IBM DATA SCIENCE PROFESSIONAL
CERTIFICATE- COURSERA
APPLIED DATA SCIENCE CAPSTONE

THE BATTLE OF THE NEIGHBORHOODS: CAPSTONE PROJECT

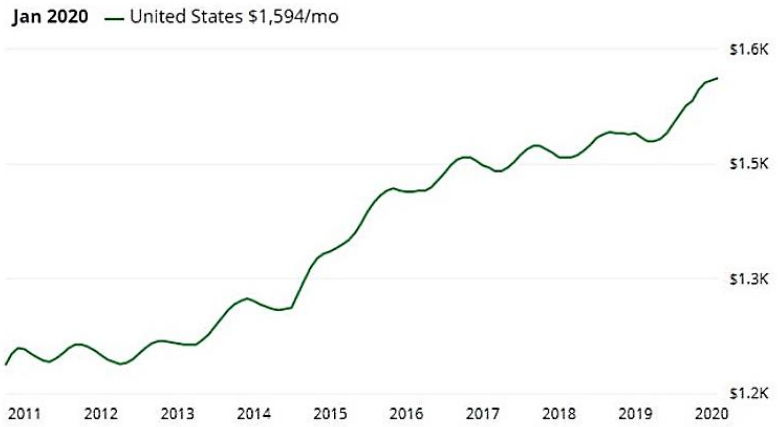
JUAN CARLOS CISNEROS

APRIL,2020

INTRODUCTION (BUSINESS PROBLEM)

In the current Covid-19 situation the world is facing, there is a lot of uncertainty in many markets. While people expect difficult times for some industries, others are more capable to reinstate their pace (such as the technology giants in the United States). Our client is a big player in the rental housing market in the United States. With interest rates going lower and lower, investing in this market seems attractive. According to a recent report, rent prices in the U.S have kept a rising trend since 2013 (Collins, 2020).

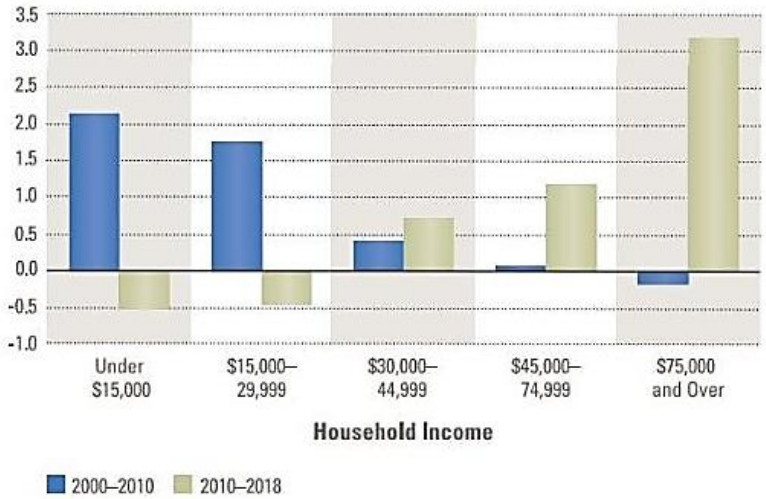
Graph 1 Rent Prices in the U.S



Source: Collins, (2020)

What is more interesting is that this growth has been mainly driven by high-income renters (Collins, 2020). As shown in the graph below, from 2010 to 2018, 3.2 million new high-income individuals and households have become renters. That is precisely why our client's business strategy is oriented towards the high-income market. It already holds some properties, mainly in California.

