Similarity Between US Counties

Similarity Between US Counties

By Ari Fleischer, Zack Larson, and Ethan Morgan

Our project is What Makes Two US Counties Similar by Ari Fleischer, Zack Larson, and Ethan Morgan.

The code and data can be found at
https://github.com/AriFleischer13/Similarities-Between-Counties

The goal of this project was to be able to classify whether two counties are similar or not and find out what features are the most important for categorizing similarity.

The real-world application of this project is that the results can be used by companies and advertisers to better reach and target their desired audiences. The results can also be used to create more targeted advertising and ad placement.

Datasets

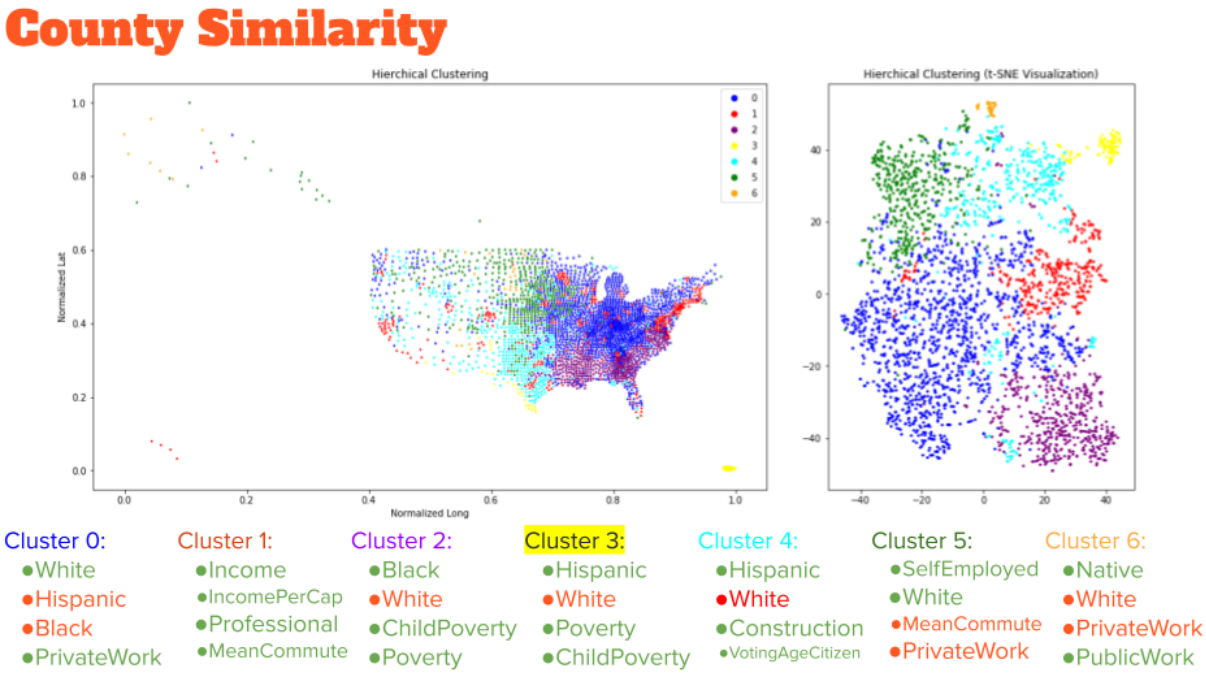
- **DP03 and DP05 tables of the 2015 and 2017 American Community Survey[1,2]**
  - **3,006 US counties, 136 county equivalents, and 78 municipalities of Puerto Rico**
  - **34 columns of economic and demographic data**
  - **2015 - Citizens & 2017 - Voting Age Citizens**
- **Latitude and Longitude Values for 3,142 US counties and county equivalents[3]**
- **Latitude and Longitude Values for 78 municipalities of Puerto Rico[4]**
  - **Values scraped from Wikipedia**
- **US Census Bureau Regions and Divisions for 50 US States Plus D.C.[5]**

[1] https://www.kaggle.com/muonneutrino/us-census-demographic-data?select=acs2015_county_data.csv
[2] https://www.kaggle.com/muonneutrino/us-census-demographic-data?select=acs2017_county_data.csv
[3] https://simplemaps.com/data/us-counties [4] https://en.wikipedia.org/wiki/Municipalities_of_Puerto_Rico
[5] https://github.com/cphalpert/census-regions/blob/master/us%20census%20bureau%20regions%20and%20divisions.csv

