Race and Gentrification within the US

Damian Franco, Nicholas Livingstone, Datenzing Tamang
Department of Computer Science
University of New Mexico
Albuquerque, NM 87106

***Abstract*—In recent years, affluent households moving into disinvested areas or neighborhood has rapidly increased, altering the racial, socioeconomic, and institutional makeup of many urban communities in the US. Gentrification is a commonly used term used to describe this migration and the the consequential effects. With the process of gentrification, there are often negative effects which can be seen in the communities affected by it. A transition of low-value neighborhoods into high-value areas tends to cause long-term residents and business to be displaced due to increasing rents, mortgages, and property taxes. In this paper, we explore various factors contributing to gentrification in the US and try to see if there exists a correlation between racial demographics and gentrification.**

***Index Terms*—gentrification, race, income, education, housing, population.**

I. Introduction

How racial demographics and gentrification of an area correlate

According to a recent White House article, 2021 has seen the greatest increase in housing prices over the past year within the US [1]. Although this is partially due to the pandemic, the rising cost of living is not a recent trend. In 2018, the Urban Institute found that millennials are becoming homeowners at lower rates than previous generations [2]. 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. However, in 2010, the National Bureau of Economic research documented a link between housing prices and gentrification [3]. 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.

In this paper, we aimed 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 models that can predict where gentrification will occur. Through this, we hope that home buyers and renters can make more informed decisions as to where they should look for homes. Additionally, legislative bodies can use this information to provide financial support to those groups which are the most susceptible to the negative impacts of gentrification.

II. Methodologies

A. Requirements for Gentrification

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 defined by the National Community Reinvestment Coalition (NCRC) [4] and retrieved relevant data-sets accordingly.

TABLE I
Requirements for Gentrification

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

Table 1 represents NCRC's requirements for an area to be considered eligible for gentrification and/or already gentrified.

Education is not considered when checking for eligibility in an area. The population is not taken into account to check for the gentrification of an area.

These criterion 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, as well as the percentile of the current median home value, median household income, and education.

B. Data Collection and Representation

Data sets were chosen based on the requirements established by the National Community Reinvestment Coalition. 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. Considering that, the data sets we did settle on that had the most complete data tables 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) [5]
- Median Home Values (US Dollars) (2010-2019) [6]
- Median Household Income (US Dollars) (2010-2019) [7]
- Educational Attainment (2010-2019) [8]

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.

C. Technologies Used

The programming language used for the project was Python, as well as common ML libraries such as sci-kit learn, numpy, pandas, and others. Google Colaboratory (Colab) was used as a platform to develop and work in.

III. Data Analysis

After merging and cleaning of the data sets, initial data analysis could now be conducted. First, counties that were eligible for gentrification were found based on the criteria in Table I. Every data instance was appended a feature to indicate the eligibility of gentrification with a 1 or a 0. 1 indicates that the county is eligible/gentrified and 0 indicates it is ineligible/not gentrified.

In both Fig. 1 and Fig. 2, each blue dot represents a county and where their current median household income and education level. Some outliers exist in both figures. In Fig. 1, the outliers here are areas that have a very large median income like San Francisco County, California which has one of the highest median incomes of all the U.S.. For Fig. 2, the outliers here are places such as counties in NM, which contain a high concentration of college educated people due to National Laboratories and the like. The percentiles for both median household income and education give insight on where most counties fit in the criteria of gentrification and eligibility of gentrification.

In Fig. 3, a visualization of each county's classification based on the criteria given in Table I is displayed.

In the visualization in Fig. 3, 35%-40% of the counties are classified as gentrified or eligible. The number of counties which are gentrified is slowly rising over time and many

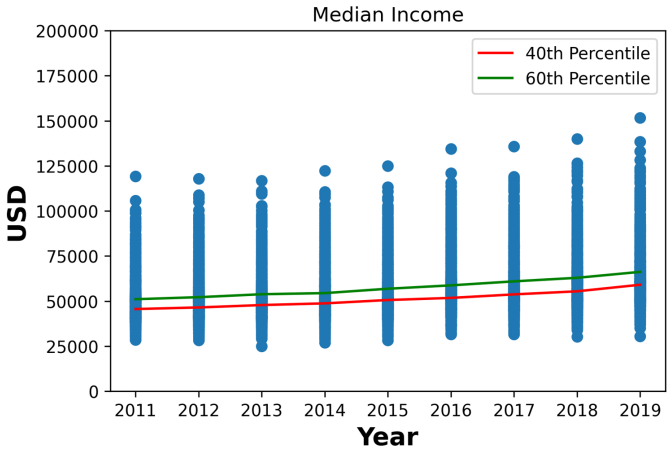

Fig. 1. Shows the 40% and 60% percentile of median household income and its density from 2011-2019.

