**Clustering Model to Help Entrepreneurs Determine the Best Location to Launch a Startup**

**IBM Coursera Applied Data Science Capstone**

**02/22/2020**

**1. Introduction**

**1.1. Background**

The San Francisco Bay Area, aptly named, “Silicon Valley,” has long been synonymous with the US software and technology industry which has been fueled by the startup culture that has developed in the region over the past three decades. However, for the next big startup deciding on an ideal location to establish its base, other cities and metropolitan regions in the country offer an attractive alternative. Business Insider’s list of the 15 best cities in the US to launch a startup reflects the changing landscape of the startup environment in the United States. Based on the analysis by a software company, Volusion that used the 2017 US Census Bureau Annual Survey of the Entrepreneurs, the list is dominated by startup friendly cities distributed throughout the country.

The cities on the list were rated on five metrices: Entrepreneurship score, startup density, percentage of firms receiving venture capital investment, percentage of self-employed workers, and the percentage of respondents who reported that the reason for starting business is not the lack of work. When deciding between the cities on this list, a startup must, not only consider the governmental incentives and the tax framework for a business in early stages, but also the amenities and resources that will help attract its most precious resource, early career employees. To convince employees to lend their talents to help innovate and build up the company, the startup must, in addition to offering a financial incentive such as an equity stake, consider other factors that are likely to be important to young professionals. These include a reasonable cost of living, good schools for their children, and cultural activities available in each city.

**1.2 Problem**

City level data such as the relative cost of living and rent, quality of the schools in the area and neighborhood level data such as home prices and culinary and other cultural attractions can help an entrepreneur in making important investment decisions in launching their startup. This project aims to categorize cities on Business Insider’s list of the 15 best cities in the US to launch a startup based on these city and neighborhood level data.

**1.3 Interest**

According to the Bureau of Labor Statistics’ Business Employment Dynamics, 70% of startups don’t survive 10 years past their establishment. For a fledgling startup, investment decisions made during its early years can mean the difference between survival and going out of business. Since the success of a startup during this crucial period depends on their employees, a location that appeals to these early career employees as well as the one that is conducive to its growth is an important consideration. Therefore, insights into the factors that affect a startup’s ability to attract talented employees are likely to prove valuable to its growth and survival.

**2. Data acquisition and cleaning**

**2.1 Data sources**

**2.1.1 Cities**

The city level data included the cost of living and the school rating data. The cost of living data were scrapped from the publicly accessible Numbeo webpage that can be accessed [here](https://www.numbeo.com/cost-of-living/region_rankings_current.jsp?region=021). The school rating data were retrieved from the greatchools.org API.

**2.1.2 Neighborhoods**

The neighborhood level data included the Zillow Home Value Index (ZHVI) for single family homes and the Foursquare venues data. The ZHVI data ranged from 1996 to 2020 for all neighborhoods in the United States in the CSV format that was read into a Pandas dataframe from the single family home ZHVIs for each neighborhood made publicly available for educational use by [Zillow Research](https://www.zillow.com/research/data/). The neighborhood venue data were retrieved via the Foursquare API for 100 venues within the 500-meter radius of each neighborhood. Latitude and longitudes for both city and neighborhoods were retrieved using the Google Maps geocoding (Google V3) API implemented via Geopy’s geocoders. Also, the latitude and longitude boundaries for cities and neighborhoods were retrieved as geojson files from [Alex Motz’s Github repository](https://github.com/AlexMotz/us-neighborhood-geojson) based on the publicly available Zillow neighborhood data which organized them by county, city and neighborhood.

**2.2 Data Cleaning**

All city and neighborhood data that were scrapped, retrieved using their APIs, and downloaded were integrated into separate city and neighborhood dataframes.

**2.2.1 Cities**

The webpage containing the Numbeo cost of living index table was scrapped using the Beautiful Soup library (bs4) from which the table was read into a list and then into a Pandas dataframe. This initial table contained the cost of living index, rent index, cost of living plus rent index, groceries index, restaurant price index, and local purchasing power index for each city, state, and country name separated by a comma in a single column for each city in North America. Please find more information about Numbeo’s methodology for deriving cost of living indices [here](https://www.numbeo.com/cost-of-living/cpi_explained.jsp). It was cleaned in multiple stages to limit the data to the cost of living plus rent index corresponding to the city names on the BI list along with a two-letter abbreviation for their corresponding states.

