Neighborhoods of Oahu, Hawaii
Trends and Insights for Incoming Residents

Brock Schmalzel
August 14, 2020

1 Introduction
1.1 Problem
Thousands of people move to Oahu each year from the U.S. mainland and from abroad. When apartment or house hunting, many people are presented with multiple options regarding which part of the island in which they would like to reside. Without having previously lived on Oahu, it can be daunting to narrow down the search.

To make the decision on which part of the island to live—and which neighborhood—people look to other factors beyond cost of rent and household amenities. Neighborhood crime rates and the quality of schools are often important factors for families, whereas proximity to restaurants and bars may be the top issue for single people.

This report seeks to provide useful insight about the different neighborhoods of Oahu. Foursquare City Guide—a mobile app that provides information and reviews on places and venues in given area—will be used to analyze differences between neighborhoods. Are there significant differences in the type and frequency of venues found around the island of Oahu?

1.2 Target Audience
Anyone moving to Oahu may be interested in this problem. Some incoming residents may even intend on opening a business on Oahu. Others move to Hawaii with their families and are interested in neighborhood schools and crime rates. In addition to general analysis of the island, two specific examples will be considered:

- Mary is moving to Oahu and is considering opening either a bakery or an Italian restaurant. Are there neighborhoods that favor one over the other?
- John is moving to Oahu with his family. He would like a neighborhood with nearby parks, but also low crime and good high schools.

2 Data
2.1 Data Sources
A list of Oahu neighborhoods and zip codes (postal codes) was found at https://www.kimicorrea.com/oahu-zip-codes/. The Python library *pgeocode* was used to look up latitude and longitude coordinates for each zip code.

Foursquare data was obtained with a Foursquare Developer Account and the Foursquare API.

Honolulu crime data was gathered from the Honolulu Police Department's 2018 Annual Report found at http://www.honolulupd.org/downloads/HPD2018annualreport.pdf.

The Strive HI Performance System results for 2017-2018 school year was the source for data on Oahu public schools. This data was released by the Hawaii State Department of Education and is available for download as an Excel sheet at http://www.hawaiipublicschools.org/VisionForSuccess/AdvancingEducation/StriveHIPerformanceSystem/Pages/2017-18-results.aspx

2.2 Data Cleaning and Processing

A few issues were encountered when gathering the neighborhood, zip code, and latitude/longitude data. Many of the Oahu zip codes did not have a unique neighborhood associated with them, which required the use of the zip codes themselves as the unique identifier. Furthermore, many of the latitude/longitude coordinates returned by *pgeocode* were incorrect or redundant. Zip codes with redundant latitude/longitude coordinates (all within the Honolulu neighborhood) were dropped from the dataframe, while incorrect latitude/longitude coordinates (associated with military bases on Oahu) were manually corrected.

For some of the latitude/longitude points on Oahu, Foursquare returned no venues within the designated radius. These "na" values were dropped from the dataframe.

The Honolulu police department does not report crime statistics based on zip code. Instead, crime is reported based on eight crime districts and their subdistricts. A map of the crime districts was used to determine in which district each zip code resides. To provide a metric of crime, the total number of crimes that occurred within a district was assigned to each zip code in that district.

The Strive HI Performance System data included test scores, graduation rates, and other metrics for K-12 public and charter schools. To simplify this analysis, only public high schools were included. To provide a single metric of school performance for a given zip code, a school's math, science, and language scores were totaled together to create a single "test score." If multiple high schools shared a zip code, the test scores were averaged.

Zip codes were used to join the Foursquare, crime, and school data into a single table for analysis.

3 Methodology

3.1 K-Means Clustering for all Zip Codes

To look at differences between Oahu neighborhoods, Foursquare was used to query venues within a given radius of a zip code. The results of the query were grouped by zip code and the top ten most common venues were calculated (mean of the frequency of occurrence). A sample of the top ten venues are shown in Table 1.

