AN EVALUATION OF 2020 CENSUS DATA

Jessica Chipera

Northcentral University: School of Technology

Course Number: TIM-8501

Dr. Nicholas Harkiolakis

October 8, 2022

Overview

This paper is a process of exploratory data analysis with data clustering. I was given primary research data from the United States Census Bureau. The foremost purpose for collecting this data was to gather information concerning employment, but a secondary purpose was to collect information on the demographics of the population (Census, 2020).

The universe, Ω, from 2020, is a population of civilian noninstitutional Americans living in housing units, as well as members of the military living externally from the base or with their families on post. All data has at least one civilian adult per household. The domain is nine noncash income sources: food stamps, school lunch programs, employer-provided group health insurance, employer-provided pensions, personal health insurance, Medicaid, Medicare, or military health care, and energy assistance. The data also contains other characteristics including age, sex, and race (Census, 2020).

A probability sample was used to collect 54,000 households, which were interviewed monthly for four consecutive months one year and then again for the same duration one year later. They were scientifically selected on the basis of their area of residence in order to be representative of the population (Census, 2020). This was verified by Figure 1 below.

Exploratory Data Analysis

After loading the libraries and data that I thought I would need, I first verified the government's claim concerning the representativeness of the sample (Census, 2020). Figure 1 shows that the sample is, in fact, representative of the population because of the uniform frequency distribution by region of residence.

Figure 1

Frequency Distribution of Census Sample by Region

```
histogram_boxplot(df2,"GEDIV")
#Recode - Census division of current residence

#Values: 1 = New England
#2 = Middle Atlantic
#3 = East North Central
#4 = West North Central
#5 = South Atlantic
#6 = East South Central
#7 = West South Central
#8 = Mountain
#9 = Pacific
```

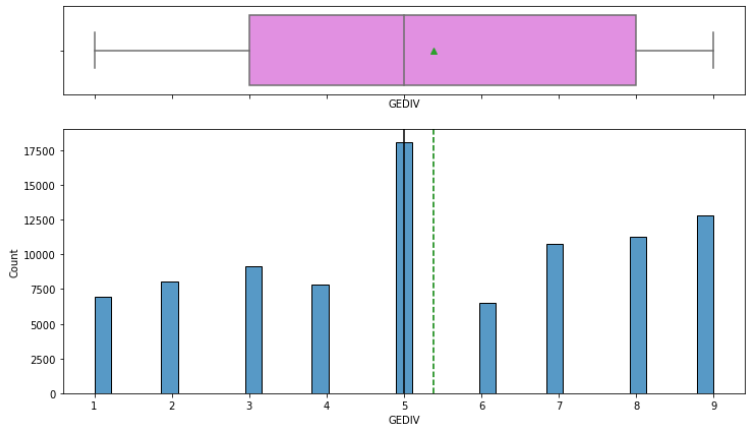
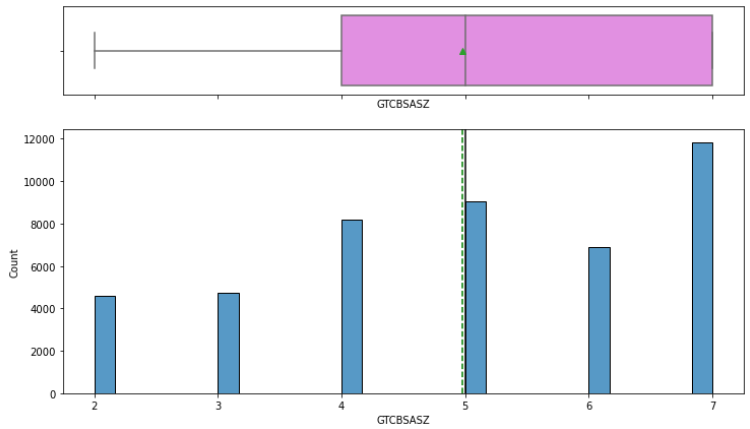


Figure 2

Frequency Distribution of Census Sample by Metropolitan Area Size

```
histogram_boxplot(df2,"GTCBSASZ")
#Metropolitan area (CBSA) size

#Values:
#2 = 100,000 - 249,999
#3 = 250,000 - 499,999
#4 = 500,000 - 999,999
#5 = 1,000,000 - 2,499,999
#6 = 2,500,000 - 4,999,999
#7 = 5,000,000+
```



The frequency distribution of Figure 2 shows that there is a skew toward larger metropolitan cities. This is also expected and likely representative of the population, given the definition of a larger metropolitan city.