Also, the XML file retrieved as the response to a successful call to the greatschools.org API was converted to a dictionary for each city using the xmltodict library. It was parsed to extract the summary rating or parent rating if no summary rating was available for both public and private schools at the elementary, middle and high schools. Then, the mean for each of these categories was calculated and stored in a dataframe along with the cost of living plus rent index and its latitudes and longitudes for each city. Please find more information about the greatschools ratings methodology [here](https://www.greatschools.org/gk/ratings/). Also, the geojson files containing the latitude and longitude boundaries for the cities on a choropleth map were cleaned to only include the city names

**2.1.2 Neighborhoods**

The CSV file containing the monthly ZHVI for single family homes in all US neighborhoods, their metropolitan areas, and two-letter abbreviation for the states from 1996 to 2020 was read into a list and then into a Pandas dataframe. This initial dataframe was cleaned to limit the data from 11-30-2019 to 11-30-2020 and any missing values were removed. Then, the dataframe was further cleaned to limit it to the city and states on the BI list. Also, latitude and longitude corresponding to each neighborhood was added to the dataframe and the any duplicate and missing values were removed. The venue data retrieved using the Foursquare API were cleaned to prepare them for frequency analysis by dummy coding the venues for each neighborhood.

**2.3 Exploratory Data Analysis**

**2.3.1 Cities**

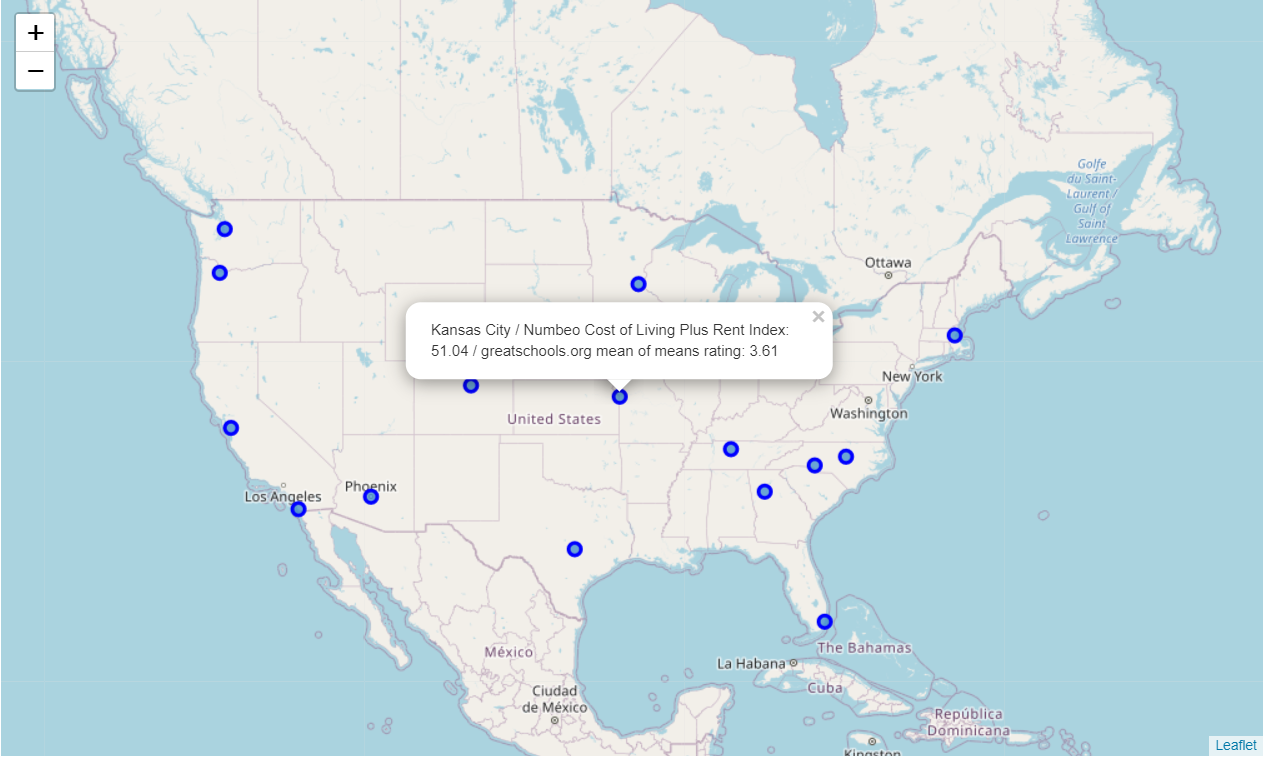


Figure 1. The selected city level data is displayed on the map when a marker is clicked.

Since school ratings were meant to summarize the quality of all schools in each city, a mean of means greatschools rating was calculated. The final cleaned city level dataframe consisted of the city name, latitudes, longitudes, cost of living plus rent index and the mean of means greatschools rating. These data were put together on a US map using the Folium mapping library where each city’s cost of living plus rent index and the mean of means great school rating could be displayed at its latitude and longitude by clicking on the corresponding marker (see Figure 1).

