IST 707: ANALYZING ACS TO IDENTIFY ZIP CODES PRONE TO POVERTY

Aruneema
Harshita Tanksali
Jaya Varshini Prabakar
Mohit Vellanki

Contents

I. PROBLEM STATEMENT

In this project, we intend to analyze the American Community Survey (ACS) data to identify the zip codes that might be prone to poverty.

The objective is to find the trend of poverty in the zip codes by analyzing various demographic variables through machine learning techniques and identify which zip codes fall 'under poverty' criteria.

RESEARCH QUESTION

To help us with the objective we framed the following research question.

- **What are the demographic characteristics of zip codes that match the criteria of 'under poverty' and what affects the poverty level?**

This will hopefully allow us to provide insights and recommendations to government organizations on the current and future poverty trends so that they can intervene and take steps to address the root cause of poverty and improve the well-being of the residents. Through our project, we are also trying to address the following sustainable development goals:

- SDG 1: "End poverty in all its forms everywhere."
- SDG 8: "Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all."
- SDG 10: "Reduce inequality within and among countries."

II. THE DATA

The data has been collected from American Community Survey (ACS) which is a demographical survey program conducted by the US Census Bureau. We obtained the data from the Social Explorer website [1] as it was easier to navigate through the website interface. This dataset is also a primary source for U.S. poverty statistics [2]. We collected data from surveys of 2017-2021, covering the following six broad categories that can help in determining poverty levels:

- Education
- Housing
- Race and Ethnicity
- Employment
- Health
- Income and Poverty

Combination of variables from these demographical categories would help in exploring how poverty status is being affected in different zip code areas.

DATA DICTIONARY

The table 1 (see Appendix) gives a snapshot of all the variables used in the analysis across 5 years. There are twenty-two variables across the five years with around 2700-2800 rows. [3]

DATA PRE-PROCESSING

Our first step in data pre-processing was filtering for zip codes since the ACS dataset had over 30,000 ZCTA zip codes. We filtered the data for the zip codes starting with 10000 to

20000 covering states of Delaware, New York, and Pennsylvania to focus on a subset of the data.

Next, we filtered the ACS data for relevant variables to be analysed since ACS dataset had more than two thousand variables from different categories. There were many sub-levels present for each variable from which we chose the relevant ones as shown in Table 1. These variables were generated by taking proportion of the sub-group to overall population. For instance, “whiteMajority” was calculated by taking proportion of “white alone” population to total population. Similar approach was taken for most of the variables.

After creating new columns, they were checked for null values, which were later dropped from the dataset, if any exi sted.

```
<class 'pandas.core.frame.DataFrame'>
Index: 2887 entries, 10001 to 19979
Data columns (total 21 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   incomeBelowPovertyLevelProportion   2887 non-null   float64
 1   incomePerCapita                     2887 non-null   float64
 2   incomeGiniIndex                     2887 non-null   float64
 3   underPoverty                        2887 non-null   int32
 4   popDensity                          2887 non-null   float64
 5   whiteMajority                       2887 non-null   float64
 6   eduHsOrLessProportion               2887 non-null   float64
 7   eduHsProportion                     2887 non-null   float64
 8   eduBachOrBetterProportion           2887 non-null   float64
 9   eduEnrollmentProportion             2887 non-null   float64
 10  eduDropoutProportion                2887 non-null   float64
 11  employedProportion                  2887 non-null   float64
 12  employedPvtSecProportion            2887 non-null   float64
 13  employedPubSecProportion            2887 non-null   float64
 14  employedSelfEmpProportion           2887 non-null   float64
 15  employedPvtNonProProportion         2887 non-null   float64
 16  employedUnpaidFamProportion         2887 non-null   float64
 17  insuredProportion                   2887 non-null   float64
 18  housingRentMoreThan30pcProportion   2887 non-null   float64
 19  housingMedianValue                  2887 non-null   float64
 20  housingMedianGrossRent              2887 non-null   float64
dtypes: float64(20), int32(1)
memory usage: 484.9+ KB
```

Table 2: Data information for 2017 dataset

After all the cleaning, we ended up with 20 variables, most of them being float type, except for “underPoverty” which is an integer type variable.

