Gentrification Within The United States

Damian Franco
Dept. of Computer Science
University of New Mexico
dfranco24@unm.edu

Meiling Traeger
Dept. of Computer Science
University of New Mexico
meilingt@unm.edu

***Abstract*—This project is intended to explore the correlation between certain population dynamics and gentrification. Areas will be investigated based on their median home value, average household income, education statistics, and racial demographics. All of these factors were used in various different analytical approaches such as information theory and machine learning in attempt to find a correlation between areas that are considered gentrified, eligible for gentrification, and non-gentrified. We found that there are many correlations between various values of Shannon entropy and the gentrification status of an area. High Shannon entropy indicated that an area is likely to be considered gentrified and low Shannon entropy corresponded with non-gentrified areas. Mutual information was calculated between two areas of gentrification, non-gentrification, and the similar correlations were found to the Shannon entropy correlations. Lastly, we forecasted gentrification status with the use of a perceptron model and only racial demographics to find a very promising correlation between ethnicity and gentrification. Overall, a very promising model for predicting gentrification of an area was produced through our research.**

***Index Terms*—Gentrification, Population Dynamics, Shannon Entropy, Mutual Information, Race, Income, Education**

I. Introduction

Recently the United States has seen the greatest increase in housing prices over the past year within the United States. Although this is partially due to the pandemic, the rising cost of living is not a new trend. Many researchers found that millennials are becoming homeowners at lower rates than previous generations. Clearly, it has become harder for the average American to find a place to live. There are a variety of factors that can impact the price of homes: available inventory, interest rates, etc. Many other researchers have documented a link between housing prices and gentrification. Gentrification can be characterized as the process of high-income households moving into low-income urban areas, which often results in existing residents, often member of marginalized groups, becoming displaced. This is just one reason that leads many to believe that there may be a correlation between racial demographics and the gentrification of an area.

We aim to explore the connection between racial demographics and socioeconomic shifts within the United States. By utilizing median housing prices, median household income, education, and racial population data we attempted to create a model that can classify areas of gentrification. Furthermore, utilize this to predict if areas will be gentrified in the future. Through this, we hope that home buyers and renters can make more informed decisions as to where they should look for homes. Additionally, human population dynamics are very complicated and there are seemingly endless factors that could contribute to the gentrification of an area. Complex systems, such as the one we are investigating, are very difficult to condense to a small set of characteristics, but we still want to attempt to do so because gentrification is a important topic within the United States.

II. Methods

Quantitatively defining gentrification is not a simple task as there are many factors contributing to it. Additionally, ones own social and political perspectives can influence its definition. Thus, we utilized a criteria for gentrification similar to the one defined by the National Community Reinvestment Coalition (NCRC) [1] and retrieved relevant data-sets accordingly. Education is not considered when checking for eligibility in an area.

TABLE I
Requirements for Gentrification and Eligibility

	Eligible	**Gentrified**
Population	above 500	above 500
Median Home Value (Percentile)	less than 40%	above 60%
Median Household Income (Percentile)	less than 40%	Increased
Education (Percentile)	N/A	above 60%

These criterion (Table I) are qualified only for areas within the U.S. excluding Puerto Rico. To maintain accuracy, the process of checking for eligibility and gentrification for each area will be conducted each year to account for changes. We examine the percentile of the current median home value, median household income, and education. Gentrification and eligibility was represented as a binary integer where 0 indicated a non-gentrified or ineligible are and a 1 represents a gentrified or eligible area. These binary integers are stored in the tuple uniquely responding to each area within the data set indicating eligibility and gentrification status.

Data sets were chosen based on the requirements established by earilier in Table I. Therefore, data sets used must include the median home value, the median household income, and educational statistics of an area. Along with these data sets, racial demographic information was was necessary to test our hypothesis.