**2.3.2 Neighborhoods**

At the neighborhood level, the initial dataframe consisted of the city, metropolitan area, state, and the ZHVI. A cleaned dataframe containing the ZHVI for cities and states on the BI list above was analyzed to calculate the mean ZHVI across the monthly ZHVIs over the one-year period from 11-30-2019 to 11-30-2020.



Figure 2. The selected neighborhood level data is displayed on the map when a corresponding neighborhood marker is clicked.

These data along with the city data above, were integrated on a US map using the Folium mapping library where each neighborhood’s mean ZHVI along with their city level data could be displayed by clicking on the corresponding marker (see Figure 2). For the cleaned venues data, the frequencies of each venue were calculated and used to construct the 10 most common venues for each neighborhood. This dataframe was then integrated with the ZHVI data including the neighborhood, city, and their corresponding latitudes and longitudes.

**3. K-means Clustering and Follow-up Analysis**

**3.1 Cities**

**3.1.1 Model pre-processing**

The cleaned city level data were prepared for k-means clustering by dropping all non-numeric and non-relevant columns: city, state and latitudes and longitudes. Also, the cost of living plus rent index and the mean greatschools rating were standardized to allow for even comparison.

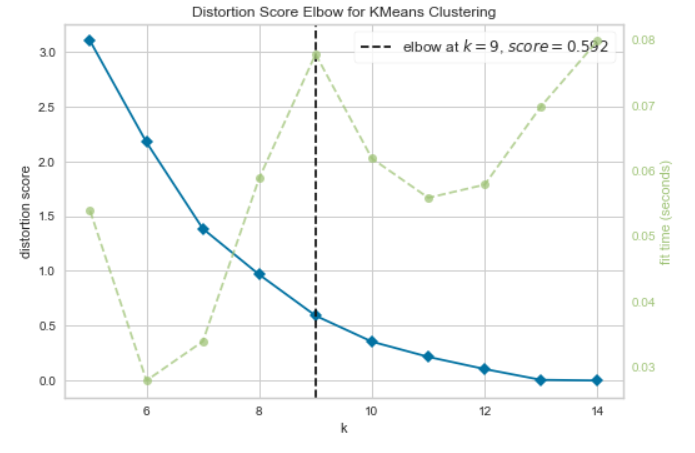


Figure 3. Elbow method to determine the optimal k means for the city data.

Since the number of means (centroids) can influence the clustering accuracy, elbow method was used to determine the k number of means. The mean squared error and the time to fit the model were plotted against the k number of clusters for the values of k ranging from 5 to 15. On this plot, the point (elbow) at which the mean squared error (distortion) was as low as possible while the time to fit the model was also as low as possible was determined to be at k = 9, the optimal value for the number of centroids to fit the model (see Figure 3).

**3.1.2 K-means clustering**

A k-means clustering model was fit to the standardized city level cost of living plus rent indices and the mean of means greatschools ratings with the number of clusters set to k = 9 according to the results of the elbow analysis described above. The resulting cluster labels were integrated back into the original dataframe which was then broken down by each cluster label to help identify similarity and differences within and between the clusters.

**3.1.3 Follow-up analysis**

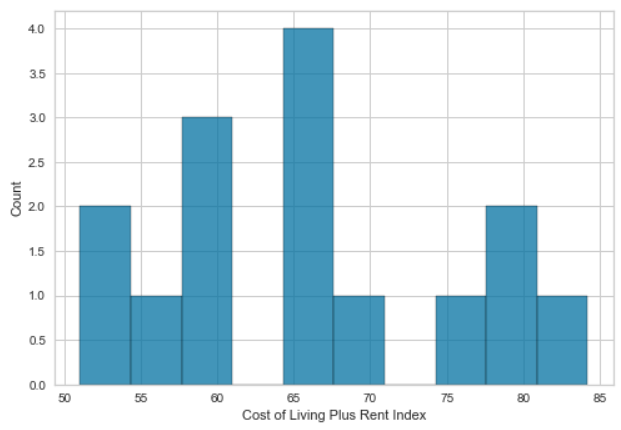


Figure 4. Distribution of the cost of living plus rent indices.

The city level data were further analyzed by examining the distribution of the cost of living plus rent indices (see Figure 4). Based on this distribution, indices between 50 and 65 (non-inclusive) were categorized as “Low”, those between 65 and 80 (non-inclusive) were categorized as “Medium”, and those at or above 80 were categorized as “High.” Also, the greatschools.org mean of means ratings were categorized according to its ratings methodology where ratings ranging from 1 to 4.50 (non-inclusive) were categorized as “Below Average,” those ranging from 4.50 to 6.50 (non-inclusive), and those at or above 6.50 were categorized as “Above Average.”

