Charting the Islands: Clustering Hawaii Geographies

Github Notebook:

https://github.com/hajnoszian/Coursera_Capstone/blob/master/Capstone%20Final%20Project.ipynb

Introduction

Despite a need to do so, utility companies across the country have faced challenges raising rates. As they turn increasingly to green energy, encounter demographic changes in communities, and old infrastructure becomes inefficient or requiring replacement, companies find themselves needing more money for their day to day needs on top of staying modern. Not surprisingly though, few customers are excited about paying more every month for gas, water, or electricity. Such increases on monthly bills are particularly impactful for those with low incomes. This means that utilities have had to balance rate increases and infrastructure investments with equitable practices that do not overburden communities. However, they can only do so if they understand what these communities look like—which ones are going to be most hit by a rate hike? Which communities rely on different utilities? And which communities are similar to each other in both of these respects?

In a state that has big wealth differences in unique segments (e.g. literally, islands of people) Hawaii poses a unique challenge in this respect. Communities can vary widely from island to island—from the urban sprawl of Honolulu to the more rural towns that dot the Big Island—and even within the same island too. Such differences are often reflected in income and infrastructure. Different kinds of buildings and services have different utility requirements. Just as, if not more, important for the sake of making equitable rate changes, the socioeconomic status of the people in these communities must also be considered. Getting even a basic understanding of the communities requires looking at different groups of nested data. In other words, we need to look at a large, medium, and small scale.

Data

The data being used in this project comes from the Census Bureau's American Community Survey estimates (Table S1901). In particular, we are interested in three levels of geographic 'scales' (i.e. zip code, county, and state levels) and their respective median household incomes. This way we can have multiple levels to understand how communities are relative to their neighbors, their island, and their entire state. Median household income is our placeholder to broadly represent socioeconomic

status. Also of note, these Census data are from 2017. While more recent estimates exist at higher levels of data (e.g. state), 2017 is the most recent available data for points as specific as zip codes for the entire state. These data help us evaluate the general socioeconomic status of communities in Hawaii, ranging from the largest scale (state) to a fine-grain scale (zip codes). I also used unitedstateszipcodes.org to doublecheck the validity of some zip codes that did not cleanly translate into the data. Geolocation data came from Hawaii's State GIS project website.

I also used data from Foursquare to look at what general kinds of businesses and venues are prominent in these communities. This is an admittedly somewhat secondary analyses to our main question of evaluating potential communities at most financial risk of increased utility costs. However, this broader venue data adds a different dimension to our question by showing us whether certain communities are more similar to others in their venues and additionally, whether such venue data serves as a relevant indicator for these communities. Such similar venues could be suggestive of similar kinds of utility needs.

Methodology

Collection and Cleaning

The primary data was downloaded from the Census Bureau and Hawaiian State GIS Program site. The data was reduced to just the variables of interest: total households, family households, nonfamily households, median household income (MHI) and geographic labels. Some MHI data was missing because the Census Bureau does not report statistics for areas in which the variance of estimates is too large to create an accurate estimate. In other words, in geographies with very few people, calculating MHI becomes very difficult and prone to error. These geographies with very few, or no, households (n=9) were removed. An additional variable was created, % of State MHI, to more simply show each geography's MHI compared to the state as a whole. This variable was calculated by diving the local geography MHI by the state MHI. The dataframe was then separated into three parts, one for each category of geography: state (n = 1), county (n = 5), and zip code (n = 85). Additional geographic data for each county and zip code were collected from the Hawaiian State GIS Program and inserted for each geography.

Venue data was collected from Foursquare, returning up to 500 venues within a five-kilometer circle around the center of each geography. Five kilometers was deemed as a compromise between being large enough to sufficiently cover each zip code area but small enough to avoid overreaching into

neighboring zip codes. A total of 1565 venues was collected across all zip codes. All zip codes had at least one venue (*M* = 26.98, s = 34.26, min = 1.00, max = 100.00,). Venues were dummy coded for each zip code and the respective variables of interest (MHI, Family Households, Nonfamily Households, Venues) were segmented out and standardized for analyses. Refer to the Notebook for a more detailed walkthrough of the cleaning process.

Analyses

The primary analysis of this project were the cluster analyses. Because we are asking the question of what areas are similar to each other, we are really looking to see how we can 'classify' them. Clustering analyses helps us evaluate what communities can be grouped together based on multiple kinds of criteria. In particular, I used two different clustering methods: K-Nearest Neighbor (KNN) and DBSCAN. KNN is a standard, versatile cluster analyses so it serves as a tried-and-true benchmark test. It is a simple, "classic" test. On the other hand, DBSCAN offers a more complex but also potentially more powerful test because it better handles ambiguous data and outliers. For this data set, which combines both venue counts and sociodemographic data of a multi-island geography, the ability to adapt to outliers is an important perspective to consider. Therefore, this project used both KNN and DBSCAN to give a more multi-faceted picture of the data.