Originally, zip code data was attempted to be pulled as it would allow for more precision in out findings. Unfortunately, data sets that have the exact features that we were seeking were not separated by zip code, instead most were separated by counties or state. The data sets we did settle on had the most complete data tables that were the size of a county. All data sets used are courtesy of the U.S. Census Bureau and are provided below:

- Population, including racial demographics (2010-2019) [2]
- Median Home Values (US Dollars) (2010-2019) [3]
- Median Household Income (US Dollars) (2010-2019) [4]
- Educational Attainment (2010-2019) [5]

This data proved to be effective for conducting research as each data set was generally consistent in the counties it contained as well as the years covered and covered all necessary features for analysis.

Data sets were merged into one, large, data set with the education, racial demographics, median household income, and median home value of a county combined. This was done by matching counties through their ID values. Other information that was also included in the merged data set is gentrification status, eligibility of gentrification status, and whether the county experienced an increase in median household income from the previous year. Due to this last feature, we had to remove all data from the year 2010 to account for a change between years.

Information Theory in the most basic sense is the link between entropy and information. Through information theory we are looking to study methods of communication and storage. Shannon Entropy, being a part of Information Theory can be defined as the amount of uncertainty involved in a random process. In Fig 1, the equation for Shannon entropy is defined. Mutual Information is the measure of mutual dependence between two variables. More specifically, meaning that with implementation we are looking to obtain information about one random variable by observing another random variable. Fig 2, below shows the relation of information between two variables, X and Y. The shared information is the overlapping portion of the two circles.

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

Fig. 1. The equation for calculating Shannon Entropy.

Our forecasting model uses a Single Perception as its core component. This model aims to simply determine whether gentrification can be identified solely on racial demographics, since we originally did not account for racial demographics within our investigation. A perceptron model is a method for the supervised learning of binary classifiers [6]. This algorithm makes predictions based on a linear predictor function to determine a set of weights for each feature of a data point,

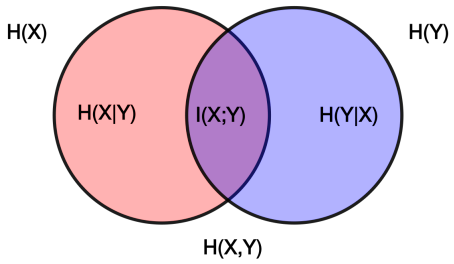

Fig. 2. The Venn diagram shows Mutual Information between two independent variables X and Y.

as shown in Fig. 3. Another primary aspect to the perceptron is a threshold or bias which gives a base value to the total predictor function. The predictor function then calculates an output value based on the weights and features and identifies it based on that value.

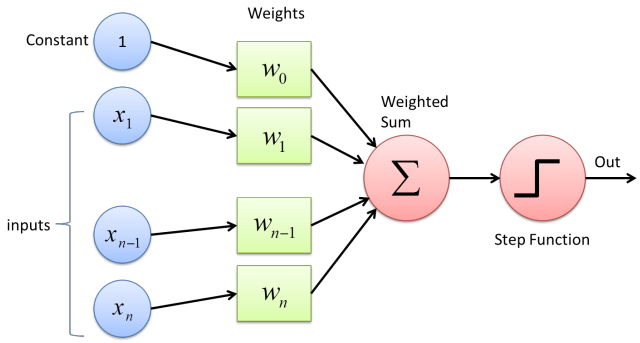

Fig. 3. Shows how the perception model makes predictions through inputs being used by a linear predictor function to determine the set of weights for each input. Then the weights will be used for a total predictor function to output a prediction based on the input.

Prior to training, our model applies polynomial feature transformation on the stratified data. This technique was applied for two reasons. The first being that we were not concerned with specific racial demographic qualities, but rather the racial makeup of a county as a whole, which meant we didn't need to analyze the resulting weights for specific features of our input. Furthermore, polynomial transformation allows the use of a simpler model, as we now have less dimensions to look at and reduces the chance the model succumbs to the curse of dimensionality. The input degree d signifies to what degree the data should be transformed. The model was tested with various values of d. After applying polynomial transformation, the Perception model could be applied. Our overall goal with this model is to use this in tangent with our information theory findings and create accurate predictions of gentrification.