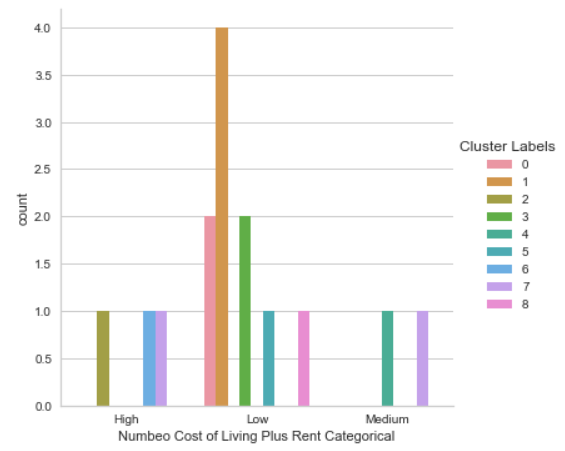


Figure 5. Distribution of the categorical Numbeo cost of living plus rent index by clusters.

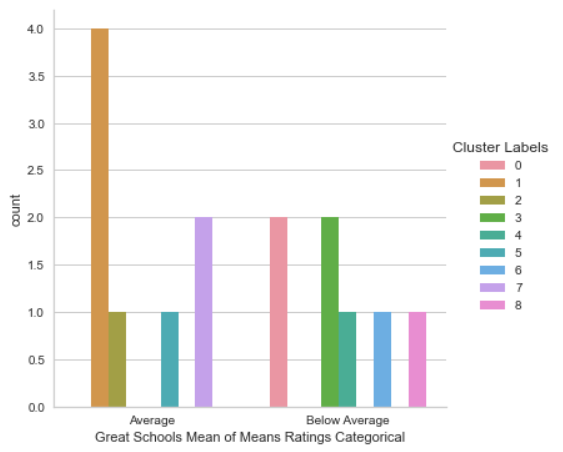


Figure 6. Distribution of greatschools.org mean of means categorical ratings by clusters.

These categorical labels for the Numbeo cost of living plus rent index and the greatschools.org mean of means rating were integrated with the dataframe containing the cluster labels for each city and were plotted on a bar plot (see Figure 5 and Figure 6). As seen above, cities in cluster 1: Austin (BI rank: 2), Charlotte (BI rank: 14), Denver (BI rank: 5), and Nashville (BI rank: 4) and in cluster 5: Portland (BI rank: 7) have both a low cost of living plus rent index and an average mean of means school rating and are therefore more desirable than other cities on the BI list. Cities in other clusters had a high cost of living or below average school ratings. The comparison between different cities based on these data can be made even more effective by integrating it on a choropleth map based on the Numbeo cost of living plus rent index (see Figure 7).