After running iterative amounts of KNN analyses (see Notebook), 11 was deemed an appropriate number of clusters. It was an appropriate compromise between accepting too many clusters, thus diluting the clustering, and accepting too few, thus oversimplifying the data. As the smallest number of clusters that had no more than half of the entire sample in a single group, 11 clusters was deemed parsimonious for the purposes of this descriptive project.

The DBSCAN analyses was done with the parameters of 6 for epsilon and 2 for minimum number of samples for a cluster. The optimal epsilon value was calculated after finding the point of maximum curvature on a plot of the nearest-neighbors distances (see Notebook).

Results

The presented results show both a small portion of the data used, as well as the final maps that were created from the analysis clusters. The maps are interactable, so refer to the Notebook for greater functionality and detail.

Table 1. Below are the first few rows of the dataframe that the previously described analyses was performed on. You can view the entire dataset, including the venues, and more details in the Notebook.

DBSCAN Cluster	KNN Cluster	MHI	Total Households	Family Households	Nonfamily Households	% of State MHI	City	County	Latitude	Longitude	Zip Code
0	1	91646	13080	9329	3751	1.22	Aiea	Honolulu	21.40605	-157.8849	96701
0	1	62500	594	404	190	0.83	Anahola	Kauai	22.14558	-159.3856	96703
1	7	51483	2578	1854	724	0.69	Captain Cook	Hawaii	19.33727	-155.8376	96704
-1	8	75938	761	595	166	1.01	Eleele	Kauai	21.89953	-159.5685	96705
-1	6	98248	18988	15708	3280	1.31	Ewa Beach	Honolulu	21.34476	-158.0222	96706
-1	6	99946	12889	10363	2526	1.33	Kapolei	Honolulu	21.36311	-158.0822	96707

KNN Results

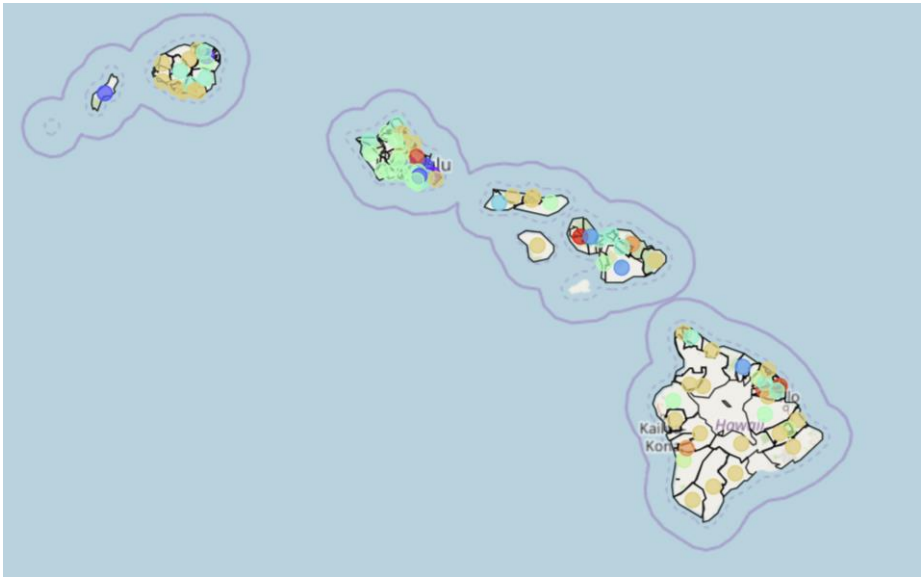

Figure 1. K-Nearest Neighbor Clusters (n = 11) marked for each zip code across the islands.