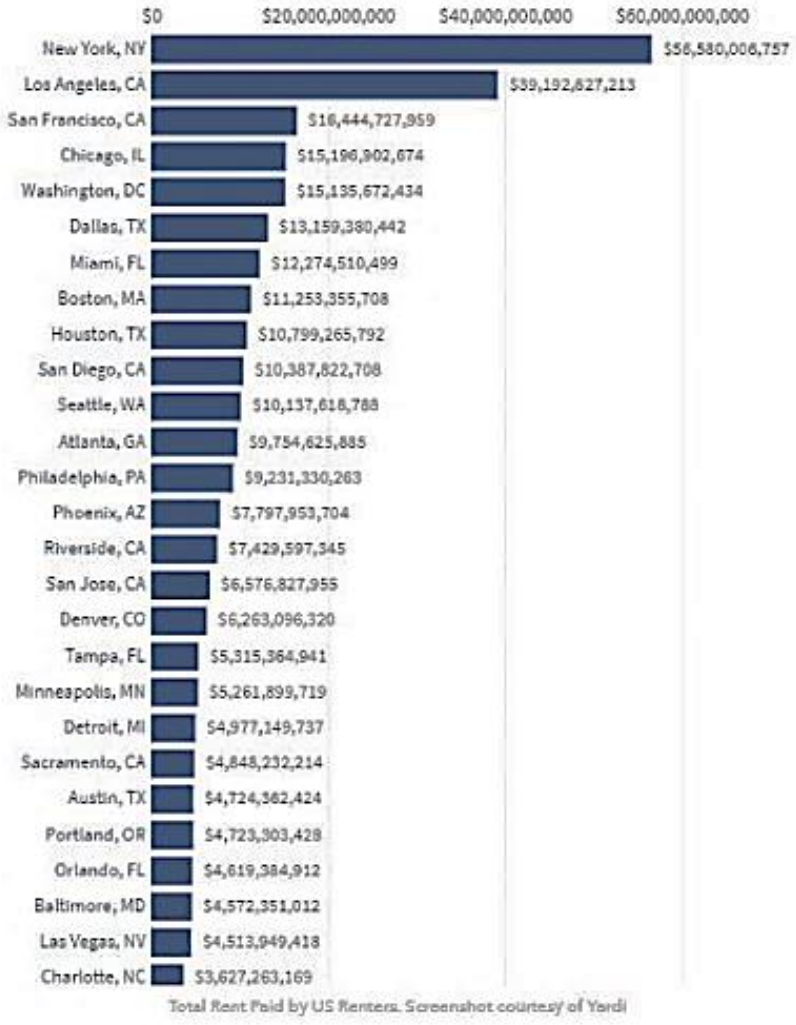
Graph 2 Net Change in Renter Households (Millions)



Source: Collins, (2020)

The most demanded metropolitan areas in the U.S rental housing market are New York, Los Angeles and San Francisco (Collins, 2020). These three top cities have amassed around $112 billion USD from 2005 to 2019.

Graph 3 Total Rent Paid (2005-2019)



Source: Collins, (2020)

Despite the considerable amount of rent the Californian cities have managed to generate, prices have remained stagnant in San Francisco and decreasing in Los Angeles (Collins, 2020). The general rising trend keeps occurring in big rental housing markets such as New York or Washington. Our client is currently planning to sell some of the properties that they own in these two cities and invest in some rising-trend markets such as the Big Apple. They are certain that they will be able to find some high-income individuals that are currently planning to leave the Golden State.

				1 Bedroom			2 Bedrooms		
Pos.	+/-	City		Price	M/M %	Y/Y %	Price	M/M %	Y/Y %
1	—	0	San Francisco, CA	$3,500	0.3%	0.0%	$4,500	0.0%	-3.2%
2	—	0	New York, NY	$3,000	1.0%	9.1%	$3,390	-1.2%	9.0%
3	—	0	Boston, MA	$2,590	3.6%	5.7%	$2,930	-0.7%	8.5%
4	—	0	Oakland, CA	$2,500	1.2%	6.4%	$3,000	0.3%	0.0%
5	—	0	San Jose, CA	$2,450	0.0%	-1.6%	$2,910	0.3%	-3.0%
6	—	0	Los Angeles, CA	$2,260	2.3%	-6.6%	$3,060	-1.3%	-5.0%
6	▲	1	Washington, DC	$2,260	2.7%	7.6%	$3,040	1.3%	15.2%
8	—	0	Seattle, WA	$1,890	1.6%	-0.5%	$2,340	1.7%	-7.9%
9	—	0	San Diego, CA	$1,790	0.6%	-8.2%	$2,390	1.7%	-3.2%
10	—	0	Miami, FL	$1,740	1.8%	-3.3%	$2,250	2.3%	-6.3%
11	▲	1	Anaheim, CA	$1,640	1.2%	-0.6%	$2,010	-1.5%	-8.2%
12	▲	1	Fort Lauderdale, FL	$1,610	0.6%	0.0%	$2,190	1.4%	9.5%
12	▼	-1	Santa Ana, CA	$1,610	-3.0%	-9.6%	$2,110	-0.5%	-6.6%
14	▲	1	Long Beach, CA	$1,600	3.2%	0.0%	$2,000	0.0%	-4.8%
14	▼	-1	Honolulu, HI	$1,600	0.0%	-5.3%	$2,150	2.4%	-2.3%
16	—	0	Denver, CO	$1,540	1.3%	0.7%	$1,980	4.2%	1.5%
17	—	0	Philadelphia, PA	$1,500	0.0%	4.2%	$1,640	0.0%	-3.5%
18	▲	1	Chicago, IL	$1,490	3.5%	-0.7%	$1,700	0.0%	-8.1%
18	—	0	Scottsdale, AZ	$1,490	2.8%	8.0%	$2,100	1.4%	11.1%
20	—	0	Atlanta, GA	$1,420	-0.7%	-4.1%	$1,790	-1.6%	-2.7%
21	▲	1	New Orleans, LA	$1,410	0.0%	-2.8%	$1,630	5.2%	6.5%
22	▲	2	Minneapolis, MN	$1,390	0.7%	-0.7%	$1,800	-1.1%	-5.3%
23	▲	2	Nashville, TN	$1,380	-2.8%	3.8%	$1,400	-4.8%	1.4%

