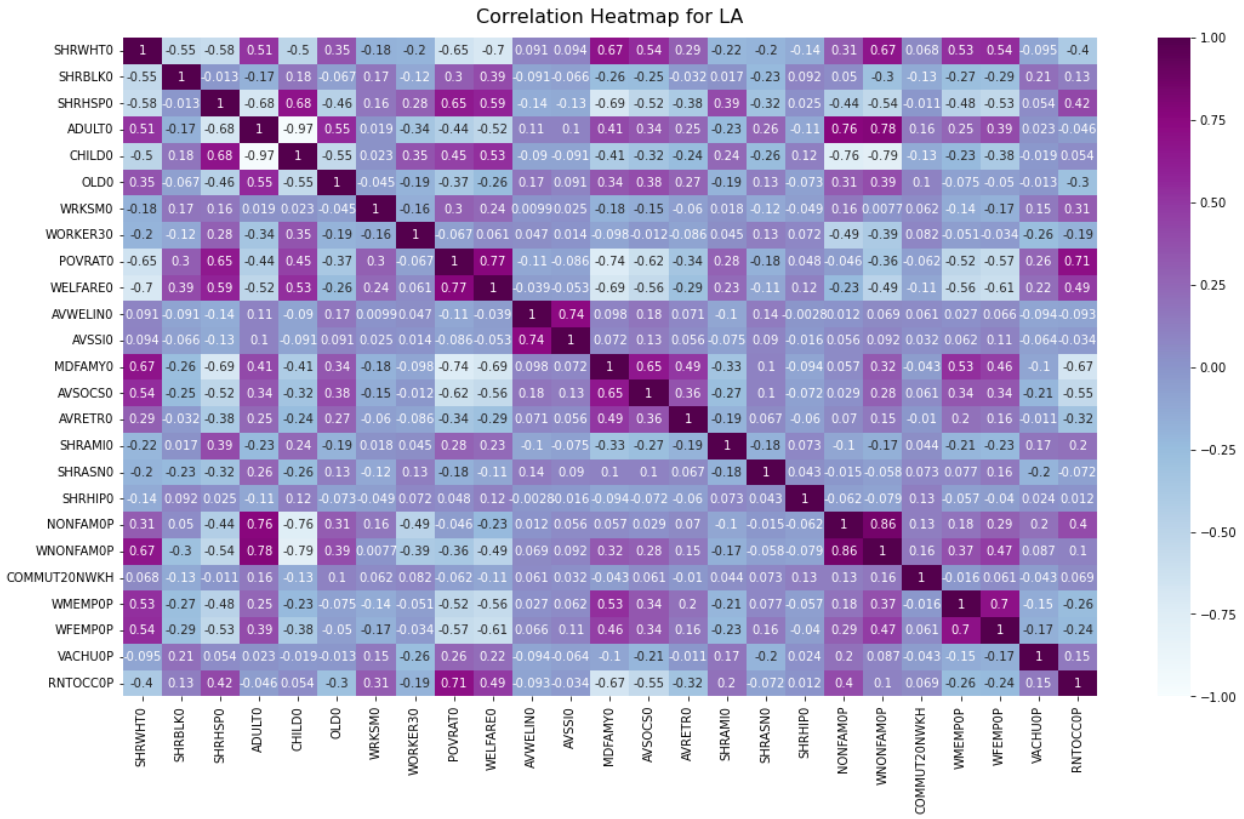


Table 5: Confusion Matrix and other metrics obtained for logistic regression model fitted on the New York MSA dataset.

		Reference	
		0	1
Prediction	0	692	184
	1	14	**13**
Accuracy		**0.78**	
Sensitivity		0.06	
Specificity		0.98	
Precision		0.48	
Recall		0.06	
F1		**0.11**	

Table 6: Confusion Matrix and other metrics obtained for logistic regression model fitted on the Los Angeles MSA dataset.

		Reference	
		0	1
Prediction	0	454	116
	1	11	**5**
Accuracy		**0.78**	
Sensitivity		0.04	
Specificity		0.97	
Precision		0.31	
Recall		0.04	
F1		**0.07**	

Table 7: Confusion Matrix and other metrics obtained for random forest classifier fitted on the New York MSA dataset.

		Reference	
		0	1
Prediction	0	681	177
	1	25	**20**
Accuracy		**0.77**	
Sensitivity		0.1	
Specificity		0.96	
Precision		0.44	
Recall		0.1	
F1		**0.16**	

Table 8: Confusion Matrix and other metrics obtained for random forest classifier fitted on the Los Angeles MSA dataset.

		Reference	
		0	1
Prediction	0	435	109
	1	30	**12**
Accuracy		**0.76**	
Sensitivity		0.09	
Specificity		0.93	
Precision		0.28	
Recall		0.09	
F1		**0.14**	

Table 9: Confusion Matrix and other metrics obtained for random forest classifier fitted on the New York MSA **balanced** dataset.

		Reference	
		0	1
Prediction	0	570	119
	1	136	**78**
Accuracy		**0.71**	
Sensitivity		0.39	
Specificity		0.8	
Precision		0.36	
Recall		0.39	
F1		**0.38**	

Table 10: Confusion Matrix and other metrics obtained for random forest classifier fitted on the Los Angeles MSA **balanced** dataset.

		Reference	
		0	1
Prediction	0	367	66
	1	98	**55**
Accuracy		**0.72**	
Sensitivity		0.45	
Specificity		0.78	
Precision		0.36	
Recall		0.45	
F1		**0.4**	

Table 11: Results obtained for logistic regression model fitted on the New York MSA dataset.

Variable	Coefficient	P-Value
SHRWHT0	-1.459	0.261
SHRBLK0	-1.395	0.291
SHRHSP0	-0.177	0.812
ADULT0	14.693	0.130
CHILD0	10.885	0.258
OLD0	**-4.334**	**5.73 x e^7**
WRKSM0	**0.8315**	**0.003**
WORKER30	0.382	0.477
POVRAT0	0.484	0.568
WELFARE0	-0.879	0.365
AVWELIN0	0.591	0.516
AVSSI0	-0.513	0.575
MDFAMY0	**-2.865**	**3.50 x e^8**
AVSOCS0	0.874	0.308
AVRETR0	0.516	0.512
SHRAMI0	-5.141	0.416
SHRASN0	-1.705	0.221
SHRHIP0	**-32.676**	**0.042**
NONFAM0P	**2.154**	**0.021**
WNONFAM0P	1.582	0.095
COMMUT20NWKH	**-1.713**	**2.01 x e^6**
WMEMP0P	0.137	0.699
WFEMP0P	**0.838**	**0.035**
VACHU0P	**1.143**	**0.046**
RNTOCC0P	**-2.821**	**1.88 x e^13**

Table 12: Results obtained for logistic regression model fitted on the Los Angeles MSA dataset.

Variable	Coefficient	P-Value
SHRNWT0	**3.855**	**0.0427**
SHRWHT0	3.089	0.091
SHRHSP0	0.958	0.371
ADULT0	-4.813	0.281
CHILD0	**-9.466**	**0.037**
OLD0	-1.999	0.146
WRKSM0	**1.465**	**0.031**
WORKER30	-0.347	0.814
POVRAT0	**-6.700**	**2.97 x e^6**
WELFARE0	-0.356	0.815
AVWELIN0	-1.065	0.270
AVSSI0	-0.112	0.935
MDFAMY0	**-4.313**	**6.7 x e^6**
AVSOCS0	-0.184	0.898
AVRETR0	-1.282	0.128
NONFAM0P	-0.969	0.505
WNONFAM0P	1.524	0.327
COMMUT20NWKH	0.714	0.266
WMEMP0P	1.229	0.127
WFEMP0P	1.229	0.412
VACHU0P	-1.000	0.513
RNTOCC0P	-0.057	0.910

Y Data Preprocessing

2022-12-22

Description: This script is used to create the Y labels (output variable) for all the tracts of New York and Los Angeles MSAs

The datasets used in this script are the following:

- Los_Angeles_2000_Y.csv - This file contains the variables mentioned in Table 1, for the year 2000 and the Los Angeles MSA
- Los_Angeles_2010_Y.csv - This file contains the variables mentioned in Table 1, for the year 2010 and the Los Angeles MSA
- New_York_2000_Y.csv - This file contains the variables mentioned in Table 1, for the year 2000 and the New York MSA
- New_York_2010_Y.csv - This file contains the variables mentioned in Table 1, for the year 2010 and the New York MSA

