Capstone Project 1: Super Zips Final Report

Problem Statement

This project intends to model which areas of the U.S. have improving (or worsening) overall standards of living over time.

My intended client would be government policymakers at all three levels (local, state, and federal) that would use my analysis and results in developing public policy. My project would serve as a reference and basis for implementing income/tax credits, subsidies, financial aid, public education programs, and other incentives supporting increases in education levels and eliminating poverty.

The source data I plan to clean and analyze are income statistics publicly provided by the Internal Revenue Service and educational attainment data publicly provided by the U.S. Census Bureau. Calculating these two together as a ratio would measure the overall standard of living.

Then, I would predict areas of growing/contracting economic inequality by plotting the education and income based descriptive statistics over time for each zip code. Additionally, using K-Means Clustering I would attempt to segment the data and draw insights from trends over a six year period.

Data Wrangling

The following steps were taken to obtain, wrangle, then clean data into a form that is suited for analysis.

1) Obtain main source data files needed from irs.gov and data.census.gov
2) Select relevant columns using documentation files provided with main source data files
3) Create subset data files from main source date files due to big source file sizes
4) Determine how to address missing data:
    a) For the IRS AGI (Adjusted Gross Income) data, the missing data was filled in thanks to cross-referencing smaller data subset files divided from the main aggregate source file and grouped by state.
    b) For the Census data, all rows containing missing data were removed from the dataset since certain estimates could not be calculated due to either no or too few observations as explained in the Symbols Dictionary
5) Re-define appropriate data types for all columns in data
6) Regroup/Tidy Data into a format appropriate for analysis using the .melt() and .pivot() dataframe methods
7) Calculate scores from raw data as described below. **EDA was not performed on the raw data due to file size and computational cost**

8) Perform EDA by the following steps and repeat for every year of data obtained (i.e. 2011-2018 for Census data and 2006-2017 for IRS AGI data)
    a) Using .info() dataframe method to confirm numeric data uniformity and no missing data as well as data types
    b) Use .describe() dataframe method to generate descriptive statistics and confirm data integrity
    c) Plot a histogram to visualize distribution of all zip code scores from all 51 states
    d) Use FacetGrid function from Seaborn to plot multiple swarmplots visualizing the score distributions by state
9) Create data analysis files suited for visualizations

<u>Calculating Education and Income Scores per Zip Code:</u>

This score is calculated using the weighted average formula below

$$Score = \sum_{n=1}^{i} x_n \frac{y_n}{p}$$

Where

- x_n = Education Attainment or IRS AGI score using the scale below (unitless)
- y_n = population satisfying the nth category (# of people)
- i = last category in either scale (i.e. 6 or 7)
- p = population for the zip code specified (# of people)

Education Scoring Scale	**IRS AGI Scoring Scale**
Categories obtained from column names in source data. ● 1 = Less than 9th grade ● 2 = 9th to 12th grade, no diploma ● 3 = High school graduate (includes equivalency) ● 4 = Some college, no degree ● 5 = Associate's degree ● 6 = Bachelor's degree ● 7 = Graduate or professional degree	Determined by Adjusted Gross Income submitted by IRS Forms 1040, 1040A, 1040EZ (column named as “AGI_STUB” or “AGI_CLASS” depending on which annual dataset) ● 1 = $1 under $25,000 ● 2 = $25,000 under $50,000 ● 3 = $50,000 under $75,000 ● 4 = $75,000 under $100,000 ● 5 = $100,000 under $200,000 ● 6 = $200,000 or more

Data Storytelling

For this section, IRS AGI and Census Educational Attainment data are analyzed to answer the following questions below. Numerous plots were generated and can be viewed in this GitHub repo link to the Jupyter Notebook containing the analysis.

1) How are states and zip codes performing over time in terms of Educational Attainment and Income?

Analysis and Outcome: After calculating the Education and Income scores for each zip code (column name is "ZipScore" in both IRS AGI and Census datasets), the scores were aggregated by arithmetic averaging on each State. Examining high-level trends of 51 states is much easier than examining thousands of zip codes at once. If upon visual inspection, I examine a state trending a certain way, I can then narrow in and look at the zip code trends within that state. As a result, this is done to answer the next question.