National Average Rents for Cities. Screenshot courtesy of Zumper

Source: Collins, (2020)

Our client has requested some insights from the upscale rental housing markets. First, they want to know what the 100 neighborhoods or areas with the highest rent prices in the United States are. Second, they want to know what type of venues are common to these areas. These venues may be highly valued by potential clients and they might expect that their new neighborhood features these as well. Finally, they want clusters of these areas so that they can identify what properties to offer to the high-income renters in California that are planning to move

DATA

The insights required by the client will be obtained from two data sources: Zillow and Foursquare. Zillow is an online real estate database that has information of the U.S rental housing market. Starting on 2012, Zillow constantly publishes the Zillow Rent Index (ZRI). The ZRI is a US Dollar valued index that tries to estimate the market rate rent across a given region/area. This database is available to download at https://www.zillow.com/research/data/. A simple peak of this database (shown below) is enough to see that it has enough data to figure out the top 100 neighborhoods we are going to be working with (they appear in the "RegionName" column).

RegionName	State	Metro	County	City	Zri
Northwest Harbor	NY	New York-Newark-Jersey City	Suffolk County	East Hampton	25372
Central Menlo Park	CA	San Francisco-Oakland-Hayward	San Mateo County	Menlo Park	9206
Presidio Heights	CA	San Francisco-Oakland-Hayward	San Francisco County	San Francisco	5721
East Village	CA	Los Angeles-Long Beach-Anaheim	Los Angeles County	San Marino	5692
North of Montana	CA	Los Angeles-Long Beach-Anaheim	Los Angeles County	Santa Monica	5623

Foursquare is a U.S-based company that participates in the location technology and information industry. Using the Foursquare API, a developer can access user-uploaded information such as venues near a specific location, trending venues near a specific location and information on those venues such as their category, their score and the users' comments.

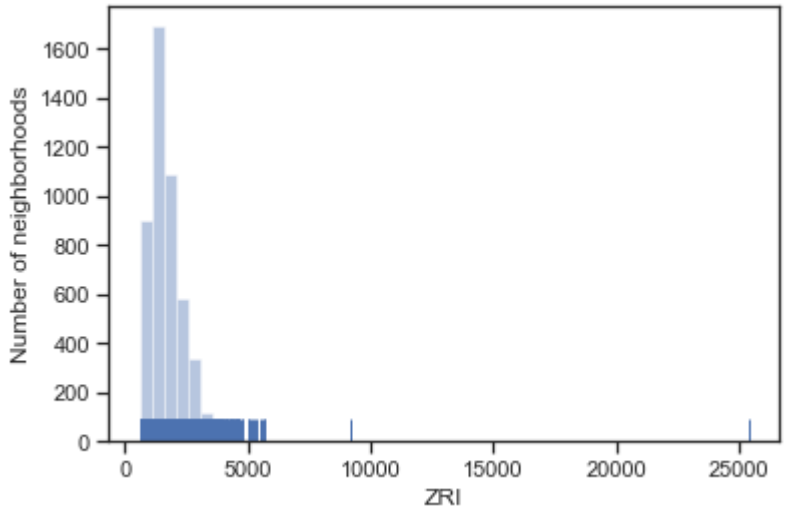
Using the Geocoder library in Python, it is feasible to find the exact latitudes and longitudes of the areas listed in the ZRI table from ArcGIS. With the latitude-longitude information from each area/neighborhood, calls to the Foursquare API will be made to obtain information about the venues around those locations.