Using this script, the following datasets are created:

- la_y.csv - This file contains the Y labels for all the tracts of Los Angeles MSA
- ny_y.csv - This file contains the Y labels for all the tracts of New York MSA

```
# clearing the workspace
rm(list=ls(all=TRUE))

# import the required libraries
library(dplyr)

#read in datasets
LA_2000 <- read.csv("Los_Angeles_2000_Y.csv")
LA_2010 <- read.csv("Los_Angeles_2010_Y.csv")
NY_2000 <- read.csv("New_York_2000_Y.csv")
NY_2010 <- read.csv("New_York_2010_Y.csv")

# summary statistics for New York MSA variables in the year 2000
write.csv(summary(NY_2000[, -which(names(NY_2000) == 'AREAKEY')], digits = 2)
,"summary1.csv")

# summary statistics for New York MSA variables in the year 2010
write.csv(summary(NY_2010[, -which(names(NY_2010) == 'AREAKEY')], digits = 2)
,"summary2.csv")

# summary statistics for Los Angeles MSA variables in the year 2000
write.csv(summary(LA_2000[, -which(names(LA_2000) == 'AREAKEY')], digits = 2)
,"summary3.csv")
```

```r
# summary statistics for Los Angeles MSA variables in the year 2010
write.csv(summary(LA_2010[, -which(names(LA_2010) == 'AREAKEY')], digits = 2)
,"summary4.csv")

#update column names for readability
colnames(LA_2000) <- c('Area Key','Total Population','Educated','Household In
come','Median Home Value','Median Gross Rent')
colnames(LA_2010) <- c('Area Key','Total Population','Educated','Household In
come','Median Home Value','Median Gross Rent')
colnames(NY_2000) <- c('Area Key','Total Population','Educated','Household In
come','Median Home Value','Median Gross Rent')
colnames(NY_2010) <- c('Area Key','Total Population','Educated','Household In
come','Median Home Value','Median Gross Rent')

#create column for % college
LA_2000$"% College" <- (LA_2000$Educated/LA_2000$`Total Population`)*100
LA_2010$"% College" <- (LA_2010$Educated/LA_2010$`Total Population`)*100
NY_2000$"% College" <- (NY_2000$Educated/NY_2000$`Total Population`)*100
NY_2010$"% College" <- (NY_2010$Educated/NY_2010$`Total Population`)*100

#change in CSA median for household income (L.A.)
CSA_hh_income_LA <- median(LA_2010$`Household Income`)-median(LA_2000$`Househ
old Income`)

#change in CSA percentage for college educated (L.A.)
LA_2000_PC_educated <- (sum(LA_2000$Educated)/sum(LA_2000$`Total Population`)
)*100
LA_2010_PC_educated <-(sum(LA_2010$Educated)/sum(LA_2010$`Total Population`))
*100
CSA_educated_LA <-LA_2010_PC_educated-LA_2000_PC_educated

#change in CSA median for gross rent (L.A.)
CSA_grent_LA <- median(LA_2010$`Median Gross Rent`)-median(LA_2000$`Median Gr
oss Rent`)

#percent increase in CSA median for home value (L.A.)
CSA_increase_HV_LA <- ((median(LA_2010$`Median Home Value`)-median(LA_2000$`M
edian Home Value`))/(median(LA_2000$`Median Home Value`)))*100

#change in CSA median for household income (N.Y.)
CSA_hh_income_NY <- median(NY_2010$`Household Income`)-median(NY_2000$`Househ
old Income`)

#change in CSA percentage for college educated (N.Y.)
NY_2000_PC_educated <- (sum(NY_2000$Educated)/sum(NY_2000$`Total Population`)
)*100
NY_2010_PC_educated <-(sum(NY_2010$Educated)/sum(NY_2010$`Total Population`))
*100
CSA_educated_NY <-NY_2010_PC_educated-NY_2000_PC_educated

#change in CSA median for gross rent (N.Y.)
```

```
CSA_grent_NY <- median(NY_2010$`Median Gross Rent`)-median(NY_2000$`Median Gr
oss Rent`)

#percent increase in CSA median for home value (N.Y.)
CSA_increase_HV_NY <- ((median(NY_2010$`Median Home Value`)-median(NY_2000$`M
edian Home Value`))/(median(NY_2000$`Median Home Value`)))*100

#create lists for columns in new dataframe (L.A.)
area_key_list_LA = LA_2000$`Area Key`
LA_hh_income_list = if_else((LA_2010$`Household Income`-LA_2000$`Household In
come`)>CSA_hh_income_LA,1,0)
LA_educated_list = if_else((LA_2010$`% College`-LA_2000$`% College`)>CSA_educ
ated_LA,1,0)
LA_grent_list = if_else((LA_2010$`Median Gross Rent`-LA_2000$`Median Gross Re
nt`)>CSA_grent_LA,1,0)
LA_HV_list = if_else((((LA_2010$`Median Home Value`-LA_2000$`Median Home Valu
e`)/LA_2000$`Median Home Value`)*100)>CSA_increase_HV_LA,1,0)

#create dataframe (LA)
LA_results <- data.frame(area_key_list_LA,LA_hh_income_list,LA_educated_list,
LA_grent_list,LA_HV_list)

#create lists for columns in new dataframe (N.Y.)
area_key_list_NY = NY_2000$`Area Key`
NY_hh_income_list = if_else((NY_2010$`Household Income`-NY_2000$`Household In
come`)>CSA_hh_income_NY,1,0)
NY_educated_list = if_else((NY_2010$`% College`-NY_2000$`% College`)>CSA_educ
ated_NY,1,0)
NY_grent_list = if_else((NY_2010$`Median Gross Rent`-NY_2000$`Median Gross Re
nt`)>CSA_grent_NY,1,0)
NY_HV_list = if_else((((NY_2010$`Median Home Value`-NY_2000$`Median Home Valu
e`)/NY_2000$`Median Home Value`)*100)>CSA_increase_HV_NY,1,0)

#create dataframe (NY)
NY_results <- data.frame(area_key_list_NY,NY_hh_income_list,NY_educated_list,
NY_grent_list,NY_HV_list)

#create lists for final gentrification score
final_scores_LA = if_else((LA_results$LA_hh_income_list==1 & LA_results$LA_ed
ucated_list==1 & (LA_results$LA_grent_list==1 | LA_results$LA_HV_list==1)),1,
0)
final_scores_NY = if_else((NY_results$NY_hh_income_list==1 & NY_results$NY_ed
ucated_list==1 & (NY_results$NY_grent_list==1 | NY_results$NY_HV_list==1)),1,
0)

#create final results dataframe
final_results_LA = data.frame(area_key_list_LA,final_scores_LA)
final_results_NY = data.frame(area_key_list_NY,final_scores_NY)
```

```
#write dataframes to .csv files
write.csv(final_results_LA, "la_y.csv")
write.csv(final_results_NY, "ny_y.csv")
```

X Data Preprocessing

2022-12-22

Description: This script is used to pre-process the X data (input variables) for all the tracts of New York and Los Angeles MSAs

The datasets used in this script are the following:

- Los_Angeles_X.csv - This file contains the variables mentioned in Table 3, for the year 2000 and the Los Angeles MSA
- New_York_X.csv - This file contains the variables mentioned in Table 3, for the year 2000 and the New York MSA

Using this script, the following datasets are created:

- la_x.csv - This file contains the X data processed for all the tracts of Los Angeles MSA
- ny_x.csv - This file contains the X data processed for all the tracts of New York MSA

```
# clearing the workspace
rm(list=ls(all=TRUE))

# loading all the required libraries
library(readr, quietly = T)
library(data.table, quietly = T)
library(ggplot2, quietly = T)
library(vtable, quietly = T)

# reading the datasets
ny_df = read.csv("New_York_X.csv")
la_df = read.csv("Los_Angeles_X.csv")

# changing the 'AREAKEY' (tract identifier variable) to factor from int data
type
ny_df$AREAKEY <- as.factor(ny_df$AREAKEY)
la_df$AREAKEY <- as.factor(la_df$AREAKEY)

# removing rows which have zeros in all the columns
ny_df <- ny_df[rowSums(ny_df[,-1]) > 0,]
la_df <- la_df[rowSums(la_df[,-1]) > 0,]
```

```
# calculating ratios for the required columns
ny_df['SHRAMI0'] = ny_df['SHRAMI0N']/ny_df['SHR0D']
la_df['SHRAMI0'] = la_df['SHRAMI0N']/la_df['SHR0D']

ny_df['SHRASN0'] = ny_df['SHRASN0N']/ny_df['SHR0D']
la_df['SHRASN0'] = la_df['SHRASN0N']/la_df['SHR0D']

ny_df['SHRHIP0'] = ny_df['SHRHIP0N']/ny_df['SHR0D']
la_df['SHRHIP0'] = la_df['SHRHIP0N']/la_df['SHR0D']

ny_df['NONFAM0P'] = ny_df['NONFAM0']/ny_df['NUMHHS0']
la_df['NONFAM0P'] = la_df['NONFAM0']/la_df['NUMHHS0']

ny_df['WNONFAM0P'] = ny_df['WNONFAM0']/ny_df['NUMHHS0']
la_df['WNONFAM0P'] = la_df['WNONFAM0']/la_df['NUMHHS0']

ny_df['COMMUT20NWKH'] = (ny_df['COMMUT20'] - ny_df['WKHOME0'])/ny_df['WRCNTY0
D']
la_df['COMMUT20NWKH'] = (la_df['COMMUT20'] - la_df['WKHOME0'])/la_df['WRCNTY0
D']

ny_df['WMEMP0P'] = ny_df['WMEMP0']/ny_df['WM16P0']
la_df['WMEMP0P'] = la_df['WMEMP0']/la_df['WM16P0']

ny_df['WFEMP0P'] = ny_df['WFEMP0']/ny_df['WF16P0']
la_df['WFEMP0P'] = la_df['WFEMP0']/la_df['WF16P0']

ny_df['VACHU0P'] = ny_df['VACHU0']/ny_df['TOTHSUN0']
la_df['VACHU0P'] = la_df['VACHU0']/la_df['TOTHSUN0']

ny_df['RNTOCC0P'] = ny_df['RNTOCC0']/ny_df['TOTHSUN0']
la_df['RNTOCC0P'] = la_df['RNTOCC0']/la_df['TOTHSUN0']
```

```
# removing the columns not required (since these columns were converted to ra
tios and new columns were created)
ny_df <- subset(ny_df, select = -c(SHR0D, SHRAMI0N, SHRASN0N, SHRHIP0N, NONFA
M0, WNONFAM0, NUMHHS0, COMMUT20, WKHOME0, WRCNTY0D, WMEMP0, WM16P0, WFEMP0, W
F16P0, VACHU0, TOTHSUN0, RNTOCC0))

la_df <- subset(la_df, select = -c(SHR0D, SHRAMI0N, SHRASN0N, SHRHIP0N, NONFA
M0, WNONFAM0, NUMHHS0, COMMUT20, WKHOME0, WRCNTY0D, WMEMP0, WM16P0, WFEMP0, W
F16P0, VACHU0, TOTHSUN0, RNTOCC0))

# replacing NANs introduced by the calculation of ratios in the above step wi
th zeros
ny_df[is.na(ny_df)] <- 0
la_df[is.na(la_df)] <- 0
```

```
# summary statistics for New York MSA variables in the year 2000
write.csv(summary(ny_df[, -which(names(ny_df) == 'AREAKEY')], digits = 2),"su
mmary5.csv")

# summary statistics for Los Angeles MSA variables in the year 2000
write.csv(summary(ny_df[, -which(names(ny_df) == 'AREAKEY')], digits = 2),"su
mmary6.csv")

# creating csv files
write_csv(ny_df, "ny_x.csv")
write_csv(la_df, "la_x.csv")
```

Creating the training data

2022-12-22

Description: This script is used to create the training data, combining the X an Y data created in the earlier steps

The datasets used in this script are the following:

- ny_x - This file contains the X data processed for all the tracts of New York MSA
- la_x - This file contains the X data processed for all the tracts of Los Angeles MSA
- ny_x - This file contains the Y labels for all the tracts of New York MSA
- la_y - This file contains the Y labels for all the tracts of Los Angeles MSA

Using this script, the following datasets are created:

- la_training_data.csv - This file is the training data for the Los Angeles MSA
- ny_training_data.csv - This file is the training data for the New York MSA

```
# clearing the workspace
rm(list=ls(all=TRUE))

# loading all the required libraries
library(readr, quietly = T)
library(data.table, quietly = T)
library(ROSE, quietly = T)

# reading all the required files
ny_x = read.csv("ny_x.csv")
ny_y = read.csv("ny_y.csv")

la_x = read.csv("la_x.csv")
la_y = read.csv("la_y.csv")
```

```
# viewing the datasets for New York MSA (X and y) to understand the layout of
the datasets before combining them (same for Los Angeles MSA)
head(ny_x)
```

```
##      AREAKEY  SHRWHT0  SHRBLK0  SHRHSP0   ADULT0   CHILD0     OLD0  WRKSM0
## 1 3.4003e+10 0.920545 0.006285 0.019752 0.698189 0.301511 0.141254 0.63661
## 2 3.4003e+10 0.765460 0.014659 0.014201 0.754924 0.245076 0.143839 0.44254
## 3 3.4003e+10 0.829314 0.016491 0.049680 0.772624 0.227376 0.171099 0.70418
## 4 3.4003e+10 0.794123 0.013389 0.026604 0.731525 0.268475 0.153365 0.66373
## 5 3.4003e+10 0.682816 0.055567 0.158197 0.766660 0.233340 0.138224 0.66743
## 6 3.4003e+10 0.649866 0.065343 0.205620 0.778100 0.221900 0.126673 0.72336
##   WORKER30  POVRAT0 WELFARE0 AVWELIN0   AVSSI0 MDFAMY0  AVSOCS0  AVRETR0
## 1      182 0.018173 0.012334 6416.346 6000.200  113379 15151.04 19367.72
## 2       85 0.061842 0.016949 6291.667 6291.667  134068 16614.55 28189.47
## 3      260 0.050593 0.062198 6000.971 6679.710   82185 14171.54 16872.78
## 4      179 0.049103 0.020076 5256.757 8300.000  100329 14240.14 20819.37
## 5      257 0.045093 0.045217 6080.769 6629.091   79287 11946.99 16126.58
## 6      201 0.045790 0.041863 3781.690 5214.286   57031 12213.81 14250.66
##       SHRAMI0    SHRASN0      SHRHIP0  NONFAM0P WNONFAM0P COMMUT20NWKH
## 1 0.000000000 0.06838246 0.0000000000 0.1518027 0.1518027    0.4722695
## 2 0.000000000 0.20338983 0.0000000000 0.1087571 0.1059322    0.2701613
## 3 0.001030715 0.13048856 0.0000000000 0.2161836 0.2053140    0.5501829
## 4 0.000000000 0.18848896 0.0001738828 0.1432447 0.1405317    0.3949681
## 5 0.005141388 0.20071188 0.0000000000 0.2486957 0.2347826    0.4884793
## 6 0.002453167 0.19603033 0.0015611062 0.3484670 0.3024764    0.5160018
##     WMEMP0P   WFEMP0P    VACHU0P   RNTOCC0P
## 1 0.7495108 0.5150250 0.01496726 0.09307764
## 2 0.7279753 0.4362819 0.03013699 0.10000000
## 3 0.6936005 0.5840555 0.01610979 0.18078759
## 4 0.7232743 0.4737401 0.01641949 0.14406780
## 5 0.7313891 0.6134398 0.02428007 0.17391304
## 6 0.7416226 0.6194690 0.01733102 0.52744079
```

```
head(ny_y)
```

```
##   X area_key_list_NY final_scores_NY
## 1 1       3.4003e+10               0
## 2 2       3.4003e+10               0
## 3 3       3.4003e+10               0
## 4 4       3.4003e+10               0
## 5 5       3.4003e+10               0
## 6 6       3.4003e+10               0
```

```
# removing the 'X' column from the 'ny_y' dataset since it is not required an
d renaming the columns for readability
ny_y <- subset(ny_y, select = -c(X))
colnames(ny_y) <- c('AREAKEY', 'Gentrified')

# removing the 'X' column from the 'la_y' dataset since it is not required an
d renaming the columns for readability
```

```
la_y <- subset(la_y, select = -c(X))
colnames(la_y) <- c('AREAKEY', 'Gentrified')

# merging the two files, by 'AREAKEY' (the unique tract identifier) and remov
ing the column 'AREAKEY' since it is not required for training
ny_training_data <- merge(ny_x, ny_y, by = 'AREAKEY', how = 'inner')
ny_training_data <- subset(ny_training_data, select = -c(AREAKEY))

# removing null values (these null values emerged due to the ratios computed
while creating the Y labels)
# since these are very small in number, we remove it
ny_training_data$Gentrified[is.na(ny_training_data$Gentrified)] = 0
ny_training_data$Gentrified <- as.factor(ny_training_data$Gentrified)

# merging the two files, by 'AREAKEY' (the unique tract identifier) and remov
ing the column 'AREAKEY' since it is not required for training
la_training_data <- merge(la_x, la_y, by = 'AREAKEY', how = 'inner')
la_training_data <- subset(la_training_data, select = -c(AREAKEY))

# removing null values (these null values emerged due to the ratios computed
while creating the Y labels)
# since these are very small in number, we remove it
la_training_data$Gentrified[is.na(la_training_data$Gentrified)] = 0
la_training_data$Gentrified <- as.factor(la_training_data$Gentrified)

# creating csv files
write_csv(ny_training_data, "ny_training_data.csv")
write_csv(la_training_data, "la_training_data.csv")
```