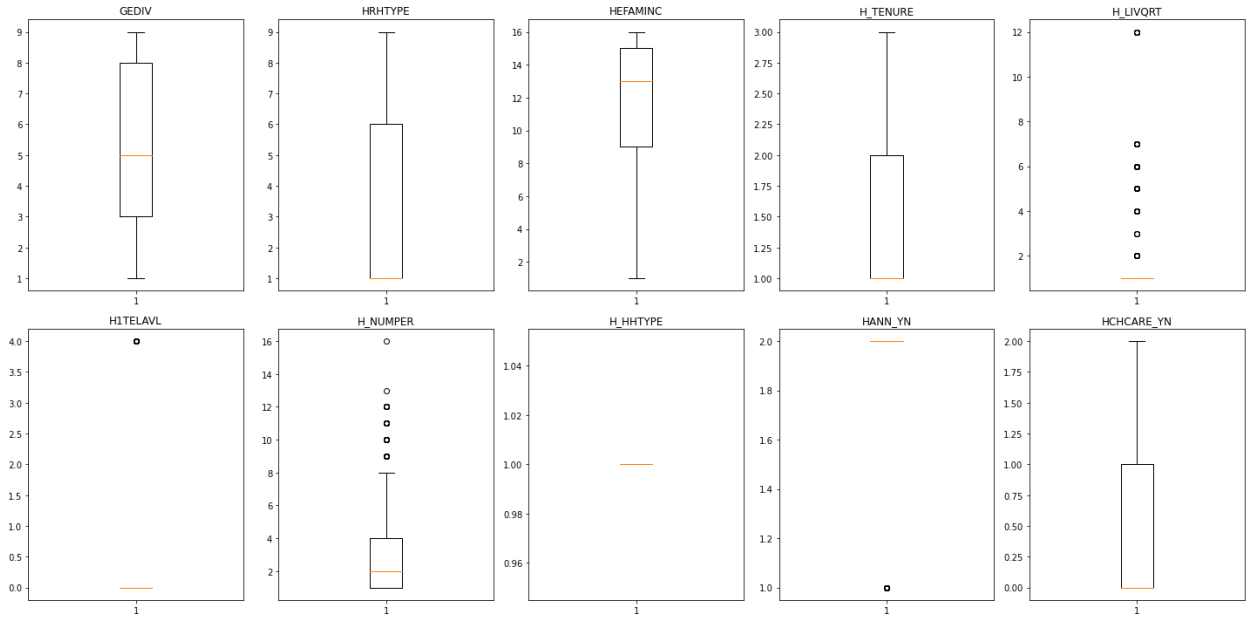
The sample is 91,500 rows by 134 columns. 133 columns belong to probability space $n \geq 0,\ n \in \mathbb{Z}\ ;\ (\Omega, \mathcal{H}, \mathbb{Z})$ where $\mathbb{Z}$ is the set of integers and $\mathscr{H}$ is a sigma algebra. A topological space (Φ, τ) bisects universe Ω. Some 133 columns contain discrete data, and the remaining column contains categorical data (Census, 2020).

After viewing the data dictionary provided by the Census Bureau (Census 2020), we can see that the entirety of $(\Omega, \mathcal{H}, \mathbb{Z}) \notin (\Phi, \tau)$. In order to do exploratory data analysis, we must delete all data that is outside the universe. It turns out that approximately 65% of the data included in the sample is either outside universe Ω or it is a subset of $\Omega \notin (\Phi, \tau)$. The entirety of this 65% of the data was deleted. The remaining subset of $n \geq 0,\ n \in \mathbb{Z}\ ;\ (\Omega, \mathcal{H}, \mathbb{Z}) \in (\Phi, \tau)$ is 91,500 rows by 53 columns. Figure A1 in the Appendix contains a table of the mean, standard deviations, min, max, and interquartile range values for $(\Omega, \mathcal{H}, \mathbb{Z}) \in (\Phi, \tau)$ (Mukhiya and Ahmed, 2020).

We can see that, concerning the isolated the data $(\Omega, \mathcal{H}, \mathbb{Z}) \in (\Phi, \tau)$, there are no subsets with nonzero numbers. All data is numerical and discrete. Outliers can be shown in the boxplots provided in Figure 3. The majority of outliers are apparent in the subsets called "Living Quarters (H_LVQRT) and the number of residents per household (H_NUMPER). Remember, we already deleted over half of the data to isolate $(\Omega, \mathcal{H}, \mathbb{Z}) \in (\Phi, \tau)$. Thus, to maintain the representativeness of the sample, we will leave the outliers untreated.

Figure 3

Outlier Tests by Column for Sample $(\Omega, \mathcal{H}, \mathbb{Z}) \in (\Phi, \tau)$



Additional abnormalities exist. One such abnormality is that the sample includes a higher proportion of individuals who are very poor, making less than $2,500 per year, and individuals who are relatively wealthy, making more than $100,000 per year, as shown by points “0” and “41” in Figure 4. It can be assumed that a more uniform distribution could potentially be shown if group 41 were broken into smaller groups. All other groups were divided into increments of $2,500.

A heatmap was generated to locate potential correlations in the probability space $(\Omega, \mathcal{H}, \mathbb{Z})$ in order to study predictability (Mukhiya and Ahmed, 2020). Due to the size of $(\Omega, \mathcal{H}, \mathbb{Z})$, the heatmap is somewhat hard to read, but it is provided in the Appendix. I created scatterplots to understand some of the correlations revealed in the heatmap (Mukhiya and Ahmed, 2020).

