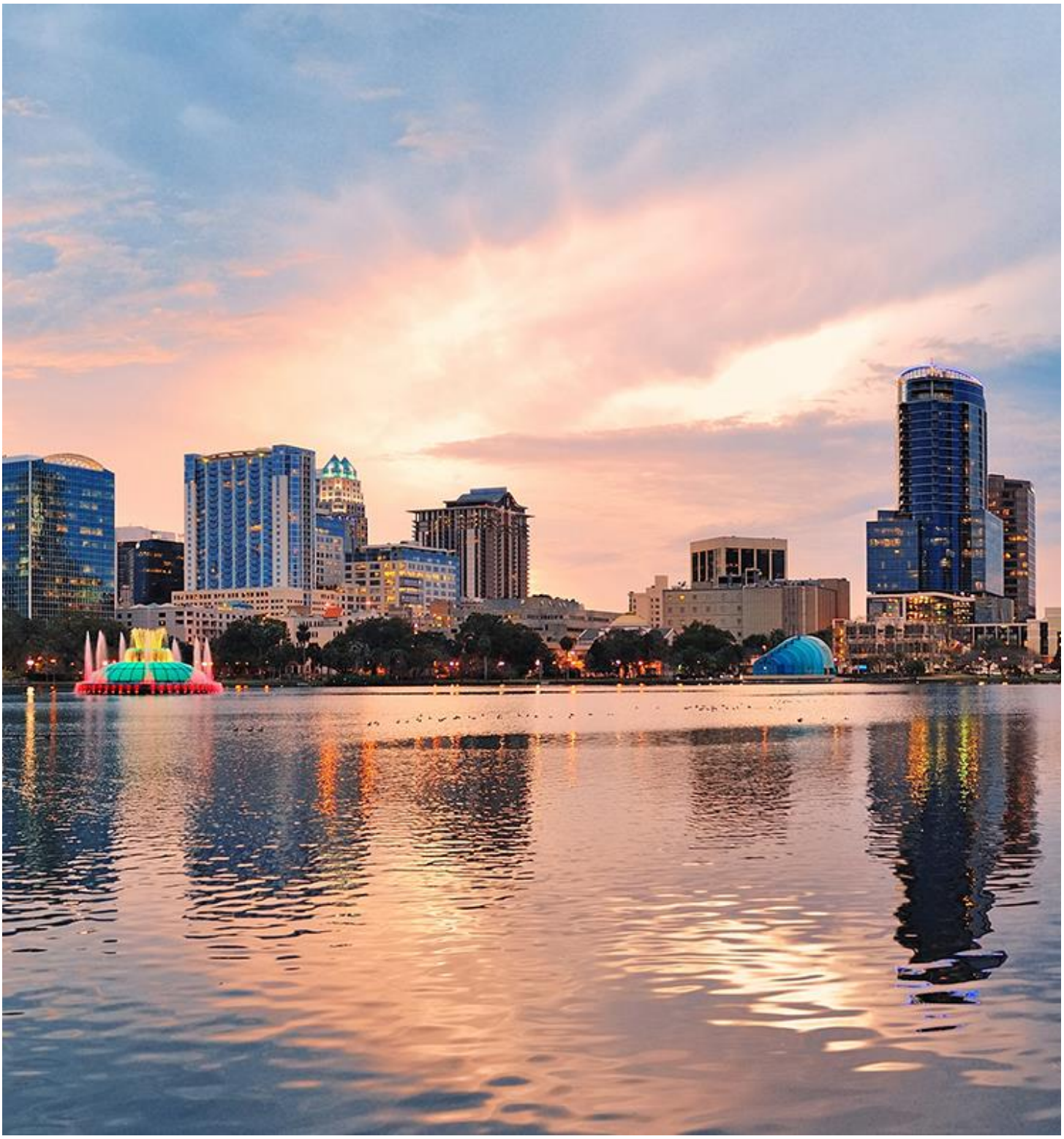
# Exploring Neighborhoods for optimum housing in Orlando, Florida

IBM Data Science Capstone Project - Suhail Jaffer - May 2020



# 1. Introduction

## 1.1 Background

Orlando is a city in Florida known for its theme parks, most notably Disney World and Universal Studios. Orlando is an attractive site for investment because of its vast availability of social venues and developing economy. It is for these reasons and more that attract individuals to settle in Orlando's neighborhoods. Because of this, the housing market in Orlando is fairly competitive. Redfin, a real estate company, rate Orlando's market a score of 77/100, a grade they deem as highly competitive. Similarly, Zillow rates the market as 'Very Hot'. According to Redfin, average housing prices have risen by 13% since last year, with the average price per square foot increasing by 8.2%. These metrics seem to back both companies assessment of Orlando's housing market.