METHODOLOGY

It is important to keep in mind what are the deliverables the client requires: a list of the top 100 neighborhoods with the highest rent prices in the United States, information of the most common venues in these set of neighborhoods and a clustering that may help them identify properties that are suitable for high-income Californians planning to move. With that in mind, it is possible to work in an orderly fashion to properly analyze with the data.

First, we import the Zillow database and analyze it. The full database has 4840 observations corresponding to the neighborhoods with highest typical rent prices in the United States. The ZRI Index (the typical rent in a particular area) for this sample has a mean value of $1712,53. As shown below, the distribution of the ZRI index is right-skewed and has two particular observations that may be outliers. Following the client request, only the top 100 neighborhoods will be kept for the rest of our analysis.

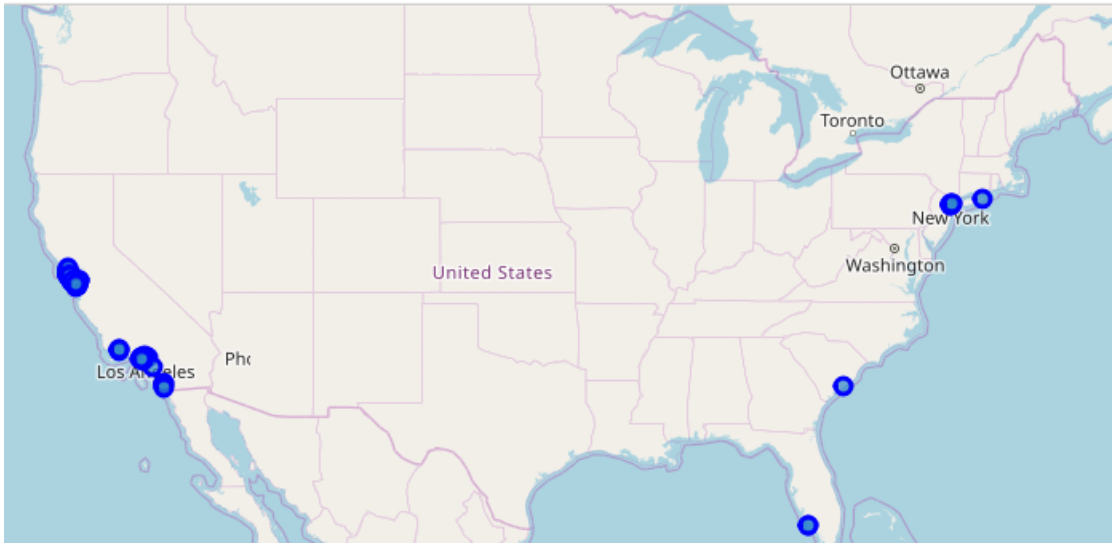
Graph 4 ZRI Distribution (Total Database)



To complement the information for the client and to be able to use the Foursquare API, the Geocoder library was used to add the Latitude and Longitude information to each neighborhood. In order to make the query more specific to Geocoder, the information on "RegionName" was combined with "City", "County" and "State". To test if the Geocoder had properly identified the position of the 100 neighborhoods, a map of the

United States with marks on these locations was created. As can be seen in the picture below, the marker positions are coherent with the data described before.

Graph 5 Map of the United Stated with the location of the top 100 more expensive areas



With the latitude and longitude data, we can run calls to the Foursquare API to obtain information about the venues near each position. We run a "explore" call for each observation and as a response we get up to 20 venues per location (the quantity of venues obtained per location depends on the information on Foursquare); aside from the name of the venue, the response also includes the location of the venue and its category.