III. Results

A. Data Categorization Analysis

First, we wanted to get a full understanding of the scope of

the project. Through calculations of percentiles, we were able to find where all counties lie regarding the median income and education data. We choose to focus on these two factors and leave out median household price because our requirements section requires that these two factors must be above a certain percentile to either be eligible or gentrified. You can see that the density falls below the 60th percentile of both education and household income, but outliers to appear to exist within this data.

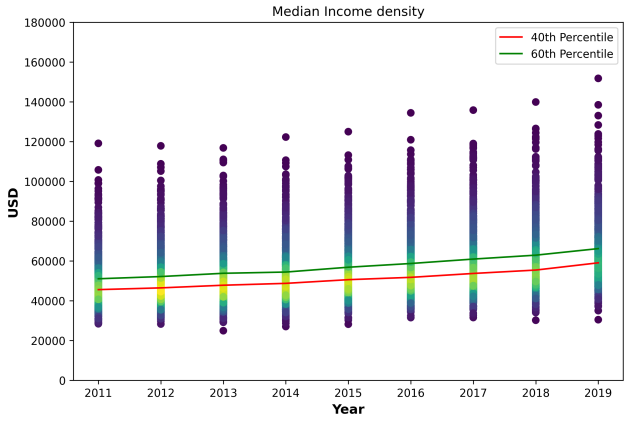

Fig. 4. All counties are plotted from 2011-2019 where they lie according to their median household income. Density colors are shown where purple is the lightest density and yellow is the highest density. The 40th and 60th percentiles of the year are plotted alongside the density.

Fig. 5 shows the median income density of all counties and Fig. 6 shows the income density for all counties education data. We believe the outlier within both of these graphs is San Francisco county because it is notoriously known for having very high education and median household income averages. It is also known to be a prime example of a gentrified area.

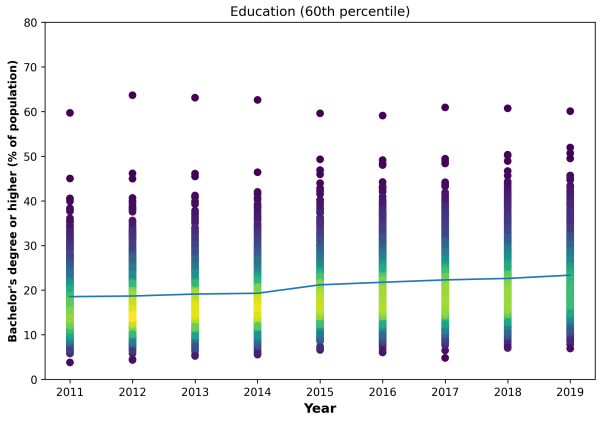

Fig. 5. All counties are plotted from 2011-2019 where they lie according to their percentage of individuals with a bachelors degree of higher within the county. Density colors are shown where purple is the lightest density and yellow is the highest density. The 60th percentile of the year are plotted alongside the density to show where most counties lie according to the percentiles.

We then categorized all of the counties based on our requirement table represented in Table I. This plot is very interesting because most counties are considered neither gentrified or eligible.These results are shown in Fig. 6. Although, there is a constant increase in counties that are considered gentrified. Meanwhile, the eligible trends bridge the middle ground of gentrified and neither categories. It is very interesting to note that all of these three lines look like they could be converging together within the future due to the neither category trending down and the gentrified category trending up. Eligible looks to be staying somewhat steady due to eligible counties being considered gentrified over time.