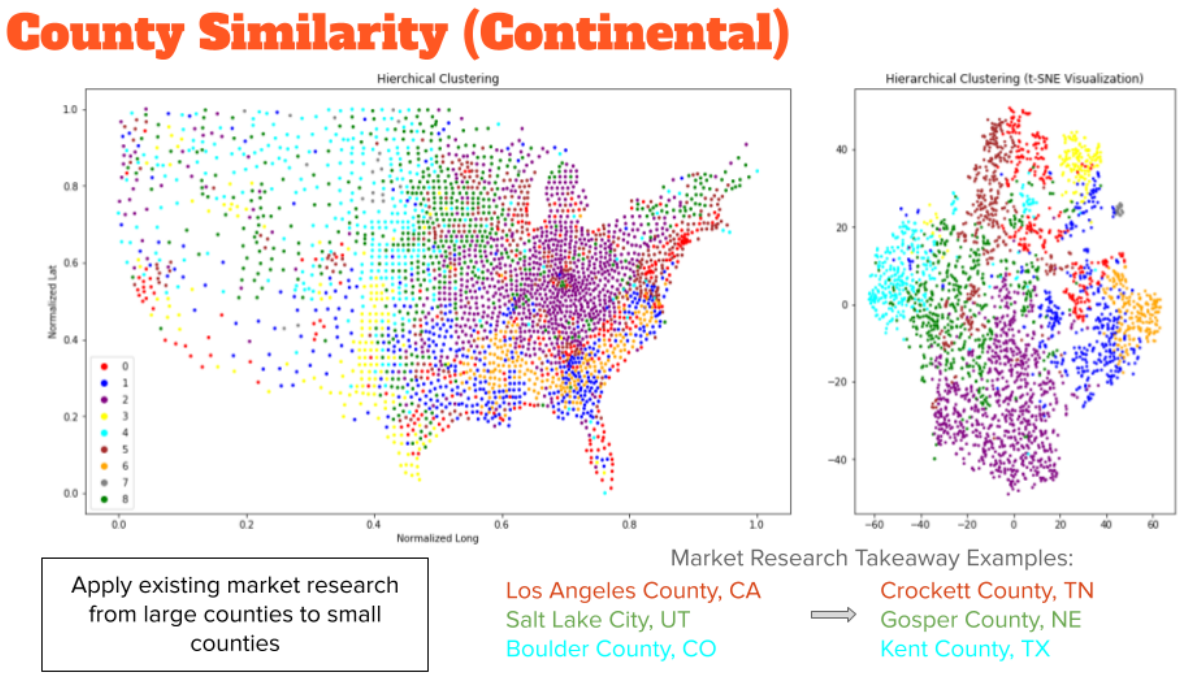
We used five separate datasets for the project. The first two are the combined results of the DP03 and DP05 tables of the American Community Survey for 2015 and 2017. The datasets contain data from the 3,006 US counties, 136 county equivalents, and the 78 municipalities of Puerto Rico. There are 34 columns of interest containing economic and demographic data related to income, race, employment, and transportation. The 2015 dataset had the number of citizens, where the 2017 dataset had the number of voting aged citizens (18+).

We also wanted to include location, which we hypothesised plays a big role in county similarity. We found a dataset with the latitude and longitude values of the 3,142 US county and county equivalents. Values for the municipalities of Puerto Rico were scraped from Wikipedia. The latitude and longitude values were joined with the economic and demographic data to create a 3,220 x 36 dataframe with every column numeric. The new dataset was then normalized, so the set of normalized feature vectors are in the unit-hypercube $[0,1]^{36}$.