## 1.2 Problem

With a sizable list of homes available for sale or rent in various neighborhoods, home buyers need access to insightful information to narrow down their options based on their preferences and budget. This brings up a relevant question: How could home buyers leverage data to find optimum neighborhoods that match their preferences and budget?

# 2. Methodology

## 2.1 Data Sources

The main data source for an overview of the housing market in each neighborhood was accessed from Realtor.com. One drawback of this dataset is that it does not include a comprehensive list of data from every neighborhood in Orlando. Rather, it only includes data gathered from listings available on Realtor in a limited number of neighborhoods. Attempts were made to find a more complete data source for the housing market, but the one available from Realtor seemed the most complete. To complete the dataset for analysis, data for nearby venues in each neighborhood was obtained using the Foursquare API. Lastly, for the map visualization, the GEOJSON file for Orlando was downloaded here.

## 2.2 Data Processing

Data from Realtor was put in a CSV file before being imported into a Jupyter notebook as a dataframe for analysis. Each entry included the name of the neighborhood, the median listing price, price per square foot, and homes available for rent or for sale. To obtain the appropriate coordinates for each neighborhood, the dataframe was run through a loop with the use of the Geopy library. The price per square foot data was

converted to integer values for later analysis. Fetching coordinates for each neighborhood from Geopy resulted in some coordinate mismatches, which were later corrected before visualizing the map.

| | Neighborhoods | Median Listing Price | $/SqFt | Homes For Sale | For Rent | Latitude | Longitude |
|---|---|---|---|---|---|---|---|
| 0 | Metro West | $178K | $136 | 146 | 263 | 28.519226 | -81.333090 |
| 1 | College Park | $399.5K | $252 | 126 | 53 | 28.574478 | -81.390258 |
| 2 | Vista East | $300K | $131 | 51 | 52 | 28.459393 | -81.245251 |

Fig.1: Dataframe after coordinate matching

Data from Foursquare was obtained using their API. Data gathered from Foursquare included nearby venues and their categories with a distance limit of 5000m and a max venue limit of 100 for each neighborhood.

| | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| 0 | Metro West | 28.519226 | -81.33309 | Pizza Bruno | 28.523953 | -81.335410 | Pizza Place |
| 1 | Metro West | 28.519226 | -81.33309 | Keke's Breakfast Café | 28.522006 | -81.329128 | Bistro |
| 2 | Metro West | 28.519226 | -81.33309 | Zaza New Cuban Diner | 28.524099 | -81.340018 | Cuban Restaurant |

Fig.2: Dataframe after Foursquare API calls

Venue names were then dropped as we are only interested in the venue category for this analysis. The data was then one hot encoded where the string values of venue categories were converted to binary values. Data was then grouped by neighborhood and the ten most frequent venue categories in each neighborhood were then extracted and put in a new dataframe for cluster analysis and then split into their respective clusters. After cluster analysis, a merged dataframe was created with insights gained after cluster analysis and earlier analysis.

| Median Listing Price | $/SqFt | Homes For Sale | For Rent | Latitude | Longitude | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 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 | Cluster Name | Price per SqFt Category |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $178K | 136 | 146 | 263 | 28.519226 | -81.333090 | Park | Convenience Store | Mexican Restaurant | Italian Restaurant | Bakery | Cosmetics Shop | Bar | Sandwich Place | Café | American Restaurant | Mixed Use | Low |
| $399.5K | 252 | 126 | 53 | 28.574478 | -81.390258 | Grocery Store | Coffee Shop | Café | American Restaurant | Asian Restaurant | Italian Restaurant | Bar | Seafood Restaurant | Park | Sandwich Place | Restaurants and Parks | High |
| $300K | 131 | 51 | 52 | 28.459393 | -81.245251 | Convenience Store | Fast Food Restaurant | Gym / Fitness Center | Pizza Place | Video Store | Pharmacy | Discount Store | Donut Shop | Bank | Department Store | Mixed Use | Low |

Fig.3: Final Dataframe (not shown: Neighborhoods column)

In addition, some neighborhood names and coordinates were corrected before visualizing the map to match those found in the GEOJSON file.