Figure 4

Frequency Distribution of Sample $(\Omega, \mathcal{H}, \mathbb{Z}) \in (\Phi, \tau)$ *by Income*

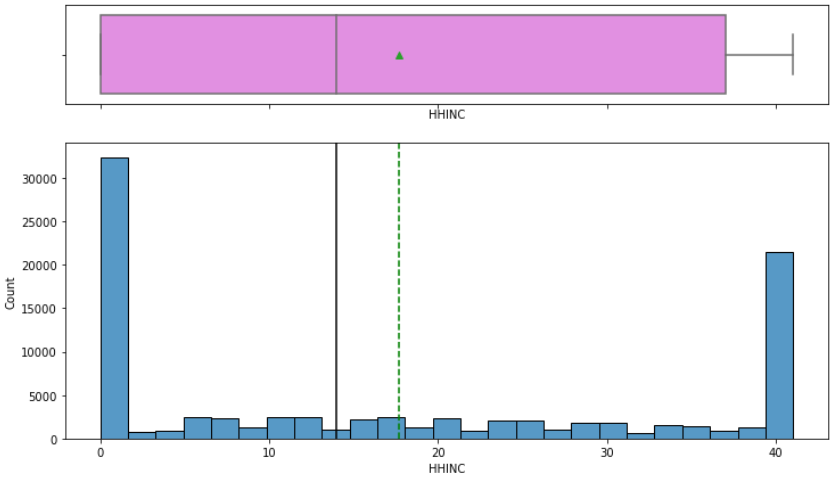


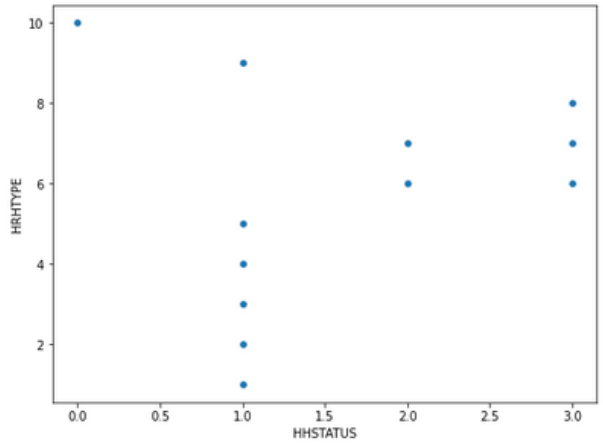
Figure 5

Correlation of Subsets $HHSTATUS \in (\Omega, \mathcal{H}, \mathbb{Z})$ *and* $HRHTYPE \in (\Omega, \mathcal{H}, \mathbb{Z})$

```
plt.figure(figsize=[8, 6])
sns.scatterplot(x=df2.HHSTATUS,y=df8.HRHTYPE)
plt.show()

#HHSTATUS = Household Status
#1 = Primary family
#2 = Nonfamily householder living alone
#3 = Nonfamily householder living with nonrelatives

#HRHTYPE = Household Type
#01 = Married couple primary family (neither spouse in Armed Forces)
#02 = Married couple primary family (one spouse in Armed Forces)
#03 = Unmarried civilian male primary family householder
#04 = Unmarried civilian female primary family householder
#05 = Primary family household - reference person in Armed Forces and unmarried
#06 = Civilian male nonfamily householder
#07 = Civilian female nonfamily householder
#08 = Nonfamily householder household - reference person in Armed Forces
# 09 = Group quarters with actual families (This is new in 1994)
#10 = Group quarters with secondary individuals only
```



One such scatterplot, Figure 5, implies a positive correlation between primary family households and other primary household classifiers in another column. This is a good sign, as it implies that isolating the data $(\Omega, \mathcal{H}, \mathbb{Z}) \in (\Phi, \tau)$ did not create incompleteness to the point of destroying correlations and neighborhoods.

As stated before, our dataset is large and multidimensional. In order to attempt to understand the variances of $(\Omega, \mathcal{H}, \mathbb{Z}) \in (\Phi, \tau)$, I ran Principal Component Analysis (PCA). PCA is a dimensionality reduction tool that simplifies complex and redundant data so that it can be analyzed. In other words, it finds projections of $\tilde{y}_n \in y_n \in (\Omega, \mathcal{H}, \mathbb{Z}) \in (\Phi, \tau)$ (Deisenroth et all, 2020). The covariance matrix "M" of our data is:

$$M = \frac{1}{N}\sum_{n=1}^{N} y_n y_n^T$$

Where y_n^T is the transpose matrix of y_n.

