Market Segmentation of Beer Focused Venues

A clustering approach encompassing most populous US neighborhoods

Victor Jacobsen

.

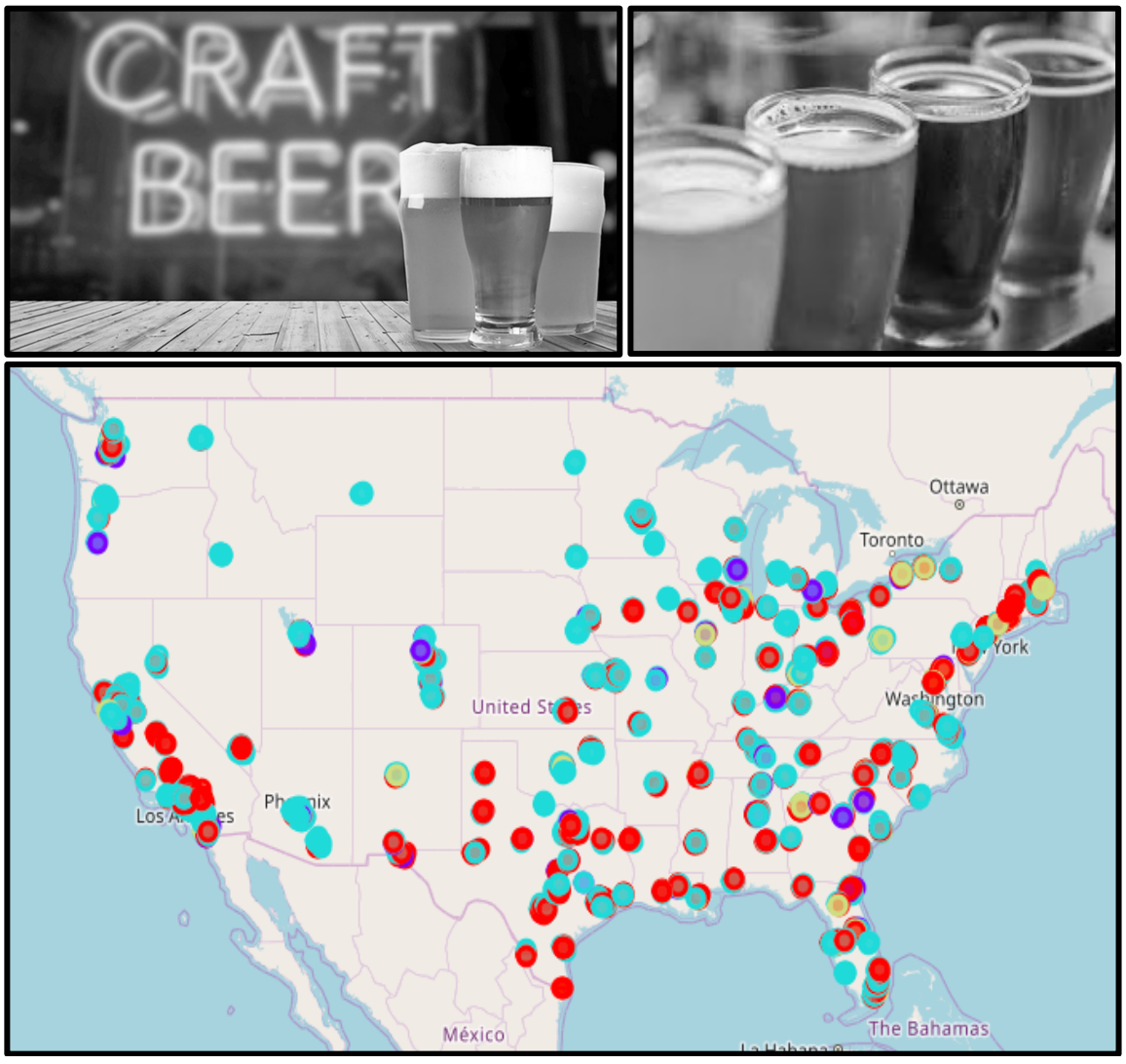


Table of Contents

Table of Figures

1. Introduction

Taprooms, beer bars, breweries and other similar themed venues have become a popular hangout spot in the US during the past years (1). Understanding the demographical commonalities around these beer focused venues can unlock key insights for breweries and venue owners around the US. These insights can be used for marketing campaign targeting, defining customer profiles or even guiding decisions of where to establish a new business.

This project will look at defining the key differences and commonalities across top US markets, based on the prevalence of beer focused entertainment venues. The focus of the project is to establish a distinction between neighborhoods within the most populous US cities, based on their demographical profile and beer venue occurrence at neighborhood level. A few potential questions this project hopes to answer are:

- Which age group is more prevalent in markets with a high ratio of beer venues?
- Which markets have a high prevalence of beer venues?
- What are potential markets for opening beer venues?

2. Data

Two main sources of data are used in this project. The first is data regarding prevalence of beer venues in the US, Foursquare API (2) will be used for this. Second, demographical data is needed to build a profile of the US neighborhoods. The source of this data will be the official United States Census API (3). Key data collected from these sources is presented in the table below:

Data Needed	**Source**
Total Population	Census
Age	Census
Sex	Census
Venue Data – Beer category	Foursquare
Educational Attainment	Census
Income level	Census
Other demographical data	Census
Zip Code geographical coordinates	

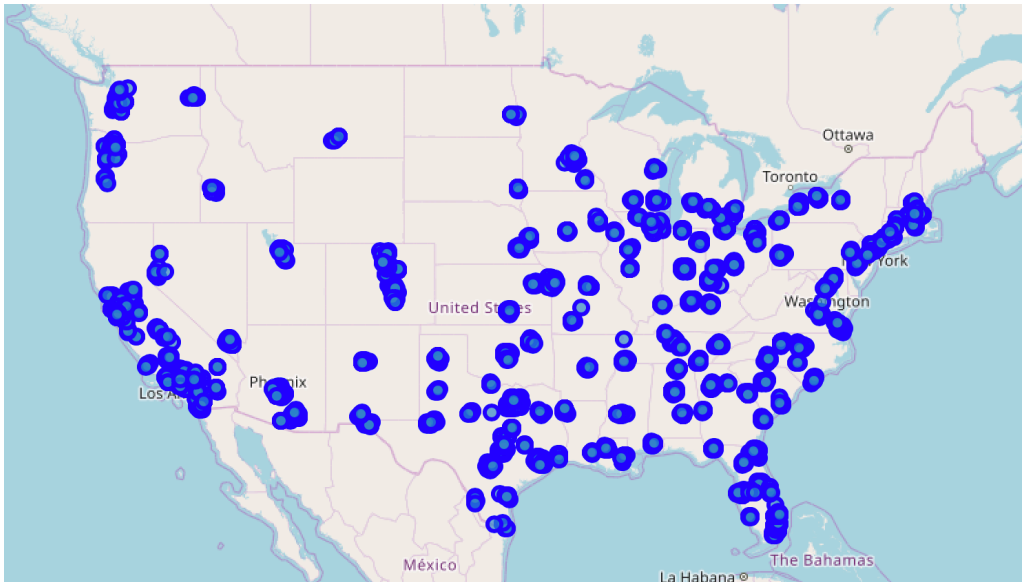
It is important to note all data needed at zip code tabulation area level. Also, percent estimates relative to total population/number of observations for a given zip code are always preferred, as they will provide a means to compare neighborhoods with different population sizes.