Table 1 – Most Common Venues by Zip Code as sourced from Foursquare

Zip Code	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common	…
96701	Mobile Phone Shop	Burger Joint	Golf Driving Range	Bank	Jewelry Store	Japanese Restaurant	
96706	Home Service	Park	Clothing Store	Golf Course	Restaurant	Grocery Store	
96707	Video Store	Coffee Shop	Supermarket	Business Service	Doctor's Office	Dog Run	
96709	Home Service	Fast Food Restaurant	Food Truck	Restaurant	Golf Course	Discount Store	
96712	Beach	Surf Spot	Construction & Landscaping	Yoga Studio	Dive Shop	Flower Shop	

K-means clustering was applied to the venue data to identify trends among the various zip codes. The clustering algorithm was tuned by varying the Foursquare query radius (250 – 1000 m) and the number of clusters (4-6). As the goal was to find distinct differences between zip code clusters, the algorithm was tuned to increase the number of zip codes in the smaller clusters. In all cases, a single dominant cluster contained most zip codes; this is discussed further in the Results section.

Once clustered, the most common venues and—more importantly—the venues not found in other clusters would be identified. For large clusters, a histogram of the most frequent venues would be created to help identify trends not apparent at first glance.

3.2 K-Means Clustering with Foursquare, Crime, and School Data

The methodology for this analysis is very similar to that of K-Means Clustering for all Zip Codes. In addition to the frequency of venues gathered from Foursquare, the K-means clustering algorithm also used normalized public high school test scores and normalized total number of district crimes. As not every Oahu zip code includes a public high school, this data was only a subset of that used in the previous section. A sample of the Foursquare, crime, and school data is shown in Table 2.

Table 2 – Most Common Venues, High School Scores (sum proficiency), and Total Crimes by Zip Code

Zip	Sum Proficiency	School(s)	Total Crimes	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	…
96701	0.068594	Aiea High	0.245348	Mobile Phone Shop	Burger Joint	Golf Driving Range	Bank	Jewelry Store	
96706	0.351234	Campbell High	0.407478	Home Service	Park	Clothing Store	Golf Course	Restaurant	
96707	-1.124775	Kapolei High	0.407478	Video Store	Coffee Shop	Supermarket	Business Service	Doctor's Office	
96731	-0.528091	Kahuku H&I	-0.918214	Airport	Trail	Yoga Studio	Dive Bar	Flower Shop	
96734	0.304127	Kailua High, Kalaheo High	-0.918214	Beach	Construction & Landscaping	Food Truck	Tennis Court	Yoga Studio	

3.3 Relationship between Crime and School Performance

Pearson's Correlation Coefficient was calculated to analyze the potential correlation between crime and school performance. Furthermore, the ordinary least squares method was used to perform multiple linear regression to see how well crimes-per-population and wealth index could predict school performance.

3.4 K-Means Clustering with Crime and Wealth Index Data

The last portion of the analysis considered the case in which an incoming Oahu resident had no interest in nearby venues. In this case, K-means clustering was accomplished using only the zip codes' crime and wealth index data.

4 Results

4.1 Foursquare Clustering Results

The K-Means cluster algorithm was tuned to have six clusters using a Foursquare query radius of 1000 meters. The result of algorithm is shown in **Error! Reference source not found.**. A histogram of the top 30 venues discovered by Foursquare is shown in Figure 2.