III. EXPLORATORY DATA ANALYSIS

On performing an exploratory data analysis across the years for each variable, following trends were observed -

VARIABLES WITH INCREASING TREND	VARIABLES WITH A DOWNWARD TREND
• eduBachOrBetterProportion	• eduDropoutProportion
• employedPubSecProportion	• eduEnrollmentProportion
• employedPvtNonProProportion	• eduHsOrLessProportion
• housingMedianGrossRent	• eduHsProportion
• housingMedianValue	• employedPvtSecProportion
• incomePerCapita	• housingRentMoreThan30pcProportion
• insuredProportion	• incomeBelowPovertyLevelProportion
• popDensity	• incomeGiniIndex
• employedProportion	• whiteMajority

Insights:

Education: While the proportion of people having bachelors or better education increased marginally across the years, the dropout, enrollment, and proportion of people having education less than high school decreased. The dropout rate increased slightly between 2019 and 2020 but it decreased drastically from 2020 to 2021.

Employment: The overall employment trend was increasing with more employment in the Public and Private Non-Profit sector. Employment decreased marginally in the private sector. Self-employment decreased from 2017-2018 but increased drastically from 2019-2020 followed by a decreasing trend in 2021.

Housing: The median rent and median value of houses increased across the years, however the trend of people spending more than 30% of their income on house rent reduced marginally.

Income and Poverty: The income per capita and Gini index have an increasing trend across the years and the income level below poverty level decreased across the year. It's interesting to notice that while income levels were increasing, the inequality of income also increased across the year (Gini index).

Health: The proportion of population having health insurance showed an increasing trend across the years

Population and Race: The population density was quite stagnant from 2017 to 2018 but was gradually increasing from 2018 to 2021. The proportion of white majority shows a decreasing trend across the years.

UNIVARIATE ANALYSIS

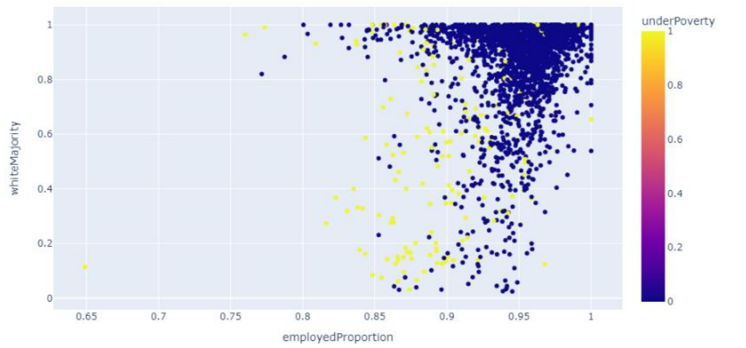
On doing univariate analysis of different variables across the years, it was found that most of them are moderately to highly skewed (both positive and negatively skewed), except for "incomeGiniIndex" and "housingRentMoreThan30pcProportion".

Additionally, most of the variables have many outliers. However, these were not removed since they were forming a significant portion of the dataset and removing them would lead to large amount of data loss. We also did not use any techniques such as Winsorization to handle the outliers, because we wanted to maintain the originality of the data and see the impact of outliers in our analysis.

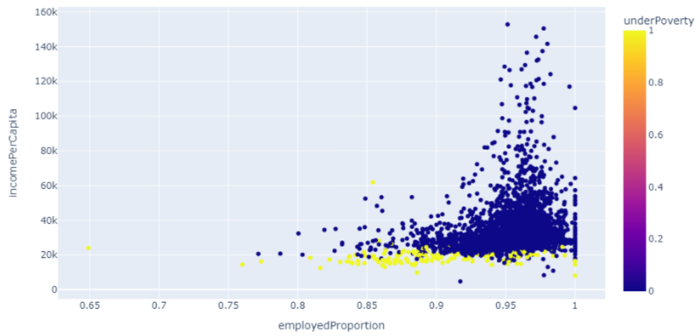
BIVARIATE ANALYSIS

Across the year bivariate analysis of different variables does not show any significant change in pattern. Some of the observations are discussed as follows:

(a)



(b)



(c)

(d)

Fig 1: Scatter plot between different variables for bivariate analysis. Blue dots are for zip codes not "underPoverty" while yellow dots are zip codes "underPoverty"

a) Zip codes "underPoverty" have per capita income in the range of 15-30k regardless of whether the race is white majority or not. However, zip codes with whiter majority race have higher income per capita range, maxing at 150k.

b) The employed proportion in mostly above 80% for all the races, regardless of if they are "underPoverty" or not. However, the white majority zip codes who are not "underPoverty" have higher employment proportion.

c) The income per capita of zip codes not "underPoverty" is higher and their employed proportion is also high as compared to zip codes not "underPoverty"

d) The range of zip codes that spend more than 30% of their income on rent is similar regardless of the fact they are "underPoverty" or not. The only difference is that per capita income of zip codes not "underPoverty" is high and some of these zip codes do not spend more than 30% of their income on rent.