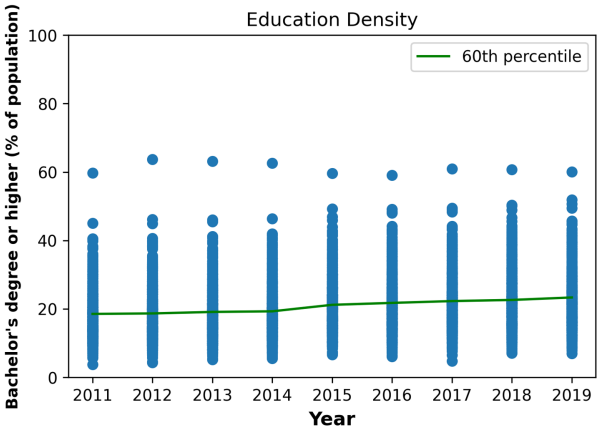

Fig. 2. Shows the 60% percentile of education and its density from 2011-2019.

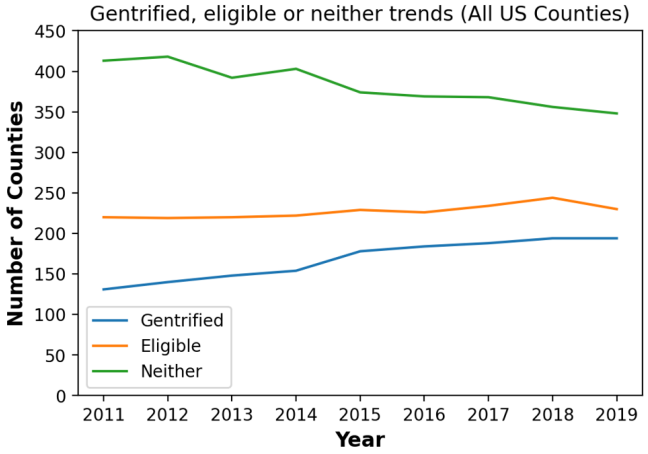

Fig. 3. Shows the number of counties eligible, gentrified or neither from 2011-2019.

counties that were neither are becoming eligible or gentrified over time. Additionally, it appears that amount of counties in each category is converging towards the same point. It does then seem possible the US will eventually reach a time where there will be more gentrified counties than counties which are neither gentrified or eligible to be gentrified.

The data visualizations are indeed useful, but none indicate a correlation between the racial demographics and gentrification. Applying Principal Component Analysis (PCA) can allow further insight towards a discovery in line with our goal. PCA reduces the dimensionality of a data and presents how strongly the features of our characterize a county. In other words, what are the most distinct aspects of any given county. Values can be generated from PCA based on a principal feature that can find a soft-correlation between that feature and other features. These values are called the loading scores, and they describe how much each variable contributes to a particular principal component. In this case, the principal component is the whether the county is considered gentrified of not.