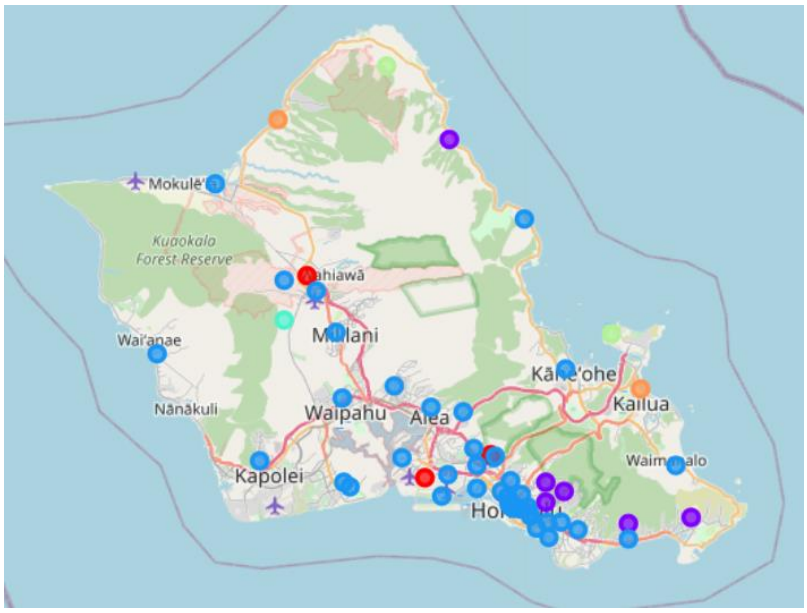

Figure 1 – Six cluster K-Means Result for Oahu Venues (1000 m radius query)

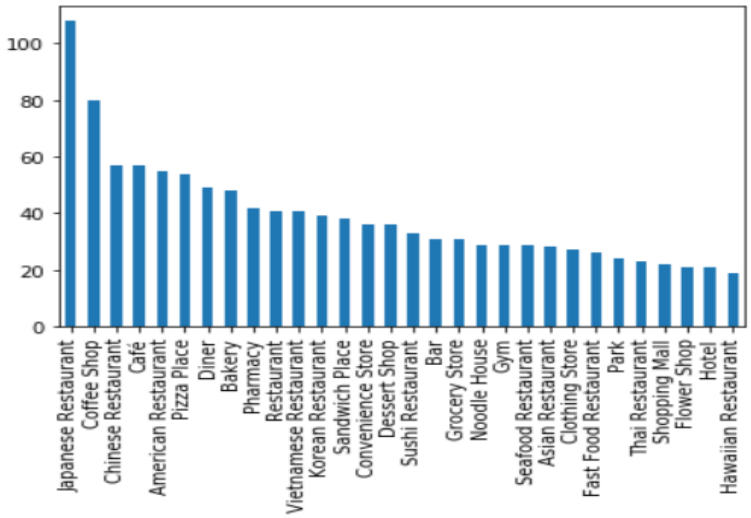

Figure 2 – Histogram of top 30 venues discovered by Foursquare

The largest cluster—the blue cluster in **Error! Reference source not found.**—contained 42 of the 56 zip codes. Many of the downtown zip codes (associated with the island's largest population center of Honolulu) belonged to this cluster. As shown in Figure 3, this cluster was dominated by eateries. Especially popular was Asian cuisine (Japanese, Chinese, and Korean Restaurants) and speedy options (Coffee Shops, Cafes, Convenience Stores, and Fast Food). Parks were also prevalent in this cluster. By and large this cluster was in line with the top venues seen in Figure 2 (as would be expected for the dominant cluster).

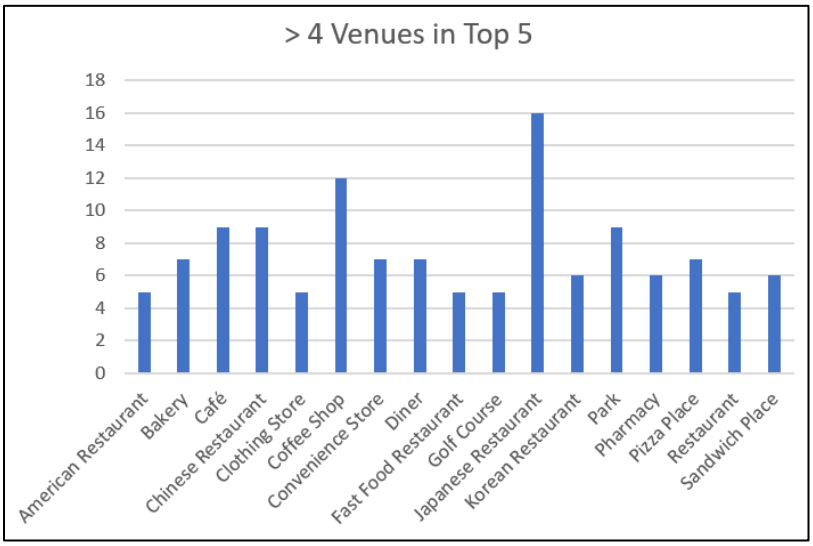

Figure 3 – Venues that appeared more than four times in the largest cluster's top 5 most common venues