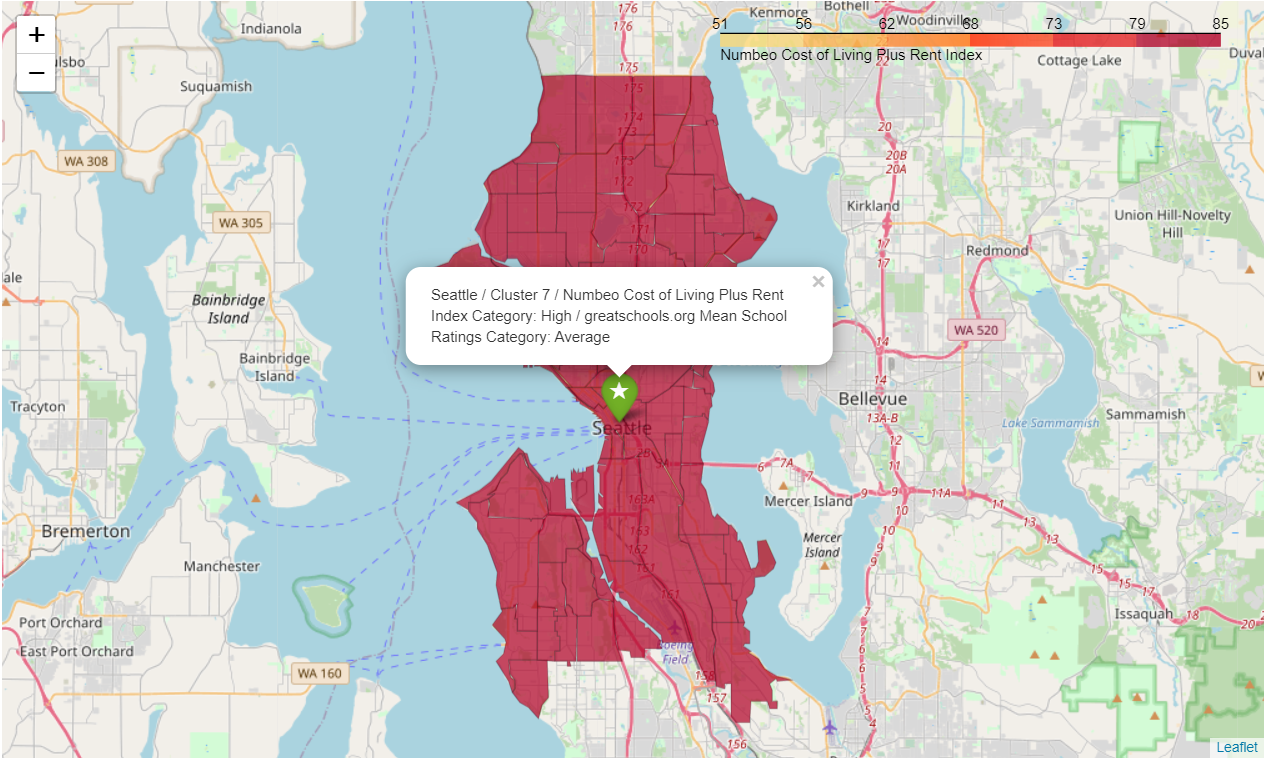


Figure 7. Choropleth map of the cities on the Business Insider’s “Top Cities to Launch a Startup” list based on Numbeo’s cost of living plus rent index.

**3.2 Neighborhoods**

**3.2.1 Model pre-processing**

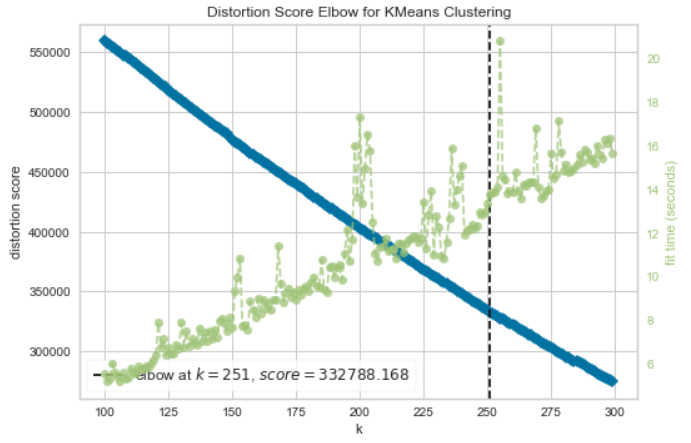


Figure 8. Elbow method to determine the optimal k means for the neighborhood data.

The cleaned neighborhood level data were prepared for k-means clustering by dropping all non-numeric and non-relevant columns: neighborhoods, and latitudes and longitudes. Also, the mean ZHVI and the dummy coded venues data were standardized to allow for even comparison. Since the number of means (centroids) can influence the clustering accuracy, elbow method was used to determine the k number of means. The mean squared error and the time to fit the model were plotted against the k number of clusters for the values of k ranging from 100 to to 300. On this plot, the optimal value for the number of centroids to fit the model was determined to be the point (elbow) at which the both mean squared error (distortion) and the time to fit the model were as low as possible which was found at k = 251 (see Figure 8).

**3.1.2 K-means clustering**

A k-means clustering model was fit to the standardized mean ZHVI and the dummy coded venues data with the number of clusters set to k = 251 according to the results of the elbow analysis described above. The resulting cluster labels were integrated back into the original dataframe which was then broken down by each cluster label to help identify similarity and differences within and between the clusters.

**3.1.3 Follow-up analysis**

The neighborhood data were further analyzed by examining the histogram of the distribution of mean ZHVI (see Figure 9). Based on this distribution, ZHVIs below $250,000 were categorized as “Low,” those between $250,000 and $500,000 were categorized as “Moderate,” those between $500,000 and $1,000,000 were categorized as “High,” and those above $1,000,000 were categorized as “Extremely High.” The clusters were analyzed in terms of the mean ZHVI categories using a bar plot of the mean ZHVI categories broken down by each cluster (see Figure 10).