3. Methodology

Preprocessing

The data will be narrowed down to neighborhoods in the US cities with more than 100000 inhabitants. To define neighborhoods, we'll use zip code tabulation areas from the Census.

The below picture shows a view of all the distinct zips in mainland US considered in this project:



Foursquare API is used to gather venue data around each neighborhood with a radius of 1000m. Due to the limit of 5000 calls per hour, the extraction was done in two parts.

The Census API is used to extract demographical data from American Community Survey and County Business Pattern. Due to the size and complexity of the API database, a flexible code was written to allow for easy variable exploration and selection.

Demographical Data Used
INCOME AND BENEFITS (IN 2018 INFLATION-ADJUSTED DOLLARS)
COMMUTING TO WORK
EMPLOYMENT STATUS
GROSS RENT
SEX AND AGE
RACE
HISPANIC OR LATINO AND RACE
PLACE OF BIRTH
MARITAL STATUS
EDUCATIONAL ATTAINMENT

When extracting Foursquare data, postal codes were used. However, these differ from the Census data ZIP Code Tabulation Areas (ZCTAs). Which the Census Bureau defines as “generalized areal representations of United States Postal Service (USPS) ZIP Code service areas.” (4). Because of this, the original list of zip codes extracted from Foursquare was filtered to ZIP codes matching ZCTAs returned in the Census API call.

A total of 3544 ZIP Codes were present in Foursquare data for US cities with more than 100k inhabitants. 33120 ZATAs were returned for the American Community Survey data and County Business Pattern, when querying the API with “*” wildcard to return all ZCTAs.

A concern when first looking at the data, is that two zip codes might be under 1000 meters of each other, resulting in overlapping venue data among them. An initial thought of removing duplicates. The duplicates were ultimately not removed, since we're classifying the data that's **near** each neighborhood and easily accessible by its inhabitants. Removing duplicate venues would hide key information about closest venues around closely located neighborhoods. Out of 171176 venues, 25217 were duplicated.

The Foursquare venue data categories have multiple values for beer centered venues. The below table shows their counts across all zips.

Venue Category	Number of Venues
Brewery	1221
Beer Garden	260
Beer Bar	208
Beer Store	124

All of these categories were replaced with a single master category: "Beer Venue".

Feature Engineering

Out of the features returned in the demographical and venue data, the below table summarizes new features created from the raw data.

Feature	Description
Venue Count	Count of Venue Category per Zip
Venue Frequency – Category Relative	Venue Category percent share of Total Venues per Zip
Beer Venues per 1000 Inhabitants	*Number of beer venues per 1000 people. Venue Count x 1000 / Total Population*
Population – Business Ratio	Number of Inhabitants per Business Establishment. This is used to determine if a neighborhood has a higher or lower business activity, relative to its residential population. Higher values indicate the neighborhood is predominantly business focused.

The full feature table can be seen in the appendix.

When plotting the percent of zip codes against the Beer Venue frequency, we can see that 74% of zip codes have 0 Beer Venues. 887 zip codes have more than one Beer Venue.

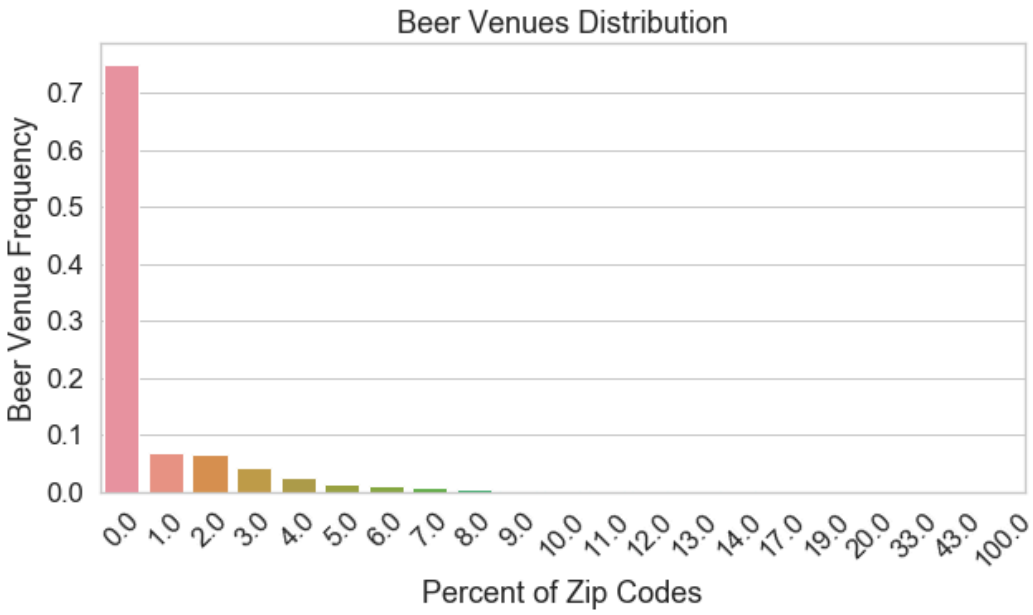


Figure 1 - Total Beer Venue Distribution

Clustering

Once the features were ready, they were scaled using SciKit Learn's Standard Scaler. This will make sure all features are within the same scale by removing mean and scaling to unit variance.

The K-Means clustering algorithm is applied to the demographical data in combination with beer venues in hopes of identifying distinct neighborhoods. The "Elbow Method" (5) was used to determine the number of clusters. This method looks at the clustering score improvement in respect to the number of clusters selected. When the marginal gain from one cluster to the next is noticeably reduced, the current cluster is considered optimal.

4. Results

Looking at the plot of Inertia by number of clusters, the curve visually changes inclination at 5 clusters. The package *KElbowVisualizer* (6) assisted with identifying the inflection point from the distortion parameter. K = 3 could potentially be used, with easier interpretability overall but would be detrimental to the analysis.