To identify unique venues, the smaller clusters are compared to the venues present in Figure 2 and Figure 3. Table 3 lists the size of the small clusters as well as their unique venues (venues that were less common in the dominant cluster).

Table 3 – Unique Venues of smaller Zip Code Clusters

Cluster	Number of Zip Codes	Unique Venues
Red	3	Pizza Place, Golf Course, Bowling Alley
Purple	6	Trail, Scenic Lookout, "Food"
Cyan	1	Post Office, Doctor's Office
Orange	2	Beach
Green	2	Airport, Yoga Studio, Donut Shop

A selection of additional clustering attempts are shown in Figure 4. Only minor differences are seen between the different clustering results; changing the query radius and number of clusters did little to change how large the dominant cluster was. Although a small query radius sometimes increased the size of the smaller clusters, it caused low population density zip codes to be dropped from the analysis. For this reason, a larger query radius was chosen (1000 m radius had slightly better diversity in clusters than the 750 m radius). Likewise, changing the number of clusters had little effect on the size of the main cluster but 6 clusters resulted in a most diverse cluster size.

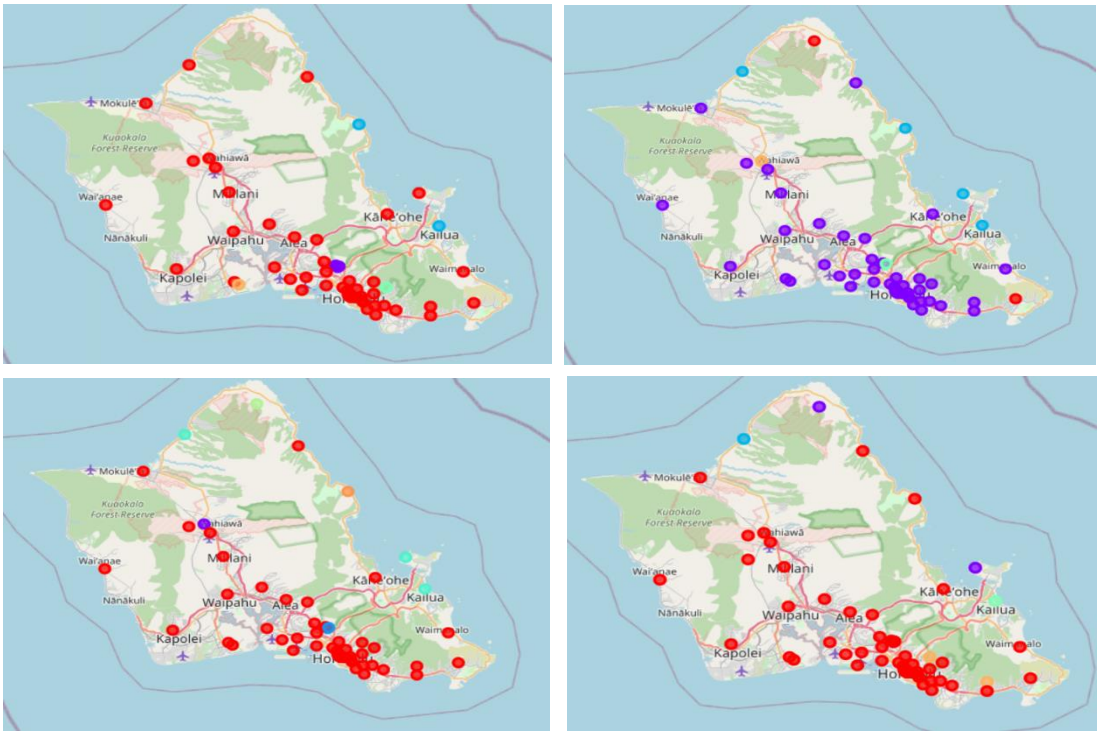

Figure 4 - Clockwise from top left: 500 m, 5 clusters; 750 m, 5 clusters; 1000 m 5 clusters; 750 m 6 clusters