IV. METHODS

To find insights into the poverty status of different zip codes we implemented unsupervised clustering techniques (K-means and Agglomerative clustering) and AR mining to find what combinations of variables influence a zip code to be "under poverty".

SCALING

Before clustering, min-max scaling was implemented on 'popDensity', 'incomePerCapita', 'housingMedianValue', 'housingMedianGrossRent'. This is because clustering is a distance-based algorithm. All distance-based algorithms are affected by the scale of the variables. The variables mentioned before are on the scale of millions while other variables are in form of proportion ranging from 0 to 1. Hence, the mentioned four variables were scaled to bring them on the same range as other variables.

K-MEANS CLUSTERING

To apply K-means clustering, the number of clusters must be predetermined to generate clusters. By making a plot of within sum of squares (wss), number of 'k' clusters can be decided. The point where elbow is formed in the plot can be chosen as number of clusters. For all years, the elbow was formed at 3 (fig 2), however, we choose '2' to form our clusters since we wanted to find only two clusters- "under poverty" and not "under poverty".

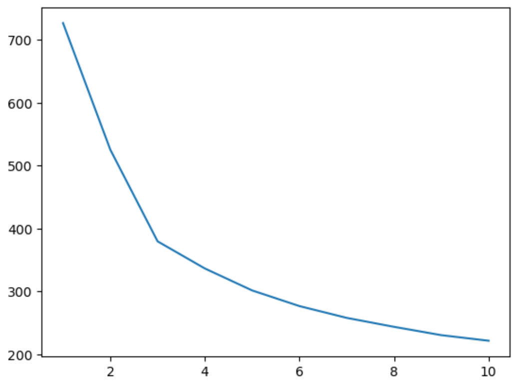


Fig 2: WSS plot for clusters, 2017

To evaluate the clusters, silhouette scored was measured to see how well the clusters were formed.

Year	Silhouette Score
2017	0.717
2018	0.712
2019	0.709
2020	0.7
2021	0.657

Table 3: Silhouette Score of clustering for different years

There were two clusters identified across the five years:

- Cluster 0: The zip codes under this cluster were identified to be not "underPoverty", since the majority of them are not below poverty level, had higher proportion of education level, more employment and higher income per capita, all characterizing being not "underPoverty".

- Cluster 1: The zip codes under this cluster were identified to be “underPoverty”, since they had opposite characteristics as compared to cluster 0.

Some interesting observations across the five years:

- The mean value of Income Gini index (which tells about income inequality) is almost similar for two clusters.
- The high school education proportion is almost similar in both the clusters.
- Employment in the private sector is more for zip codes in cluster '0'.
- The zip codes in cluster '1' spend more proportion of their income on housing rent.
- The proportion of zip codes “underPoverty” in 2021, increased drastically to 28.9%

The cluster proportion across the years are visualized as follows:

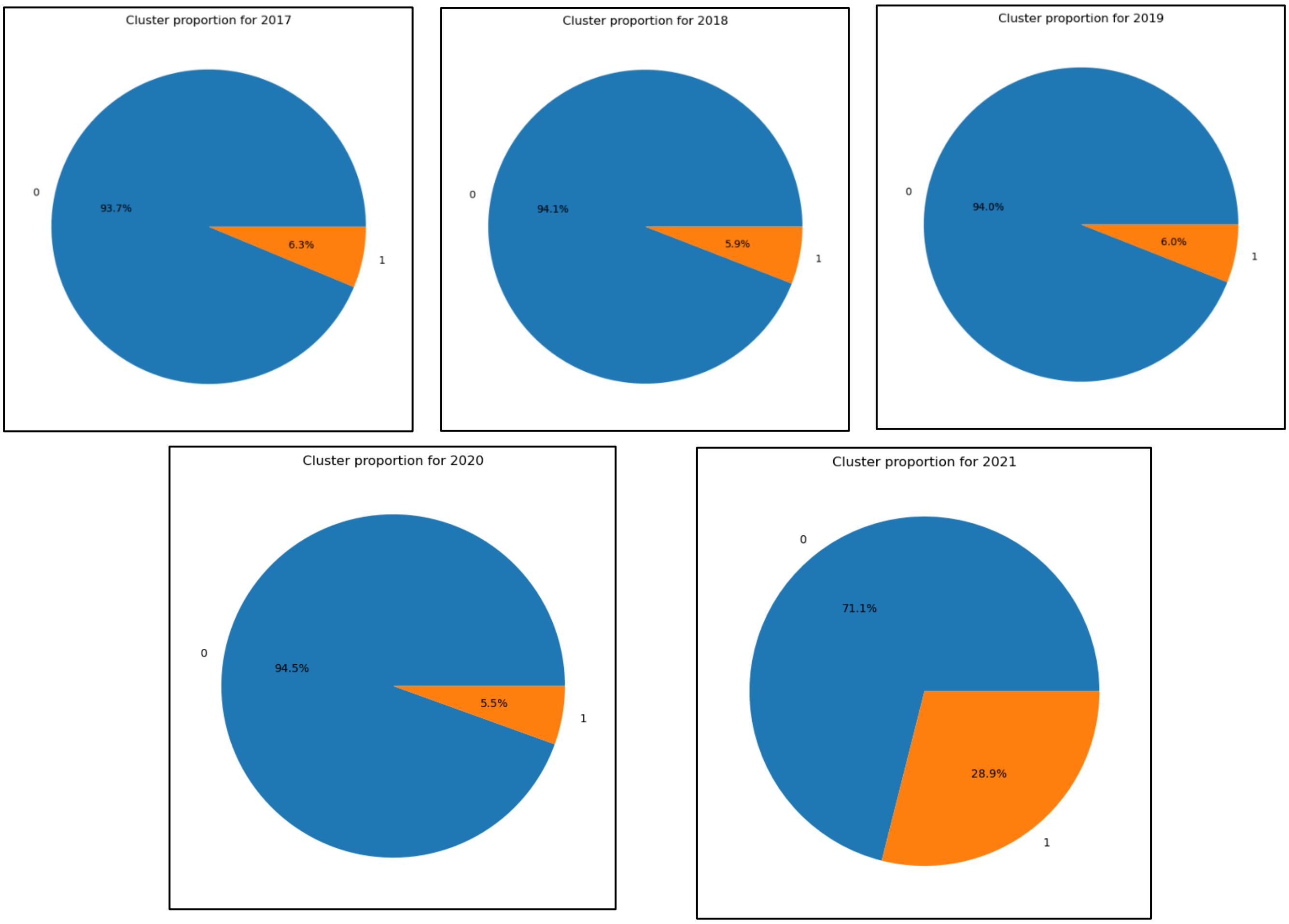


Fig 3: Cluster proportion across the years

AGGLOMERATIVE CLUSTERING

Agglomerative clustering was also implemented on the scale data to find clusters. The parameters chosen were as follows:

n_cluster =2, affinity= Euclidean, linkage= ward

These parameters were giving the most discreet clusters hence, they were taken as parameters. The results of agglomerative clustering were coming same as k-means clustering hence the insights for agglomerative clustering was coming same as the k-means clustering.

AR MINING
After implementing clustering, it was found that there is addition of new set of zip codes that come under the category of poverty every year. To identify the reason driving the transition of these zip codes to poverty status, AR mining rules were applied to these newly identified zip codes found in the years spanning from 2018-2021.

After converting all variables in categorical format through encoding, these variables were converted into factors and transaction tables in R. Then apriori rules were applied with support of 0.90, confidence of 0.80, and rhs= "underPoverty=1". Under these conditions the AR rules were able to generate an optimum number of meaningful results. Note should be taken that rhs= "underPoverty=0" was taken for the year of 2021 because even though these zip codes are labelled not "underPoverty" they were clustered "underPoverty" by k-means clustering. Applying these combinations of parameters fetched a set of:

- 755 rules for 2018
- 2024 rules for 2019
- 2789 rules for 2020
- 1022 rules for 2021

Upon inspecting the rules (see appendix) with highest support and confidence (see appendix), the common theme among the zip codes transitioning to poverty each year was that the majority of the houses have their rent and value more than the median.

The zip codes with less white majority have low education enrollment and low public sector employment.

