**The Battle of the Neighborhoods, but make it Miami**

**IBM/Coursera Applied Data Science Capstone Project**

**Submitted by Carolina Belmonte**

1. **Introduction**

Before we dive in to what you're really here for, I wanted to share my thoughts on my journey through this course. I came in with basically zero knowledge of any of these concepts, and much worse...zero programming (read: Python) experience. I am very happy to report that this class has made me come a long way, but I am walking away (especially after all of the struggles it took get this capstone done!!) knowing where to focus myself - I look forward to deep diving into Python now that this course is coming to a close for me. I feel that to fully take advantage of the concepts & libraries presented in this course, I really need to sharpen my Python. Thank you instructors & fellow students for an amazing journey!

##### ***Description of the problem and discussion of the background***

*“Miami, officially the City of Miami, is the seat of Miami-Dade County, and the cultural, economic and financial center of South Florida in the United States.”*  
If you think I'm making up the bit above, I invite you to navigate to Google and enter "Miami, Florida" in the search bar :) As a Miami native, I've had the privilege of always being able to boast that I live where people vacation.  
However, no place is perfect - Miami real estate is almost prohibitively expensive for your average citizen. According to [Zillow](https://www.zillow.com/miami-fl/home-values/): *The median home value in Miami is . Miami home values have gone up 0.6\% over the past year and Zillow predicts they will rise 3\% within the next year. The median list price per square foot in Miami is , which is higher than the Miami-Fort Lauderdale-West Palm Beach Metro average of . The median price of homes currently listed in Miami is while the median price of homes that sold is .*  
This conundrum has left me wondering - for many years, in fact - how can we look for reasonably priced real estate in an area of Miami that still has a lot to offer? Thus, our target audience is middle class potential home buyers looking to maximize their ability to enjoy Miami’s venue offerings while still paying a reasonable price for a property. Given that Miami is an ever growing metropolis that continuously attracts people, potential home buyers, investors and city planners would be interested in this data.

##### ***Description of the data and how it will be used to solve the problem***

Luckily, we live in a world (world-wide-web) full of open data sets. I knew I wanted to segment the city of Miami by its Zip Codes, so I first went on the hunt for a data set that contained all of the zip codes in Miami. I was able to find and download one from [ODS](https://public.opendatasoft.com/explore/?sort=modified). Next, I needed to locate data on property values in the City of Miami. This data was actually quite easy to locate and extract, as [Miami-Dade County (where the city sits) has an open data hub as well](https://gis-mdc.opendata.arcgis.com/). The final component would be data on nearby venues for each zip code. Based on the labs done with Foursquare, we will use the Foursquare API as the source of our venue data based on our zip codes. Putting all of these pieces together, we'll be able to explore Miami's zip codes to figure out where we can buy real estate sensibly but still have access to the amazing things Miami has to offer.

1. **Methodology**

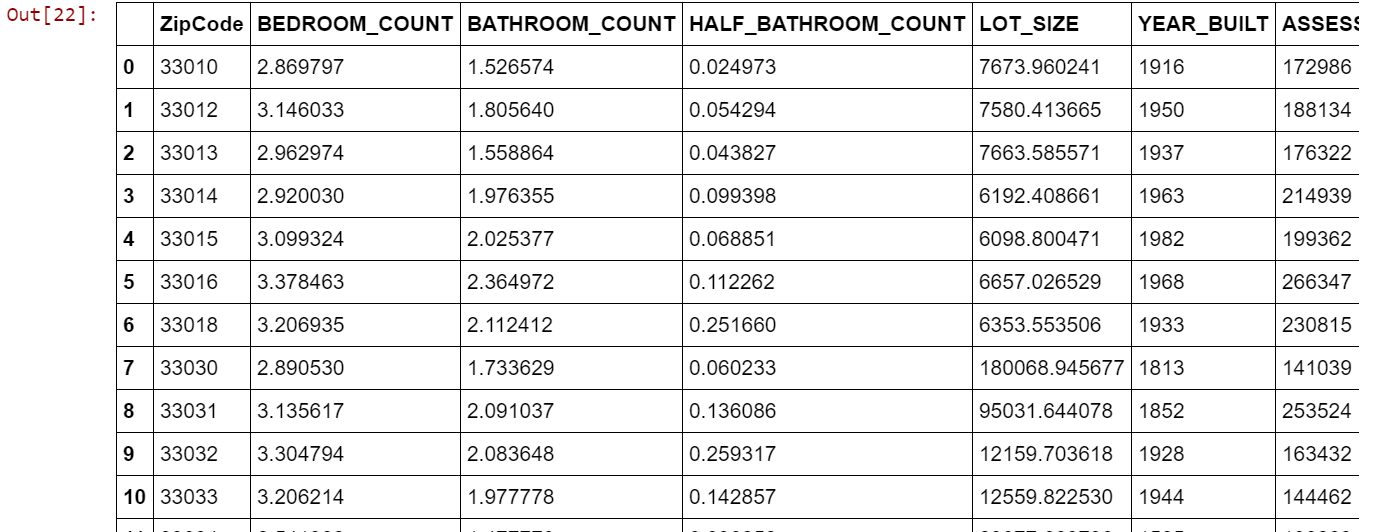
The first step in this journey was to make sure all the libraries necessary for data prep/processing, modeling and visualization were imported. The following libraries were imported (and/or installed):