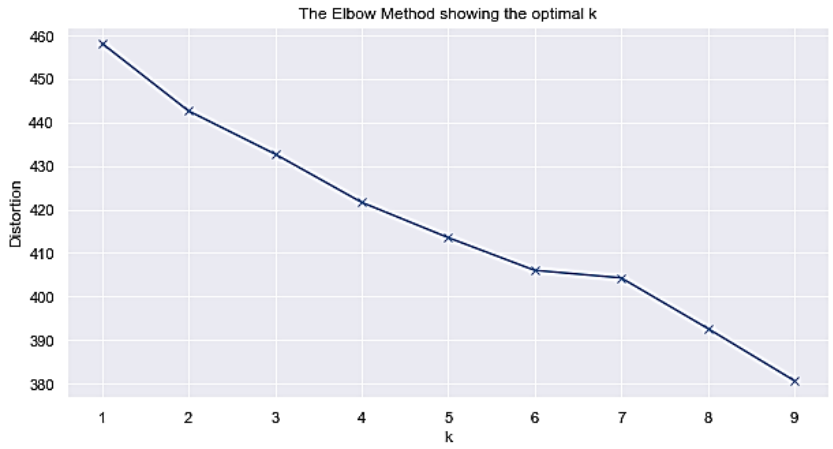
Now we can analyze this information to give the client the insight requested on the most common venues of this set of neighborhoods. The rent might be a market proxy for the subjective value of a neighborhood; thus, the venues located in the neighborhoods with the highest typical rent might give an idea of what type of venues are valued by high-income clients. The Foursquare response included 249 different venue categories that are among the top 20 venues for each neighborhood position. The second deliverable for the client is obtained by grouping this data by venue category.

For the last request, the strategy is to cluster the neighborhoods from this set based on the ZRI and the 10 most common venue categories for each observation. To prepare the data for the clustering, two previous steps are required. First, the categorical variables (the venue categories) are transformed to dummy variables; this is because the machine learning technique that will be used only uses numerical variables. The second step is to normalize the ZRI variable. If the ZRI variable is not normalized from 0 to 1, the algorithm is going to assign considerably more weight to this variable, as its values are in hundreds or even thousands in magnitude.

The machine learning technique that will be used for the clustering process is k-means. K-means is a technique of unsupervised learning that minimizes within-cluster variances until reaching a local optimum. Precisely the fact that it converges towards a local optimum, and not a global optimum, may be an issue with this method. To obtain a clustering with less error, optimal k tests will be performed. The first test is the Elbow Method. The Elbow Method is a heuristic method of choosing the optimal number of clusters. More clusters mean less error, but a cost in generalization error. The Elbow

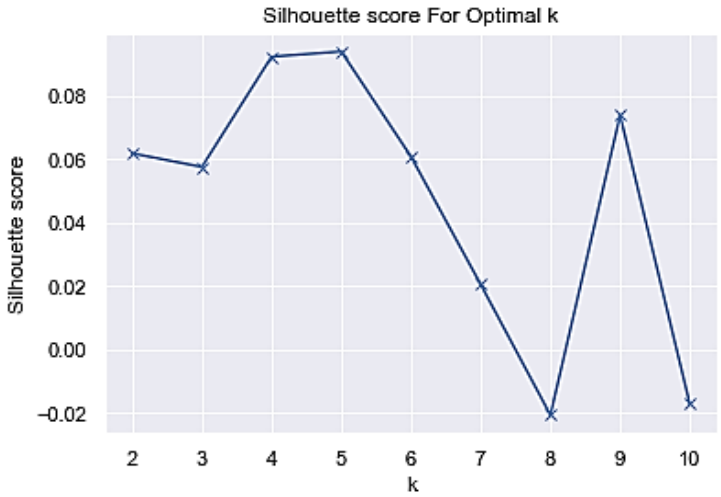
Method consists in identifying the k in which a k+1 clustering will not contribute much error reduction. The following plot shows that the elbow method for this case does not show a clear choice for k.

Graph 6 Elbow Method



To further confirm this choice, another testing method is used. The silhouette method measures how similar each observation is to its own cluster and how different it is to other clusters. The silhouette scores range from -1 to 1, and higher values mean that the observation is well matched to its cluster and poorly matched to neighboring clusters. The silhouette method suggests k=5 would be good choice for the k-means algorithm.