ABOUT THE MODELS
For the unsupervised learning, we have decided to use Kmeans and agglomerative clustering. We decided to use Kmeans because it was an easy to interpret and a simple yet powerful algorithm. From the scikit-learn library documentation we also learnt that transductive models are ideal for cases where they would not be applied on new or unseen data. For this reason, we chose agglomerative clustering as it is good at clusters which might possibly have connectivity constraints between them.

We determined that the performance of Kmeans and Agglomerative clustering were very similar. The only difference was that agglomerative clustering had a slight less silhouette score indicating that the clusters formed by Kmeans were well formed compared to agglomerative. For this reason, we proceeded with using Kmeans in the further analysis.

Associative rule mining was also chosen because we wanted to find rules that were leading to poverty.

V. RESULTS AND DISCUSSION
As mentioned in the AR mining section it was found that not all zip codes were "underPoverty" status across the five years. In addition to that, table 4 (see appendix) further summarizes the descriptive statistics information for those zip codes. It seems that Zip codes that have median income of around 20k and median house value below 100k and Gini index moving towards 0.40 has the potentiality to move towards poverty status.

Perhaps, the concerned social bodies and government authorities can focus on neighborhoods with households with median income of less than equal to 20k and median house value below 100k and work in those communities to improve their poverty status.
Additionally, providing opportunities and easier ways to establish businesses might help the residents in poverty prone zip codes to uplift their economic status.
The inequality towards other races in regions with white majority indicates that there's a need for policies and interventions that promote economic upliftment of zip codes with diverse races. This could help reduce the disparities in poverty rates across different races and promote social equity, covering the issue of SDG 10 on reducing inequality.

VI. APPENDIX

Variable Category	Variables Name	Variable Description
	popDensity	Population Density (Per Sq. Mile)
Race & Ethnicity	whiteMajority	Proportion of population with white people majority
Income & Poverty	incomeBelowPovertyLevelProportion	Proportion of population whose income is below poverty level in last 12 months
	incomeGiniIndex	Gini Index of income for that zip code
	underPoverty	Proportion of population who are under poverty. 0 - if more than 50% of families in the zip code have a ratio over 2 (not under poverty) 1 - if 50% or more of families in the zip code are having ratio under 2 (under poverty)
	incomePerCapita	Income per capita for that zip code
	incomeMedian	Median income for the zip code
Housing	housingRentMoreThan30pcProportion	Proportion of population who spend more than 30% of their income on house rent
	housingMedianValue	Median value of all occupied houses
	housingMedianGrossRent	Median Gross rent of all occupied houses
Education	eduHsOrLessProportion	Proportion of population having education less than equal to HighSchool
	eduHsProportion	Proportion of population having education equal to HighSchool
	eduBachOrBetterProportion	Proportion of population having education equal to bachelor's or better
	eduDropoutMajority	Dropout proportion

	eduEnrollmentMajority	Enrolment Proportion
Employment	employedMajority	Proportion of employment
	employedPvtSecProportion	Proportion of employment in private sector
	employedPubSecProportion	Proportion of employment in public sector
	employedSelfEmpProportion	Proportion of self-employment
	employedPvtNonProProportion	Proportion of private non-profit employment
	employedUnpaidFamProportion	Proportion of unpaid families employees
Health	insuredProportion	Proportion of population having health insurance

Table 1: Data Dictionary

	LHS	RHS	support	confidence	coverage	lift	count
	All	All	All	All	All	All	All
[68]	{housingMedianValue=1,housingMedianGrossRent=1}	{underPoverty=1}	1.000	1.000	1.000	1.000	39.000
[69]	{employedUnpaidFamProportion=0,housingMedianGrossRent=1}	{underPoverty=1}	1.000	1.000	1.000	1.000	39.000
[70]	{employedPvtNonProProportion=0,housingMedianGrossRent=1}	{underPoverty=1}	1.000	1.000	1.000	1.000	39.000
[71]	{employedSelfEmpProportion=0,housingMedianGrossRent=1}	{underPoverty=1}	1.000	1.000	1.000	1.000	39.000
[72]	{employedPubSecProportion=0,housingMedianGrossRent=1}	{underPoverty=1}	1.000	1.000	1.000	1.000	39.000
[73]	{employedProportion=1,housingMedianGrossRent=1}	{underPoverty=1}	1.000	1.000	1.000	1.000	39.000
[74]	{eduEnrollmentProportion=0,housingMedianGrossRent=1}	{underPoverty=1}	1.000	1.000	1.000	1.000	39.000
[75]	{eduBachOrBetterProportion=0,housingMedianGrossRent=1}	{underPoverty=1}	1.000	1.000	1.000	1.000	39.000
[76]	{employedUnpaidFamProportion=0,housingMedianValue=1}	{underPoverty=1}	1.000	1.000	1.000	1.000	39.000
[77]	{employedPvtNonProProportion=0,housingMedianValue=1}	{underPoverty=1}	1.000	1.000	1.000	1.000	39.000
[957]	{incomeBelowPovertyLevelProportion=0,eduBachOrBetterProportion=0,eduEnrollmentProportion=0,housingMedianGrossRent=1}	{underPoverty=1}					0.949
[601]	{incomePerCapita=incomePerCapitaQ1,employedProportion=1,employedSelfEmpProportion=0,housingMedianValue=1}	{underPoverty=1}					0.923
[668]	{whiteMajority=1,eduEnrollmentProportion=0,employedPubSecProportion=0,housingMedianGrossRent=1}	{underPoverty=1}					0.923