* Numpy
* Pandas
* Requests
* Matplotlib
* Folium
* Geopy
* SciKit Learn

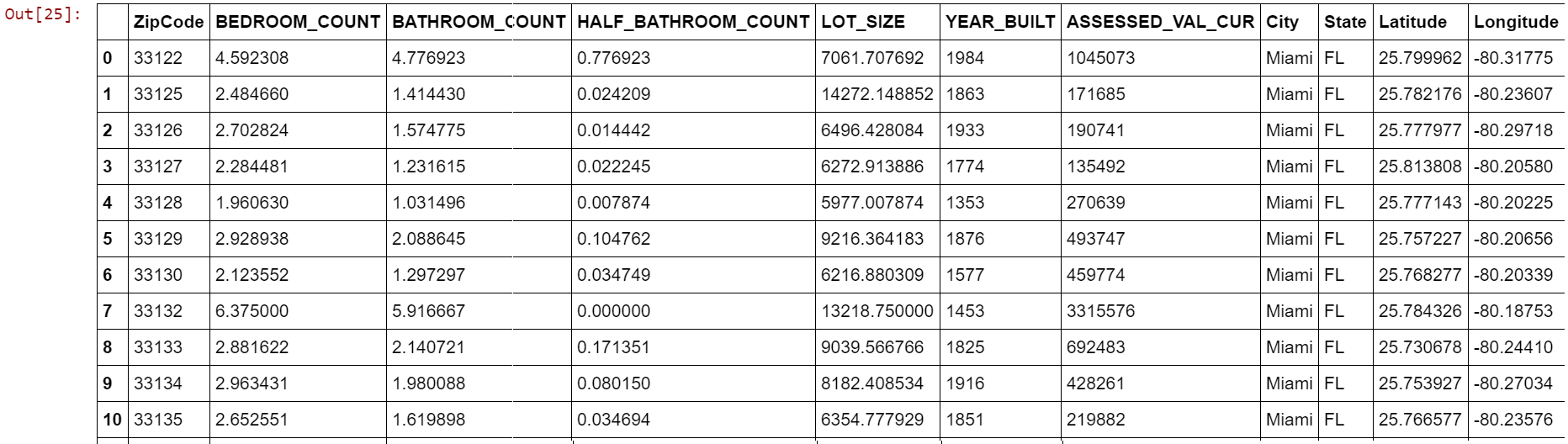
I had never taken advantage of IBM Watson’s storage capabilities within Project instances. Since I was able to download all my data as CSV files, I decided to try this out and insert my code directly into the cells. It was rather nifty to be able to do this and create a pandas dataframe directly from my files. I started with zip code data. After several pre-processing steps (detailed in the notebook), I arrived at the dataframe below (only the first ten rows are pictured below):



Then, I proceeded to do the same with my property data. There was even more preprocessing involved with this data set, which is also outlined in detail in the notebook. The data contained many features that are not relevant for this analysis. The key features that were kept included: property zip code, number of bedrooms and bathrooms, lot size, year built and current assessed value. A snapshot of this frame is below, picturing the first ten rows:



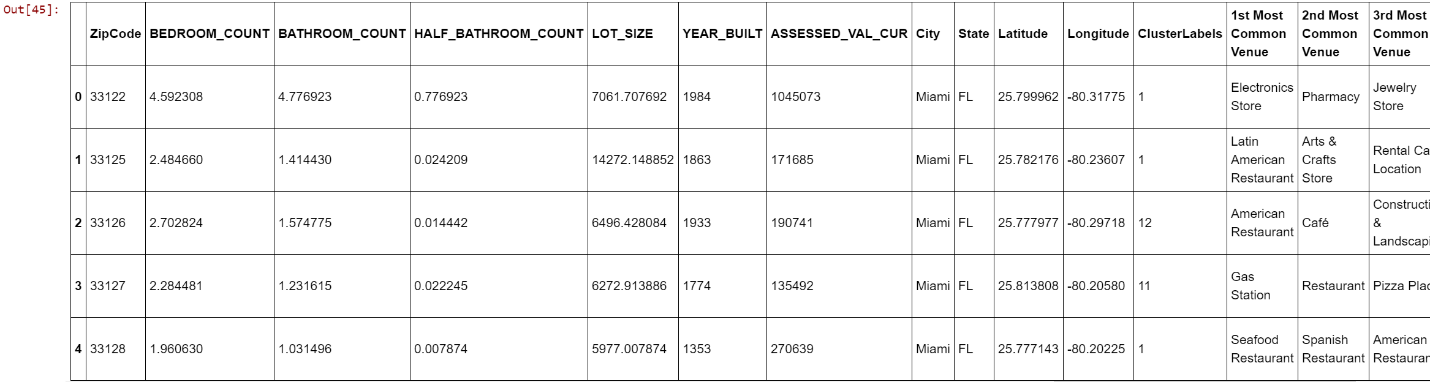
At this point, the next step was to merge these two data frames by using the Zip Codes as the key.



With this merged data frame based on Miami’s zip codes, we are ready to start exploring nearby venues using the Foursquare API. I used my merged table to pass all the relevant zip codes to see nearby venues, setting the search parameters to a radius of 500 meters and limiting the results to 100 venues. I grouped the results by zip code and arrived at the conclusion that there are 181 unique categories.

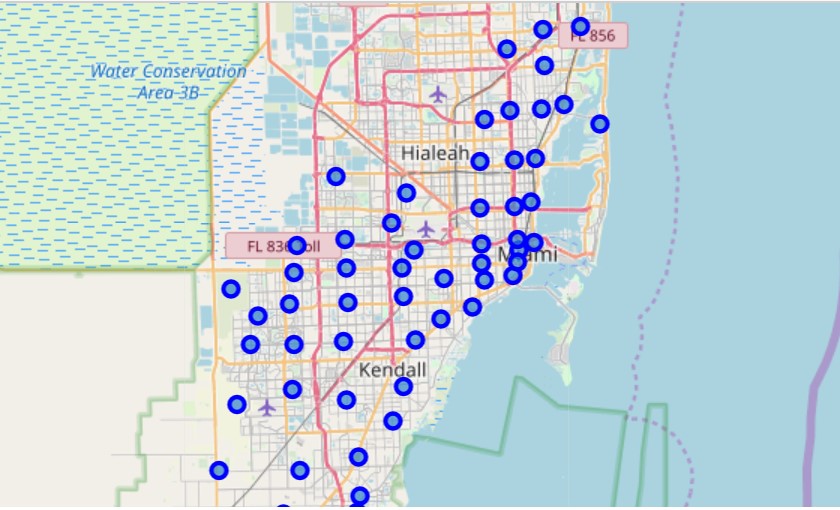
We need to be able to cluster our zip codes by venue – so we need to convert all the categorical data into numerical data. I used one-hot encoding to do so, and then I grouped rows by zip code and by taking the mean of the frequency of occurrence of each venue category. Finally, I sorted the venues into the 10 most common per zip code. A snapshot of this data frame is below:

Now the fun really begins, we are ready to start clustering our Miami zip codes! I used K Means Clustering on this data and segmented the zip codes into 15 clusters. I am using this unsupervised method because I want to let the data guide my exploration (rather than me guiding my data). I arrived at Kclusters = 15 after several rounds of trial and error. Finally, I merged the clusters generated with the property info data frame:

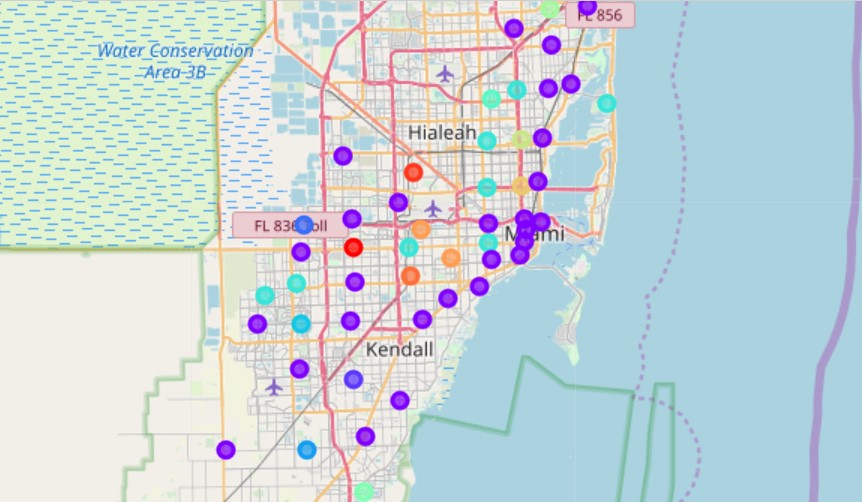


1. **Results**

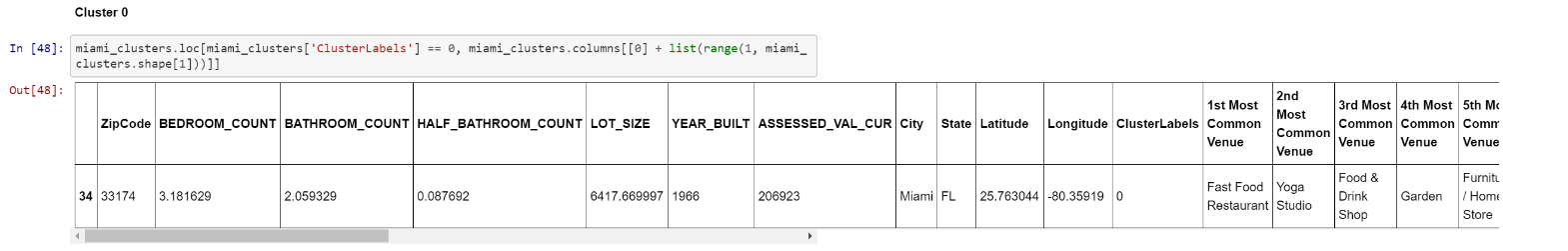
The first step toward understanding the results is to visualize the clusters on a map. I did so by leveraging Folium. The map below simply shows all of Miami’s zip codes as markers on the map.