Modeling for New York MSA dataset

2022-12-22

Description: This script is used to train, test, and interpret the machine learning models for the tracts of the New York MSA.

The datasets used in this script are the following:

- ny_training_data.csv - This file is the training data for the New York MSA

```
# clearing the workspace
rm(list=ls(all=TRUE))

# loading all the required libraries
library(readr, quietly = T)
library(data.table, quietly = T)
library(dplyr, quietly = T)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.table':
##
##     between, first, last

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(caret, quietly = T)
library(randomForest, quietly = T)

## randomForest 4.7-1.1

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':
##
##     margin

## The following object is masked from 'package:dplyr':
##
##     combine

library(smotefamily, quietly = T)

# reading in the training dataset
ny_data = read.csv("ny_training_data.csv")

# converting the output variable to a factor
ny_data$Gentrified <- as.factor(ny_data$Gentrified)

# viewing the training dataset
head(ny_data)

##    SHRWHT0  SHRBLK0  SHRHSP0   ADULT0   CHILD0     OLD0   WRKSM0 WORKER30
## 1 0.920545 0.006285 0.019752 0.698189 0.301511 0.141254 0.636611      182
## 2 0.765460 0.014659 0.014201 0.754924 0.245076 0.143839 0.442540       85
## 3 0.829314 0.016491 0.049680 0.772624 0.227376 0.171099 0.704185      260
## 4 0.794123 0.013389 0.026604 0.731525 0.268475 0.153365 0.663738      179
## 5 0.682816 0.055567 0.158197 0.766660 0.233340 0.138224 0.667435      257
## 6 0.649866 0.065343 0.205620 0.778100 0.221900 0.126673 0.723367      201
##    POVRAT0 WELFARE0 AVWELIN0   AVSSI0 MDFAMY0  AVSOCS0  AVRETR0  SHRAMI0
## 1 0.018173 0.012334 6416.346 6000.200  113379 15151.04 19367.72 0.00000000
```

```
## 2 0.061842 0.016949 6291.667 6291.667  134068 16614.55 28189.47 0.00000000
## 3 0.050593 0.062198 6000.971 6679.710   82185 14171.54 16872.78 0.00103071
## 4 0.049103 0.020076 5256.757 8300.000  100329 14240.14 20819.37 0.00000000
## 5 0.045093 0.045217 6080.769 6629.091   79287 11946.99 16126.58 0.00514138
## 6 0.045790 0.041863 3781.690 5214.286   57031 12213.81 14250.66 0.00245316
##      SHRASN0      SHRHIP0  NONFAM0P WNONFAM0P COMMUT20NWKH WMEMP0P WFEMP0P
## 1 0.06838246 0.000000000 0.1518027 0.1518027  0.4722695   0.7495108 0.5150
## 2 0.20338983 0.000000000 0.1087571 0.1059322  0.2701613   0.7279753 0.4362
## 3 0.13048856 0.000000000 0.2161836 0.2053140  0.5501829   0.6936005 0.5840
## 4 0.18848896 0.000173882 0.1432447 0.1405317  0.3949681   0.7232743 0.4737
## 5 0.20071188 0.000000000 0.2486957 0.2347826  0.4884793   0.7313891 0.6134
## 6 0.19603033 0.001561106 0.3484670 0.3024764  0.5160018 0.7416226   0.6194
##      VACHU0P   RNTOCC0P Gentrified
## 1 0.01496726 0.09307764          0
## 2 0.03013699 0.10000000          0
## 3 0.01610979 0.18078759          0
## 4 0.01641949 0.14406780          0
## 5 0.02428007 0.17391304          0
## 6 0.01733102 0.52744079          0
```

```
# normalizing the variables which are not in proportions (since proportions a
re already between 0 - 1)
minMax <- function(x) {
  (x - min(x)) / (max(x) - min(x))
}

ny_data[c('WORKER30', 'AVWELIN0', 'AVSSI0', 'MDFAMY0', 'AVSOCS0', 'AVRETR0')]
<-
                                  lapply(ny_data[c('WORKER30', 'AVWELIN0', 'AV
SSI0', 'MDFAMY0', 'AVSOCS0', 'AVRETR0')], minMax)
head(ny_data)
```

```
##    SHRWHT0  SHRBLK0  SHRHSP0   ADULT0   CHILD0     OLD0   WRKSM0  WORKER30
## 1 0.920545 0.006285 0.019752 0.698189 0.301511 0.141254 0.636611 0.19654
## 2 0.765460 0.014659 0.014201 0.754924 0.245076 0.143839 0.442540 0.09179
## 3 0.829314 0.016491 0.049680 0.772624 0.227376 0.171099 0.704185 0.28077
## 4 0.794123 0.013389 0.026604 0.731525 0.268475 0.153365 0.663738 0.19330
## 5 0.682816 0.055567 0.158197 0.766660 0.233340 0.138224 0.667435 0.27753
## 6 0.649866 0.065343 0.205620 0.778100 0.221900 0.126673 0.723367 0.21706
##    POVRAT0 WELFARE0  AVWELIN0    AVSSI0   MDFAMY0   AVSOCS0    AVRETR0
## 1 0.018173 0.012334 0.2138782 0.1670434 0.5668922 0.4957300 0.10195685
## 2 0.061842 0.016949 0.2097222 0.1751578 0.6703366 0.5436148 0.14839691
## 3 0.050593 0.062198 0.2000324 0.1859608 0.4109229 0.4636815 0.08882281
## 4 0.049103 0.020076 0.1752252 0.2310690 0.5016425 0.4659261 0.10959870
## 5 0.045093 0.045217 0.2026923 0.1845515 0.3964330 0.3908959 0.08489463
## 6 0.045790 0.041863 0.1260563 0.1451639 0.2851536 0.3996262 0.07501925
##       SHRAMI0    SHRASN0      SHRHIP0  NONFAM0P WNONFAM0P COMMUT20NWKH
## 1 0.000000000 0.06838246 0.0000000000 0.1518027 0.1518027    0.4722695
## 2 0.000000000 0.20338983 0.0000000000 0.1087571 0.1059322    0.2701613
## 3 0.001030715 0.13048856 0.0000000000 0.2161836 0.2053140    0.5501829
```

```
## 4 0.000000000 0.18848896 0.0001738828 0.1432447 0.1405317    0.3949681
## 5 0.005141388 0.20071188 0.0000000000 0.2486957 0.2347826    0.4884793
## 6 0.002453167 0.19603033 0.0015611062 0.3484670 0.3024764    0.5160018
##     WMEMP0P   WFEMP0P    VACHU0P   RNTOCC0P Gentrified
## 1 0.7495108 0.5150250 0.01496726 0.09307764          0
## 2 0.7279753 0.4362819 0.03013699 0.10000000          0
## 3 0.6936005 0.5840555 0.01610979 0.18078759          0
## 4 0.7232743 0.4737401 0.01641949 0.14406780          0
## 5 0.7313891 0.6134398 0.02428007 0.17391304          0
## 6 0.7416226 0.6194690 0.01733102 0.52744079          0

# count of 0s and 1s of the output variable in the training data
print(table(ny_data$Gentrified))

##
##    0    1
## 3522  990

# randomly splitting the data into training and testing sets
set.seed(12345)
tr <- sample.int(n = nrow(ny_data), size = floor(.8*nrow(ny_data)), replace =
FALSE)
ny_data_train <- ny_data[tr, ]
ny_data_test <- ny_data[-tr, ]

# count of 0s and 1s of the output variable in the training and testing sets
print(table(ny_data_train$Gentrified))

##
##    0    1
## 2816  793

print(table(ny_data_test$Gentrified))

##
##   0   1
## 706 197
```