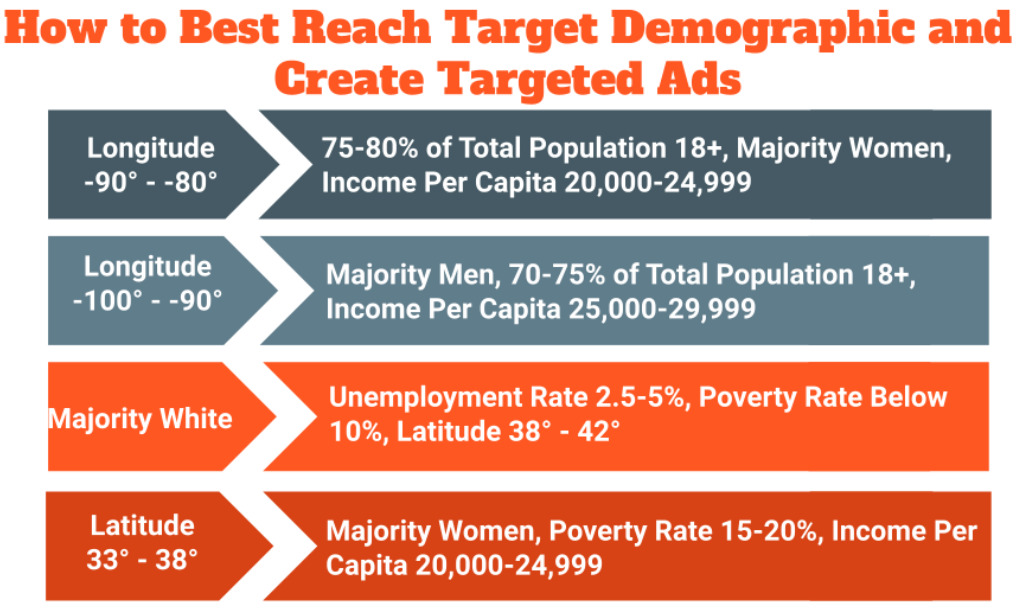
The fifth dataset (Regions and Divisions of states + D.C.) was just used as a way to potentially color plots. Puerto Rico was added into the South region and South Atlantic division.



Though several clustering methods resulted in good separation, Hierarchical Clustering provided the best results on this dataset. Using 7 clusters, we were able to separate the country into 6 distinct geographical regions, with one distinct cluster representing large cities. These results were achieved without incorporating any locational data into the model. The coloring here represents the generated clusters overlaid onto the unlabeled data points, graphed by latitude and longitude. For each cluster, we calculated a list of features, ranked by the difference between the population means and the cluster means. For the 'regional' clusters, we found that the racial, poverty, and commute features were most important. However, for the 'big cities' cluster, the most important features tended to relate to income and employment. The first few top features for each cluster are shown here, colored by whether they are positively or negatively correlated.



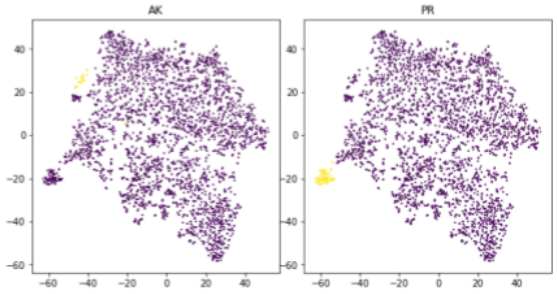
Removing counties belonging to Alaska, Hawaii, and Puerto Rico leaves us with only the continental US. We found that although the clustering solution here looks similar to the previous slide, most of the removed data points served as outliers in the dataset. Performing clustering on this set gives a cleaner result, with less confusion between clusters in the t-SNE plots. We suggest utilizing these clustering results to expand existing marketing research performed on large cities, to a range of smaller counties. For instance, insights from surveys more commonly conducted on large metropolitan areas, such as LA County, could be adapted to a much smaller county from the same cluster, such as Crockett County, TN, where less research has been conducted. Since the input data was stripped of all locational identifiers, and each column was normalized by its county's total population, we interpret these results as showing that the counties in a certain cluster will be more likely to have higher similarities between each other than with counties belonging to other clusters.



Next, we wanted to look into how companies and advertisers can best reach their targeted demographics and create properly targeted ads. We did this through Market Basket Analysis with one antecedent and one consequent to see what demographics a county is likely to have, given that it also contains a particular demographic of interest. We used the full unnormalized 2017 dataset and first replaced the categories for race, gender, transportation, employment type, and job sector with the max value in each row for those categories. Also, groups were made for the remaining columns and if columns were highly correlated, only one was kept. Each row had a basket size of 15 items with 70 total items. A min support of 0.1 and a min lift of 1.15 were used.

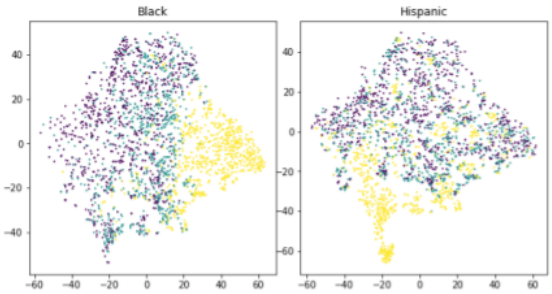
We found that if a company or advertiser is looking to market to those located between -90° and -80° longitude, Mississippi to Florida, then to reach the most people, they should advertise in places frequented by those 18+ (such as businesses) and women (such as nail salons) in that area. Also, they should target their ads to those demographics as well as those who make between $20,000 and $25,000 a year. Whereas if you are looking to market to those located just to the west, middle of Texas to Louisiana, you should market in places frequented by men (such as sports bars) and target those who make between $25,000 and $30,000 a year in that area.