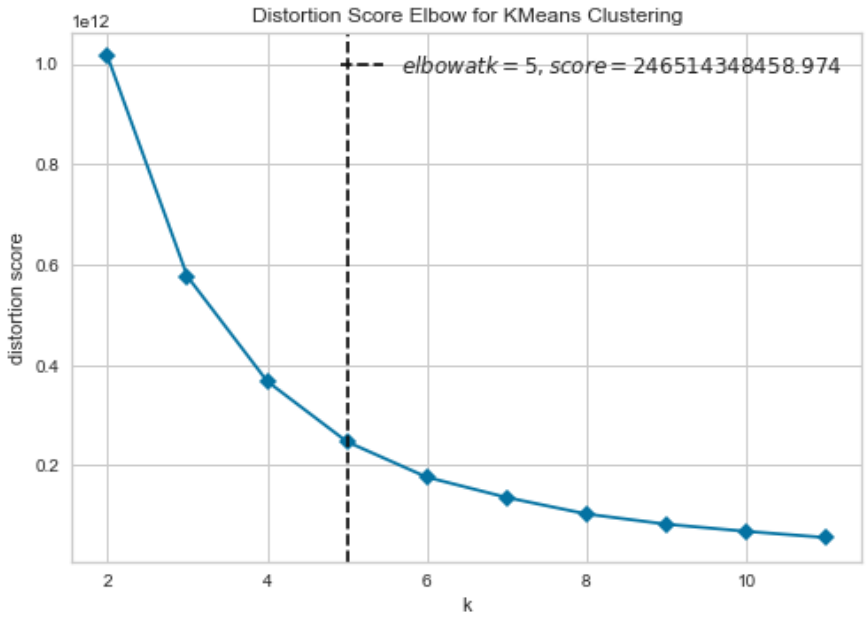


Figure 2 - Elbow Plot

The table below shows the number of ZCTAs per cluster. Cluster 0 has the highest number of ZIPs, while cluster 4 has the lowest at 129. The distribution breakdown of data under each cluster will be looked at next.

cluster	Total Zips
0	1623
1	589
2	799
3	404
4	129

The cluster map in Figure 3 and Figure 4 shows the results of the clusters in the US, a view of South Florida and New York City.

Legend:

Red – Cluster 0 **Purple – Cluster 1** **Blue – Cluster 2** **Green – Cluster 3** **Orange – Cluster 4**

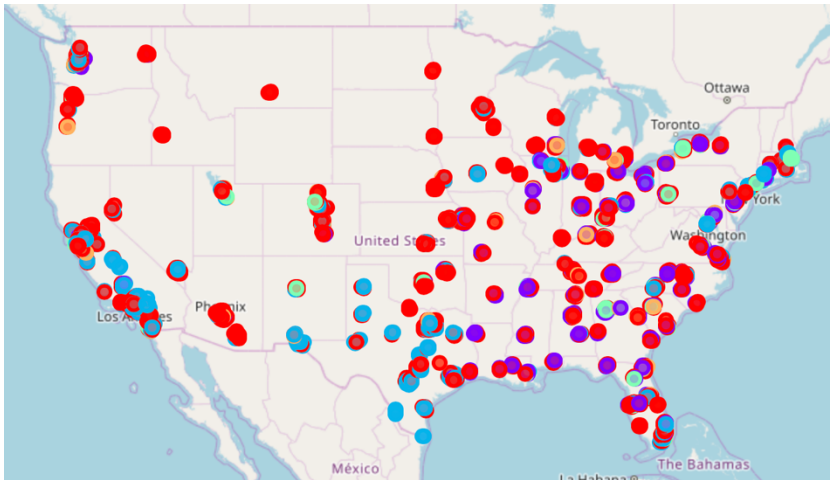


Figure 3 - Clustering Results

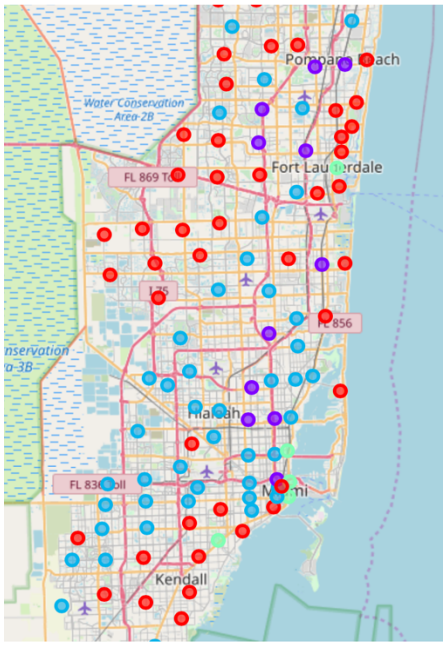


Figure 4 - South Florida Clusters

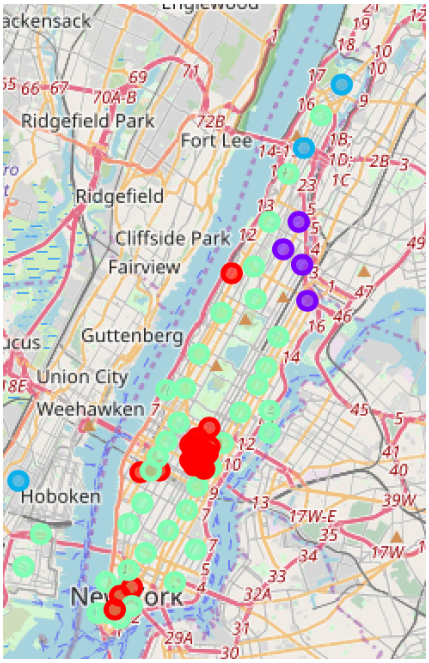


Figure 5 - New York City

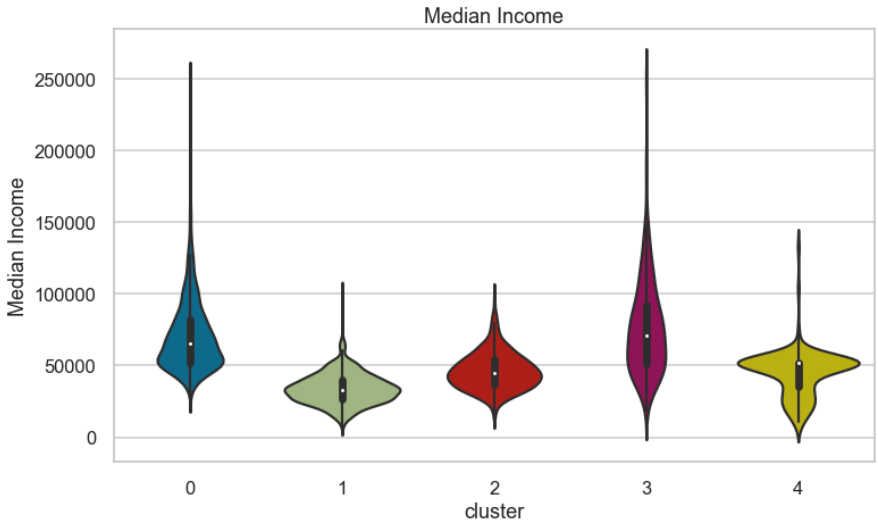


Figure 6 - Income Distribution by Cluster

Cluster 1 is made up lower incomes, cluster 2 can be defined as middle income. While cluster 3 and 0 are high income. Cluster 4 has a peculiar distribution consisting of mainly middle income level, with some lower income neighborhoods. The graph below shows a similar view of cluster mean median income.