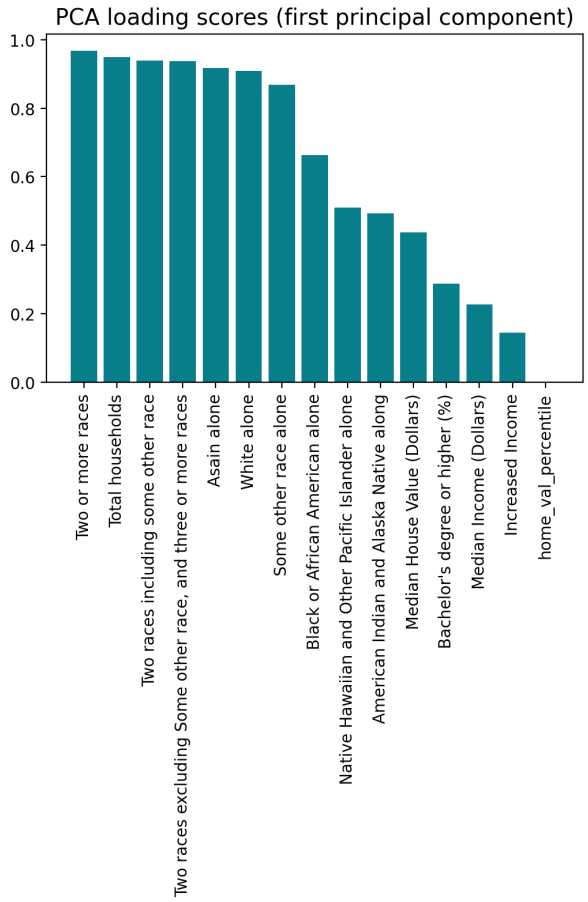

Fig. 4. Loading Scores from Principal Component Analysis using the first principal component.

In Fig. 4. the top three features that correlate with gentrified areas are the number individuals who identify as two or more races, the number of total households in the county, and individuals who identify as two races including some other race. Many other racial demographic features have a high loading scores such as individuals who identify as Asian alone or White alone. Principal Component Analysis gives an initial hint as to how racial demographics play a role in gentrification, but is not solidified proof to demonstrate the correctness of our hypothesis. With this soft-correlation in mind, models can now be implemented to help provide further evidence that racial demographics of an area could correlate to gentrification.

IV. Models

Two models were formed to test the hypothesis, model 1 and model 2. Each approach takes a different perspective of the data and utilizes different components in its architecture. Model 1 considers each county and year pairing as a separate data point, model 2 considers the entire timeline of a county as a singular data point.

A. Data Setup

Before the implementation of the either Model, the data must be setup for training and testing. First the merged data set was split using an 80/20 split, 80% of the data would be set as the training data set and 20% was used as the testing data set. The training data set was partitioned using Stratified Sampling based off of the median home value of a county. The data points were binned into four categories based on 25%, 50%, 75%, and 100% percentiles. Model 1 took this stratification as-is. For model 2, a slight adaptation had to be made, the data points of model 2 were binned based on the *average median home value over each year from 2011-2019* and were binned based on that value. Next, the removal of every feature that defined our criteria for gentrification was removed, which left only racial demographic data and the counties gentrification status. Next, the features of the data were scaled. In this model, normalization is necessary as it prevents the skewing of the data from different features. As an example, it'd be unfair to compare housing price and cost of living without scaling them as the values would vary so drastically. The data was transformed using standard scaling to have a mean of 0 and a standard deviation of 1.

B. Model 1

Model 1 uses a Single Perception as its core component. This model aims to simply determine whether gentrification can be identified solely on racial demographics.

A perceptron Model is a method for the supervised learning of binary classifiers. This algorithm makes predictions based on a linear predictor function to determine a set of weights for each feature of a data point, as shown in Fig. 5. 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.

Prior to training, model 1 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

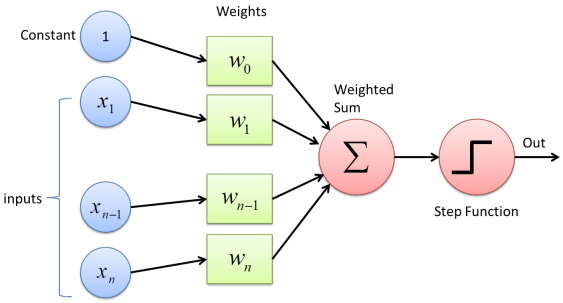

Fig. 5. Diagram of the a Single Perceptron Model.

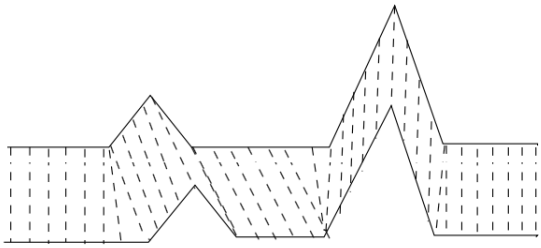

Fig. 6. Diagram demonstrating how Dynamic Time Warping operates.

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.

For results, see the Results section.