AR Rules for 2018

	LHS	RHS	support	confidence	coverage	lift	count
	All	All	All	All	All	All	All
[361]	{employedProportion=1,housingMedianValue=1,housingMedianGrossRent=1}	{underPoverty=1}	1.000	1.000	1.000	1.000	28.000
[362]	{eduBachOrBetterProportion=0,housingMedianValue=1,housingMedianGrossRent=1}	{underPoverty=1}	1.000	1.000	1.000	1.000	28.000
[363]	{eduHsProportion=1,housingMedianValue=1,housingMedianGrossRent=1}	{underPoverty=1}	1.000	1.000	1.000	1.000	28.000
[364]	{eduHsOrLessProportion=0,housingMedianValue=1,housingMedianGrossRent=1}	{underPoverty=1}	1.000	1.000	1.000	1.000	28.000
[365]	{employedUnpaidFamProportion=0,insuredProportion=1,housingMedianGrossRent=1}	{underPoverty=1}	1.000	1.000	1.000	1.000	28.000
[366]	{employedPvtNonProProportion=0,insuredProportion=1,housingMedianGrossRent=1}	{underPoverty=1}	1.000	1.000	1.000	1.000	28.000
[367]	{employedSelfEmpProportion=0,insuredProportion=1,housingMedianGrossRent=1}	{underPoverty=1}	1.000	1.000	1.000	1.000	28.000
[368]	{employedPubSecProportion=0,insuredProportion=1,housingMedianGrossRent=1}	{underPoverty=1}	1.000	1.000	1.000	1.000	28.000
[369]	{employedProportion=1,insuredProportion=1,housingMedianGrossRent=1}	{underPoverty=1}	1.000	1.000	1.000	1.000	28.000
[370]	{eduBachOrBetterProportion=0,insuredProportion=1,housingMedianGrossRent=1}	{underPoverty=1}	1.000	1.000	1.000	1.000	28.000
[159]	{incomeBelowPovertyLevelProportion=0,eduEnrollmentProportion=0,housingMedianValue=1}	{underPoverty=1}			0.929		1.000

AR Rules for 2019

	LHS	RHS	support	confidence
	All	All	All	All
[520]	{employedProportion=1,housingMedianValue=1,housingMedianGrossRent=1}	{underPoverty=1}	1.000	1.000
[521]	{eduHsOrLessProportion=0,housingMedianValue=1,housingMedianGrossRent=1}	{underPoverty=1}	1.000	1.000
[522]	{incomeBelowPovertyLevelProportion=0,housingMedianValue=1,housingMedianGrossRent=1}	{underPoverty=1}	1.000	1.000
[523]	{employedPvtNonProProportion=0,employedUnpaidFamProportion=0,housingMedianGrossRent=1}	{underPoverty=1}	1.000	1.000
[524]	{employedSelfEmpProportion=0,employedUnpaidFamProportion=0,housingMedianGrossRent=1}	{underPoverty=1}	1.000	1.000
[525]	{employedPubSecProportion=0,employedUnpaidFamProportion=0,housingMedianGrossRent=1}	{underPoverty=1}	1.000	1.000
[526]	{employedProportion=1,employedUnpaidFamProportion=0,housingMedianGrossRent=1}	{underPoverty=1}	1.000	1.000
[527]	{eduHsOrLessProportion=0,employedUnpaidFamProportion=0,housingMedianGrossRent=1}	{underPoverty=1}	1.000	1.000
[528]	{incomeBelowPovertyLevelProportion=0,employedUnpaidFamProportion=0,housingMedianGrossRent=1}	{underPoverty=1}	1.000	1.000
[529]	{employedSelfEmpProportion=0,employedPvtNonProProportion=0,housingMedianGrossRent=1}	{underPoverty=1}	1.000	1.000
[369]	{incomeBelowPovertyLevelProportion=0,incomePerCapita=incomePerCapitaQ1,employedPvtNonProProportion=0}	{underPoverty=1}	0.967	1.00
[241]	{incomePerCapita=incomePerCapitaQ1,eduDropoutProportion=0,employedUnpaidFamProportion=0}	{underPoverty=1}	0.900	1.00
[671]	{whiteMajority=1,employedPubSecProportion=0,employedPvtNonProProportion=0,housingMedianValue=1}	{underPoverty=1}	0.900	1.00
[745]	{incomePerCapita=incomePerCapitaQ1,eduHsProportion=1,eduEnrollmentProportion=0,employedSelfEmpProportion=0}	{underPoverty=1}	0.900	1.00