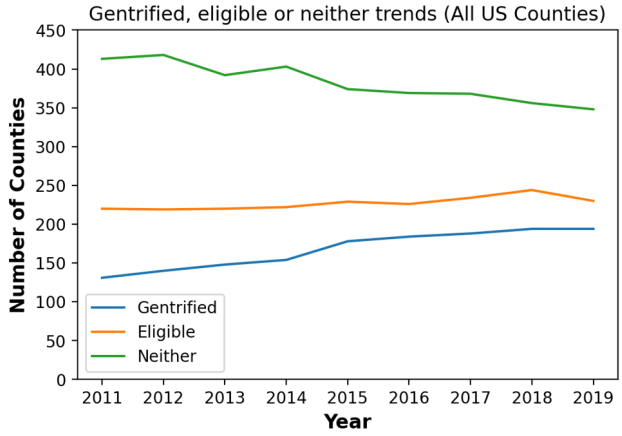

Fig. 6. Plot above shows all trends for all gentrified, eligible for gentrification or neither counties. Convergence here is something interesting to note. All counties were placed in a category based on the requirements table (Table I).

B. Information Theory Analysis

After getting a better view of the data we are dealing with, we decided to know apply Shannon entropy calculations to all counties and make comparisons based on gentrification status. We wanted to find a correlation through numerical values of uncertainty and relate that to their status. First, we found wanted to select a small sample of counties based on their gentrification status. Within this sample case we look into four distinct counties. Two of these counties are considered gentrified (Los Angeles & Denver) and the other two are considered non-gentrified (Milwaukee & Bernalillo).

You might be able to make an assumption on which counties are considered gentrified just by viewing Fig. 7. The areas that are considered gentrified are Los Angeles and Denver counties. As you might be able to see, both of the gentrified counties have a much higher median home value and median household income Shannon entropy than the other two non-gentrified counties. Milwaukee and Bernalillo counties are considered non-gentrified and the uncertainty calculated within those two counties is much lower than the two counties. The extremely interesting note to make between the two non-gentrified counties entropy's is that Milwaukee has a much higher degree of uncertainty compared to Bernalillo county. This is partially due to the fact that Milwaukee county is considered eligible for gentrification and Bernalillo is not. We

conclude that this sample case can maybe conclude that if the Shannon entropy and the level of uncertainty is higher within counties, then the more likely they are to be gentrified.

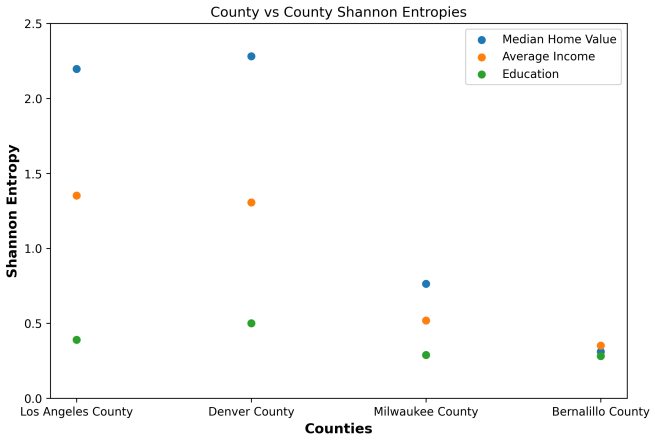

Fig. 7. This plot shows the Shannon entropy values for median house value, average income and education calculated for counties considered gentrified (Los Angeles & Denver) and two considered non-gentrified (Milwaukee & Bernalillo). Although Milwaukee county is considered eligible for gentrification which makes the uncertainty values very different from Bernalillo county.

To expand on our conclusion from the sample case, we decided to look into the uncertainty of all counties and try to find a correlation between all counties and Shannon entropy. Calculating all the Shannon entropy's and placing them on a density plot could then be compared to the status of each county as shown in Fig. 6. The three characteristics that we investigated the uncertainty were the median home value (Fig. 8), average household income (Fig. 9) and the average percentage of individuals with a bachelor's degree (Fig. 10) of a the year 2019. We plotted all Shannon entropies on heat map which show some interesting results. We can see that within Fig. 8, there is much more uncertainty within the median home value of each county. Meanwhile, there is less uncertainty within the median household income and even lesser uncertainty of education data within these areas. All plots indicate that the more uncertainty of a variable may be correlated to the gentrification which is an assumption that will be important when we investigate this further.