Low Dimension Projection with t-SNE



(Above) Points associated with Alaska and Puerto Rico are distinctly separated from the continental US.

(Below) Points are colored by decile, with yellow being highest concentration. Black and Hispanic populations were the most evidently delineated.

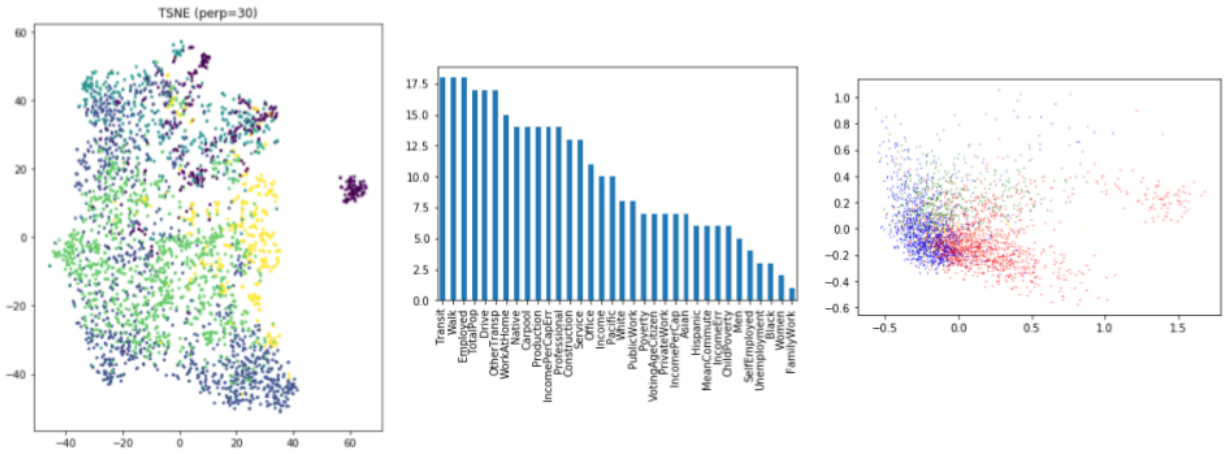


To view the data in lower dimensions while attempting to maintain pertinent information, we applied PCA and t-SNE algorithms. In doing this, we also obtain some insight into which features are most important when determining the distance between points in the lower-dimensional space. While both PCA and t-SNE separated points associated with Puerto Rico well, t-SNE provided the more cleanly visible result of the two. Additionally, t-SNE offered some strong insight when labels signifying deciles of each feature were laid over the result.

As seen in the slide, when labeled by state, the best separated clusters were those of non-continental states (Hawaii's cluster was also well separated, though there were only four points). This is what we hypothesized, since these are the most isolated territories, and gives validity to marketing and business strategies that have been specialized for these areas. Out of the 34 features, Black and Hispanic ethnicities appeared to be the strongest indicators of similarity according to t-SNE, as their deciles were the most strongly delineated. We recommend using these features as a proxy for a more complicated set of features, and prioritize collection of data on these demographics.

Failed Experiments

GMM Clustering	Ridge/Lasso	PCA

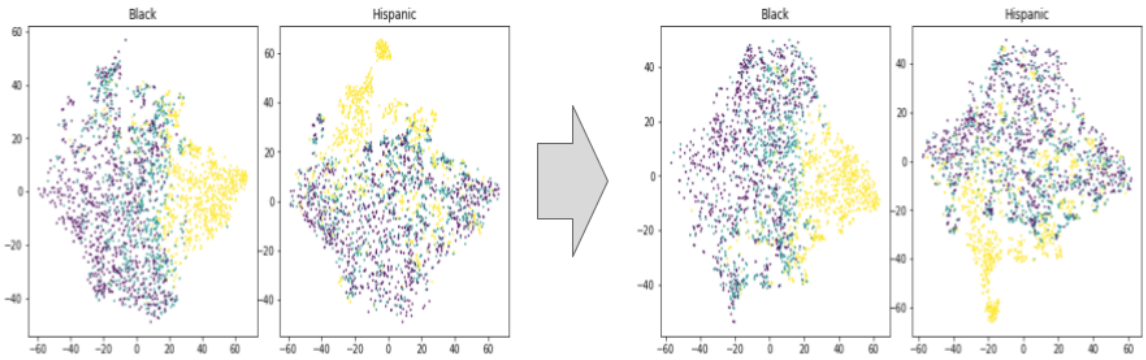


Though we have presented the best results here, there were several avenues that did not return any compelling results. First, Gaussian Mixture Models, one of the initially proposed clustering methods, resulted in bad separation between clusters. For instance, the dark blue cluster is interspersed throughout the entire t-SNE visualization shown on the left.