2) Which states and zip codes perform the best/worst over time in terms of Educational Attainment and Income?

Analysis and Outcome: Upon visual inspection, I noticed the following:

- Education:
    - District of Columbia (DC) has the highest Education scores and the steepest positive trend, thus performing the best. Narrowing in, we see that most zip codes are generally trending above a Education Score of 5 which is outside the range of the states (around 4-5).
    - Louisiana (LA) has the lowest Education scores and a gradual positive trend, but overall perform the worst. Narrowing in, we see that most zip codes are generally trending below a Education Score of 4 which is outside the range of the states (around 4-5).
- Income:
    - District of Columbia (DC) has the highest Income scores and the steepest positive trend, thus performing the best. Narrowing in, we see that most zip codes are generally trending above a Income Score of 3 which is outside the range of the states (around 2-3).
    - Mississippi (MS) has the lowest Education scores and a gradual positive trend, but overall performs the worst. Narrowing in, we see that most zip codes are generally trending a little over a Income Score of 2 which is in the lower half of the range of the states (around 2-3).

3) Is there a relationship/correlation between Educational Attainment and Income?

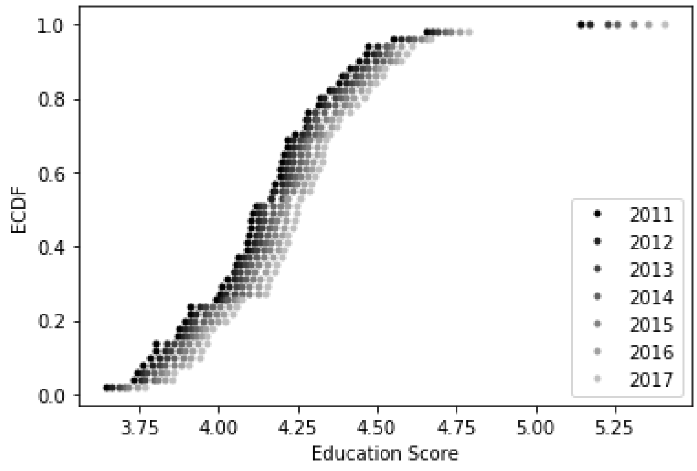
Analysis and Outcome: For this case, again, data was aggregated by State and not by Zip Code due to too many Zip Code data points resulting in a "blobplot" and not a scatter plot. By combining the Census and IRS AGI data frames then using a scatter plot for the Education and Income Scores by State, we can see that there is certainly a positive relationship between the two. In other words, the higher your education, the higher your income.

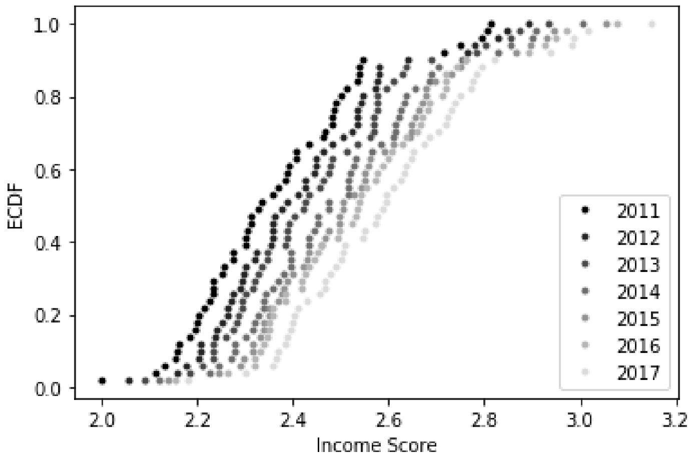
Statistical Data Analysis

This section summarizes the steps and corresponding results of performing Statistical Data Analysis on Education (obtained from the US Census) and Income (obtained from the IRS) data. For this analysis, these two variables are defined to be the primary factors of an area’s overall standard of living. Also, they’re shown to be linearly related and vary together. **Only 2011 and 2017 data are used due to intersection of annual releases of datasets**

Steps and Findings

1. Conduct EDA: Plot ECDFs for all years available for Education and Income Scores with color variation to show change over time.
   Results and Findings: The ECDF plots for Education and Income Scores show gradual shifts to the right over time as indicated by the color variation from black to light gray. This is a good sign that overall people are gradually becoming more educated and earning more money to live better.