AR Rules for 2020

	LHS	RHS	support	confidence
	All	All	All	All
[299]	{eduHsOrLessProportion=0,housingMedianValue=1,housingMedianGrossRent=1}	{underPoverty=0}	0.992	0.992
[300]	{incomeBelowPovertyLevelProportion=0,housingMedianValue=1,housingMedianGrossRent=1}	{underPoverty=0}	0.992	0.992
[301]	{employedUnpaidFamProportion=0,insuredProportion=1,housingMedianGrossRent=1}	{underPoverty=0}	0.992	0.992
[302]	{employedSelfEmpProportion=0,insuredProportion=1,housingMedianGrossRent=1}	{underPoverty=0}	0.992	0.992
[303]	{employedPubSecProportion=0,insuredProportion=1,housingMedianGrossRent=1}	{underPoverty=0}	0.992	0.992
[304]	{employedProportion=1,insuredProportion=1,housingMedianGrossRent=1}	{underPoverty=0}	0.992	0.992
[305]	{eduHsOrLessProportion=0,insuredProportion=1,housingMedianGrossRent=1}	{underPoverty=0}	0.992	0.992
[306]	{incomeBelowPovertyLevelProportion=0,insuredProportion=1,housingMedianGrossRent=1}	{underPoverty=0}	0.992	0.992
[307]	{employedSelfEmpProportion=0,employedUnpaidFamProportion=0,housingMedianGrossRent=1}	{underPoverty=0}	0.992	0.992
[308]	{employedPubSecProportion=0,employedUnpaidFamProportion=0,housingMedianGrossRent=1}	{underPoverty=0}	0.992	0.992
[916]	{incomeBelowPovertyLevelProportion=0,eduDropoutProportion=0,insuredProportion=1,housingMedianValue=1}	{underPoverty=0}	0.990	0.992

AR Rules for 2021

Year	Number of new zip codes in "underPoverty" status	Insights on these zip codes
2018	39	• The median income per capita is 20340. • The median population density is 215. • The median housing median value is 95k. • median employed proportion is 92% • the median Gini index is 0.41
2019	28	• the median income per capita is 21697. • the median population density is 383. • the median housing median value is 99k. • median employed proportion is 93%

		• the median Gini index is 0.43
2020	30	• The median income per capita is 20606. • The median population density is 170. • The median housing median value is 94k. • Median employed proportion is 95% • The median Gini index is 0.41
2021	713	• The median income per capita is 49096 • The median population density is 2837 • Median employed proportion is 95% • The median gini index is 0.44

Table 4: Descriptive statistics for new zip codes "underPoverty" for different years

VII. REFERENCES

[1] "Social Explorer," [Online]. Available: https://www.socialexplorer.com/.

[2] "Social Explorer Report URL," [Online]. Available: http://www.socialexplorer.com/pub/reportdata/HtmlResults.aspx?reportid=R13352078.

[3] L. Kilduff, "How Poverty in the United States Is Measured and Why It Matters," [Online]. Available: https://www.prb.org/resources/how-poverty-in-the-united-states-is-measured-and-why-it-matters/.

[4] "Scikit-learn Clustering," [Online]. Available: https://scikit-learn.org/stable/modules/clustering.html.