The Five Clusters of Orlando, Florida

A project by Joseph Gibbons for a friend

1 Problem Introduction

My friend indicated that she has long considered living in the Orlando area but knows very little about the area. I wanted to group the neighborhoods of Orlando into 5 different groups based on the types of venues that are within a 500m radius from the center.

2 Dataset

I started with a list of about 120 neighborhoods which are technically considered to be in the Orlando city limits from Wikipedia. I cross-checked the neighborhoods with the official government site to make sure there were no errors. I deleted about 40 of these neighborhoods because they weren't geographically relevant (they were really in another county but only considered Orlando because there was an airport near them) and/or had less than 3 venues returned from the Foursquare API.

I ran each of the neighborhoods through a python library which finds and saves latitudes and longitudes and then created a map (right) with markers on each of the neighborhoods to get a better sense of the area. I then ran each of the coordinates through the Foursquare API in order to get information about venues in the area. I used a radius of 500m and a maximum number of venues per neighborhood of 100 (although the highest that was returned was only 74). This yielded a table of 85 neighborhoods and a total of 232 distinct venue categories.

VenueCategory	
BakeriesAndDesserts	47
BooksAndGames	9
Cafes	33
CraftAlcohol	15
CulturalCenters	42
Hotels	30
Misc	62
Nightlife	60
OutdoorLeisure	90
RestaurantsAmerican	149
RestaurantsAsian	44
RestaurantsEuropean	67
RestaurantsLatin	34
Services	101
Shops	196
ShopsFood	31
Sports	53

I noticed that many categories overlapped (eg. 'Asian Restaurant', 'Japanese Restaurant', 'Sushi Restaurant', 'Ramen House') so I manually went through all 232 categories and split them into 17 categories based on similarity (also interest since I know my friend is not a big clubber but does enjoy video games and books so I separated 'nightlife' from more relaxing breweries/wineries and also separated bookstores/libraries and video game stores from other shops). This resulted in the venue table above, which also shows the amount of venues which fall into each of the individual categories.

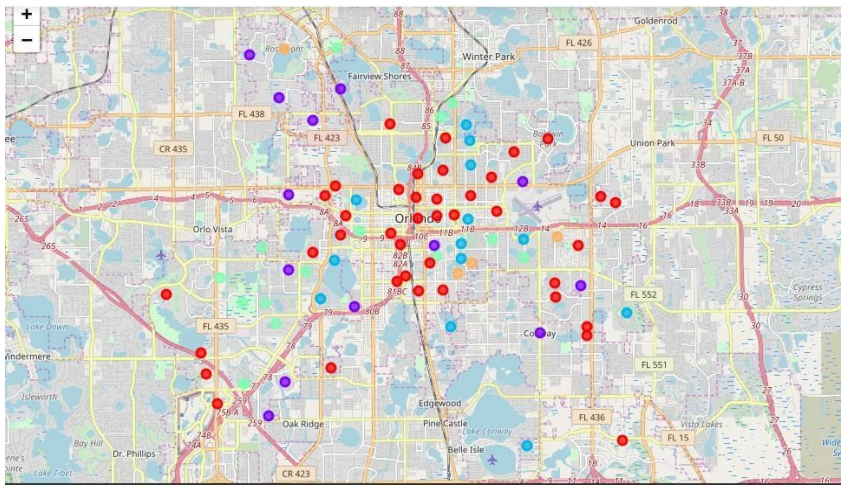
Finally, I looked at the top 3 venue categories for each neighborhood to see if anything stood out before running my KMeans algorithm. I noticed that just about every neighborhood had Shops, Services and RestaurantsAmerican in their top 3 categories. This makes sense because these are also the most frequent categories in general, making up about ⅓ of the total venues in Orlando. In the future, I would probably split those up even more, but due to the deadline, I had to accept it as it was.

3 Methodology

I used the KMeans++ clustering algorithm from scikit-learn in order to group up the neighborhoods by similarity. This algorithm plots all of the categories as dimensions on a 17-dimension scatter plot (because there were 17 categories). The rest gets kind of complicated, but basically I told the algorithm that I want 5 clusters so it picked 5 random centerpoints for the clusters and then grouped the clusters based on nearness to the random centerpoint. I then told the algorithm to run 100 times, thus creating the 5 centerpoints 100 times, and it recorded the cluster number each time it ran. The final result of the algorithm is the cluster number for each neighborhood in which the neighborhood appeared the most in those 100 times.