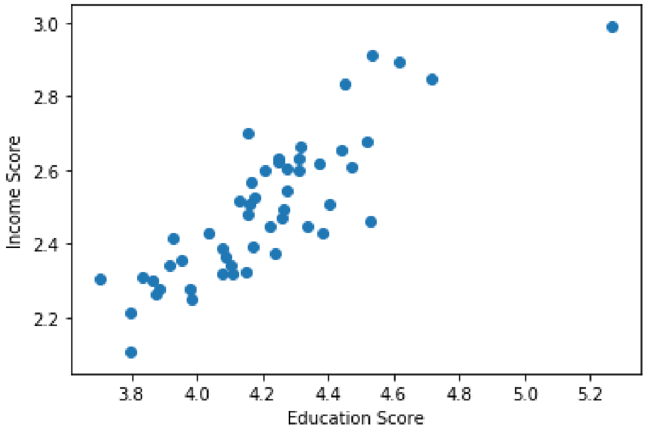
2. Parameter Estimation: Estimate the *difference* of the mean Education and Income Scores of all zip codes from 2011-2017.
   Results and Findings: Difference of Education Score means = 0.126 and Difference of Income Score means = 0.231, which shows, on average, how much the Education and Income scores have changed in 7 years (including the year of 2017).

3. Confidence Interval Calculation: Use Bootstrap replicates method for 2011 and 2017 datasets only for both Education and Income scores to report a 95% bootstrap confidence interval.
   Results and Findings:
   - 95% bootstrap Conf. Int. bounds of Education Scores = [0.112 0.140]
   - 95% bootstrap Conf. Int. bounds of Income Scores = [0.221   0.241]

   These bounds show the Confidence Interval containing the differences of means of Education and Income estimated in Step 2.

4. Hypothesis Testing: Answer the question, "Have Education and Income Scores improved?" (i.e. Null Hypothesis: $\mu_{2017} - \mu_{2011} = 0$, Alternate Hypothesis: $\mu_{2017} - \mu_{2011} > 0$) by performing a bootstrap permutation test by shifting the two data sets so that they have the same mean and then use bootstrap sampling to compute the difference of means.
   Results and Findings: The p-values for both bootstrap hypothesis tests are so small that Python essentially returns zero resulting in statistical signifying that the null hypotheses for both Education and Income can be rejected. Thus, the two means are different which indicate change between 2011 and 2017.

5. Correlation and Covariance Analysis: Calculate Pearson correlation coefficient and covariance between Education Scores and Income Scores from 2011-2017 (confirm relationship between Education and Income).
   Results and Findings: The scatterplot suggests, at first glance, that there is a positive relationship between Education and Income plus the two variables vary together. This is confirmed by calculating the Pearson Correlation Coefficient (r = 0.845, confirm the strong relationship) and Covariance (cov = 0.044) between the two sets of scores.