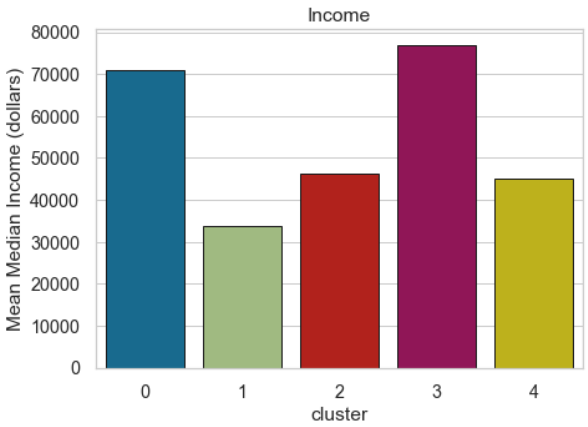
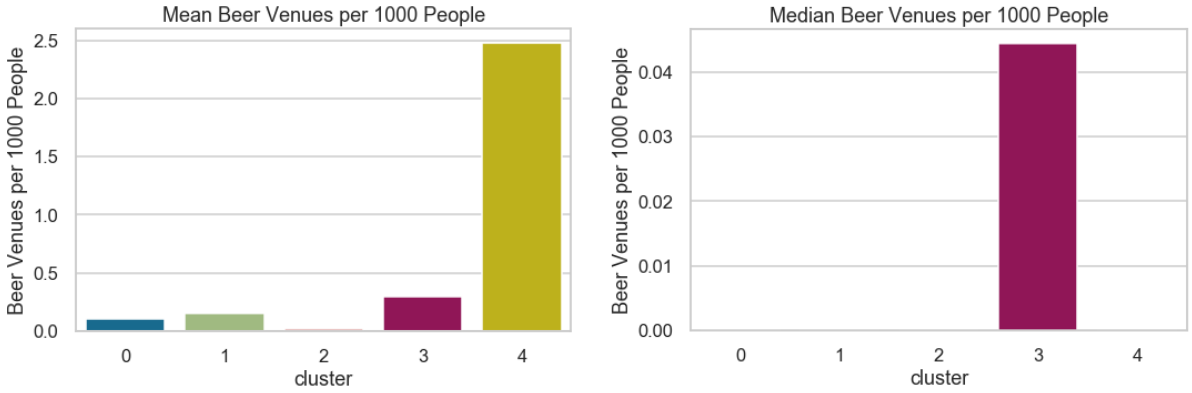


Figure 7 - Mean Median Income by Cluster

Looking at the venue specific features, cluster 4 and 3 have higher numbers of venues per 1000 people. It's worth noting, however, grouping by mean can be deceiving. When the median is also analyzed, only cluster 3 has noticeably higher rate of beer venues than others.



The violin plot in Figure 8, with high outliers (values > 0.5) filtered out shows that clusters 3 and 4 indeed have a higher number of beer venues per 1000 people at lower rates (less than 0.1)

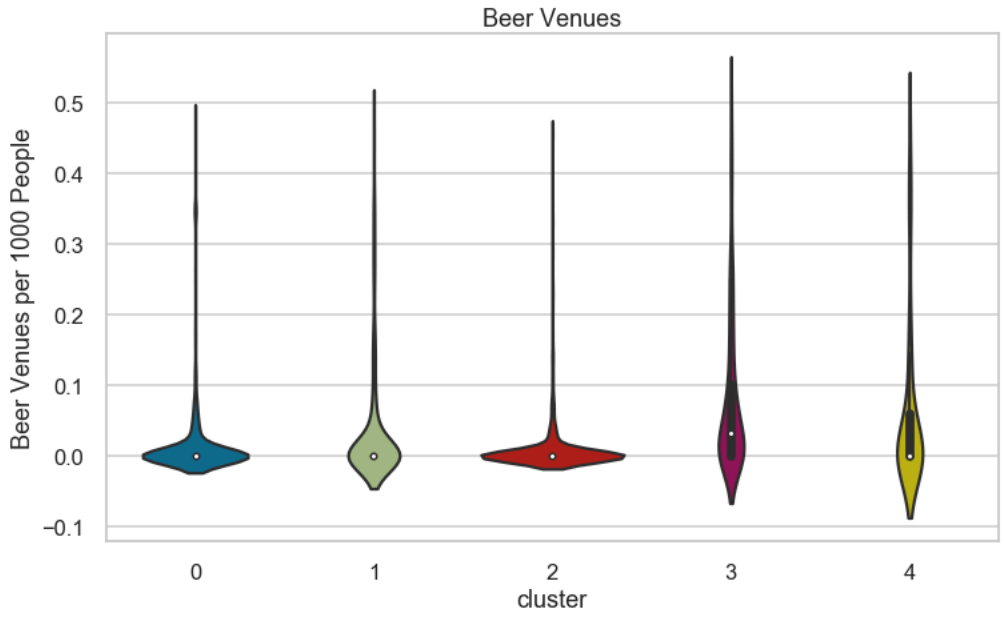


Figure 8 - Distribution of Beer Venues per 1000 People

When looking at the proportion of beer venues to total number of venues, clusters 3 and 4 also show higher mean beer venue frequency in their neighborhoods. Median frequencies for ZCTAs only show cluster 3 as having median neighborhood frequency higher than 0 (Figure 9).