The projection matrix of y_n can be defined as:

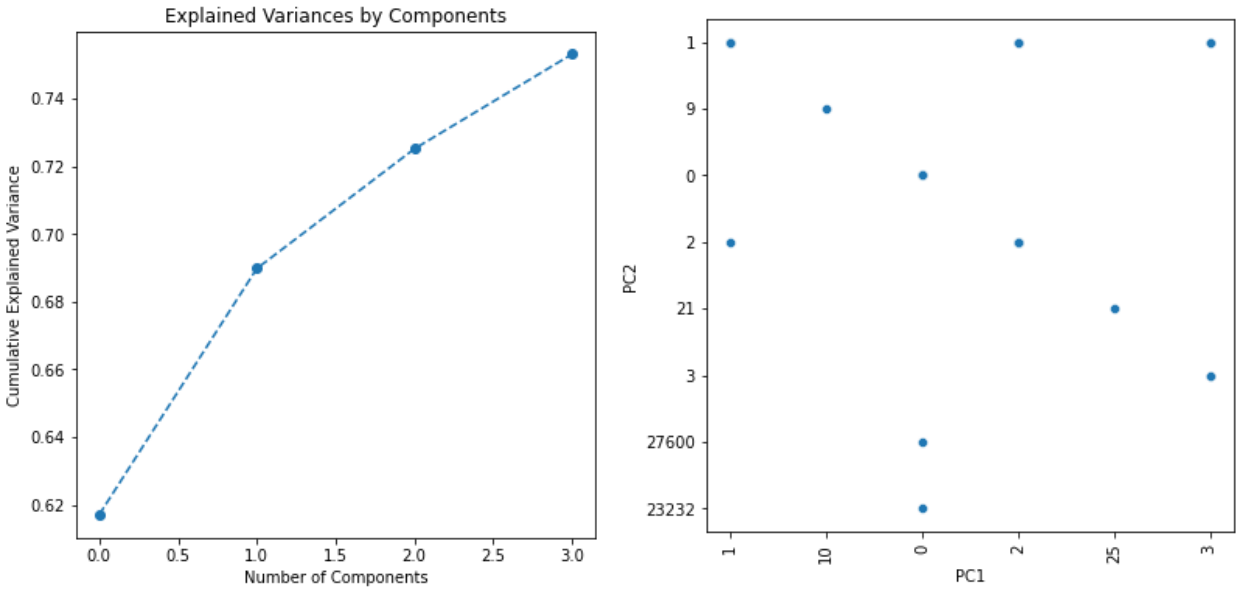
$$B := [b_1, \dots b_M] \in \mathbb{R}^{DM}$$

Where the columns of "B" are orthogonal such that $b_i^T b_j = 0$ if and only if $i \neq j$ and $b_i^T b_i = 1$. We can then find the M-dimensional subspace $U \subseteq \mathbb{R}^D$, $\dim(U) = M < D$, which is the space onto which we will try to project the data, denoted as $\tilde{y}_n \in U$ (Deisenroth et all, 2020).

The scree plot in Figure 6 reveals that we can use PCA to group or cluster all of $(\Omega, \mathcal{H}, \mathbb{Z}) \in (\Phi, \tau)$ such that it is projected onto a metric or normed space because reducing the dimensionality by this method maintains 69 percent of the data. This is usually acceptable for analysis; however, we have already deleted 65 percent of the data prior to this 69 percent reduction (Deisenroth et all, 2020; Gut, 2013; Mukhiya and Ahmed, 2020). As shown in Figure 6, the relationship between the data in PC1 and PC2 is not completely linear. Thus, a nonlinear clustering algorithm must be used.

Figure 6

Principal Component Analysis of $(\Omega, \mathcal{H}, \mathbb{Z}) \in (\Phi, \tau)$



T-Distributed Stochastic Neighbor Embedding (TSNE) was used to generate a nonlinear cluster group for $(\Omega, \mathcal{H}, \mathbb{Z}) \in (\Phi, \tau)$. TSNE is a dimension reduction technique that preserves local neighborhood structure and can help when PCA does not work or when the triangle property is violated (VanderMaaten, 2008).

TSNE maintains nearest neighbors by constructing a lower dimensional map and converting pair-wise Euclidean distances between points into a probability density function $\mathbb{P}(j|i)$ where "j" is a neighbor point of "i" and $i, j \in (\Omega, \mathcal{H}, \mathbb{Z})$ (Deisenroth et all, 2020; VanderMaaten, 2008). In other words, TSNE maps $\mathbb{P}(i)$ to $P(i)$ as shown below:

$$\mathbb{P}(j|i) = \frac{\frac{exp(-||x_i - x_j||^2}{2\sigma_i^2}}{\frac{\sum_{k \neq 1} exp(-||x_i - x_j||^2}{2\sigma_i^2}} \rightarrow P(j|i) = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq 1}(1 + ||y_i - y_j||^2)^{-1}}$$