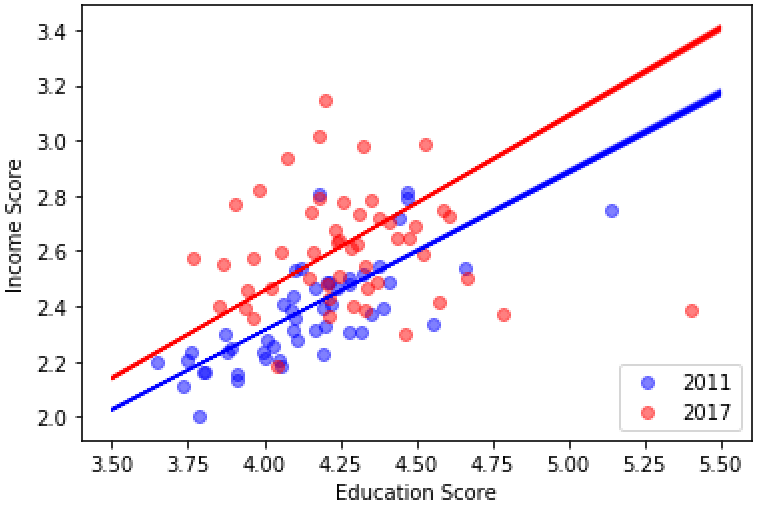


6. Linear Regression: Perform a linear regression for both the 2011 and 2017 data. Then, perform pairs bootstrap estimates for the regression parameters. Report 95% confidence intervals on the slope and intercept of the regression line (show how Education and Income change together over time).

   Results and Findings:

   - 2011: slope = 0.573, conf int = [0.565 0.580]
   - 2011: intercept = 0.020, conf int = [-0.007 0.052]
   - 2017: slope = 0.634, conf int = [0.626 0.642]
   - 2017: intercept = -0.079, conf int = [-0.112 -0.047]

   Calculating these bootstrap results quantify how Education and Income change together over time. These are shown visually from the combined line and scatter plot below as the regressing line changes from 2011 to 2017.



7. Calculate 95% bootstrap confidence interval for overall standard of living: Compare the *mean ratio* of Education Scores to Income Scores.

   Results and Findings:

   - 2011: mean ratio = 1.750, conf int = [1.747 1.753]
   - 2017: mean ratio = 1.644, conf int = [1.641 1.647]

   These results suggest changes in overall standard of living over 7 years (including the year of 2017) and even the confidence intervals overlap. To confirm, bootstrap hypothesis testing is done in the next and final step.

8. Bootstrap hypothesis testing on ratios: Perform a bootstrap permutation test (like in Step 4) by shifting the two data sets so that they have the same mean. Then use bootstrap sampling to compute the difference of means.

   Results and Findings: The p-value for the bootstrap hypothesis tests are so small that Python essentially returns zero resulting in statistical signifying that the two means are different. This indicates a change in overall standard of living between 2011 and 2017.

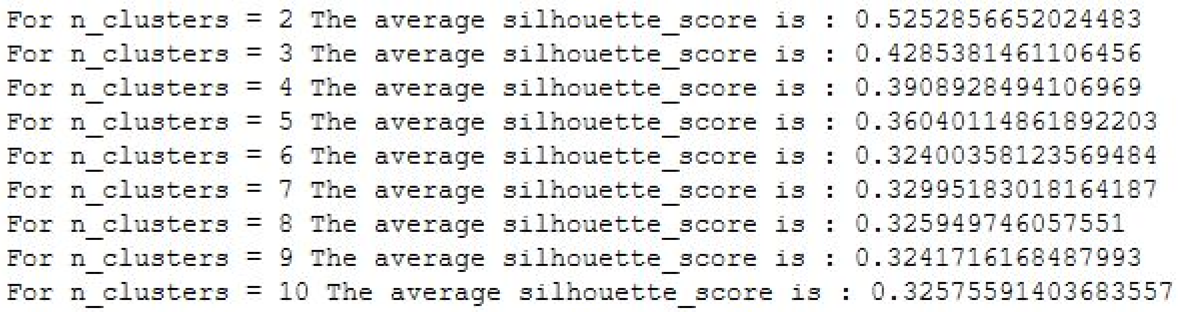
In-Depth Analysis

This section summarizes the steps and corresponding results of determining the right value of K using Silhouette Score analysis then performing K-Means Clustering that will best segment the data and draw insights. My hypothesis was that the data can be clearly categorized into clusters similar to the economic classes (i.e. lower class, middle class, upper class, etc.). However, after going through the analysis shown below, I found out instead that all the data forms a single cluster (this is clearly shown visually below), but I also propose an opinion on how clustering may be used and interpreted.