Logistic Regression

```
# fitting a logistic regression model and printing the summary of the model
logistic_model <- glm(Gentrified ~ ., data = ny_data_train, family = "binomia
l")
summary(logistic_model)

##
## Call:
## glm(formula = Gentrified ~ ., family = "binomial", data = ny_data_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5087  -0.7344  -0.5859  -0.3723   2.7578
```

```
##
## 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 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3800.7  on 3608  degrees of freedom
## Residual deviance: 3553.4  on 3583  degrees of freedom
## AIC: 3605.4
##
## Number of Fisher Scoring iterations: 6
```

```
# obtaining the predictions on the testing set using the above model and prin
ting the confusion matrix
predlg1 <- predict(logistic_model, ny_data_test, type = 'response')
predlg1 <- ifelse(predlg1 > 0.5, 1, 0)
predlg1 <- as.factor(predlg1)
confusionMatrix(predlg1, ny_data_test$Gentrified, mode = 'everything', positi
ve = '1')
```

```
## Confusion Matrix and Statistics
##
```

```
##           Reference
## Prediction   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
##
```

```
# sub-setting the training and testing sets with only the variables which are inferred as statistically significant from the logistic model
ny_data_train <- subset(ny_data_train, select = c(OLD0, WRKSM0, MDFAMY0, SHRHIP0, NONFAM0P, COMMUT20NWKH, WFEMP0P, VACHU0P, RNTOCC0P, Gentrified))
ny_data_test <- subset(ny_data_test, select = c(OLD0, WRKSM0, MDFAMY0, SHRHIP0, NONFAM0P, COMMUT20NWKH, WFEMP0P, VACHU0P, RNTOCC0P, Gentrified))
```

Random Forest Classifier

```
# fitting a random forest classifier and printing the confusion matrix
rf1 <- randomForest(Gentrified ~ ., data=ny_data_train)
predrf1 <- predict(rf1, ny_data_test)
confusionMatrix(predrf1, ny_data_test$Gentrified, mode = 'everything', positive = '1')
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 683 178
##          1  23  19
##
```

```
##                Accuracy : 0.7774
##                  95% CI : (0.7488, 0.8042)
##     No Information Rate : 0.7818
##     P-Value [Acc > NIR] : 0.644
##
##                   Kappa : 0.0892
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.09645
##             Specificity : 0.96742
##          Pos Pred Value : 0.45238
##          Neg Pred Value : 0.79326
##               Precision : 0.45238
##                  Recall : 0.09645
##                      F1 : 0.15900
##              Prevalence : 0.21816
##          Detection Rate : 0.02104
##    Detection Prevalence : 0.04651
##       Balanced Accuracy : 0.53193
##
##        'Positive' Class : 1
##
```

Re-sampling data to balance the two classes in the output variable

```
# using ADASYN (Adaptive Synthetic Sampling) technique to generate samples of
the minority class
set.seed(1234)
ny_data_train_balanced <- ADAS(ny_data_train[,-ncol(ny_data_train)], ny_data_
train$Gentrified)
ny_data_train_balanced <- ny_data_train_balanced$data
names(ny_data_train_balanced)[names(ny_data_train_balanced) == 'class'] <- 'G
entrified'

# converting the output variable to a factor
ny_data_train_balanced$Gentrified <- as.factor(ny_data_train_balanced$Gentrif
ied)

# count of 0s and 1s of the output variable in the new training set
table(ny_data_train_balanced$Gentrified)
```

```
##
##    0    1
## 2816 2857
```

Random Forest Classifier (using the new training dataset)

```
# fitting a random forest classifier and printing the confusion matrix
rf2 <- randomForest(Gentrified ~ ., data=ny_data_train_balanced)
predrf2 <- predict(rf2, ny_data_test)
```

```
confusionMatrix(predrf2, ny_data_test$Gentrified, mode = 'everything', positi
ve = '1')
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 581 121
##          1 125  76
##
##                Accuracy : 0.7276
##                  95% CI : (0.6973, 0.7564)
##     No Information Rate : 0.7818
##     P-Value [Acc > NIR] : 0.9999
##
##                   Kappa : 0.2072
##
##  Mcnemar's Test P-Value : 0.8483
##
##             Sensitivity : 0.38579
##             Specificity : 0.82295
##          Pos Pred Value : 0.37811
##          Neg Pred Value : 0.82764
##               Precision : 0.37811
##                  Recall : 0.38579
##                      F1 : 0.38191
##              Prevalence : 0.21816
##          Detection Rate : 0.08416
##    Detection Prevalence : 0.22259
##       Balanced Accuracy : 0.60437
##
##        'Positive' Class : 1
##
```

Modeling for Los Angeles MSA dataset

2022-12-22

Description: This script is used to train, test, and interpret the machine learning models for the tracts of the Los Angeles MSA.

The datasets used in this script are the following:

- la_training_data.csv - This file is the training data for the Los Angeles MSA

```
# clearing the workspace
rm(list=ls(all=TRUE))
```

```r
# loading all the required libraries
library(readr, quietly = T)
library(data.table, quietly = T)
library(dplyr, quietly = T)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.table':
##
##     between, first, last

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(caret, quietly = T)
library(randomForest, quietly = T)

## randomForest 4.7-1.1

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':
##
##     margin

## The following object is masked from 'package:dplyr':
##
##     combine

library(smotefamily, quietly = T)

# reading in the training dataset
la_data = read.csv("la_training_data.csv")

# converting the output variable to a factor
la_data$Gentrified <- as.factor(la_data$Gentrified)

# creating a new column as the sum of multiple columns and dropping the indiv
idual columns
# combining the proportions of Black, American Indians, Asians, Native Hawaii
an, and other Pacific Islander population.
la_data['SHRNWT0'] = la_data['SHRBLK0'] + la_data['SHRAMI0'] + la_data['SHRAS
N0'] + la_data['SHRHIP0']
la_data <- subset(la_data, select = -c(SHRBLK0, SHRAMI0, SHRASN0, SHRHIP0))
```

```r
la_data <- select(la_data, 'SHRNWT0', everything())

# viewing the training dataset
head(la_data)
```

```
##      SHRNWT0  SHRWHT0  SHRHSP0   ADULT0   CHILD0     OLD0  WRKSM0 WORKER30
## 1 0.12111067 0.747333 0.281333 0.739556 0.260444 0.088000 0.97664      150
## 2 0.08985525 0.843478 0.147826 0.773913 0.224638 0.113043 0.97791       20
## 3 0.15067558 0.635742 0.407809 0.697480 0.302520 0.047055 0.97482      158
## 4 0.07223455 0.674621 0.408255 0.729765 0.270235 0.068688 0.97547       48
## 5 0.09818302 0.867239 0.090611 0.792277 0.207219 0.164816 0.98183      165
## 6 0.09731430 0.827174 0.178676 0.769210 0.230790 0.125764 0.95082      117
##    POVRAT0 WELFARE0 AVWELIN0    AVSSI0 MDFAMY0  AVSOCS0  AVRETR0  NONFAM0P
## 1 0.131091 0.095907 7300.000  8338.710   48221 10574.14  7473.333 0.339645
## 2 0.039039 0.083004 7621.333  8648.625   61714 11733.62 33019.167 0.272727
## 3 0.262712 0.182216 5999.467  8650.820   33094  9192.35  7755.844 0.295918
## 4 0.204979 0.162828 5272.195  7951.667   32121 10376.13  7578.400 0.482128
## 5 0.034881 0.071477 9707.236 10994.950   67293 13909.02 25312.252 0.252191
## 6 0.034036 0.031915 6657.867  7189.250   54875 11066.47 17671.254 0.326241
##   WNONFAM0P COMMUT20NWKH   WMEMP0P   WFEMP0P   VACHU0P RNTOCC0P Gentrified
## 1 0.2840562    0.4771372 0.6360032 0.5398230 0.02576393 0.48172          0
## 2 0.2608696    0.3753943 0.7021277 0.5265487 0.02702703 0.13899          0
## 3 0.2405248    0.4477549 0.6244411 0.5103550 0.03592814 0.85168          0
## 4 0.4209690    0.4930499 0.6568758 0.4913218 0.04605777 0.59250          1
## 5 0.2258935    0.4668803 0.6194424 0.5210450 0.02496715 0.14388          0
## 6 0.3163121    0.4375353 0.6766667 0.5435334 0.03988996 0.25653          0
```

```r
# normalizing the variables which are not in proportions (since proportions a
re already between 0 - 1)
minMax <- function(x) {
  (x - min(x)) / (max(x) - min(x))
}

la_data[c('WORKER30', 'AVWELIN0', 'AVSSI0', 'MDFAMY0', 'AVSOCS0', 'AVRETR0')]
<-
                                lapply(la_data[c('WORKER30', 'AVWELIN0', 'AV
SSI0', 'MDFAMY0', 'AVSOCS0', 'AVRETR0')], minMax)
head(la_data)
```

```
##      SHRNWT0  SHRWHT0  SHRHSP0   ADULT0  CHILD0   OLD0    WRKSM0  WORKER30
## 1 0.12111067 0.747333 0.281333 0.739556 0.260444 0.088000 0.97664 0.08370
## 2 0.08985525 0.843478 0.147826 0.773913 0.224638 0.113043 0.97791 0.01116
## 3 0.15067558 0.635742 0.407809 0.697480 0.302520 0.047055 0.97482 0.08816
## 4 0.07223455 0.674621 0.408255 0.729765 0.270235 0.068688 0.97547 0.02678
## 5 0.09818302 0.867239 0.090611 0.792277 0.207219 0.164816 0.98183 0.09207
## 6 0.09731430 0.827174 0.178676 0.769210 0.230790 0.125764 0.95082 0.06529
##    POVRAT0 WELFARE0  AVWELIN0    AVSSI0   MDFAMY0   AVSOCS0    AVRETR0
## 1 0.131091 0.095907 0.3882979 0.2214062 0.2411038 0.2510062 0.06302993
## 2 0.039039 0.083004 0.4053901 0.2296349 0.3085685 0.2785297 0.27848295
## 3 0.262712 0.182216 0.3191206 0.2296932 0.1654692 0.2182057 0.06541263
```

```
## 4 0.204979 0.162828 0.2804359 0.2111295 0.1606042 0.2463058 0.06391607
## 5 0.034881 0.071477 0.5163423 0.2919336 0.3364633 0.3301688 0.21348300
## 6 0.034036 0.031915 0.3541418 0.1908862 0.2743736 0.2626931 0.14903898
##    NONFAM0P WNONFAM0P COMMUT20NWKH   WMEMP0P   WFEMP0P   VACHU0P  RNTOCC0P
## 1 0.3396457 0.2840562    0.4771372 0.6360032 0.5398230 0.02576393 0.481725
## 2 0.2727273 0.2608696    0.3753943 0.7021277 0.5265487 0.02702703 0.138996
## 3 0.2959184 0.2405248    0.4477549 0.6244411 0.5103550 0.03592814 0.851681
## 4 0.4821287 0.4209690    0.4930499 0.6568758 0.4913218 0.04605777 0.592505
## 5 0.2521915 0.2258935    0.4668803 0.6194424 0.5210450 0.02496715 0.143889
## 6 0.3262411 0.3163121    0.4375353 0.6766667 0.5435334 0.03988996 0.256533
##   Gentrified
## 1          0
## 2          0
## 3          0
## 4          1
## 5          0
## 6          0
```

```
# count of 0s and 1s of the output variable in the training data
print(table(la_data$Gentrified))
```

```
##
##    0    1
## 2343  583
```

```
# randomly splitting the data into training and testing sets
set.seed(12345)
tr <- sample.int(n = nrow(la_data), size = floor(.8*nrow(la_data)), replace =
FALSE)
la_data_train <- la_data[tr, ]
la_data_test <- la_data[-tr, ]
```

```
# count of 0s and 1s of the output variable in the training and testing sets
print(table(la_data_train$Gentrified))
```

```
##
##    0    1
## 1878  462
```

```
print(table(la_data_test$Gentrified))
```

```
##
##   0   1
## 465 121
```

Logistic Regression

```
# fitting a logistic regression model and printing the summary of the model
logistic_model <- glm(Gentrified ~ ., data = la_data_train, family = "binomia
l")
summary(logistic_model)
```

```
##
## Call:
## glm(formula = Gentrified ~ ., family = "binomial", data = la_data_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1501  -0.7230  -0.4869  -0.2241   2.7219
##
## 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
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2325.2  on 2339  degrees of freedom
## Residual deviance: 2056.6  on 2317  degrees of freedom
## AIC: 2102.6
##
## Number of Fisher Scoring iterations: 5

# obtaining the predictions on the testing set using the above model and prin
ting the confusion matrix
predlg1 <- predict(logistic_model, la_data_test, type = 'response')
predlg1 <- ifelse(predlg1 > 0.5, 1, 0)
predlg1 <- as.factor(predlg1)
```

```
confusionMatrix(predlg1, la_data_test$Gentrified, mode = 'everything', positi
ve = '1')
```

```
## 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
##
```

```
# sub-setting the training and testing sets with only the variables which are
inferred as statistically significant from the logistic model
la_data_train <- subset(la_data_train, select = c(SHRNWT0, CHILD0, WRKSM0, PO
VRAT0, MDFAMY0, Gentrified))
la_data_test <- subset(la_data_test, select = c(SHRNWT0, CHILD0, WRKSM0, POVR
AT0, MDFAMY0, Gentrified))
```

Random Forest Classifier

```
# fitting a random forest classifier and printing the confusion matrix
rf1 <- randomForest(Gentrified ~ ., data=la_data_train)
predrf1 <- predict(rf1, la_data_test)
confusionMatrix(predrf1, la_data_test$Gentrified, mode = 'everything', positi
ve = '1')
```

```
## Confusion Matrix and Statistics
##
```

```
##           Reference
## Prediction   0   1
##          0 434 109
##          1  31  12
##
##                Accuracy : 0.7611
##                  95% CI : (0.7244, 0.7951)
##     No Information Rate : 0.7935
##     P-Value [Acc > NIR] : 0.9752
##
##                   Kappa : 0.0427
##
##  Mcnemar's Test P-Value : 7.632e-11
##
##             Sensitivity : 0.09917
##             Specificity : 0.93333
##          Pos Pred Value : 0.27907
##          Neg Pred Value : 0.79926
##               Precision : 0.27907
##                  Recall : 0.09917
##                      F1 : 0.14634
##              Prevalence : 0.20648
##          Detection Rate : 0.02048
##    Detection Prevalence : 0.07338
##       Balanced Accuracy : 0.51625
##
##        'Positive' Class : 1
##
```

Re-sampling data to balance the two classes in the output variable

```
# using ADASYN (Adaptive Synthetic Sampling) technique to generate samples of
the minority class
set.seed(1234)
la_data_train_balanced <- ADAS(la_data_train[,-ncol(la_data_train)], la_data_
train$Gentrified)
la_data_train_balanced <- la_data_train_balanced$data
names(la_data_train_balanced)[names(la_data_train_balanced) == 'class'] <- 'G
entrified'

# converting the output variable to a factor
la_data_train_balanced$Gentrified <- as.factor(la_data_train_balanced$Gentrif
ied)

# count of 0s and 1s of the output variable in the new training set
table(la_data_train_balanced$Gentrified)
```

```
##
##    0    1
## 1878 1964
```

Random Forest Classifier (using the new training dataset)

```
# fitting a random forest classifier and printing the confusion matrix
rf2 <- randomForest(Gentrified ~ ., data = la_data_train_balanced)
predrf2 <- predict(rf2, la_data_test)
confusionMatrix(predrf2, la_data_test$Gentrified, mode = 'everything', positi
ve = '1')
```