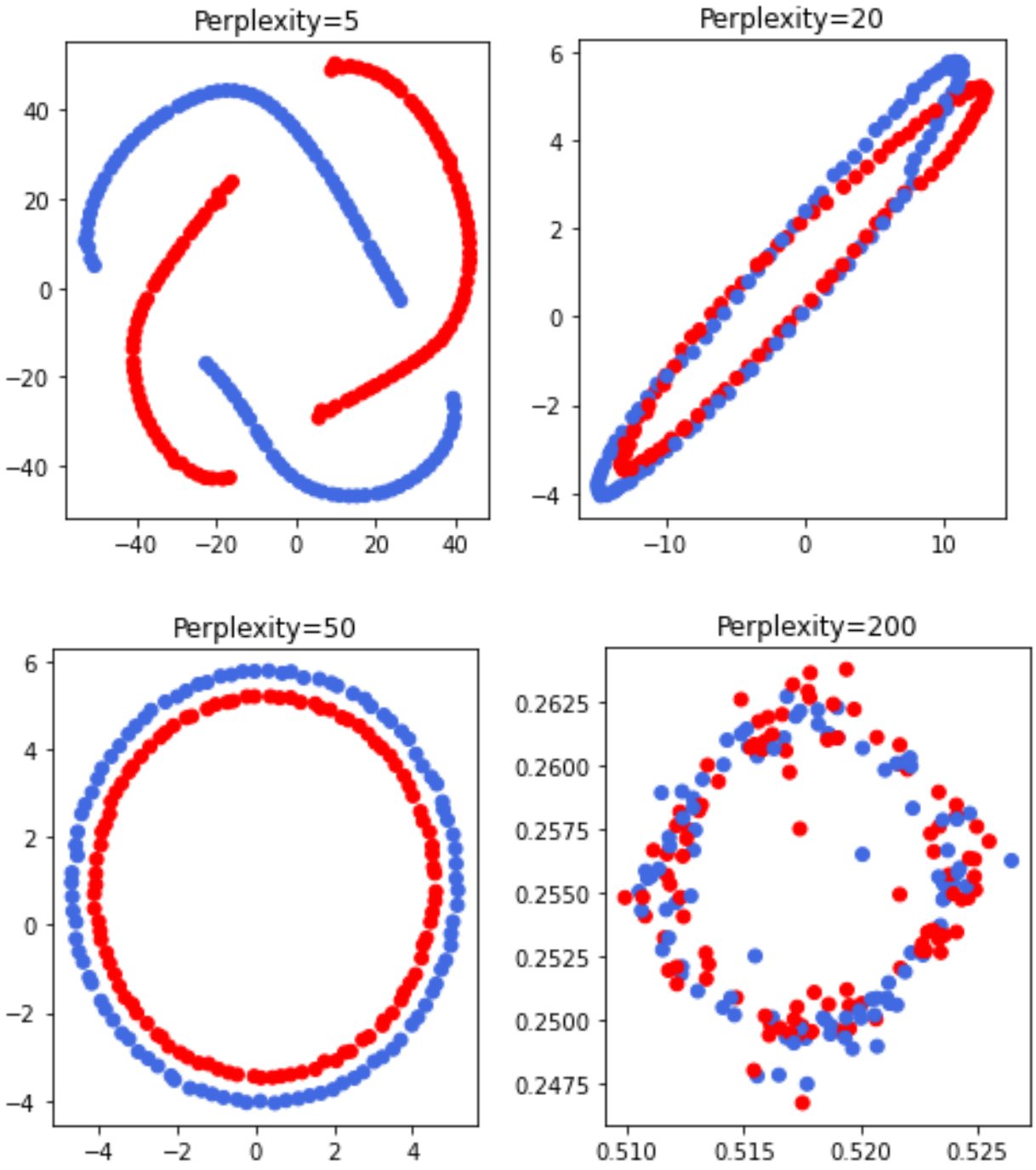
The algorithm then uses Kullback-Leibler divergence loss function to find the gradient descent and quantify the differences between the two probability distributions $\mathbb{P}(j|i)$ and P($j|i$) for each datum in order to cluster the data (Gut, 2013; VanderMaaten, 2008). The formula for Kullback-Leibler divergence is as follows:

$$KL(\mathbb{P}||P) = \sum_i \sum_{j \neq 1} \mathbb{P}_{ij} log \frac{\mathbb{P}_{ij}}{P_{ij}}$$

Thus, high-dimensional similarities belong to $\mathbb{P}(j|i)$ and lower-dimensional similarities belong to $P(j|i)$. Figure 7 shows that the data clusters nicely, as expected. The "perplexity" value is the vector denoting how many neighbors should be considered in the grouping. As expected, the higher perplexity values created data that clustered together. The blue group and red group are $\mathbb{P}(j|i)$ and $P(j|i)$ respectively, and the kernel is chosen adaptively to achieve the desired perplexity (number of neighbors) (VanderMaaten, 2008).

Figure 7

T-Distributed Stochastic Neighbor Embedding of $(\Omega, \mathcal{H}, \mathbb{Z}) \in (\Phi, \tau)$ *with Various Perplexity*