DBSCAN Results

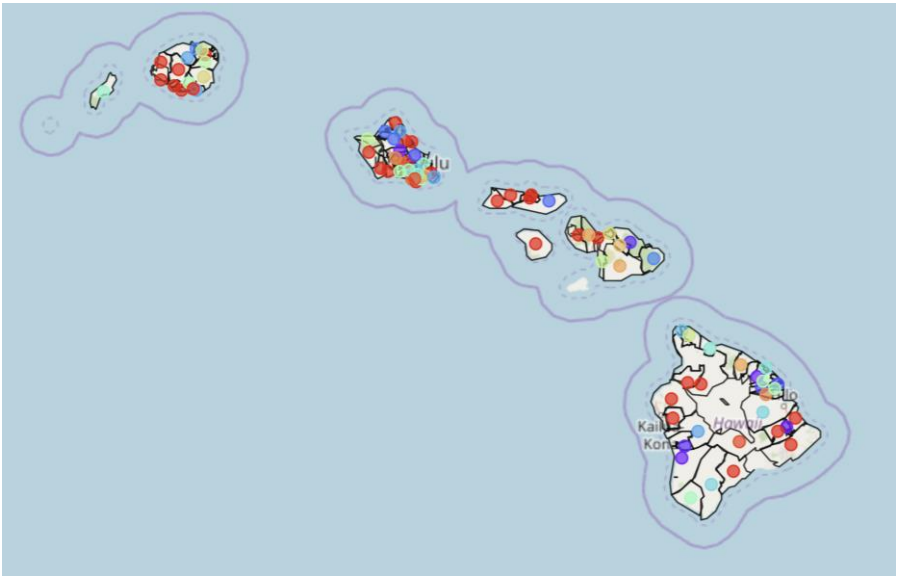

Figure 2. DBSCAN clusters (n = 22) marked for each zip code across the islands

Additionally, to compare how similar or dissimilar the clusters were between these two analyses types, I ran a Pearson correlation test between both clusters. The relationship between the KNN and DBCSCAN clusters was significant and negative ($r = -.24$, $p < 0.05$).

Discussion

Clustering analysis revealed several major differences between geographies across the state of Hawaii. Most notably, there were big differences in the communities between Oahu and the Big Island. This difference is apparent in both clustering analyses. It is the starkest example of how dense, more urban areas of the state are differentiated from its less populous, rural areas. The KNN analysis for example, shows this distinction very clearly—Cluster 6 (colored in light green) appears as an artery down Oahu's main metropolitan area around Honolulu, as well as other cities like Hilo and Kona. Meanwhile Cluster 5 (light teal) consistently marks the other smaller cities (e.g. Lihu'e and Hanalei on Kauai, or Kahului on Maui). However, the rest of the Big Island and most of Kauai are cluster 8 (light brown). While the distinctions between urban and rural is a national narrative, it is unique to see such a similar narrative even in a far smaller context—both in area and population—such as Hawaii.

The clusters from the DBSCAN however, hint at how pluralistic Hawaii is. Nearly a quarter of the zip codes are considered outlier to the rest of the data points, with no clear rule of thumb emerging what made them so different. For example, much of the outlier zip codes are areas dominated by state

parks and/or relative geographical remoteness. However, the city of Kahului on Maui is also considered an outlier. In other words, that there is such a variety of what is or is not an outlier, on top of the over 20 groups created by the DBSCAN analysis, suggests that Hawaii is incredibly varied in its people and infrastructure. Indeed, that is part of what makes density-based cluster analysis such a helpful analyses is that it is not "afraid" to outcast points that do not agree with others (whereas KNN forces all points to be in *some* group). This means DBSCAN can attend to less obvious connections between data. Hawaii evades simple broad strokes when considering its social, economic, and geographic features.

Interestingly, we do see a relationship between each clustering type. The significant negative correlation between the clusters types suggests that both do pick up on some similar main features that distinguish certain zip codes from others. In effect, the negative correlation means as, for example, KNN clusters with big numbers are similar to DBSCAN clusters with low numbers. The negative direction of the correlation likely an artifact of the order in which each analysis labeled the zip codes. What is notable, is that the correlation was non-zero—this means that there is a relationship at all and the p-value was below 0.05, making it statistically significantly so. This means that it is very unlikely that we would see expect to see this relationship if it was purely up to chance alone. In other words, there does seem to be some underlying similarities between each cluster, which is a good thing considering that they are both trying to group the same data (just in different ways).

What could be driving this main thread that the clusters are both picking up on? It is likely the socioeconomic differences between all of these geographic areas. Look at the figure below.