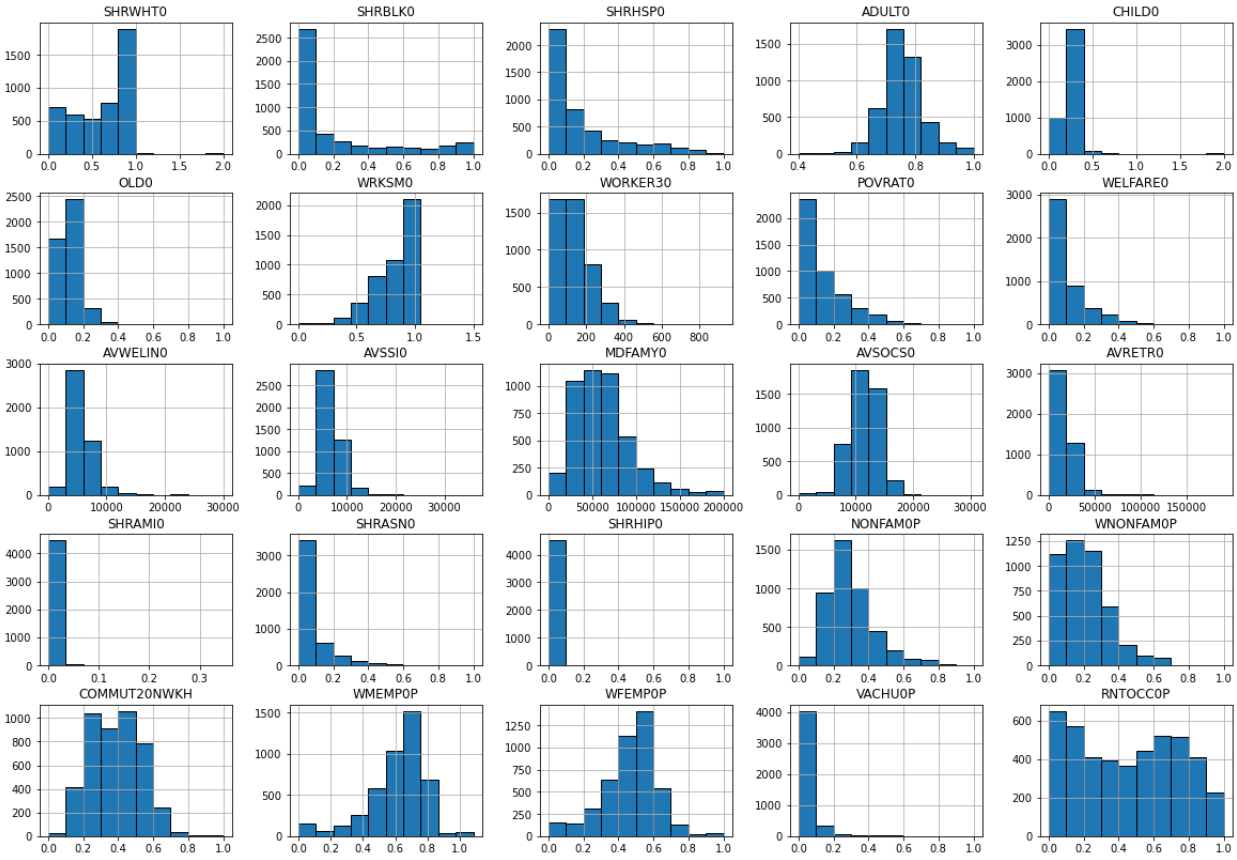
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 367  67
##          1  98  54
##
##                Accuracy : 0.7184
##                  95% CI : (0.6801, 0.7545)
##     No Information Rate : 0.7935
##     P-Value [Acc > NIR] : 0.99999
##
##                   Kappa : 0.2151
##
##  Mcnemar's Test P-Value : 0.01952
##
##             Sensitivity : 0.44628
##             Specificity : 0.78925
##          Pos Pred Value : 0.35526
##          Neg Pred Value : 0.84562
##               Precision : 0.35526
##                  Recall : 0.44628
##                      F1 : 0.39560
##              Prevalence : 0.20648
##          Detection Rate : 0.09215
##    Detection Prevalence : 0.25939
##       Balanced Accuracy : 0.61776
##
##        'Positive' Class : 1
##
```

Data Visualizations for Gentrification Model

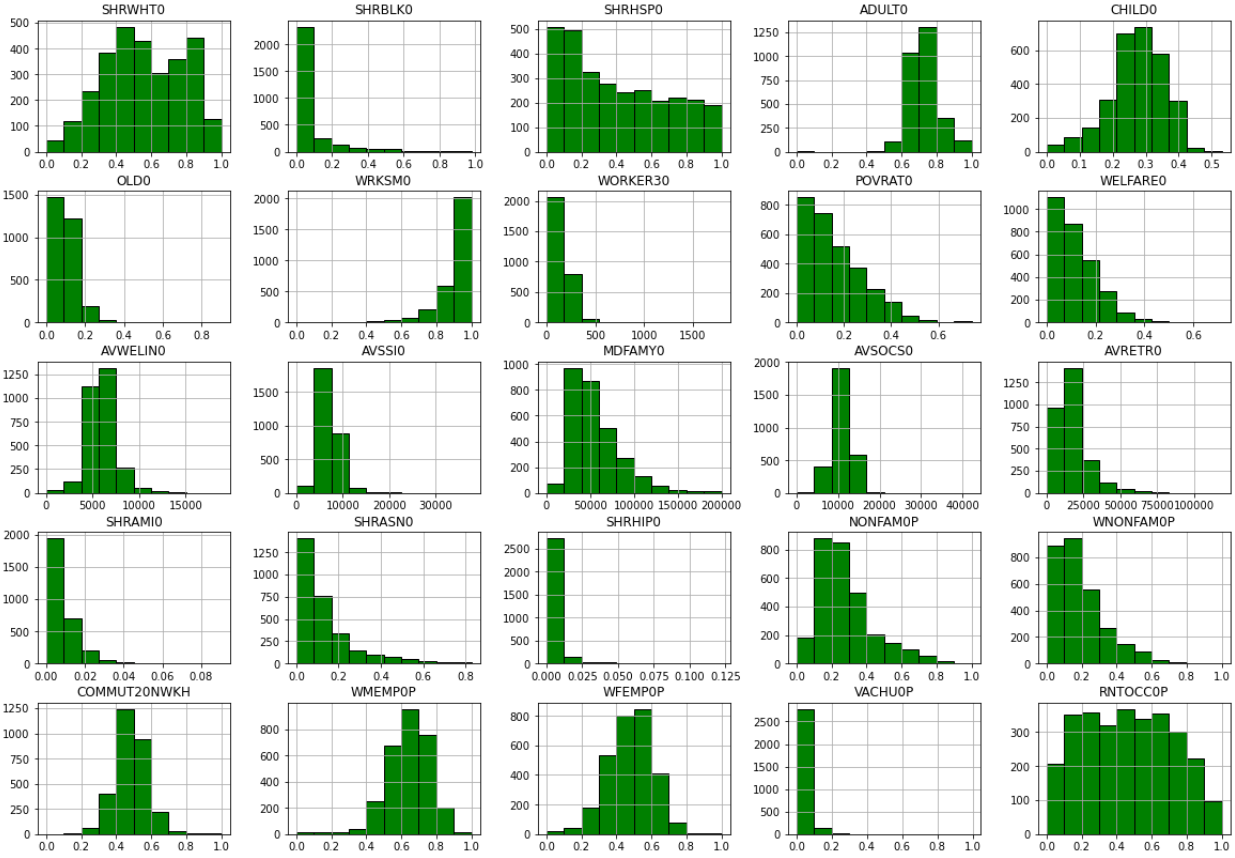
This notebook contains code that creates histograms/heatmaps for the independent variables of both the New York MSA and Los Angeles MSA datasets

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

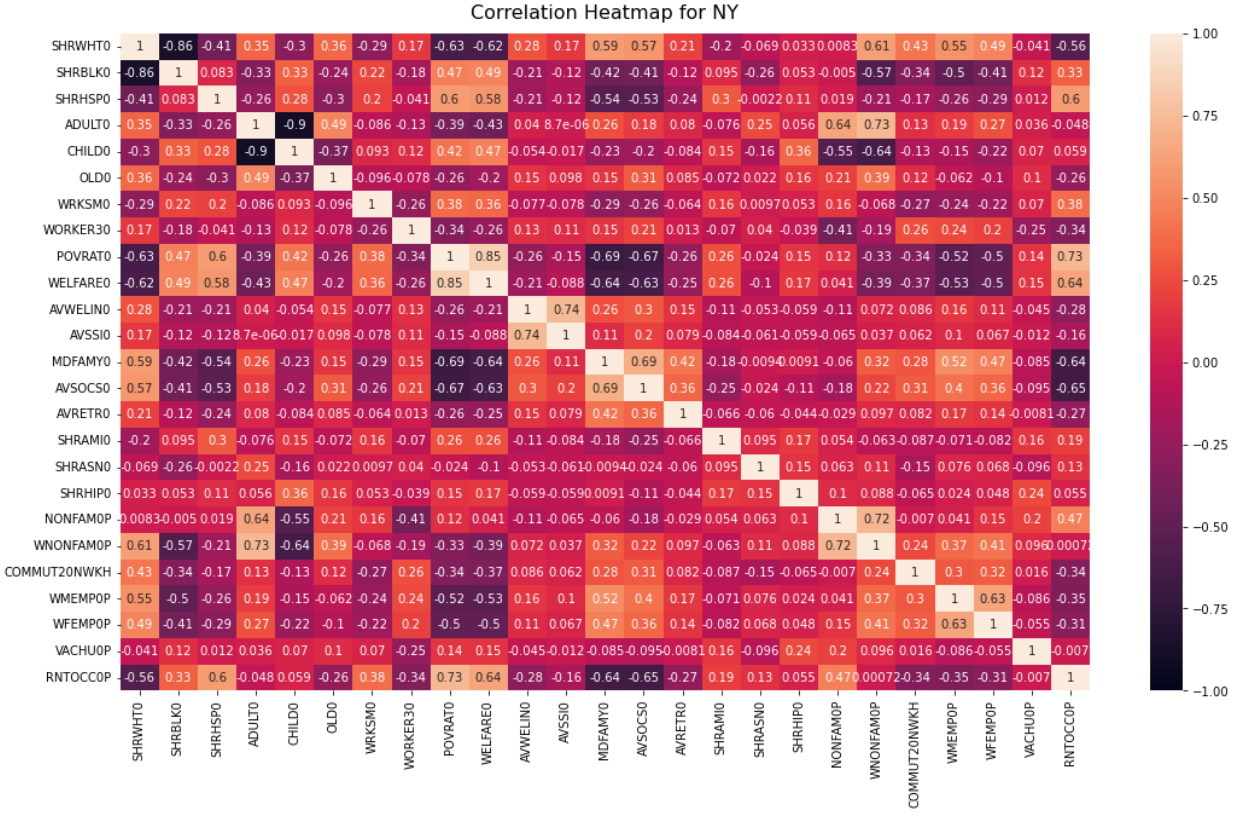
In [5]:
```
#read in the NY MSA file
NY = pd.read_csv("/Users/vanessavenkataraman/Desktop/DATA-200/DATA200 Group Pro
#read in the LA MSA file
LA = pd.read_csv("/Users/vanessavenkataraman/Desktop/DATA-200/DATA200 Group Pro