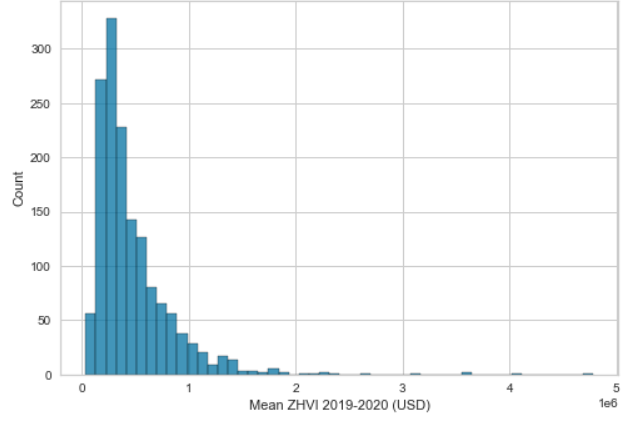


Figure 9. Distribution of mean ZHVI for the one-year period from 11-30-2019 to 11-30-2020.

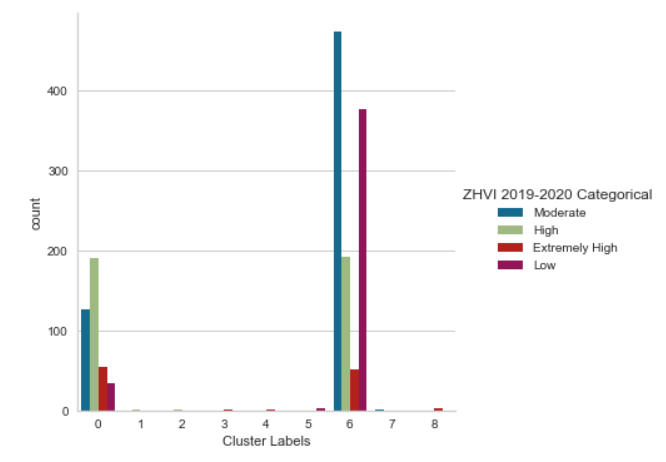


Figure 10. The distribution of mean ZHVI categories by clusters.

Most of the single-family homes seem to be in neighborhood clusters 0 and 6, and there were more homes in cluster 6 than in cluster 0. Specifically, greater proportion of homes in cluster 6 had moderate and low ZHVI than those in cluster 0 while a smaller proportion of homes in cluster 6 had high ZHVI than those in cluster 0. So, while cluster 0 was dominated by high ZHVI homes, low and moderate ZHVI homes made up a greater proportion of cluster 6. The clusters were further characterized by venue categories that are most likely to appeal to young professionals according to the [list](https://www.niche.com/places-to-live/search/best-cities-for-young-professionals/) compiled by niche.com which included “Restaurant,” “Pizza,” “Burger,” “Bar,” “Pub,” “Brewery,” “Music,” “Comedy Club,” “Coffee,” “Tea,” “Museum,” “Park,” “Theater,” “Art,” “Entertainment,” “Gym,” “Café,” and “Yoga.” Then, the total frequencies of each of these categories as the first most common venue was calculated for each cluster and they were plotted using a category plot (see Figure 11).

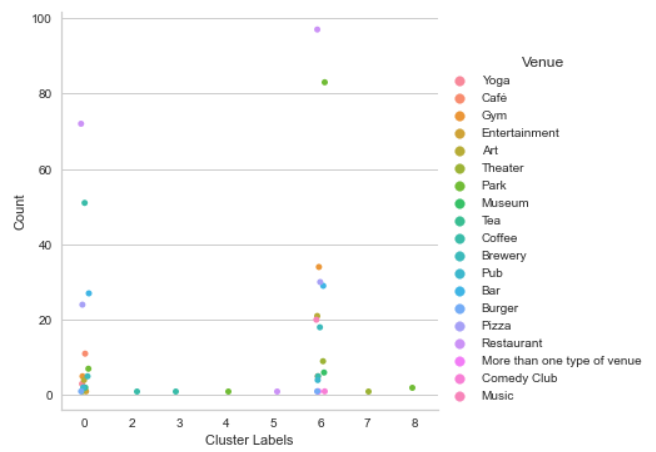


Figure 11. Category plot of the 1st most common venues likely to appeal to young professionals broken down by clusters.

As seen above, the venues were distributed mostly between clusters 0 and 6. While cluster 6 had more restaurants, more parks, gyms, music venues, comedy clubs, pubs, bars, breweries, and other entertainment venues, cluster 0 had more coffee places, some yoga studios, fewer bars, pubs and breweries and a few other entertainment venues. So, cluster 6 seems to contain more venues that are likely to appeal to young professionals than cluster 0. To gain further insight into the variety of restaurants available in neighborhoods in clusters 0 and 6, the total number of each type of restaurants were broken down by those clusters. As suspected, cluster 6 contained both greater number and a greater variety of restaurants. These insights were integrated with other neighborhood data on a choropleth map where each neighborhood was color-coded by the mean ZHVIs (see Figure 12 and Figure 13).