# 3. Results and Discussion

## 3.1 Cluster Analysis

K-means Clustering is an unsupervised machine learning algorithm that partitions samples into a specified number of clusters. K-means clustering helps group similar data points together to discover underlying patterns. Neighborhoods in the dataset were clustered into 5 clusters based on the ten most common venues. The appropriate number of clusters was checked using the elbow method of minimizing distortion.

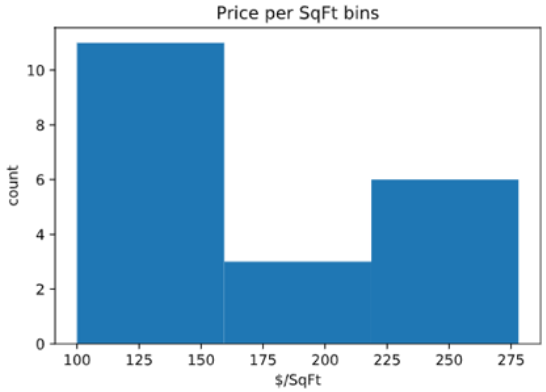
| | Neighborhoods | Median Listing Price | $/SqFt | Homes For Sale | For Rent | Latitude | Longitude | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Metro West | $178K | 136 | 146 | 263 | 28.519226 | -81.333090 | Park | Convenience Store | Mexican Restaurant | Italian Restaurant | Bakery | Cosmetics Shop | Bar | Sandwich Place | Café | American Restaurant |
| 1 | College Park | $399.5K | 252 | 126 | 53 | 28.574478 | -81.390258 | Grocery Store | Coffee Shop | Café | American Restaurant | Asian Restaurant | Italian Restaurant | Bar | Seafood Restaurant | Park | Sandwich Place |
| 2 | Vista East | $300K | 131 | 51 | 52 | 28.459393 | -81.245251 | Convenience Store | Fast Food Restaurant | Gym / Fitness Center | Pizza Place | Video Store | Pharmacy | Discount Store | Donut Shop | Bank | Department Store |
| 3 | Florida Center North | $150K | 130 | 90 | 106 | 28.484958 | -81.447864 | Theme Park Ride / Attraction | Clothing Store | Hotel | Pizza Place | Furniture / Home Store | Theme Park | Coffee Shop | Mexican Restaurant | Fast Food Restaurant | Sporting Goods Shop |
| 4 | Lake Nona South | $499.9K | 199 | 79 | 48 | 28.377325 | -81.261730 | Fast Food Restaurant | Sandwich Place | Golf Course | Convenience Store | New American Restaurant | Pizza Place | Hardware Store | Donut Shop | Shipping Store | Mexican Restaurant |

Fig.4: A preview of Cluster 0 (Restaurants and Parks)

The clusters were then evaluated, and the following cluster labels were created:

- Cluster 0: Restaurants and Parks
- Cluster 1: Restaurants and Community Services
- Cluster 2: Theme Parks
- Cluster 3: Restaurants and Stores
- Cluster 4: Mixed Use

This categorization would help home seekers narrow down their choices by looking into neighborhoods that appeal to their nearby venue preferences. Furthermore, Neighborhoods in each cluster were further categorized by price range using price per square foot as a measure. Price per square foot was chosen over median listing price as it is a better measure to use when comparing prices of houses in different neighborhoods. The price range was split into three categories:




- Low ($100 < X < $159.3)
- Average ($159.3 < X < $218.67)
- High ($218.67 < X < $278)

Fig.5: Price per square foot being categorized into 3 distinct categories

| Price per SqFt Category | Neighborhoods |
|---|---|
| Low | 2 |
| Average | 0 |
| High | 3 |

| Price per SqFt Category | Neighborhoods |
|---|---|
| Low | 1 |
| Average | 0 |
| High | 0 |

| Price per SqFt Category | Neighborhoods |
|---|---|
| Low | 3 |
| Average | 2 |
| High | 1 |

| Price per SqFt Category | Neighborhoods |
|---|---|
| Low | 1 |
| Average | 0 |
| High | 1 |