Graph 7 Silhouette Method

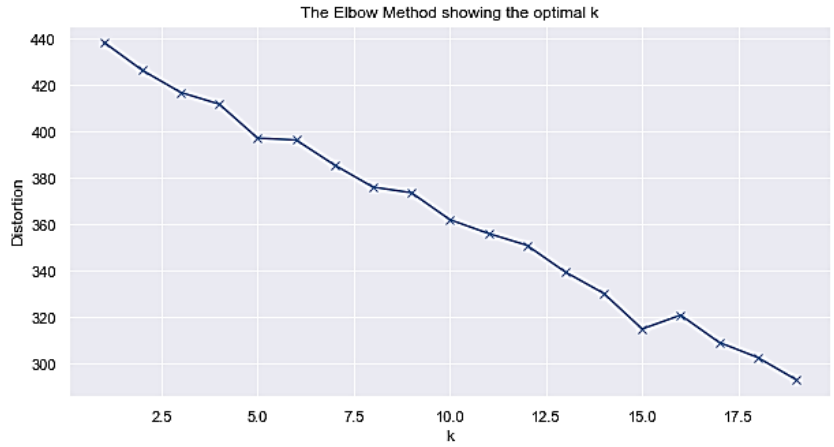


The clustering algorithm was run and made clear that something was wrong with the data being used. Some of the clusters were perfect matches of neighborhoods with identical information. Examples as the one in the table below made clear that the Geocoder was unable to differentiate between neighboring locations with similar names.

Region	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Zri	Latitude	Longitude
Inner Sunset, San Francisco, San Francisco Cou...	Mexican Restaurant	Spa	Bank	Sporting Goods Shop	Video Store	Sushi Restaurant	Clothing Store	4092	37.729634	-122.493391
Outer Sunset, San Francisco, San Francisco Cou...	Mexican Restaurant	Spa	Bank	Sporting Goods Shop	Video Store	Sushi Restaurant	Clothing Store	3933	37.729634	-122.493391

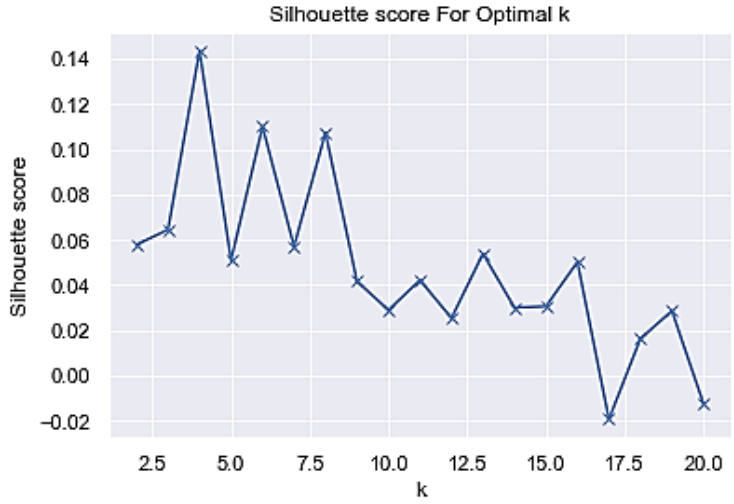
This was the case for six of the neighborhoods on the database. To overcome this clustering bias, the observation with the highest ZRI from each similar pair was kept. Once this was corrected, the clustering process was repeated. The Elbow Method to choose the optimal k for the k-means algorithm was repeated. As in the previous case, there does not seem to be a clear choice for an "elbow point".

Graph 8 Elbow Method for Corrected Database



The silhouette method suggests k=34might be a good choice for the clustering process, as it possesses the highest silhouette score. Consequently, we run the k-means algorithm for 4 clusters. After we get the final cluster labels for each location, we add that information to the database and filter each cluster. The clusters identified by the k-means technique might suggest possible neighborhoods our client might offer potential high-income renters currently in California that wish to move to other state.