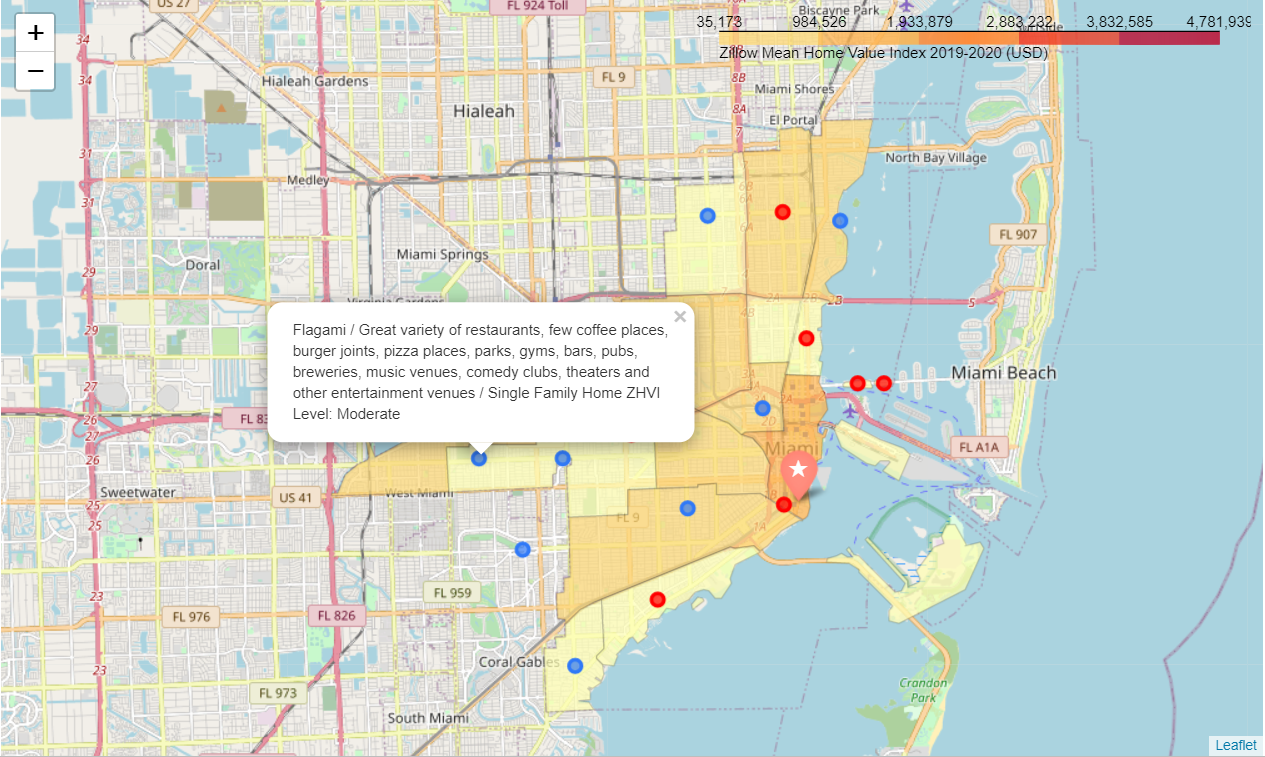


Figure 12. Choropleth map of a neighborhood in cluster 6 for a city on the Business Insider’s “Top Cities to Launch a Startup” list based on the mean ZHVI.