Findings

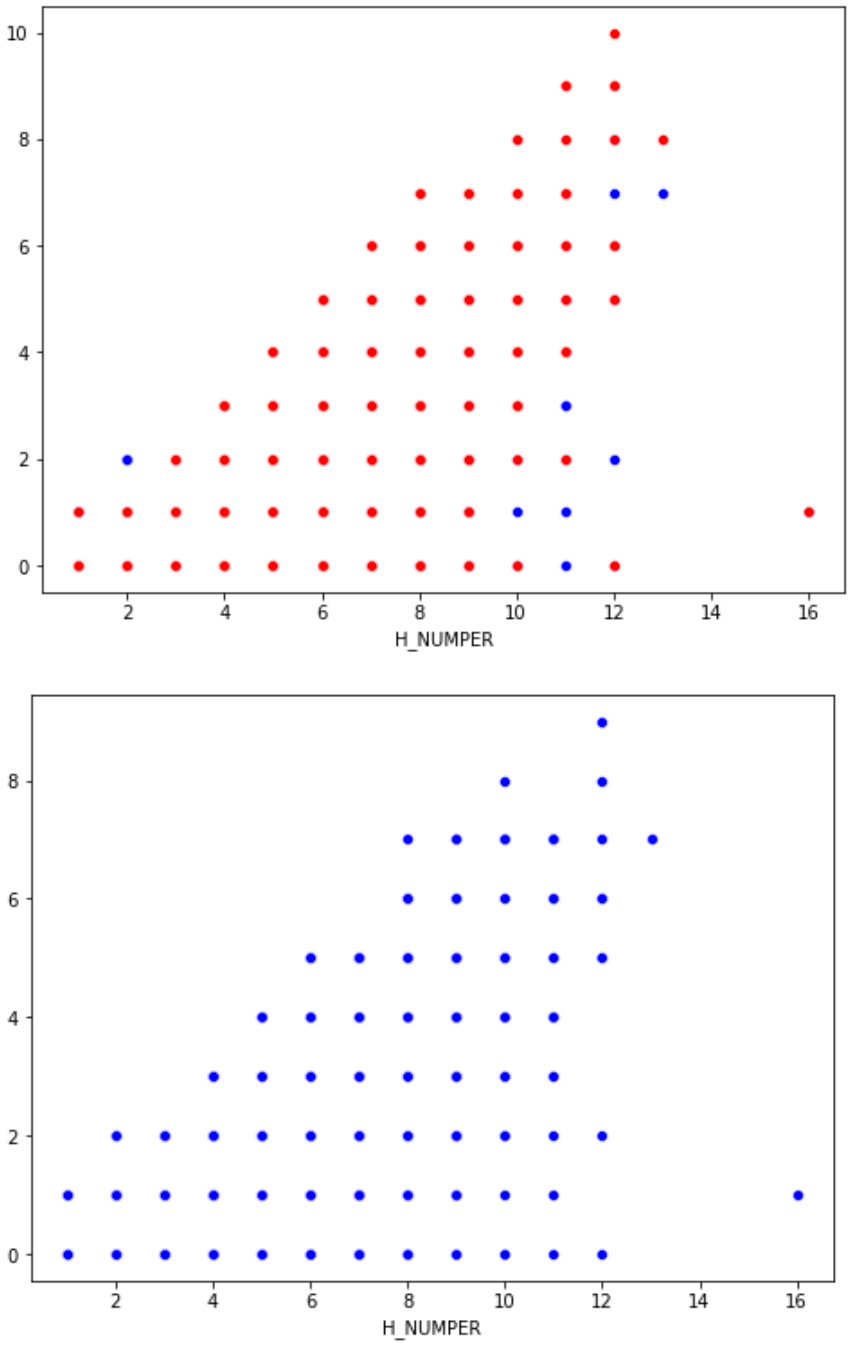
The combination of exploratory data analysis and dimension reduction led to some inferences, some of which were unexpected. There is a strong positive correlation between the status of the house (owned, rented, or no-cash rent) and whether or not the house is a public housing project owned by a housing authority or public agency. Similarly, there is a strong correlation between whether the house is owned or rented and the question pertaining to rent assistance from the government. This is to be expected, as people who live in public housing are not homeowners. Additionally, there is a strong negative correlation between the presence of a home loan or mortgage and whether the house is owned or rented. This indicates that the majority of the people surveyed either rent or they live in a house that is paid off or not borrowed against.

Concerning unemployment itself, we can see from the data that 2.1 percent of the sample indicated that they had a disability that prohibited them from working. These individuals collecting disability income received a median amount of $9,000 per year or a mode value of $14,400 per year. In addition, 2.5 percent of the sample indicated that they were receiving unemployment insurance.

There is a significant overlap in subsets HHUNDER18 $\cap$ H_NUMPER $\in (\Omega, \mathcal{H}, \mathbb{Z})$ and HUNDER18^C $\cap$ H_NUMPER $\in (\Omega, \mathcal{H}, \mathbb{Z})$ where HHUNDER18 is the number of households that had a child age younger than 18 and HUNDER18^C is the complement of households that had a child under age 18 (so, households that did not have such a child.) As shown in Figure 8, the subsets nearly overlap. This is unexpected and shows a potential underlying pattern of familial structure in the United States where people tend to have the same quantity of children at the same age that their parents did.