4 Results

I took the results from my KMeans++ algorithm and plotted them onto the same map as above (in the Dataset section) but this time they were color-coded based on cluster.



Cluster 0 -Red
Cluster 1 -Purple
Cluster 2 -Blue
Cluster 3 -Green
Cluster 4 -Orange

On the following page is a table of the neighborhoods in each cluster (I will explain the similarities / differences in the Discussion section below).

0	1	2	3	4
'Airport North' 'Audubon Park' 'Azalea Park' 'Baldwin Park' 'Callahan' 'Central Business District' 'College Park' 'Colonialtown South' 'Coytown' 'Dover Estates' 'East Central Park' 'Florida Center' 'Haralson Estates' 'Hibiscus' 'Holden Heights' 'Holden/Parramore' 'Kirkman South' 'Lake Copeland' 'Lake Dot' 'Lake Eola Heights' 'Lake Formosa' 'Lake Fredrica' 'Lake Sunset' 'Mariners Village' 'MetroWest' 'Monterey' 'North Orange' 'Park Central' 'Park Lake/Highland' 'South Division' 'South Eola' 'South Orange' 'South Semoran' 'Thornton Park' 'Uptown' 'Wadeview Park' 'Washington Shores' 'West Colonial' 'Westfield' 'Windhover'	'Camellia Gardens' 'Colonial Town Center' 'Conway' 'Dixie Belle' 'Lake Cherokee' 'Lake Shore Village' 'Mercy Drive' 'Millenia' 'Princeton/Silver Star' 'Rio Grande Park' 'Roosevelt Park' 'Seaboard Industrial' 'Signal Hill'	'Catalina' 'Clear Lake' 'Colonialtown North' 'Crescent Park' 'Lake Davis/Greenwood' 'Lake Underhill' 'Lake Weldona' 'Lawsona/Fern Creek' 'Rock Lake' 'Rose Isle' 'Rowena Gardens' 'Southern Oaks' 'Ventura'	'33rd Street Industrial' 'Delaney Park' 'Dover Shores West' 'Eagles Nest' 'Florida Center North' 'Johnson Village' 'Kirkman North' 'Lake Como' 'Lake Fairview' 'Lake Richmond' 'Lake Terrace' 'Lorna Doone' 'New Malibu' 'Orwin Manor' 'Vista Park'	'Bel Air' 'Lancaster Park' 'Rosemont' 'The Dovers'

5 Discussion

Okay, so now we know which neighborhoods are in what cluster. Cool. BUT WHY? Well, here is a quick breakdown of each of the clusters.

Cluster 0 - This is by far the largest cluster and I think it is due to the wide variety of venue categories in each of these neighborhoods. None of these neighborhoods cleanly fit one type of style so they were all grouped together. I call this cluster the amalgamation cluster since it is a big group of neighborhoods with all sorts of different styles of venues. While a majority of the venues here are shops, American restaurants, and services, it can't be said that these neighborhoods are solely based around those.

Cluster 1 - This cluster is dominated by shops. Every neighborhood in this group has its number 1 most common venue category as 'shops', while the 2nd and 3rd most common are a wide variety of restaurants.

Cluster 2 - This cluster is all about outdoor activities. The number 1 venue category for all of their neighborhoods is 'outdoor leisure,' making it the perfect place to live if you love parks, trails, lakes, and overlooks.

Cluster 3 - This cluster has a healthy mix of sports, outdoor leisure, and cafes. This means parks, stadiums, gyms, coffee, and tea rooms. The most important neighborhood of note here is '33rd Industrial Street' which is the only neighborhood to have the category 'Books and Games' in its top 3 categories (it sits at number 2). This cluster also has a few cultural centers which sounds nice.

Cluster 4 - This is the smallest cluster with only four neighborhoods. The top category for each of these neighborhoods is 'misc' which ranges from bank to storage facility to construction & landscaping companies. These neighborhoods also have some outdoor leisure venues, but it seems that more than anything they just have random places.

6 Conclusion

I feel like clusters 2 and 3 are the best fit for my friend because they have a lot of outdoor activities which will help when she is in a writing slump. Also, these neighborhoods are just outside of the city center so it looks like they have more of a suburban vibe which would be nice since my friend is looking to buy a house in the future.