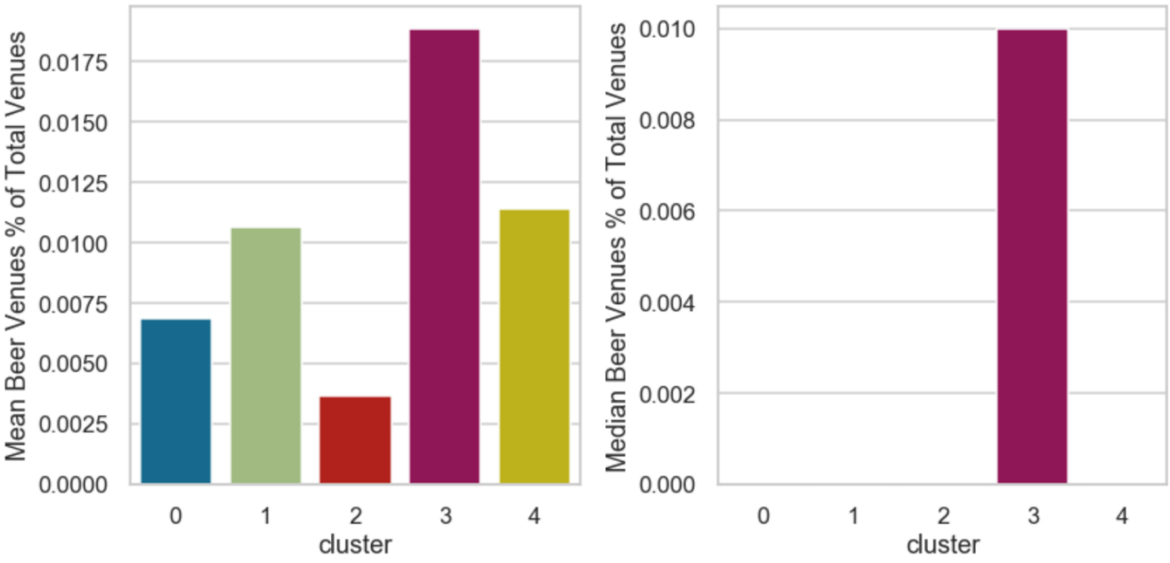


Figure 9 - Median and Mean Beer Venue Frequencies

Higher focus will be given to understand the demographical data of population in clusters 3 and 4, since they have higher number of beer venues. Continuing the analysis by looking at race, clusters 3 and 4 have predominantly White population. Cluster 2 has a higher percentage of Hispanics/Latinos and cluster 1 has a higher percentage of Black or African Americans (Figure 10).

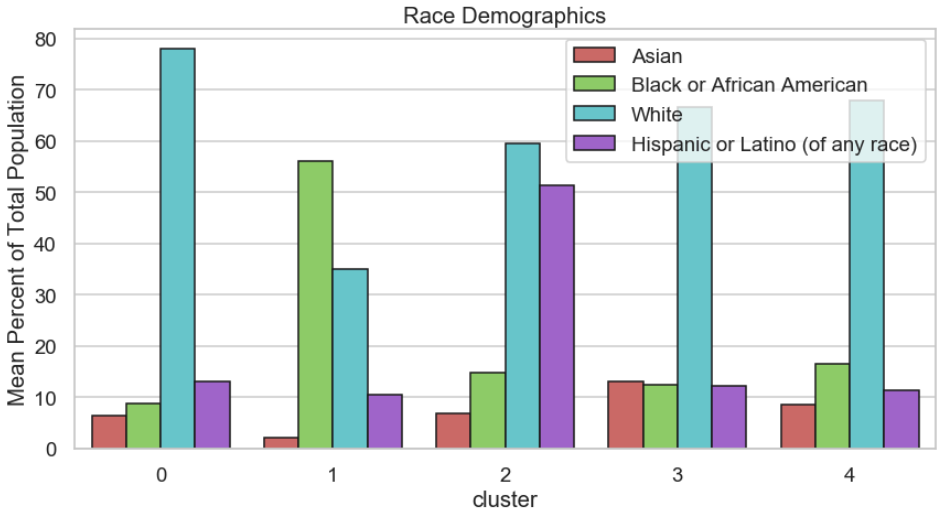


Figure 10 - Race Demographics

In terms of age, clusters 3 and 4 have a younger population than other clusters, with a large portion of population between 15 and 34 years (Figure 11).

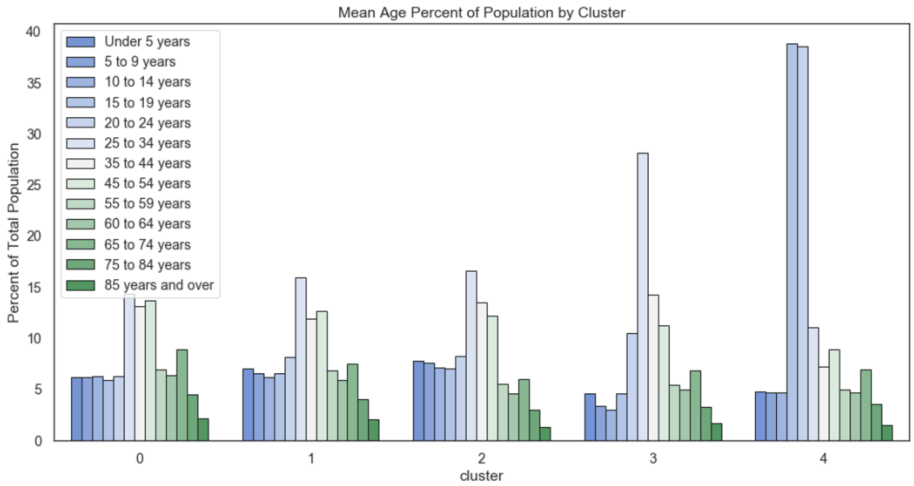


Figure 11 - Mean Age Share of Population

Cluster 4 has a higher rate of population that was never married, most likely due to age distribution (Figure 12).

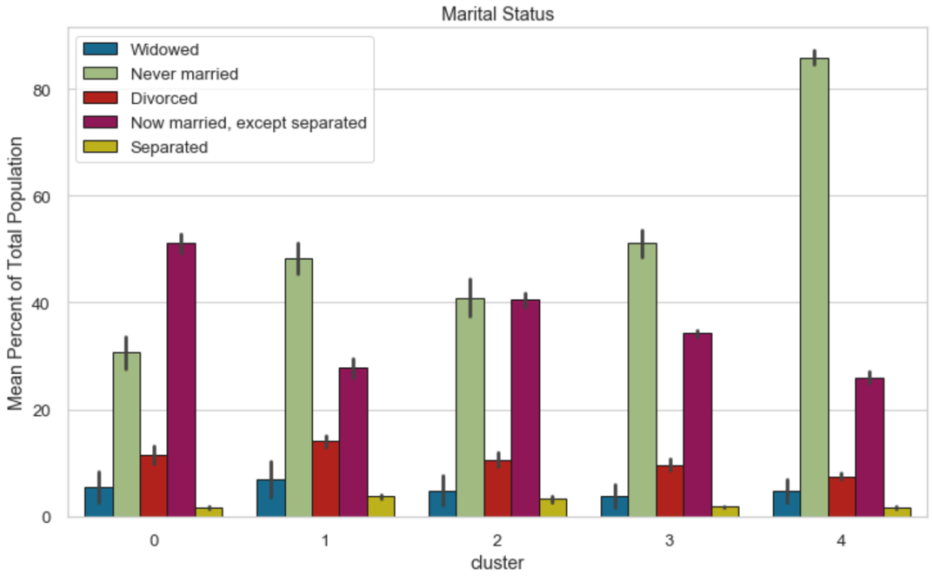


Figure 12 - Marital Status by Cluster

Figure 13 shows educational attainment is higher in clusters 3 and 4, with cluster 3 having on average more than 60% of the population with bachelor's degree or higher.