4.2 Foursquare, Crime, and School Clustering Results

The K-Means cluster algorithm was tuned to have five clusters using a Foursquare query radius of 1000 meters. The result of algorithm including crime and public high school data is shown in Figure 5. Unlike the venues-only result, here there was no dominant cluster.

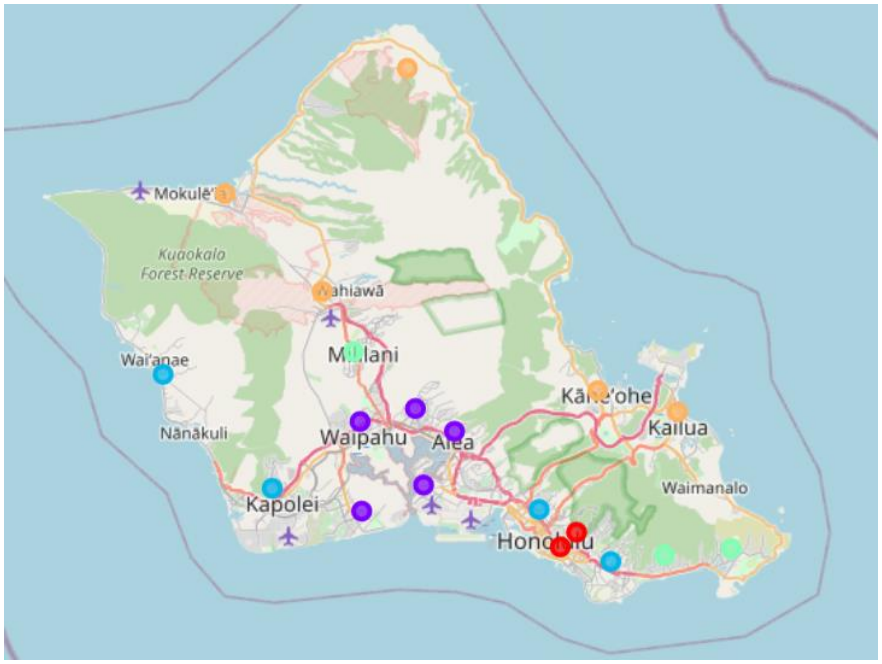

Figure 5 – Five Cluster K-Means Result for Oahu Venues, Total Crimes, and Public High School Scores

Common venues for each cluster, as well as trends in school scores and crime are provided in Table 4. Venues were considered common if they appeared in most zip codes within a cluster. If a given cluster's zip codes shared a lower-than-average, average, or higher-than-average school score or amount of crime, that was considered a trend.

Table 4 – Common Venues and School Score/Crime Trends by Cluster

Cluster	Number of Zip Codes	Common Venues	School Scores	Crime
Red	2	Cafe	-	Very Higher
Purple	5	Gold Course, Park	Higher	Higher
Cyan	3	Park, "Food', Trail	Higher	Lower
Orange	5	Farmer's Market, Café, Yoga Studio, Electronics	Lower	Lower
Blue	4	Coffee Shop, Pizza Shop, Diner	Very Lower	Higher

4.3 Crime and School Performance Results

The relationship between a zip code's number of crimes (determined by the total number of crimes in that zip code's crime district) and its high school(s)' average test scores was considered first. A Pearson R coefficient of -0.0187 with a p-value of 0.939 indicates there was no significant correlation.

What if population and wealth were accounted for? A multiple linear regression model (using ordinary least squares) with the average test scores as a function of a zip code's wealth index and crimes per population yielded a model score of only 0.244.