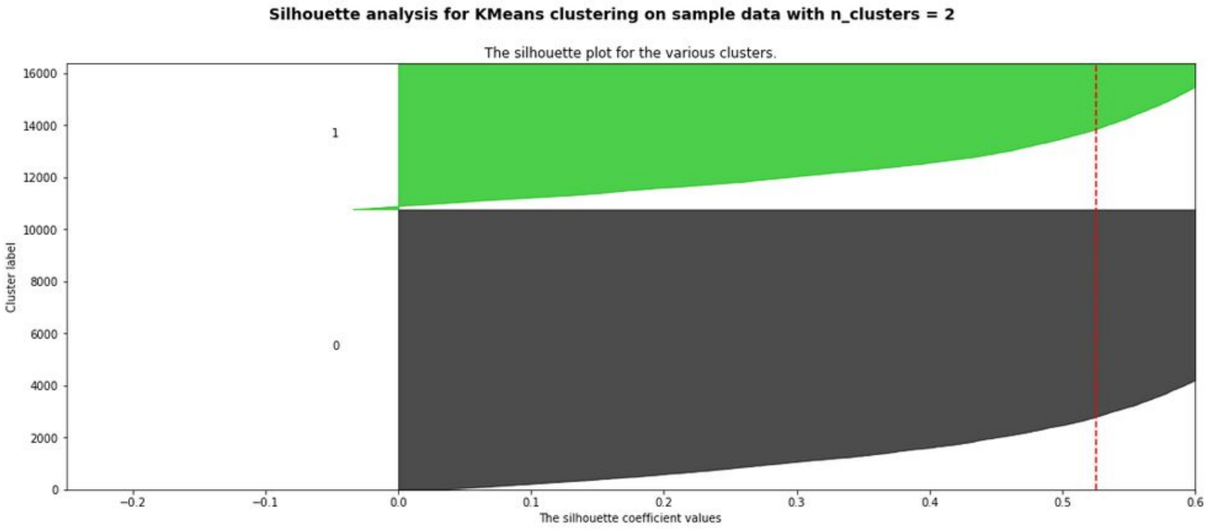
Since Education and Income Scores are the two primary variables of interest, these will be the two main features and no PCA/Dimensionality Reduction is required. To reiterate, only the 2011 and 2017 annual datasets are used due to intersection of annual releases of both IRS AGI and Census data. Additionally, solely analyzing these two datasets will show the changes/trends over a period of six years.