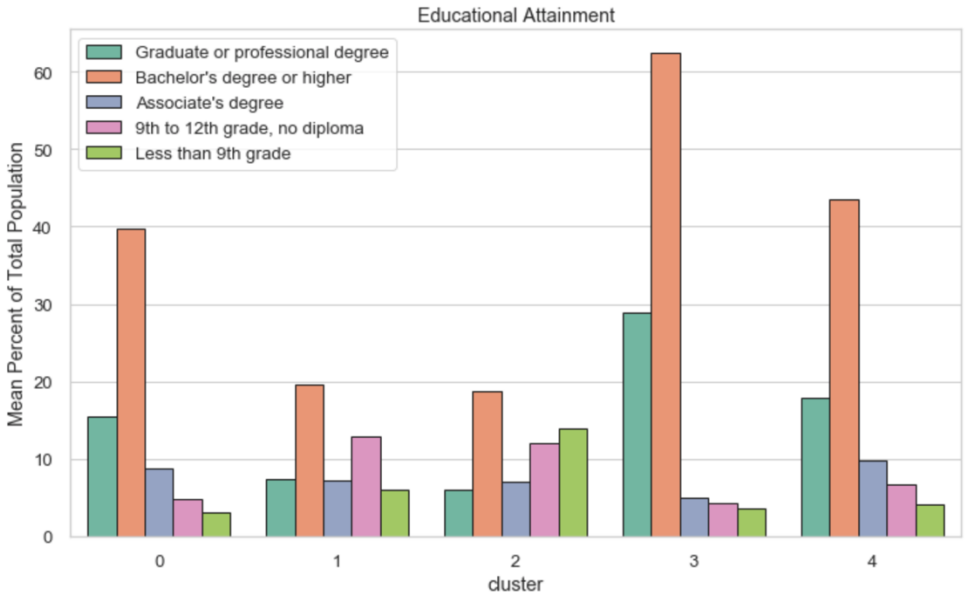


Figure 13 - Educational Attainment

In terms of commute method, cluster 4 has a visibly higher rate of people walking to their jobs instead of driving Figure 14.

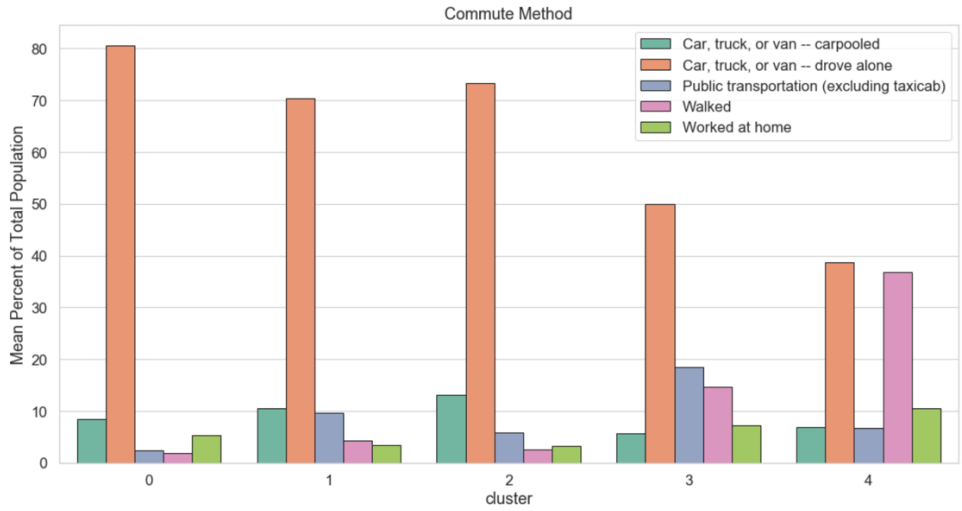


Figure 14 - Commute Method

From the table below, cluster 4 also has a mean – median commute time to work as much as 10 minutes shorter than other clusters.

Cluster	Mean Median Travel Time to Work (min)
0	24
1	24
2	26
3	25
4	15

Looking at rent in Figure 15, median rent is higher in cluster 3 than other clusters

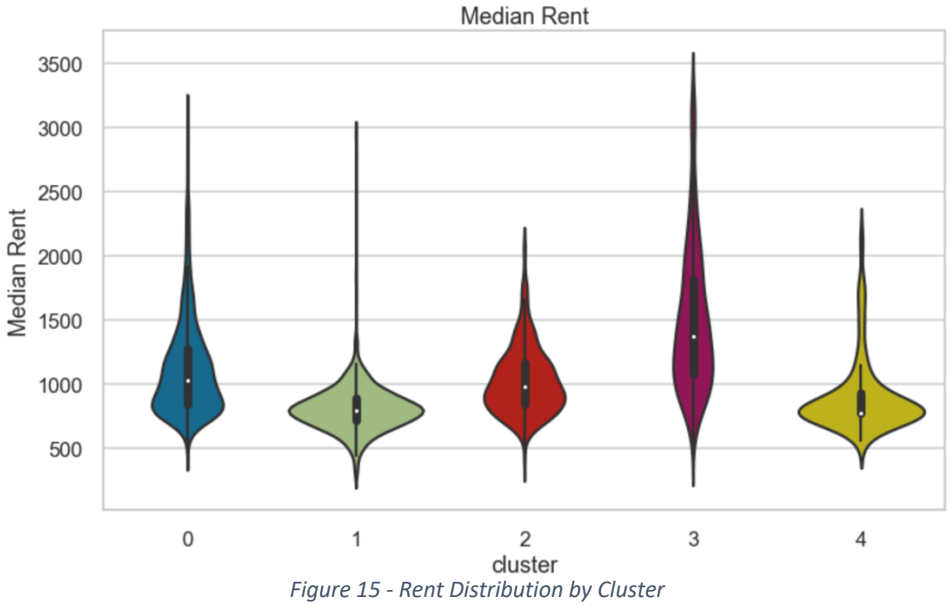


Figure 15 - Rent Distribution by Cluster

Lastly, looking at the business establishments in the neighborhoods, we see that all clusters have fairly similar distributions (Figure 16) although they have different medians (Figure 17), affected by outliers. It's possible to notice cluster 3 has a higher number of establishments, a possible explanation is that the algorithm clustered city centers and commercial neighborhoods in cluster 3.