Finally, the relationship between a zip code's wealth index and its high school(s)' average test scores was considered. A Pearson R coefficient of 0.494 with a p-value of 0.032 indicates there was a significant positive correlation between the two.

4.4 Crime and Wealth Index Clustering Results

The final K-Means cluster algorithm sorted the zip codes with high schools into five clusters using crime and wealth index data. Figure 6 displays the result of this clustering.

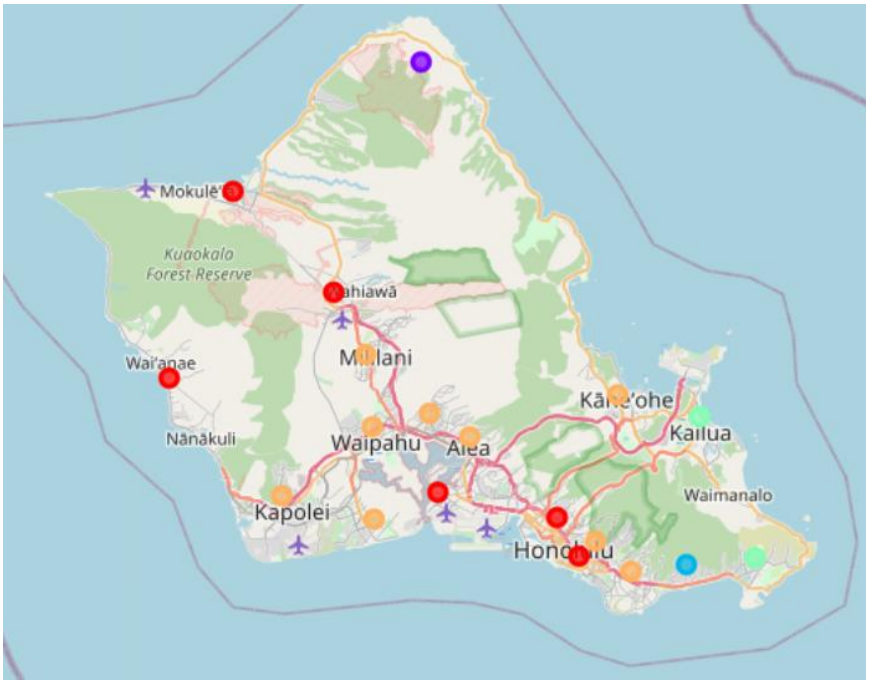

Figure 6 - Five Cluster K-Means Result for Crimes per Population and Wealth Index

Trends in wealth and crime are provided in Table 5. If a given cluster's zip codes shared a lower-than-average, average, or higher-than-average wealth index or amount of crime, that was considered a trend.

Table 5 – Wealth and Crime Trends by Cluster

Cluster	Number of Zip Codes	Wealth Index	Crimes Per Population
Red	6	Lower	Mixed
Blue	1	Very Higher	Higher
Purple	1	Lower	Much Higher
Green	2	Higher	Lower
Orange	9	Mixed	Lower

4.5 Example 1: Mary

Mary is moving to Oahu and is considering opening either a bakery or an Italian restaurant. Are there neighborhoods that favor one over the other?

As seen in Section 4.1, Italian restaurants were not common on Oahu (although pizza places were common) but bakeries were. If Mary wished to open a bakery in neighborhoods similar to those in which bakeries were plentiful, she should look at zip codes belonging to the blue cluster in Figure 1.

4.6 Example 2: John

John is moving to Oahu with his family. He would like a neighborhood with nearby parks, but also low crime and good high schools.

As seen in Section 4.2, the clustering algorithm did in fact generate a cluster of zip codes with parks, lower-than-average crime, and higher-than-average high school scores. John should try and move his

family into one of the cyan zip codes shown in Figure 5. Trails—as well as parks—were common in those zip codes. Would John be unable to find a house in those neighborhoods, he should next look to purple cluster (parks, good schools, but higher crime). If John were less concerned about proximity to parks, he could consider the green clusters shown in Figure 6 where crime was lower-than-average and the wealth index was higher-than-average; there is a significant positive relationship between a higher wealth index and higher-than-average high school test scores.