We attempted to use ridge/lasso regression and PCA in order to determine whether we could accurately capture most of the dataset's variance in a small subset of features. For the ridge/lasso experiment, we trained classifiers to predict each of the individual columns, and parsed the results to find the most commonly important features. The middle plot here shows that there is no small list of features that are ubiquitously important in prediction.

Finally, we performed PCA and plotted the features of max variance for each of the principal components. Each time, the principal component axes were composed of a mixture of features, rather than one or two individual features. The plot offered little clarification either, with different labels falling close together and no clear trends appearing.

Comparing 2015 vs 2017



2015 **2017**

Finally, we compared the two datasets that we used to see if there is any significant difference to be found across time. While there may have been small differences in particular counties, much of the data in the census is predicated on the movement of people - where they live and work. These forces move and change slowly, so perhaps we would not expect counties to change from similar to dissimilar so rapidly. To compare, we used t-SNE again to visualize similar counties based on feature deciles. Shown in the slide are our two most distinguished features in 2015 and 2017; clearly they did not change drastically, meaning these patterns hold over time and can be meaningful in determining a multi-year business plan.

An interesting extension to this result would be to now compare the same features from the year 2020 or 2021, after the pandemic. This unforeseen and ubiquitous phenomenon forced a vast portion of the population to change how they live and work over the course of mere weeks. People moved out of cities and began working from home much more frequently, forces that could greatly change many of our features. While counties maintained similarity from 2015 to 2017, a fundamental shift we have all seen could provide new similarities in 2020.

Works Cited

1. MuonNeutrino, "US Census Demographic Data," *Kaggle*, 03-Mar-2019. [Online]. Available: https://www.kaggle.com/muonneutrino/us-census-demographic-data?select=acs2015_county_data.csv. [Accessed: 11-Nov-2021].
2. MuonNeutrino, "US Census Demographic Data," *Kaggle*, 03-Mar-2019. [Online]. Available: https://www.kaggle.com/muonneutrino/us-census-demographic-data?select=acs2017_county_data.csv. [Accessed: 11-Nov-2021].
3. "United States Counties Database," *simplemaps*, 11-Aug-2021. [Online]. Available: https://simplemaps.com/data/us-counties. [Accessed: 11-Nov-2021].
4. "Municipalities of Puerto Rico," *Wikipedia*, 11-Nov-2021. [Online]. Available: https://en.wikipedia.org/wiki/Municipalities_of_Puerto_Rico. [Accessed: 12-Nov-2021].
5. C. Halpert, "Census-regions/US census bureau regions and divisions.csv at master · CPHALPERT/Census-Regions," *GitHub*, 23-Jun-2014. [Online]. Available: https://github.com/cphalpert/census-regions/blob/master/us%20census%20bureau%20regions%20and%20divisions.csv. [Accessed: 13-Nov-2021].

Works Cited

1. MuonNeutrino, "US Census Demographic Data," *Kaggle*, 03-Mar-2019. [Online]. Available: https://www.kaggle.com/muonneutrino/us-census-demographic-data?select=acs2015_county_data.csv. [Accessed: 28-Nov-2021].
2. MuonNeutrino, "US Census Demographic Data," *Kaggle*, 03-Mar-2019. [Online]. Available: https://www.kaggle.com/muonneutrino/us-census-demographic-data?select=acs2017_county_data.csv. [Accessed: 28-Nov-2021].
3. "United States Counties Database," *simplemaps*, 11-Aug-2021. [Online]. Available: https://simplemaps.com/data/us-counties. [Accessed: 28-Nov-2021].
4. "Municipalities of Puerto Rico," *Wikipedia*, 11-Nov-2021. [Online]. Available: https://en.wikipedia.org/wiki/Municipalities_of_Puerto_Rico. [Accessed: 28-Nov-2021].
5. C. Halpert, "Census-regions/US census bureau regions and divisions.csv at master · CPHALPERT/Census-Regions," *GitHub*, 23-Jun-2014. [Online]. Available: https://github.com/cphalpert/census-regions/blob/master/us%20census%20bureau%20regions%20and%20divisions.csv. [Accessed: 28-Nov-2021].