These three figures initially may seem to not relay much information to our goal because the comparison between Shannon entropies and the gentrification status of a county is not present. Knowing that, we separated data of all the gentrified counties, eligible counties and all non-gentrified counties and calculated the average Shannon entropy for all areas and status of each area in 2019. From the each counties categories, we found some very interesting stats. We see that the average uncertainty for median home and average household income value is very high compared to other entropy values calculated (Table II & III). In fact, all of the values this occurs within the education data as well (Table IV). We can also see that the non-gentrified counties display the lowest amount of Shannon entropy and eligible county are

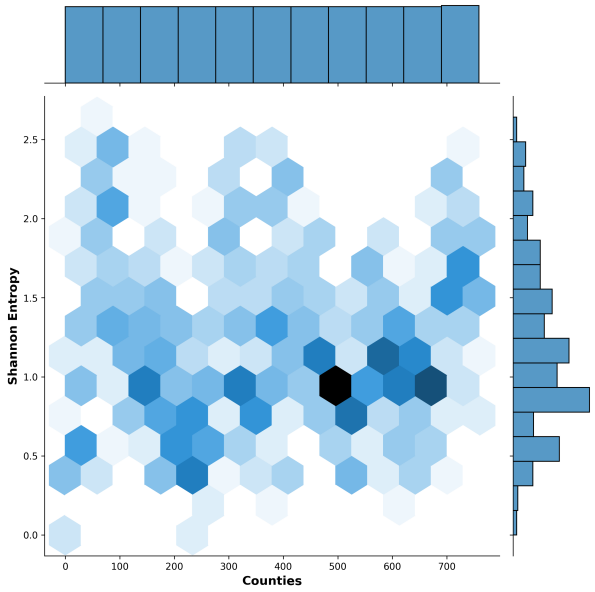

Fig. 8. Heat map of Shannon entropies that are calculated with the use of the average home value data. This plot shows a strong concentration within the 1.0 area with a large amount of higher levels entropies ($>$ 1.5) of uncertainty when compared to the other heat maps.

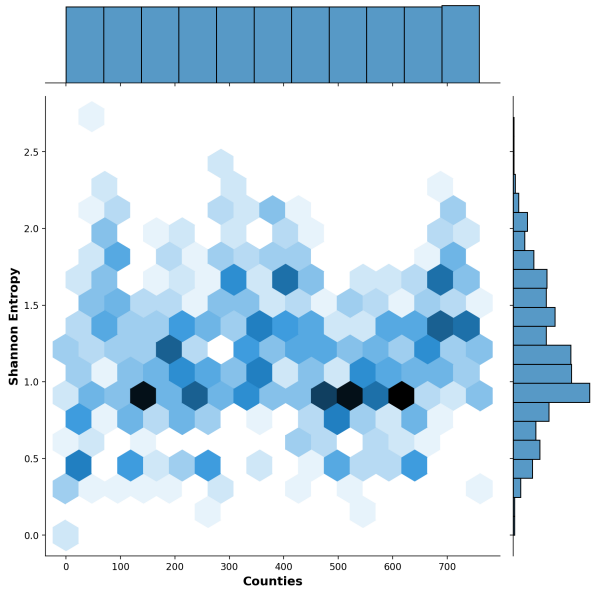

Fig. 9. Heat map of Shannon entropies that are calculated with the use of the median household income data. This plot shows a strong concentration within the 0.9-1.0 area with some high level entropies ($>$ 1.5) of uncertainty when compared to the other heat maps, but not as much as the average home value data.

somewhat of a "middle ground" that lays between gentrified and non-gentrified counties. This indirectly correlates with the three heat map figures and the assumption we made about those figures as well. Shannon entropy values for home value are much higher than the other two values calculated for the other variables.