C. Model 2

Model 2 utilizes two core components: Multidimensional Dynamic Time Warping and DBSCAN clustering. This model attempts to find an underlying structure within the data. First by comparing similarities of time-series, and then attempting to cluster the time-series based on that similarity.

Dynamic Time Warping (DTW) is used to determine the distance for a time-series pair for various values of epsilon (EPS), which is the primary parameter of Dynamic Time Warping. It allows similarity between any two given time-series to be measured. The implementation of Multidimensional DTW involves finding the difference between each feature in every potential county pairing. These differences are then summed. Additionally, DTW attempts to find consistent patterns in a time-series by utilizing a 'one-to-many' matching of features e.g., a 2012 value of feature 1 of county A could be matched to the 2012, 2013, and 2014 value of feature 1 of county B . An example of this is shown in in Fig. 6. A smaller distance calculated between two counties time-series then implies that the counties have behaved similarly overtime and vice-versa. The model then uses the calculated distances in DBSCAN clustering.

Clustering is a very essential technique to machine learning that allows the partitioning of large volumes of structured and unstructured data/observations into logical groupings. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is an algorithm states that a point that belongs to a cluster if it is close to many points from that cluster, as shown in Fig. 7. The main idea behind DBSCAN is that it is used to find arbitrary shaped clusters and clusters with noise, which is a main reason for implementing this model. With a data set of many different counties with wildly differing feature values, it's almost certain that outliers will be present. Furthermore, DBSCAN was used as it allows to clusters to be generated based on *pre-calculated* distances, which we have here, as opposed to a method like k-means clustering. DBSCAN maintains an epsilon parameter, which is the distance that specifies the boundaries a neighborhood/cluster. Two points are considered to be neighbors if the distance between them is less than or equal to the epsilon value.

The training data was passed through DTW then DBSCAN to cluster. These outputs would provide insight on which counties are mostly similar by racial demographics. Once clustered, we can identify if there is a consistency among the features of a gentrified county.

For results, see the Results section.

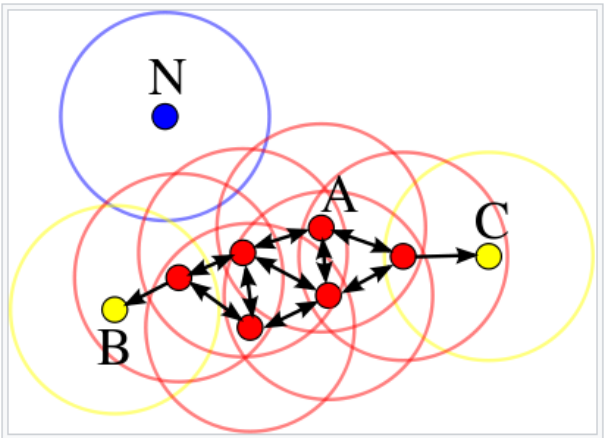

Fig. 7. Diagram demonstrating how DBSCAN clusters data.

V. Results

A. Model 1

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 1 would be a perfect

TABLE II
MODEL 1 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

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 II, you can see each output of model 1 for various values of d. These outputs are the average testing score over 50 iterations. The testing scores 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.

B. Model 2

Model 2 attempted to find meaningful clusters within our data.

In Fig. 8, a heat-map of the distances calculated by Dynamic Time Warping between every county is represented by a single point/square. A square's color represents the size of the distance. A yellow color represents a large distance and dissimilarity between two counties time-series and a purple color represents a very short distance between two counties time-series which indicates a very strong relationship between those counties.

There is clear banding visible in the heatmap, which suggests that some counties are more average or more of an outlier than others. For example, a stripe of mostly yellow suggests that this county is an outlier in comparison to all others. This matches the data analysis finds earlier that there are some clear and consistent counties that exhibit anomalies in their feature values.