Figure 8

A Comparison of $HHUNDER18 \cap H\_NUMPER \in (\Omega, \mathcal{H}, \mathbb{Z})$ *and* $HUNDER18^{C} \cap H\_NUMPER \in (\Omega, \mathcal{H}, \mathbb{Z})$*:* Top chart shows overlap in red. Both charts show only $HUNDER18^{C} \cap H\_NUMPER \in (\Omega, \mathcal{H}, \mathbb{Z})$ in blue



The majority of people in the sample lived alone. Similarly, the majority of people in the sample live in a house, apartment or flat. Far fewer live in other residential situations: hotels, mobile homes, or otherwise. Most people included in the sample make $150,000 or more annually, a figure that could be used to define the "middle class." There is a large portion of people who do not make any income. The results of the residential situation data subset implies that the majority of people who do not have an income are likely living with another person who does have an income (for example, a spouse). Also applicable to the data, most of the interviewees identified as married couples. This is contradictory to the data subset for household size, which had a modality of one.

Subsequent Actions

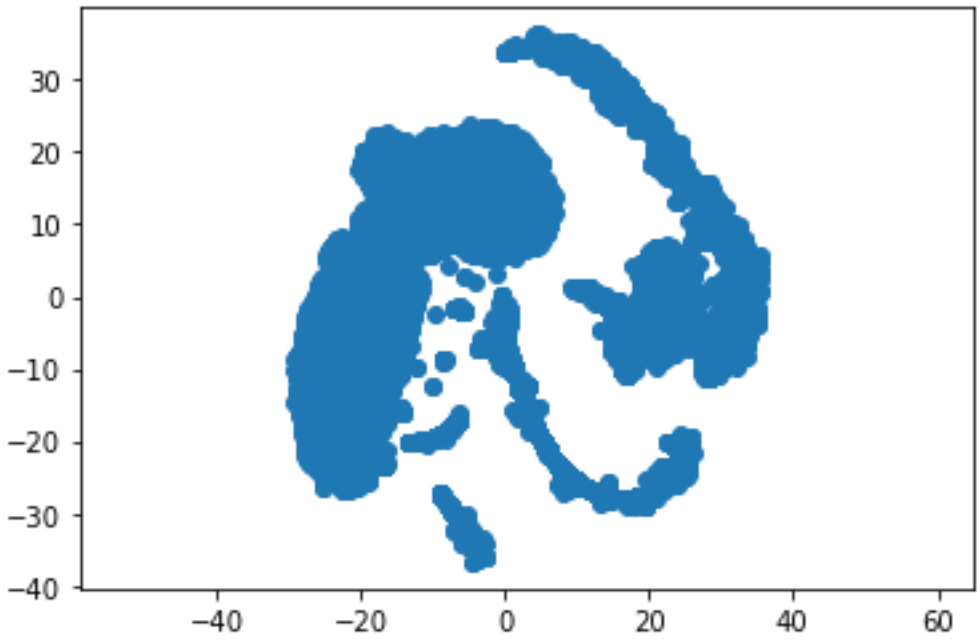
Obviously, the situation in the United States has potentially become drastically different than when this census was taken, given that immediately afterward the country began mitigation programs for the Covid-19 pandemic. It could prove truly fascinating to compare the results of the 2020 census with the next sample. In addition, it would be interesting to compare this census sample with the previous one. Both activities could show or predict some potential otherwise-unexpected trend changes in the population.

I was able to obtain a TSNE cluster graph of the entire dataset that would show groups that the heatmap did not find. Unfortunately, my computer does not have the 19 GB of RAM necessary to add color to show the separate subsets in the data. (However, I included a picture of the single-color graph in Figure 9). It would be nice to try this again in the future or look for a shortcut that might not be so calculation-heavy for my computer.

The clustering of the data would be useful for comparison with datasets from future and past censuses, in order to search for trends.

Figure 9

TSNE Clustering of Entire Dataset



References

Census (2020). Current Population Survey: 2020 Annual Social and Economic (ASEC) Supplement. https://www2.census.gov/programs-surveys/cps/techdocs/cpsmar20.pdf

Deisenroth et all (2020). Mathematics for Machine Learning. Cambridge University Press.

Gut, A. (2013). Probability: A Graduate Course. Second Ed. Springer Science+Business Media.

Mukhiya, S. K., & Ahmed, U. (2020). Hands-on exploratory data analysis with Python. Packt Publishing.

VanderMaaten, L. (2008). Visualizing Data Using t-SNE. Journal of Machine Learning Research.

Appendix

Figure A1

Mean, Standard Deviation, Minimum Value, Maximum Value and Interquartile Range Values for $(\Omega, \mathcal{H}, \mathbb{Z}) \in (\Phi, \tau)$