5 Discussion

5.1 Observations

Over the course of this project, several observations were made concerning the nuances of the data and applicability of the results.

- **Unequal Zip Codes.** Due to different population density across the island of Oahu, it was difficult to find the most effective query radius for Foursquare. If the query radius were too small, Foursquare would return little-to-no venues for rural zip codes. If the query radius were too large, the search areas would overlap in closely spaced urban zip codes
- **Generic Venues.** Foursquare returned generic venues such as "Food" and "Restaurants". This likely worsened the performance of the clustering algorithm
- **Crime, Schools, and Wealth.** It is hypothesized that these societal metrics are related to each other. If the target audience is interested in one (such as the "John" example in Section 4.6), they should consider the others to get the full picture. However, the analysis in this report was rather cursory and likely missed important factors
- **Identifying Venue Trends.** There was some difficulty in identifying venue trends in one cluster compared to another. The use of histograms for large clusters was helpful, but the analysis lacked a consistent strategy. For instance, the method used in to generate the common trends in Table 4 required that a majority (>50%) of zip codes in a cluster contained the same venue. However, there was no consideration for the venue being the 1st most common in one cluster and 10th most common in another
- **Dominant Venue Cluster.** The Section 4.1 results found a large, dominant cluster that contained most zip codes on Oahu. These zip codes were consistent with the greater Honolulu downtown area (the most populous part of the island) and could indicate a distinct blend of venues compared to the rest of the island. Conversely, it could be the case that there was overlap in the search query areas as mentioned in **Unequal Zip Codes**. This overlap could have caused individual venues to be counted multiple times, artificially increasing the frequency of that type of venue
- **Successful Business Examples.** The results—namely the K-means clustering algorithms—provided useful insights for the two given examples of incoming Oahu residents
- **Questionable Venue-School-Crime Clustering**. It was unclear how much impact the venues had on the clusters compared to the school and crime data
- **Missing Factors in House-hunting**. It is very likely the case that extraneous factors are at play when incoming residents are looking for a place to live. Residents may consider their commute time to work to be very important. Furthermore, certain parts of the island have more housing opportunities at any given point and time while others have higher-than-average property values that may restrict lower-income residents

5.2 Recommendations

The author of this study makes the following recommendations for future work:

- Adjust the Foursquare query radius according to the population density of target area
- Drop or classify Foursquare data that is generic e.g. “restaurant”
- Consolidate Foursquare data into larger categories. There may be value gained by creating categories such as Asian cuisine (Japanese, Chinese, Korean, etc), outdoor activities (trail, beach, park, etc), or white-collar (event planning, doctor’s office, etc
- Develop a consistent technique for identifying venue trends within a cluster
- Two recommendations related to improving the data behind the school-crime-wealth analysis
    - More comprehensive assessment of schools beyond test scores such as attendance, graduation, and college acceptance rates
    - Break down the amount of crime by zip code rather than lumping them into a single crime district

6 Conclusion

This study attempted to provide useful insight about the different neighborhoods of Oahu to incoming residents. Moving to a new locale can be daunting, and incoming residents may be overwhelmed by the different neighborhoods available to them. When a rental unit costs the same in Pearl City as another unit in Hawaii-Kai, what other factors can be considered?

The K-means cluster algorithms used in this study found that distinct trends could be found between neighborhoods on Oahu. Pearl City belongs to the cluster known for its diverse restaurant options and parks, whereas Hawaii-kai’s cluster is set apart by its trails and scenic lookouts. When looking at crime, school performance, and the wealth index, differences between neighborhoods were also apparent. While Pearl City and Hawaii-Kai both had lower-than-average crimes per population, Hawaii-Kai had a notably higher wealth index.

However, it is important to realize that the differences between clusters were not always clear. When considering other extraneous factors—such as the desire for a short work commute or proximity to friends and family—the results of this study may be of limited value to some incoming residents. Much work remains to improve the clarity and utility of the approach laid out in this report, but initial results are promising.