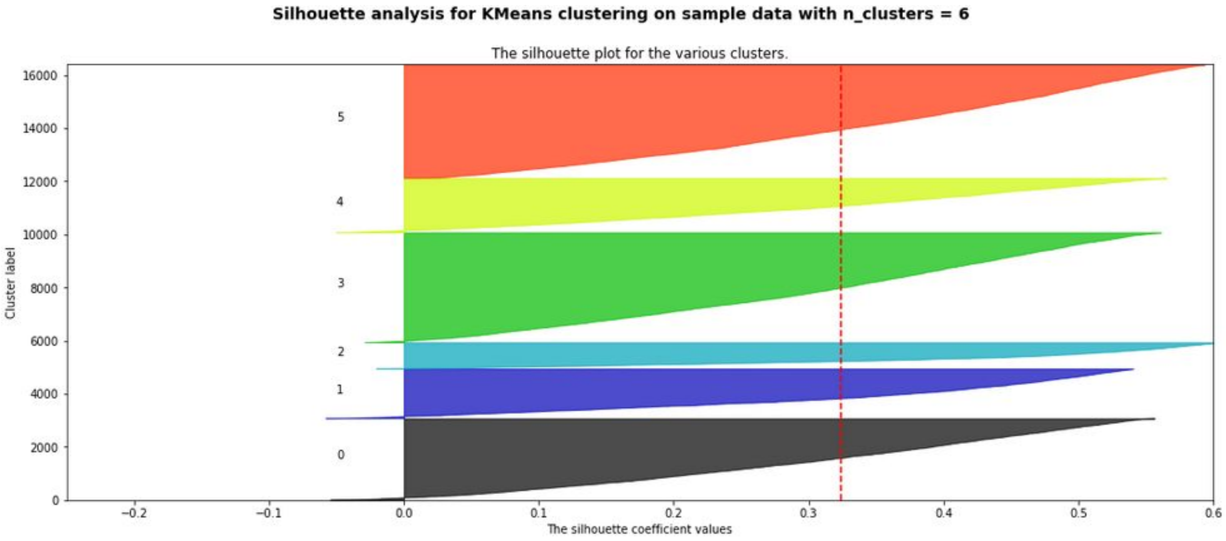
Steps and Findings

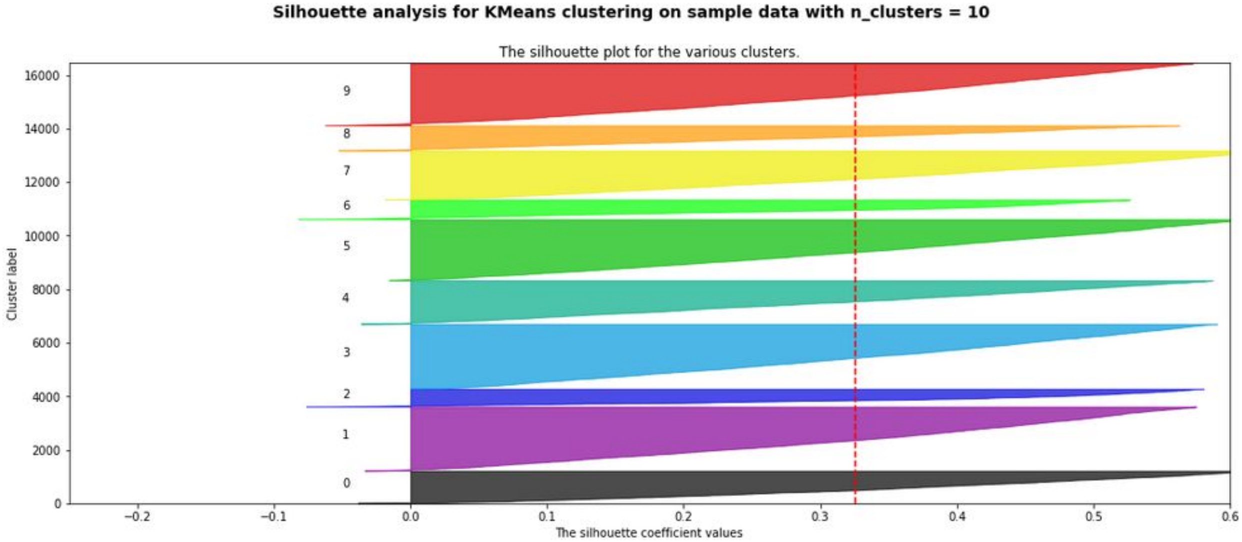
1. Conduct Silhouette Score Analysis: Calculate Silhouette Scores and generate Silhouette Samples to determine best value of K for K-Means clustering.
   Results and Findings: As shown below, the silhouette score decreases then settles around 0.325 with increasing K which is undesirable. However, I personally chose K = 5 as explained in the next section. The silhouette plots for 2, 6, and 10 clusters are shown below while the remaining plots are displayed in the accompanying presentation and Jupyter notebook.



```
For n_clusters = 2 The average silhouette_score is : 0.5252856652024483
For n_clusters = 3 The average silhouette_score is : 0.4285381461106456
For n_clusters = 4 The average silhouette_score is : 0.3908928494106969
For n_clusters = 5 The average silhouette_score is : 0.36040114861892203
For n_clusters = 6 The average silhouette_score is : 0.32400358123569484
For n_clusters = 7 The average silhouette_score is : 0.32995183018164187
For n_clusters = 8 The average silhouette_score is : 0.325949746057551
For n_clusters = 9 The average silhouette_score is : 0.3241716168487993
For n_clusters = 10 The average silhouette_score is : 0.32575591403683557
```


Silhouette analysis for KMeans clustering on sample data with n_clusters = 2
The silhouette plot for the various clusters.
Cluster label
16000
14000
12000
10000
8000
6000
4000
2000
0
1
0
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
The silhouette coefficient values