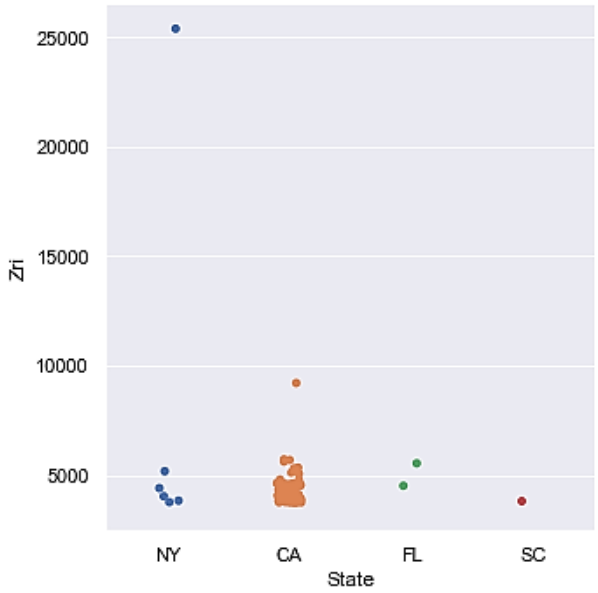
Graph 9 Silhouette Method for Corrected Database



The client wants deliverables that will allow the optimization of its current operation in the upscale rental housing markets. Its current business strategy is not in harmony with the current growth trends that can be seen in this market in the U.S. Following the methodology described above, it is possible to give the client the insights that were demanded.

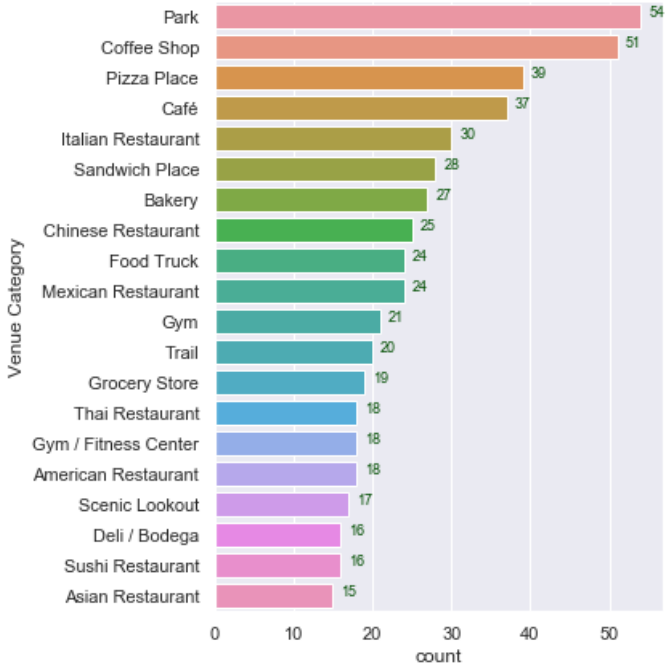
The first request was a list of the 100 neighborhoods with the highest rent prices in the United States. The top 100 neighborhoods are in just four states: New York, California, Florida and South Carolina. California amasses most of the neighborhoods, but they could potentially be matched in a cluster with a neighborhood from another state. Aside from the two most expensive neighborhoods in the set (one in New York and the other in California), all the observations are in a similar ZRI range. The database will be made available for the client in a .csv format.

Graph 10 ZRI Distribution (Top 100 by State)