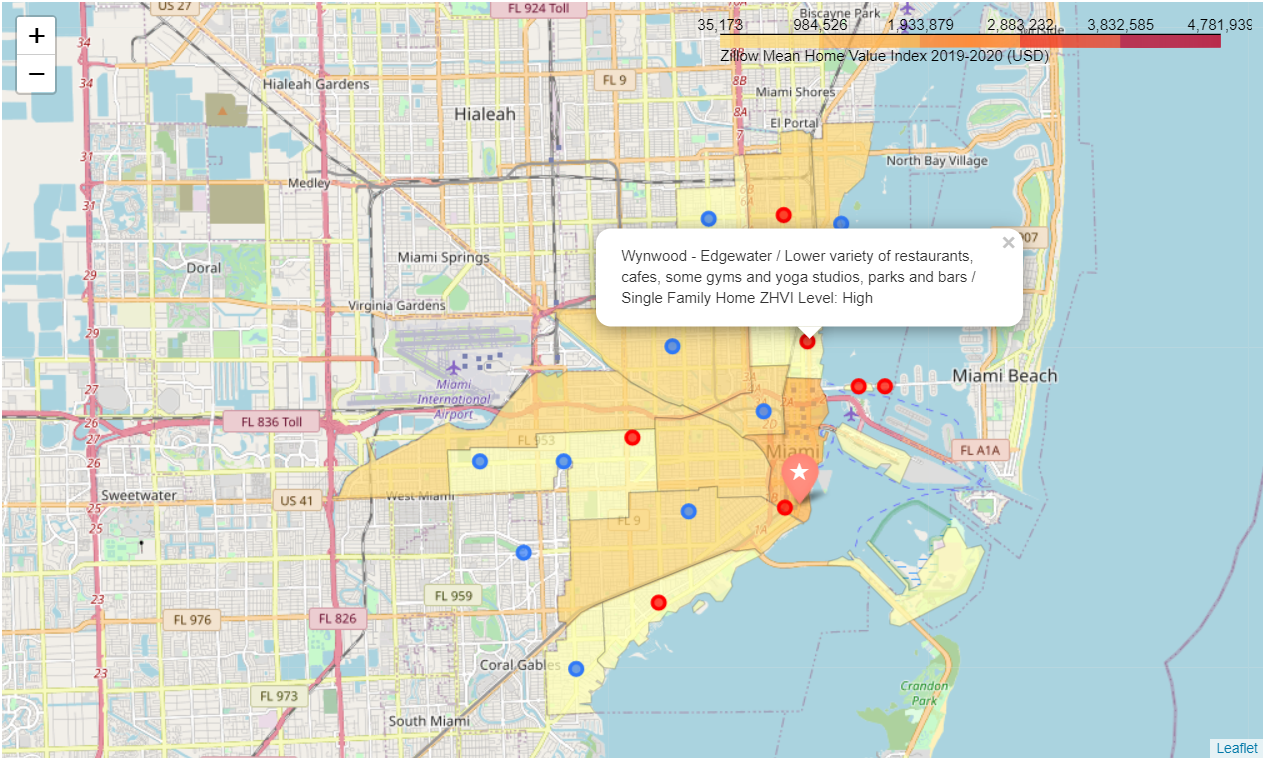


Figure 13. Choropleth map of a neighborhood in cluster 0 for a city on the Business Insider’s “Top Cities to Launch a Startup” list based on the mean ZHVI.

**4. Conclusion**

This project explored the cities and neighborhoods on the Business Insider’s (BI) list of 15 best US cities to launch a startup. The Numbeo cost of living index, greatschools.org school ratings, Zillow single family home values, and the Foursquare venues data were used to help entrepreneurs determine the cities that have both governmental frameworks most amenable to a startup and neighborhoods that are most likely to appeal to young professionals. The city level analysis revealed that Austin (BI rank: 2), Nashville (BI rank: 4), Denver (BI rank: 5), Portland (BI rank: 7), and Charlotte (BI rank: 14) have both a low cost of living including rent and the highest school rating range (average) among all of the cities on the list.

A cluster analysis of neighborhoods in these cities provided insights into venues that are likely to be attractive to young professionals. Most neighborhoods were divided into two clusters while the remaining ones either did not contain any neighborhoods with venues that are likely to appeal to young professionals or were spread out across the other seven clusters. While cluster 6 contained a great variety of restaurants, a few coffee places, gyms, parks, bars and pubs, breweries, comedy clubs, theaters and other entertainment venues, cluster 0 contained more coffee places, yoga studios, few parks, no comedy clubs and a few other entertainment venues. Therefore, within the cities mentioned above, the neighborhoods in cluster 6 with low or moderate single-family home ZHVIs appear to be ideal for young professionals.

**5. Future directions**

Since this project mainly used secondary data, the validity of the results depends on the validity and reliability of the primary data on which these secondary data were based. For example, although Numbeo’s cost of living indices are widely used by individual consumers and businesses, they are likely to be affected by seasonal trends in cost of fuel, groceries, and rent and may be unequally represented across different neighborhoods in their respective cities. Also, greatschools.org makes best efforts to provide parents with the most widespread and reliable information about the quality of the schools in their cities, but it still was not consistently available across cities. Specifically, in the school ratings data retrieved for this project, some of the unavailable school summary ratings were substituted with parental ratings (less reliable), but this was not possible in all cases where neither the summary nor the parental ratings were available. Therefore, notwithstanding some obvious difficulties with increasing the scope and availability of data and using primary data where possible due to reliance on volunteer input for data sourcing and bias that is inherent in user generated content depending on the sample size, taking such measures will likely lead to greater confidence in the results.