Silhouette analysis for KMeans clustering on sample data with n_clusters = 6
The silhouette plot for the various clusters.
Cluster label
16000
14000
12000
10000
8000
6000
4000
2000
0
5
4
3
2
1
0
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
The silhouette coefficient values



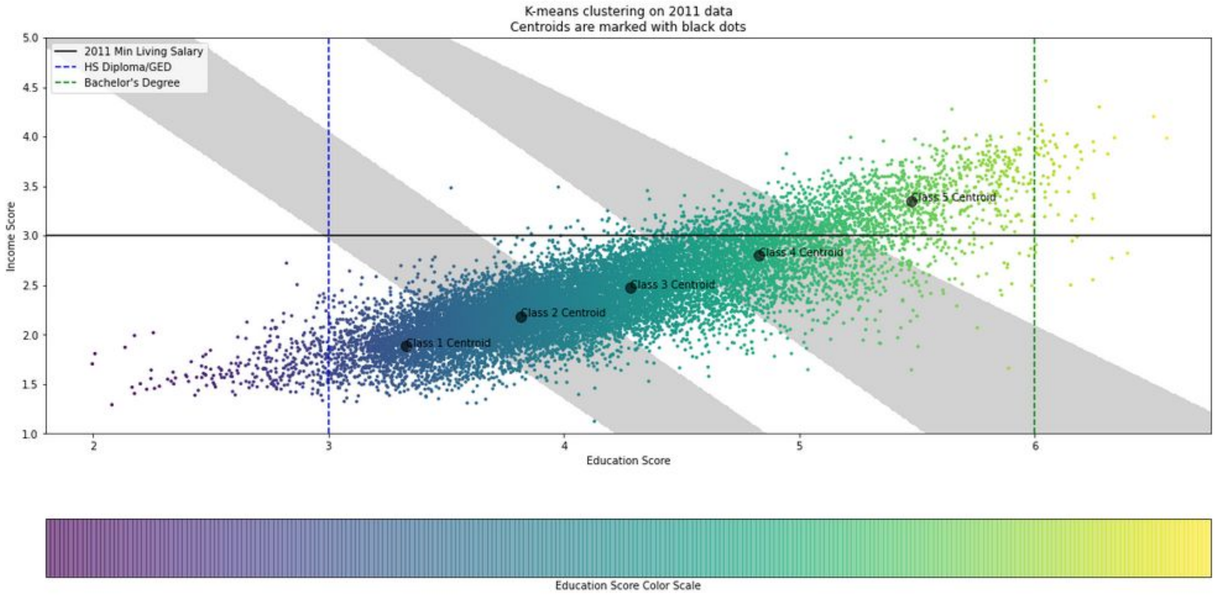
2. <u>Conduct K-Means Clustering with Five Clusters for 2011 and 2017 data:</u> K = 5 was chosen as analogous labeling categories to the five economic classes: Lower, Lower-Middle, Middle, Upper-Middle, and Upper. Similarly, these clusters are labelled Classes 1-5 with Class 1 having the lowest Standard of Living (SoL) and Class 5 having the highest SoL. These clusters are displayed in the scatter plots below where a Voronoi diagram (sectioned by white and gray zones) displays the cluster boundaries and the clusters themselves can "roughly" be distinguished by the following colors: Class 1 - Blue/Purple, Class 2 - Blue, Class 3 - Turquoise, Class 4 - Green, Class 5 - Yellow/Green. The color scale is based on Education Score since, in my opinion, peoples' general perception of success is based on education level.

   I considered this labelling scheme not from an objective, logical perspective where there are concrete, easily definable boundaries but from a subjective, emotional perspective based on peoples' broad, spectrum-like perception of how much a person should be paid based on the same person's level of education. This accounts for the "gray" areas of understanding based on the common notions of "getting a degree as the golden ticket to success" and "obtaining a higher degree along with years of experience results in earning a six-figure salary" as the popular standards to achieve the American Dream. In today's economy, there are people that graduated with Master's and PhD degrees, but are paid the minimum (if not, lower) livable salary. Additionally, a relatively small portion of the population don't possess Bachelor's degrees but earn higher than average salaries.