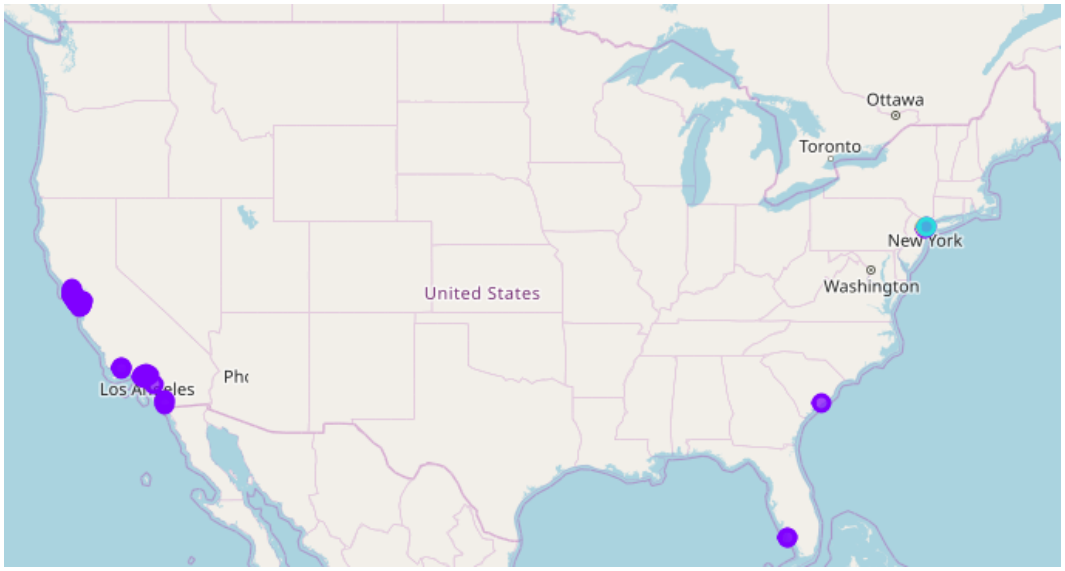
The second request was to obtain information about the types of venues that are common in this set of neighborhoods. Using both Geocoder and the Foursquare API, the information regarding all the venues in each of these neighborhoods was added to the database. The graphic below shows the 20 most common venues in the neighborhoods with higher rent prices in the U.S. Parks and coffee shops are the most common venues located in these neighborhoods. This list of venue might give a hint of the lifestyle that upscale renters might want. The venues in these neighborhoods cater to the comfort and well-being of its residents, with places such as restaurants, gyms, groceries stores, trails or scenic lookouts; there is no sight of places related to business activities.

Graph 11 20 most common venues in the Top 100 set



The third deliverable was the clustering of these neighborhoods with the purpose of identifying suitable properties to offer high-income renters in California that are looking to rent a property elsewhere in the United States. After selecting k=4 for the k-means algorithm, the clusters formed were quite particular. Three of the four clusters featured just one element that could not be matched with the rest; these unmatched neighborhoods are Downtown Menlo Park in San Mateo, CA, Upper East Side and Midtown in New York, NY. That means the other cluster includes all the possible matches of Californian neighborhoods with out of the state options. The other 82 Californian neighborhoods included in the dataset can be matched with 6 neighborhood options outside of California: Charlestowne in Charleston, SC, Old Naples and Royal Harbor in Naples, FL, and Flatiron District, Little Italy, and Roosevelt Island in New York, NY. This cluster will be made available to the client in a .csv file.

Graph 12 Clusters in the U.S map



DISCUSSION

There are many things to consider before using this report as a base for decision making. The method used can have some limitations; these limitations will be described in this section. Moreover, suggestions for future works will be made considering the limitations mentioned.

The first point to take into account is that the clustering algorithm is classifying the neighborhoods using information of nearby venues and ZRI index. A potential issue with the Foursquare API is that data of the venues might be incomplete or outdated varying by neighborhood. Additionally, the model is not considering actual clients' preference when looking for a place to rent. This decision might not be based entirely on amenities nearby or rent price, but also on other facts such as weather, availability of relevant jobs, traffic or acquaintances in the neighborhood or nearby. A suggestion for future projects would be to obtain a database from actual clients that have moved from one place to another (and base the model on that real choice) and to run qualitative surveys that try to uncover the reasons behind each decision.

The second limitation is that using the k-means algorithm does not guarantee reaching a global optimum for our classification task. The cluster classification by k-means largely depends on the point where the centroids start. To make sure the clustering results are robust, there are two possible methods to undertake. The first is to loop through the clustering algorithm varying the starting points in each run. The results of each run can be stored and then the most common classification result can be used. Furthermore, other machine learning classification technique such as decision trees can be used for this task. If varying the classification technique does not change result much, it would suggest our classification is reliable.

CONCLUSION

Businesses in the rental housing market can optimize their choices by using relevant data appropriately. In this case, our client, if the strategy of operating mostly in California is kept, might fail to take advantage of the considerable growths this industry is going through in other markets. Using machine learning classification techniques such as k-means can help the client identify what products to offer each of its customers. Other techniques such as regression, together with proper data collection, could also contribute with insights about the factors that drive the prices of the properties or to find the determinants of the choices the customers in this market make.

REFERENCES

Collins, G. (03/21/2020). The US Rental Property Investment Market 2020. https://managecasa.com/es/articles/us-rental-property-market/

Zillow Rent Index. (2020). https://www.zillow.com/research/data/