Now, a proper conclusion Shannon entropy values can

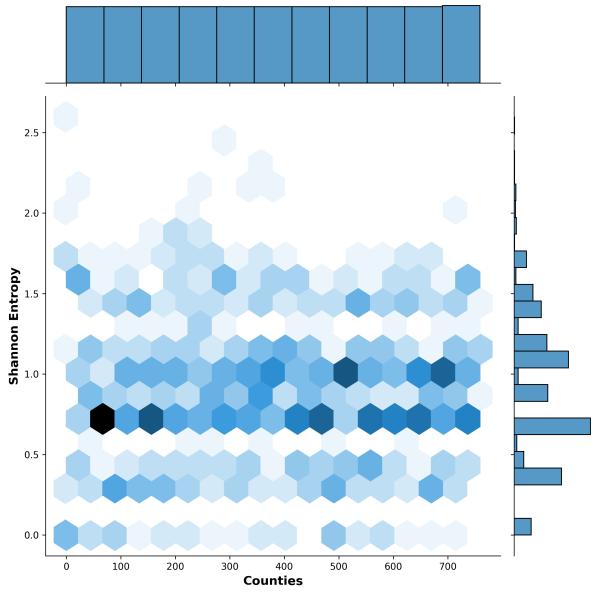

Fig. 10. Heat map of Shannon entropies that are calculated with the use of the education data. This plot shows a strong concentration within the 0.7-0.8 area with a smaller amount of high level entropies (> 1.5) of uncertainty when compared to the other heat maps, but not nearly as much as the other two heat maps. This is the only graph that has strange behavior to us. We believe that these values should be lower and the concetration should lie between 0.2-0.3.

TABLE II
SHANNON ENTROPY OF MEDIAN HOME VALUE (2019)

Status	Average Shannon Entropy
Gentrified	2.09715972
Eligible	1.35162892
Non-Gentrified	0.28199895

be made. Through all of research there has been a strong indication that most areas that have a high uncertainty within the three variables of gentrification we investigated seem to be categorized as gentrified. Most areas that have low Shannon entropy values do end up being categorized as non-gentrified. Although this is a great finding for us, there are many outliers that exist where this categorization completely fails. Lets take a look at San Francisco county. This county is notoriously known for being a highly gentrified area, but the Shannon entropy calculations showed up as very low for each variable when we took a deeper look into it. We believe since it has been gentrified and the median home values, average income

TABLE III
SHANNON ENTROPY OF AVERAGE HOUSEHOLD INCOME (2019)

Status	Average Shannon Entropy
Gentrified	1.30552884
Eligible	0.65164412
Non-Gentrified	0.352211389

TABLE IV
SHANNON ENTROPY OF EDUCATION (2019)

Status	Average Shannon Entropy
Gentrified	0.499975
Eligible	0.389975
Non-Gentrified	0.289975

TABLE V
MUTUAL HOME VALUE INFORMATION OF AREAS

Comparison	Average Mutual Information
Gentrified vs. Gentrified	0.37241
Non-Gentrified vs. Non-Gentrified	0.17875
Gentrified vs. Non-Gentrified	0.01643

values, and education averages have somewhat "leveled out" and converged to a more stable point. This can also occur in other areas that are already know to be gentrified over a long period of time. In order to compute higher accuracy results and mediate this problem we can always choose a specific time period to investigate further into.

Taking the data we used for computing Shannon entropies, we then set our eyes on investigating mutual information between all the counties. We wanted to calculate how much mutual information exists between counties considered gentrified with other counties considered gentrified and non-gentrified. Non-gentrified counties were also compared with other counties that also are categorized as non-gentrified. Values were generated by the JIDT toolkit [8] and are displayed in Table V. Results shown here are only for home value and not displayed for the other two variable due to the similar results in those variables. Many of the values generated displayed more mutual information was present between the areas that were categorized in the same group and gentrified compared to0 non-gentrified counties appear to have lower mutual information. We were expecting these results due to the findings presented in the previous tables, but it is always great to double check to see if any other approaches can be used to produce the same results. These results only solidifies our findings within our Shannon entropy research.