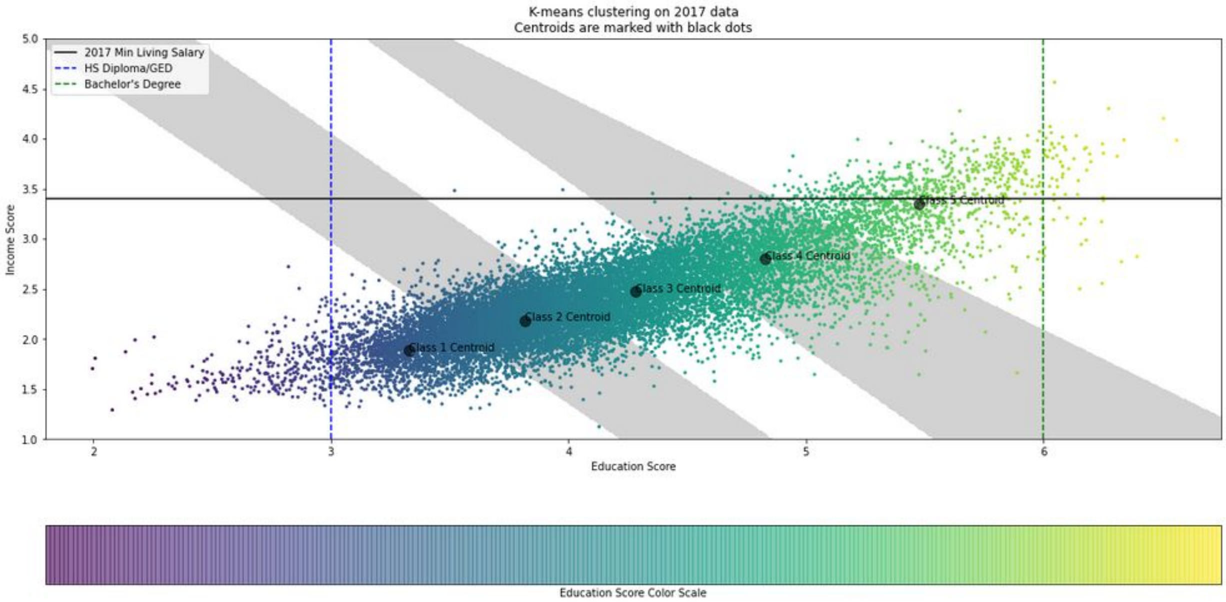
   Another way of thinking about this is what is the expected sentimental value of a zip code (Education Score) vs. what is the actual sentimental value of a zip code (Income Score).

Graphically over time, a key goal of public policymakers in all levels of government is for all zip codes to move up and to the right as shown in the scatter plots below.

Results and Findings for Clustering 2011 Data: In 2011, we see that the majority of zip codes are earning less than the 2011 Minimum Living Salary (approx. $50,000 which is equivalent to a Income Score of 3, Source: Statista - US Annual Consumer Expenditures). What is also troubling to see is a small portion of zip codes in the upper right sextant of the scatter plot. Personally, I never realized growing up that the majority of Americans are being paid and educated lower than the common standards stated earlier. I just presumed that everyone was more/less achieving the common standards since the U.S. is a developed country and considered a superpower.



Results and Findings for Clustering 2017 Data: In 2017, we see that even more zip codes are earning less than the 2017 Minimum Living Salary which is $10,000 higher (approx. $60,000 which is equivalent to a Income Score of 3.4, Source: Statista - US Annual Consumer Expenditures). We also see that the clusters and centroids themselves changed only a little but although people are earning more and becoming more educated, this rate of increasing income is slower than the rate of increasing minimum livable salary. These small changes in clusters will unfortunately reinforce the common notions even though drastic change is needed.



3. Draw conclusions: Even though the previous notebooks of this project show optimism, everyone must understand the reality of the situation the everyday American is facing now and not turn a blind eye. These findings show evidence of growing economic inequality in this country and should inform everyone how our general view of success and the American Dream will change.