An unconventional approach was used to test with DBSCAN. Based off the training set distances, clusters were generated over different epsilon values. Then the model attempted to match each county from the test set to one these clusters. The core idea is that if the clusters generated are meaningful, we should be able to successfully match a county to an existing cluster. Counties will be successfully matched with a cluster

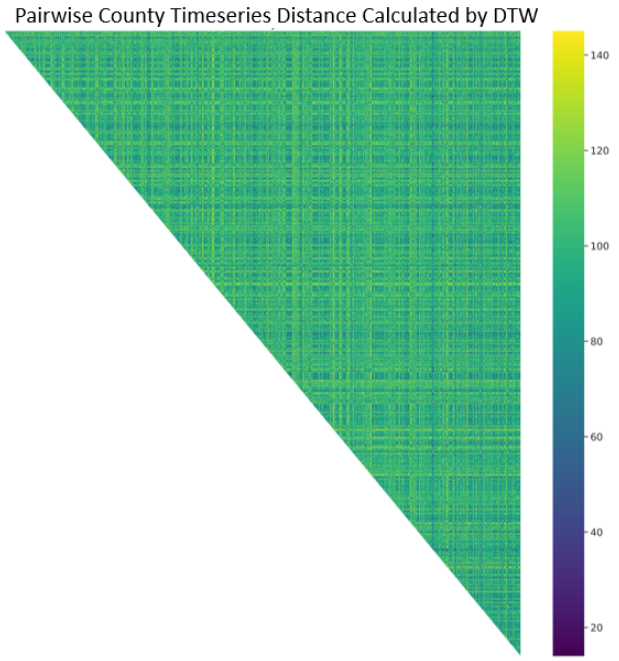

Fig. 8. Heat-map of the results of the Dynamic Time Warping Model.

TABLE III
MODEL 2 RESULTS

EPS	Counties Labeled	Unique Labels
2.0	1.0	1.0
13.0	3.0	2.0
15.0	3.0	1.0
28.0	10.0	2.0
33.0	14.0	1.0
38.0	14.0	1.0
39.0	15.0	1.0
58.0	98.0	1.0
62.0	142.0	1.0

if they match the characteristics of that cluster and are not just noise. Additionally, we should expect multiple clusters to exist, as there is such a wide variety of data points.

Table III shows the results of the DBSCAN clustering model. *Counties Labeled* represents the number of counties that were successfully matched to an existing cluster from the test set. There were about 200 counties used in the test set so a perfect score would be near 200. *Unique Labels* represents the number of clusters that were generated. As epsilon increases, the number of counties recognized and labeled should increase as well. Unique labels represents the number of clusters that was generated from the labeled counties.

As the value for epsilon increased, the number of counties that were recognized also increased, which is what our team predicted. What we did not predict was the number of counties

still not being labeled and considered noise was very high. For a value of 39.0 for epsilon input, the number of counties labeled was only 15, which is very low. With a high enough epsilon, the model recognized and labeled most counties, but the number of clusters that were generated is significantly lower than expected. It implies that the model was essentially adding every county to the same group. The maximum number that the model was able to find was only 2, which is too low to provide any evidence for our hypothesis.

VI. Conclusion

In this paper, we attempted to determine whether there exists a correlation between racial demographics and gentrification. We developed two different machine learning models and trained them on US Census data of every US county from 2010-2019. Although not perfect, model 1 certainly hints that race and gentrification are related to some degree. This model could likely be improved by increasing the complexity to utilize multiple perceptions and additional data transformations. Our approach in Model 1 could also potentially be used to identify other unknown correlated features in a different context, whether that be social, biological, or anything beyond. Model 2 proved to be unsuccessful in providing support to hypothesis. It is hypothesized that Model 2 failed because the time-series data was short and varied. It considered only 10 data points and considered a significant number of data values. Additionally, it's likely the parameters of this model were not tuned well enough. In the future, improvements on the parameters of this model could be made, as well as further processing of the data beforehand to generate better results, such as another clustering method. Beyond the models, different data sources could be utilized like real estate data or using zip-codes instead of counties. Nonetheless, It's clear that there are a lot of different factors which impact gentrification, and predicting it well is an incredibly difficult task.

Acknowledgment

References

[1] Bernstein, Jared, et al. "Housing Prices and Inflation." The White House, The United States Government, 9 Sept. 2021, https://www.whitehouse.gov/cea/blog/2021/09/09/housing-prices-and-inflation/

[2] Choi, Jung Hyun, et al. "The State of Millennial Homeownership." Urban Institute, 18 July 2018, https://www.urban.org/urban-wire/state-millennial-homeownership.

[3] Guerrieri, Veronica, et al. "Endogenous Gentrification and Housing Price Dynamics." NBER, National Bureau of Economic Research, 27 July 2010, https://www.nber.org/papers/w16237.

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

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

[6] 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

[7] 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

[8] 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