C. Racial Demographic Prediction Analysis

The perceptron model was trained to identify whether a given county was gentrified based on the transformed racial demographic data. The perceptron model will then try to predict gentrification status of counties in the test set. A test score is assigned from 0 to 1, representing the portion of counties which were correctly labeled. A value of 1 would be a perfect test score and represent that our model correctly labeled every county of the test set. Testing scores were generated for different values of d of the polynomial feature transformation. In Table VI, you can see each output of the perceptron model for various values of d. These outputs are the average testing score over 50 iterations. The testing scores

TABLE VI
PERCEPTRON MODEL RESULTS

Degree (d)	Testing Score
1	0.6724
2	0.6990
3	0.6573
4	0.7175
5	0.6752
6	0.6903
7	0.6609
8	0.6581
9	0.6767
10	0.6344

that were produced averaged a value of 0.67 over every d. This implies that our model was able to predict with fair accuracy whether a county was gentrified based purely off of racial demographic data, and that there are some consistencies when looking at the racial demographics of gentrified counties. We can then conclude that that there does exist a correlation between racial demographics of an area and gentrification.

IV. DISCUSSION AND CONCLUSIONS

In this paper, we attempted to determine whether there exists a correlation between various factors of an area and gentrification. We developed a model that investigated this correlation and used US Census data of every United States county from 2010-2019. Although not perfect, this model certainly hints that race, median home value, average household income and education has some degree of influence in the gentrification of an area. This model could likely be improved by in various ways. We believe that we have a very strong start to a great model for predicting gentrification within an area. Through our model, many other may be able to take smaller area sizes and more closely predict gentrification within those areas rather than full counties like how we did. Our approach in model could also potentially be used to identify other unknown correlated features in a different context, whether that be social, biological, or anything beyond. Beyond the model, different data sources could be utilized like real estate data or using zip-codes instead of counties like we stated before. Nonetheless, It's clear that there are a lot of different factors which impact gentrification, and predicting it well is an incredibly difficult task and we are happy to contribute a bit to the research of this very complex system.

V. CONTRIBUTIONS

The information from this project was gathered jointly between both partners. Each group member researched the data from the US Census to determine the best data sets to analyze. After each member collected their desired data sets we worked together to decide which ones would be the best representation for our data set and that would fit into our model the best. To facilitate collaboration on the document and code we focused majority of our efforts on Google Docs and Google Colab with the use of a Jupyter notebook. This allowed us both to be able to write and edit with easier access. After sections where written, we transferred them over too Overleaf. Damian worked to implement Shannon Entropy, while Meiling worked to implement Mutual Information. Together, Meiling and Damian interpreted the findings and results.

REFERENCES

[1] Richardson, Jared, et al. "*Gentrification and Disinvestment 2020: Do Opportunity Zones benefit or gentrify low-income neighborhoods?*" NCRC, National Community Reinvestment Coalition

[2] U.S. Census Bureau (2020). *Race*, 2010-2019 American Community Survey 1-year estimates*. Retrieved from https://data.census.gov/cedsci/table?q=race

[3] U.S. Census Bureau (2020). *Median Home Value (Dollars)*, 2010-2019 American Community Survey 1-year estimates. Retrieved from https://data.census.gov/cedsci/table?q=B25077%3A MEDIAN VALUE (DOLLARS)

[4] U.S. Census Bureau (2020). *MEDIAN INCOME IN THE PAST 12 MONTHS**, 2010-2019 American Community Survey 1-year estimates*. Retrieved from https://data.census.gov/cedsci/table?q=Income

[5] U.S. Census Bureau (2020). *Educational Attainment**, 2010-2019 American Community Survey 1-year estimates*. Retrieved from https://data.census.gov/cedsci/table?q=education

[6] L. Kanal, "*Perceptron*", Encyclopedia of Computer Science, Jan, 2003

[7] M. Mitchell, "*Complexity: A Guided Tour*", Oxford: Oxford University Press, 2011

[8] J. T. Lizier, "*JIDT: An Information-Theoretic Toolkit for Studying Dynamics of Complex Systems*", Frontiers in Robotics and AI 1:11, 2014, https://github.com/jlizier/jidt