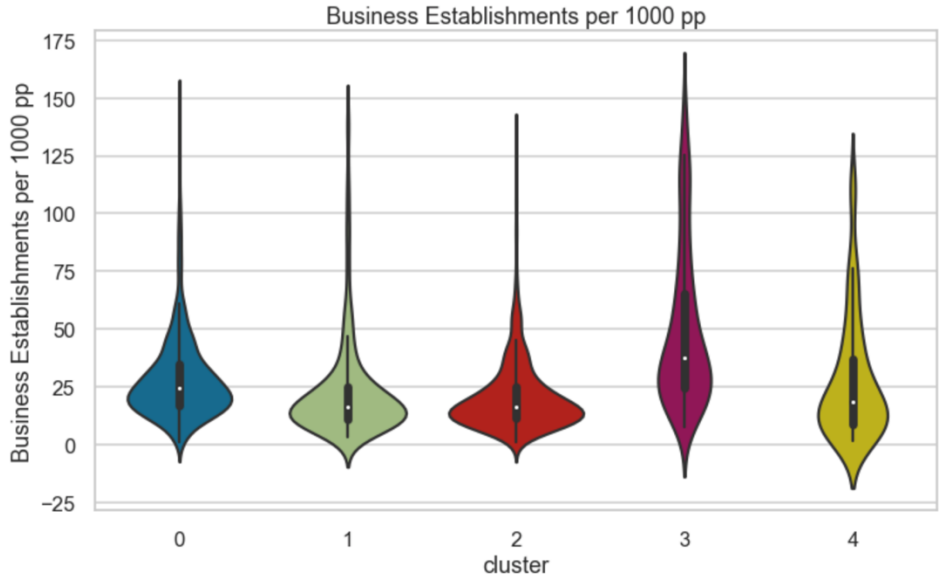


Figure 16 - Business Establishment per 1000 People Distribution

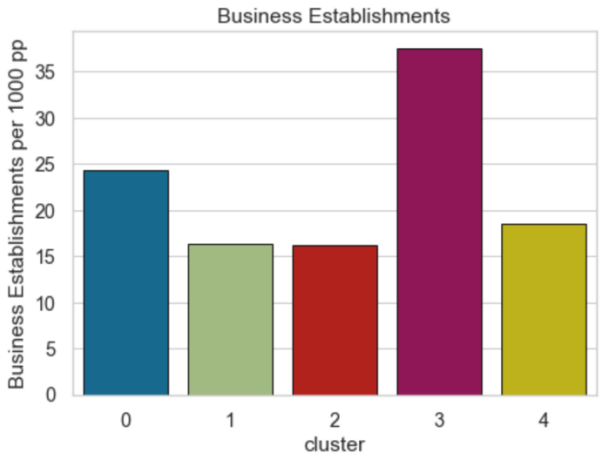


Figure 17 - Median Business Establishments per 1000 pp by Cluster

5. Discussion

After looking at the data from the clusters, the below profile can be constructed for cluster 3 neighborhoods, which has the one with higher rates of Beer Venues.

- Population
  - High income
  - Young (large amount of people ages 25-44)
  - Highly educated
- More expensive real estate
- Higher number of businesses

The prevalence of beer venues in this cluster isn't surprising, as it most likely confirms that beer venues attract customers with higher income and are located in city centers or places with higher economic activity that search for happy hour places after work, for example.

The last part of the work, a list of neighborhoods that show potential for opening new Beer Venues is suggested. The selection logic was done by filtering the data with the following:

- Cluster 3 neighborhoods
- Beer Venue per 1000 People **is lower** than MEDIAN value for the cluster
- Beer Venue Frequency People **is lower** than MEDIAN value for the cluster

The Median filter was used to consider that neighborhoods with lower frequencies/occurrences of Beer Venues had markets potentially less saturated of these venues, hence would be better options when opening a new venture.

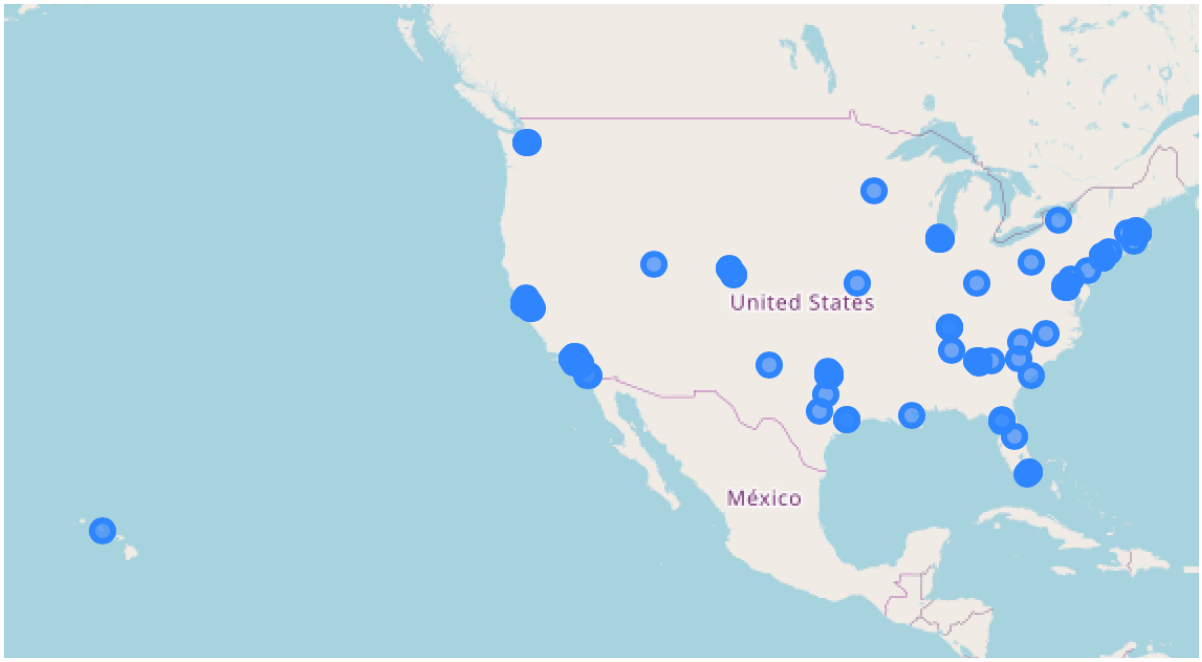


Figure 18 - Suggested New Beer Venue Neighborhoods

6. Conclusion

This report detailed the work done of identifying the differences in neighborhoods with higher rate of Beer Venues, with the goal of determining potential neighborhoods for opening new venues and also segmenting the markets for these venues. While this project only looked at beer themed venues, this analysis can be potentially done for many other categories.