#drop dependent variable
NY.drop(['Gentrified'], axis="columns", inplace=True)
#histogram for NY MSA variables
NY.hist(edgecolor='black', linewidth=1, figsize=(20, 14))
plt.show()
```



In [8]:
```
#drop dependent variable
LA.drop(['Gentrified'], axis="columns", inplace=True)
#histogram for LA MSA variables
LA.hist(color='green', edgecolor='black', linewidth=1, figsize=(20, 14))
plt.show()
```

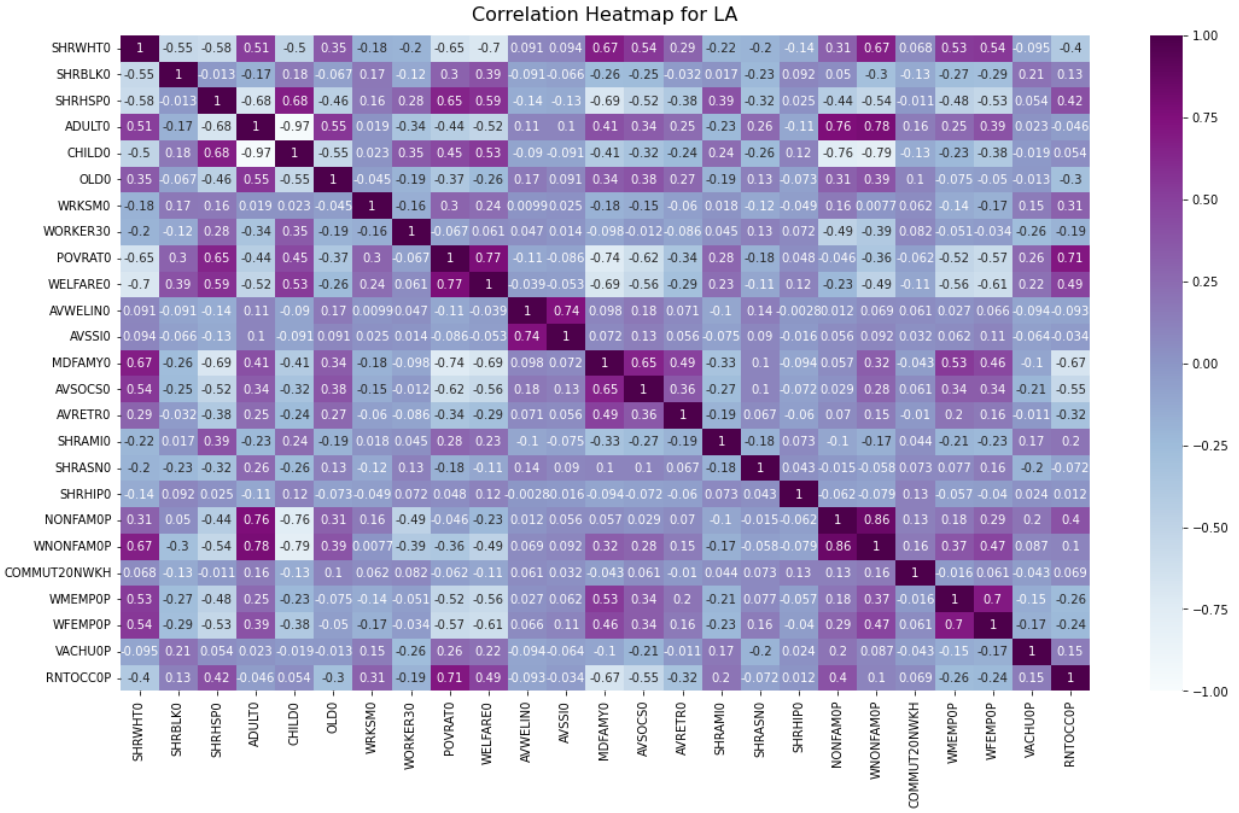


In [6]:
```
#Correlation heatmap for NY MSA variables
plt.figure(figsize=(18, 10))
heatmap = sns.heatmap(NY.corr(), vmin=-1, vmax=1, annot=True)
heatmap.set_title('Correlation Heatmap for NY', fontdict={'fontsize':16}, pad=1
```



In [10]:
```
#Correlation heatmap for LA MSA variables
```

```
plt.figure(figsize=(18, 10))
heatmap = sns.heatmap(LA.corr(), cmap = 'BuPu',vmin=-1, vmax=1, annot=True)
heatmap.set_title('Correlation Heatmap for LA', fontdict={'fontsize':16}, pad=1
```



Author Contributions

Vishwesh Srinivasan, Vanessa Venkataraman, and Hanzhen Wang read past research papers on using the census database, and Hanzhen Wang came up with the idea presented. Vishwesh Srinivasan, Vanessa Venkataraman, and Hanzhen Wang formalized and refined the concept and defined the project's scope.

Vishwesh Srinivasan and Vanessa Venkataraman cleaned and formulated the dataset to match the requirements. Vanessa Venkatraman created the visualizations presented. Vishwesh Srinivasan developed the modeling process presented and built the models with the help of Hanzhen Wang.

Vishwesh Srinivasan, Vanessa Venkataraman, and Hanzhen Wang contributed to preparing the presentation and writing the report.