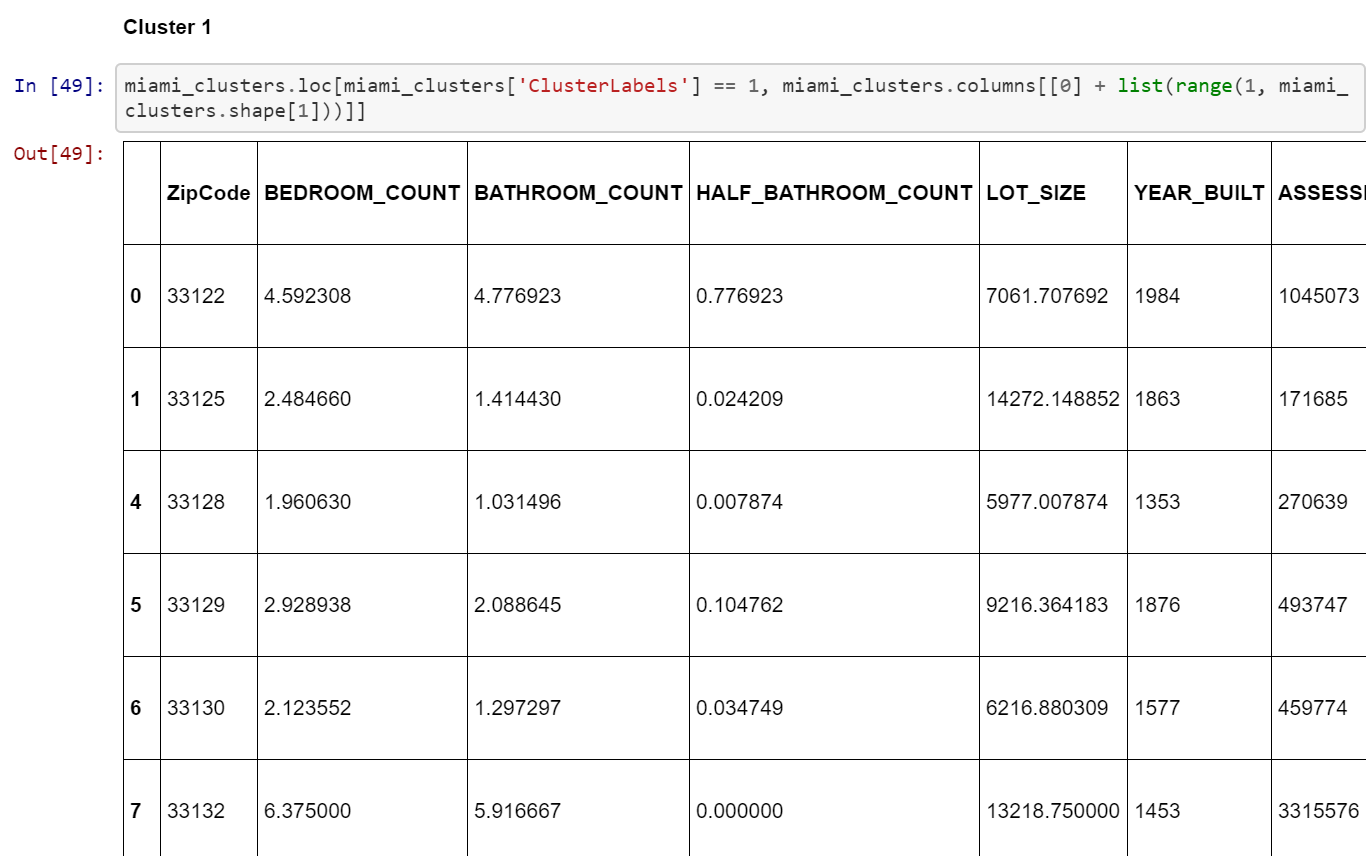


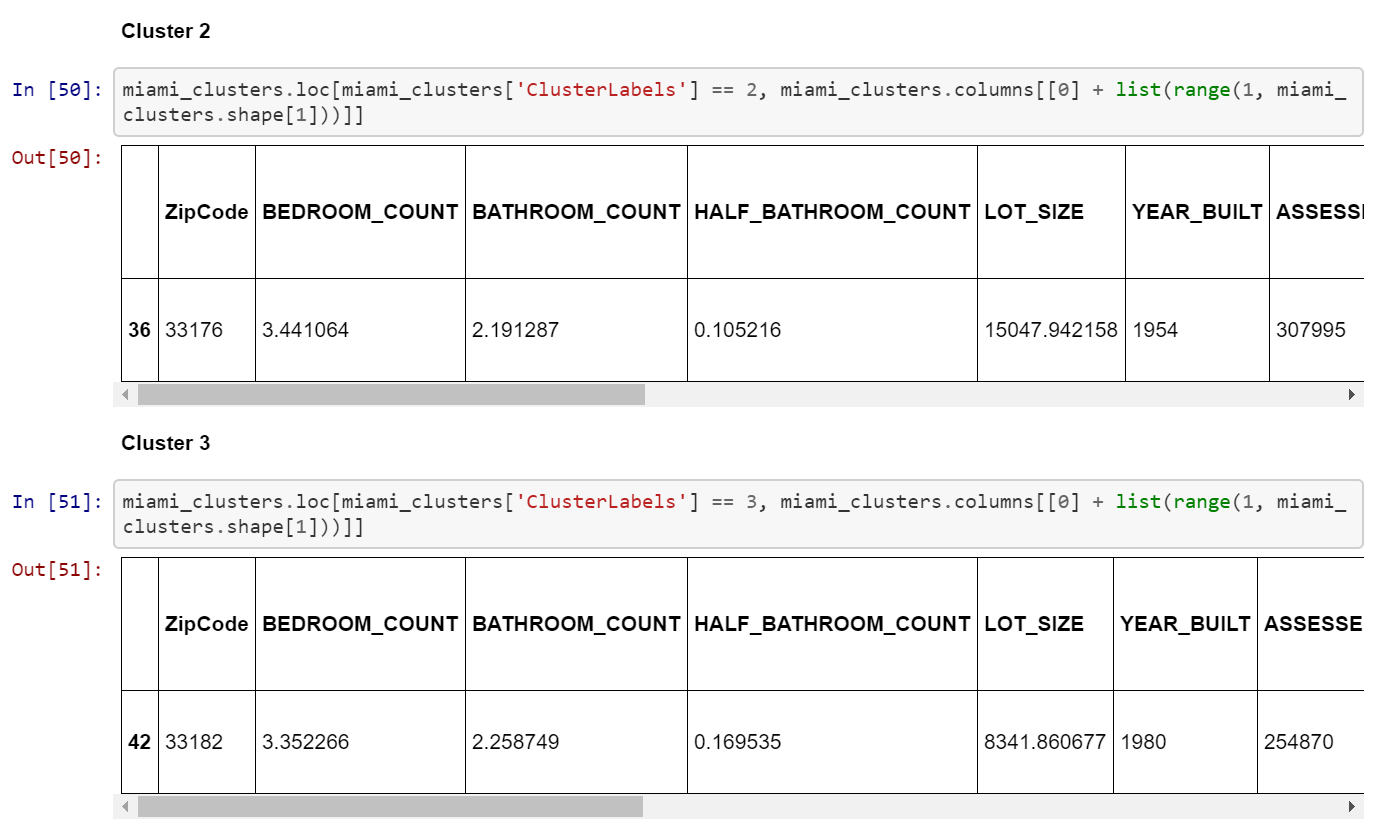
After clustering our zip codes, the map looks like this:



Just by looking at the map, we can see that one of the clusters (purple) has an abnormal looking density and contains most of the zip codes. As a result, I wanted to take a closer look at the data in each cluster. It seems that Cluster 1 (purple) is not very indicative, but Cluster 6 is a potential gold mine. The rest of the clusters are far too sparse to be useful. Below are a few images of the cluster’s associated data (which you can explore in more detail in my notebook):







1. **Discussion**

After closely looking at each cluster, it seems there is much room for improvement. Zip codes were densely clustered between Clusters 1 and 6. Cluster 1 is all over the place and doesn't provide much insight. However, Cluster 6 is our gold mine. We can see that the zip codes in Cluster 6 boast very reasonable property prices, fair bed/bath distribution and access to a plethora of awesome nearby venues. Based on this finding, I would recommend that any average buyer looking to purchase real estate in Miami, Fl further explore the zip codes in Cluster 6.

In terms of where to steer these efforts in the future, my thoughts are as follows: Given the complexity of a city like Miami, very different approaches can be tried in clustering and classification studies. However, not every method can yield the same high quality results for this metropolis. I used the KMeans Clustering because of its speed and efficacy. However, given the results it is clear that for more in-depth guidance, the data set should be expanded and the details of the neighborhood and/or street should be further looked into. I hope that as this information is refined, it would be subsequently turned into an easy to access web app that can be accessed via mobile or even a chat-based interface!

1. **Conclusion**

In conclusion, we set out to recommend Miami viable zip codes for average home buyers who want to pay reasonable prices but still have access to a good variety of Miami's venues. We did so by gathering zip code data, property data and Foursquare location data on the city of Miami. We used this information to cluster Miami zip codes and then further explore those clusters. As time goes on, I hope to refine this project so that I may share it with my friends and network of real estate agents in Miami - we have all been curious about this for as long as I can remember and currently only have anecdotal data to serve us.