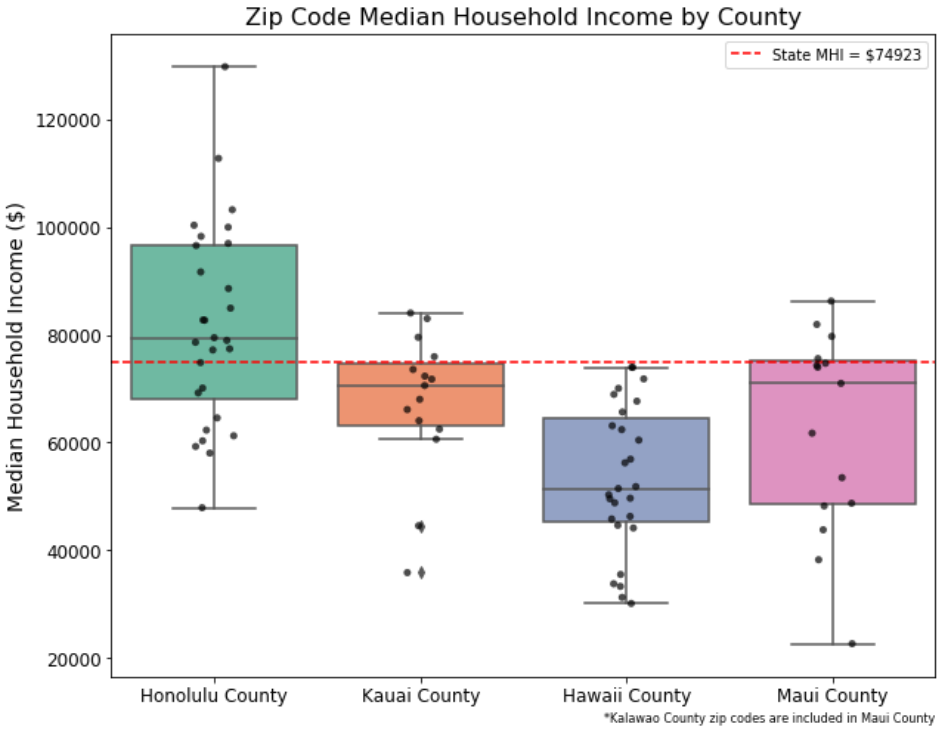

Figure 3. Boxplot representing quartiles of each county's MHI, with respective zip codes MHI nested within.

The big island is certainly distinct from the other islands. Both cluster analyses show this. However, that difference is even more pronounced when looking at the sociodemographic data (Figure 3). The big island is overall far poorer relative to the rest of the islands, with even the wealthiest zip codes *below* the state median income level. On the other hand, the weight of Oahu's larger population and wealth drives the state MHI higher than all of the other islands. Alternatively, while Kauai and Maui both have relatively the same "rich areas" and a similar overall MHI, they are not spread the same way. Kauai has relatively little gap between its wealth quartiles while Maui has a larger disparity. It is likely that the cluster analyses tracked some of these major differences between zip codes in each island, hence their somewhat similar clusters. However, this suggests that the venue data does not help in identifying certain communities relative to others in Hawaii. While further analyses should be run in the future for more confidence, the presented analysis lends credence to the notion that venue infrastructure is not a robust marker for identifying communities. Take a look at the Notebook for this interactive map (Figure 4) that shows how each county compares to each other and the MHI for each zip code within the county.

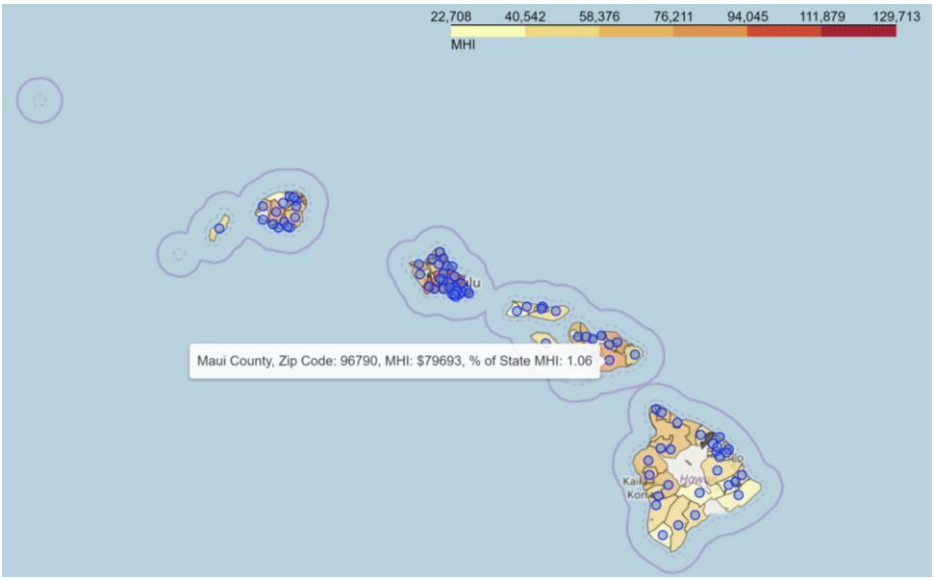

Figure 4. MHI data shaded for each zip code, with additional details on hover. Check the online Notebook for more detailed information and interaction.

Conclusion

The question of identifying similar communities in Hawaii should lean more on socioeconomic data to inform its policymaking, as those factors seem to be a more consistent identifier of communities than its venue infrastructure. Moreover, considering such diversity in its socioeconomics and geographies, policymaking must be flexible to adapt to attend to counties and cities are a more specific basis in order to a provide equitable service to residents of Hawaii.