In the end, out of more than 3000 neighborhoods analyzed, 133 are considered suitors for opening up a new beer venue.

Data integrity and accuracy is an important factor in this work. If Foursquare venue data is missing or miscategorized, the analysis can be skewed. Such might be the case for the number of beer venues, which might be higher than reported.

It's worth noting that the population segment defined in the discussion isn't necessarily the people who are most likely to frequent Beer Venues. There can be no assumption of that in this report since the data only looked at characteristics of population **within the neighborhoods** and not the specific characteristics of customers of Beer Venues. This means that these venues might see customer from other demographics and other neighborhoods frequenting these places. This work only gives an insight about the surroundings of the Beer Venues.

Customer segmentation for Beer Venues is a potential future work that would require data collected from people who visited these venues. Also, the suggestion of potential neighborhoods for new venues can be further improved by looking at venue specific demographics and matching that to neighborhoods demographical data.

7. Appendix A – Feature Table

Type	Measure	Restriction	Sub Group 1	Sub Group 2
Estimate	INCOME AND BENEFITS (IN 2018 INFLATION-ADJUSTED DOLLARS)	Total households	Median household income (dollars)	
Estimate	COMMUTING TO WORK	Workers 16 years and over	Mean travel time to work (minutes)	
Percent Estimate	EMPLOYMENT STATUS	Civilian labor force	Unemployment Rate	
Estimate	GROSS RENT	Occupied units paying rent	Median (dollars)	
Percent Estimate	SEX AND AGE	Total population	Male	
Estimate	SEX AND AGE	Total population	Median age (years)	
Percent Estimate	SEX AND AGE	Total population	Under 5 years	
Percent Estimate	SEX AND AGE	Total population	5 to 9 years	
Percent Estimate	SEX AND AGE	Total population	10 to 14 years	
Percent Estimate	SEX AND AGE	Total population	15 to 19 years	
Percent Estimate	SEX AND AGE	Total population	20 to 24 years	
Percent Estimate	SEX AND AGE	Total population	25 to 34 years	
Percent Estimate	SEX AND AGE	Total population	35 to 44 years	
Percent Estimate	SEX AND AGE	Total population	45 to 54 years	
Percent Estimate	SEX AND AGE	Total population	55 to 59 years	
Percent Estimate	SEX AND AGE	Total population	60 to 64 years	
Percent Estimate	SEX AND AGE	Total population	65 to 74 years	
Percent Estimate	SEX AND AGE	Total population	75 to 84 years	
Percent Estimate	SEX AND AGE	Total population	85 years and over	
Percent Estimate	RACE	Total population	One race	Asian
Percent Estimate	RACE	Total population	One race	Black or African American
Percent Estimate	RACE	Total population	One race	White
Percent Estimate	HISPANIC OR LATINO AND RACE	Total population	Hispanic or Latino (of any race)	
Percent Estimate	PLACE OF BIRTH	Total population	Foreign born	
Percent Estimate	MARITAL STATUS	Males 15 years and over	Widowed	
Percent Estimate	MARITAL STATUS	Males 15 years and over	Never married	
Percent Estimate	MARITAL STATUS	Males 15 years and over	Divorced	
Percent Estimate	MARITAL STATUS	Males 15 years and over	Now married, except separated	
Percent Estimate	MARITAL STATUS	Males 15 years and over	Separated	
Percent Estimate	MARITAL STATUS	Females 15 years and over	Widowed	
Percent Estimate	MARITAL STATUS	Females 15 years and over	Separated	
Percent Estimate	MARITAL STATUS	Females 15 years and over	Now married, except separated	
Percent Estimate	MARITAL STATUS	Females 15 years and over	Never married	
Percent Estimate	MARITAL STATUS	Females 15 years and over	Divorced	

Percent Estimate	EDUCATIONAL ATTAINMENT	Population 25 years and over	Graduate or professional degree	
Percent Estimate	EDUCATIONAL ATTAINMENT	Population 25 years and over	Bachelor's degree or higher	
Percent Estimate	EDUCATIONAL ATTAINMENT	Population 25 years and over	Associate's degree	
Percent Estimate	COMMUTING TO WORK	Workers 16 years and over	Car, truck, or van -- carpooled	
Percent Estimate	COMMUTING TO WORK	Workers 16 years and over	Car, truck, or van -- drove alone	
Percent Estimate	COMMUTING TO WORK	Workers 16 years and over	Public transportation (excluding taxicab)	
Percent Estimate	COMMUTING TO WORK	Workers 16 years and over	Walked	
Percent Estimate	COMMUTING TO WORK	Workers 16 years and over	Worked at home	
Percent Estimate	EDUCATIONAL ATTAINMENT	Population 25 years and over	9th to 12th grade, no diploma	
Percent Estimate	EDUCATIONAL ATTAINMENT	Population 25 years and over	Less than 9th grade	
Beer Venue_per1000	*Number of beer venues per 1000 people. Venue Count x 1000 / Total Population*			
Beer Venue_freq	Venue Category percent share of Total Venues per Zip			
pop_estab_ratio	Number of Inhabitants per Business Establishment			

Bibliography

1. **Doering, Christopher.** Tapped out? Brewpubs, taprooms inundate beer scene as brewers aim to stand out. *Food Dive.* [Online] Jan 2019. https://www.fooddive.com/news/tapped-out-brewpubs-taprooms-inundate-beer-scene-as-brewers-aim-to-stand/544103/.
2. **Foursquare.** Foursquare Developers. *Foursquare.* [Online] https://developer.foursquare.com.
3. **Bureau, United States Census.** Developers. *United States Census Bureau.* [Online] https://www.census.gov/developers/.
4. —. Zip Code Tabulation Areas. *Census Bureau.* [Online] 2020. https://www.census.gov/programs-surveys/geography/guidance/geo-areas/zctas.html.
5. **AlindGupta.** Elbow Method for optimal value of k in KMeans . *GeeksforGeeks.* [Online] https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/.
6. **developers, The scikit-yb.** Elbow Method . *Yellowbrick.* [Online] https://www.scikit-yb.org/en/latest/api/cluster/elbow.html.