	count	mean	std	min	25%	50%	75%	max
GEDIV	91500.0	5.382656	2.497849	1.0	3.0	5.0	8.0	9.0
HRHTYPE	91500.0	2.124678	2.519624	0.0	0.0	1.0	4.0	10.0
HEFAMINC	91500.0	7.260481	6.691254	-1.0	-1.0	9.0	14.0	16.0
H_TENURE	91500.0	1.035268	0.724797	0.0	1.0	1.0	2.0	3.0
H_LIVQRT	91500.0	1.262612	1.155883	1.0	1.0	1.0	1.0	12.0
H1TELAVL	91500.0	0.004066	0.127460	0.0	0.0	0.0	0.0	4.0
H_NUMPER	91500.0	1.726328	1.730261	0.0	0.0	1.0	3.0	16.0
HANN_YN	91500.0	1.303257	0.943277	0.0	0.0	2.0	2.0	2.0
HCHCARE_YN	91500.0	0.353093	0.729884	0.0	0.0	0.0	0.0	2.0
HCOV	91500.0	0.764590	0.662614	0.0	0.0	1.0	1.0	3.0
HCSPVAL	91500.0	135.069093	1417.263625	0.0	0.0	0.0	0.0	99999.0
HCSP_YN	91500.0	1.298962	0.942371	0.0	0.0	2.0	2.0	2.0
HDIS_YN	91500.0	1.307956	0.944245	0.0	0.0	2.0	2.0	2.0
HDIVVAL	91500.0	890.436448	11112.500978	0.0	0.0	0.0	0.0	1009999.0
HDIV_YN	91500.0	1.196546	0.914546	0.0	0.0	2.0	2.0	2.0
HDST_YN	91500.0	1.275355	0.937024	0.0	0.0	2.0	2.0	2.0
HED_YN	91500.0	1.286831	0.939701	0.0	0.0	2.0	2.0	2.0
HENGAST	91500.0	1.300339	0.942664	0.0	0.0	2.0	2.0	2.0
HFINVAL	91500.0	107.550852	2548.150986	0.0	0.0	0.0	0.0	500000.0
HFIN_YN	91500.0	1.311038	0.944867	0.0	0.0	2.0	2.0	2.0
HFOODSP	91500.0	1.263071	0.933994	0.0	0.0	2.0	2.0	2.0
HH5TO18	91500.0	0.357049	0.811811	0.0	0.0	0.0	0.0	9.0
HHINC	91500.0	17.656907	16.770146	0.0	0.0	14.0	37.0	41.0
HHSTATUS	91500.0	0.906219	0.809742	0.0	0.0	1.0	1.0	3.0
HH_HI_UNIV	91500.0	1.012404	1.015949	0.0	0.0	1.0	1.0	3.0
HINC_FR	91500.0	1.311016	0.944862	0.0	0.0	2.0	2.0	2.0

	count	mean	std	min	25%	50%	75%	max
HINC_SE	91500.0	1.261552	0.933607	0.0	0.0	2.0	2.0	2.0
HINC_UC	91500.0	1.304973	0.943633	0.0	0.0	2.0	2.0	2.0
HINC_WC	91500.0	1.316055	0.945856	0.0	0.0	2.0	2.0	2.0
HINC_WS	91500.0	0.825202	0.687839	0.0	0.0	1.0	1.0	2.0
HINT_YN	91500.0	0.865366	0.725062	0.0	0.0	1.0	1.0	2.0
HMCAID	91500.0	1.780230	1.377420	0.0	0.0	3.0	3.0	3.0
HNUMFAM	91500.0	0.764579	0.657651	0.0	0.0	1.0	1.0	9.0
HOI_YN	91500.0	1.304984	0.943636	0.0	0.0	2.0	2.0	2.0
HPAW_YN	91500.0	1.314831	0.945617	0.0	0.0	2.0	2.0	2.0
HPENVAL	91500.0	2251.072918	16761.612545	0.0	0.0	0.0	0.0	1065075.0
HPEN_YN	91500.0	1.243891	0.928917	0.0	0.0	2.0	2.0	2.0
HPRIV	91500.0	1.095191	1.048788	0.0	0.0	1.0	2.0	3.0
HPUB	91500.0	1.464929	1.273205	0.0	0.0	1.0	3.0	3.0
HRNT_YN	91500.0	1.270492	0.935845	0.0	0.0	2.0	2.0	2.0
HSSI_YN	91500.0	1.294492	0.941406	0.0	0.0	2.0	2.0	2.0
HSSVAL	91500.0	4491.124678	10645.758429	0.0	0.0	0.0	0.0	126080.0
HSS_YN	91500.0	1.115126	0.883374	0.0	0.0	1.0	2.0	2.0
HSUR_YN	91500.0	1.306383	0.943924	0.0	0.0	2.0	2.0	2.0
HUNDER15	91500.0	0.362328	0.834027	0.0	0.0	0.0	0.0	10.0
HUNDER18	91500.0	0.439301	0.930660	0.0	0.0	0.0	0.0	10.0
HUNITS	91500.0	1.162557	1.427268	0.0	0.0	1.0	1.0	5.0
HVET_YN	91500.0	1.298230	0.942215	0.0	0.0	2.0	2.0	2.0
NOW_HCOV	91500.0	0.774699	0.677095	0.0	0.0	1.0	1.0	3.0
NOW_HMCAID	91500.0	1.787246	1.378774	0.0	0.0	3.0	3.0	3.0
NOW_HPRIV	91500.0	1.112951	1.062489	0.0	0.0	1.0	2.0	3.0
NOW_HPUB	91500.0	1.467934	1.274434	0.0	0.0	1.0	3.0	3.0
GTCBSASZ	91500.0	3.667574	2.616627	0.0	0.0	4.0	6.0	7.0

Figure A2

Heatmap of $(\Omega, \mathcal{H}, \mathbb{Z}) \in (\Phi, \tau)$