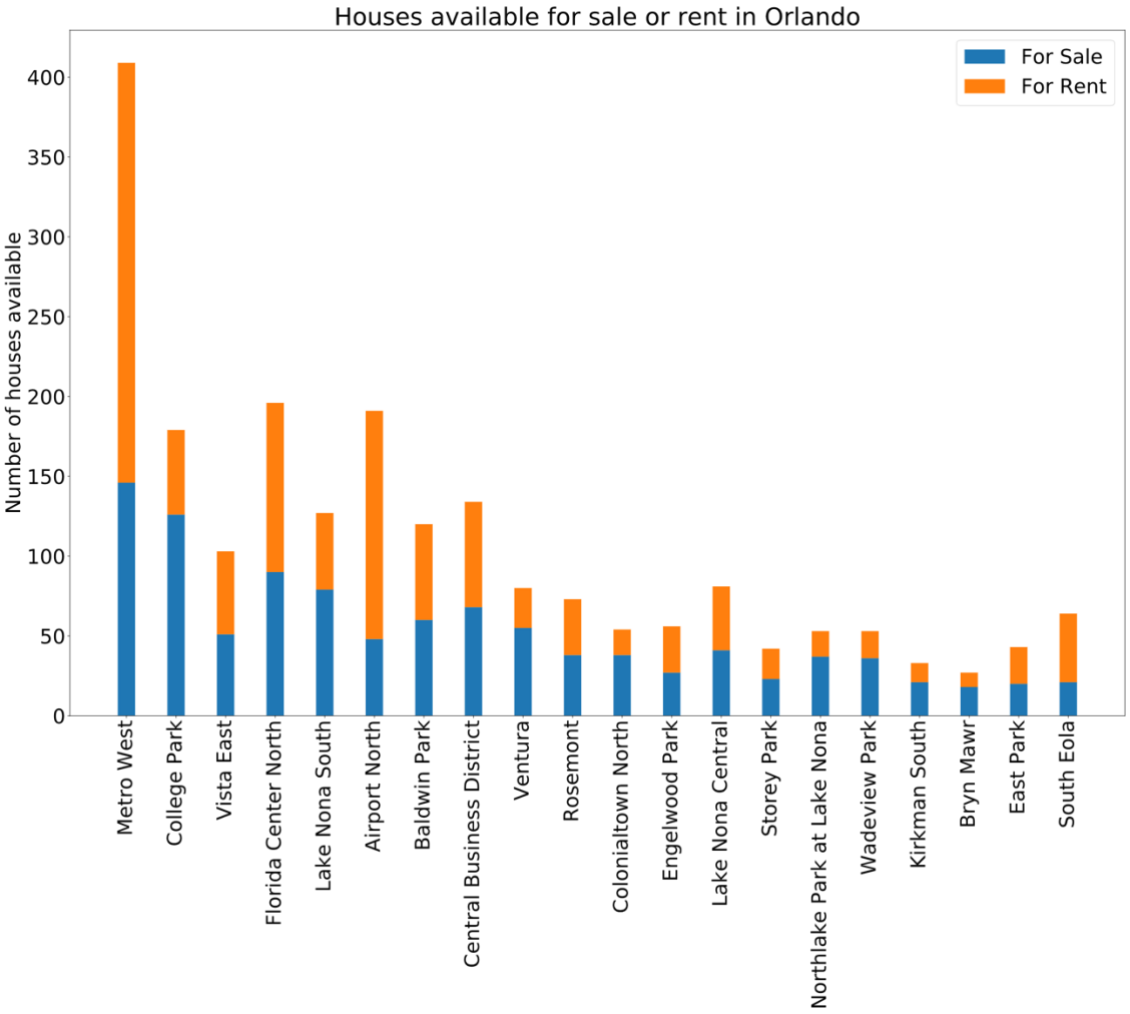
| Price per SqFt Category | Neighborhoods |
|---|---|
| Low | 4 |
| Average | 1 |
| High | 1 |

Fig.6-10: Breakdown of neighborhoods by price category in each cluster (0-4, left to right, top to bottom)

From this price categorization, we can examine the underlying patterns of each cluster in terms of price points. Cluster 0 (Restaurants and Parks) seems to have neighborhoods that are split between low and high price categories. Cluster 1 (Restaurants and Community Services) has one neighborhood in the low-price category. Cluster 2 (Theme Parks) has its neighborhoods in the low-average price category. Cluster 3 (Restaurants and Stores) is split between the low and high price categories, similar to Cluster 0. Lastly, Cluster 4 (Mixed Use) has the majority of its neighborhoods in the low-price category. It is important to note that the clusters might have had a different composition if all 121 neighborhoods in Orlando were included.

## 3.2 Availability of housing

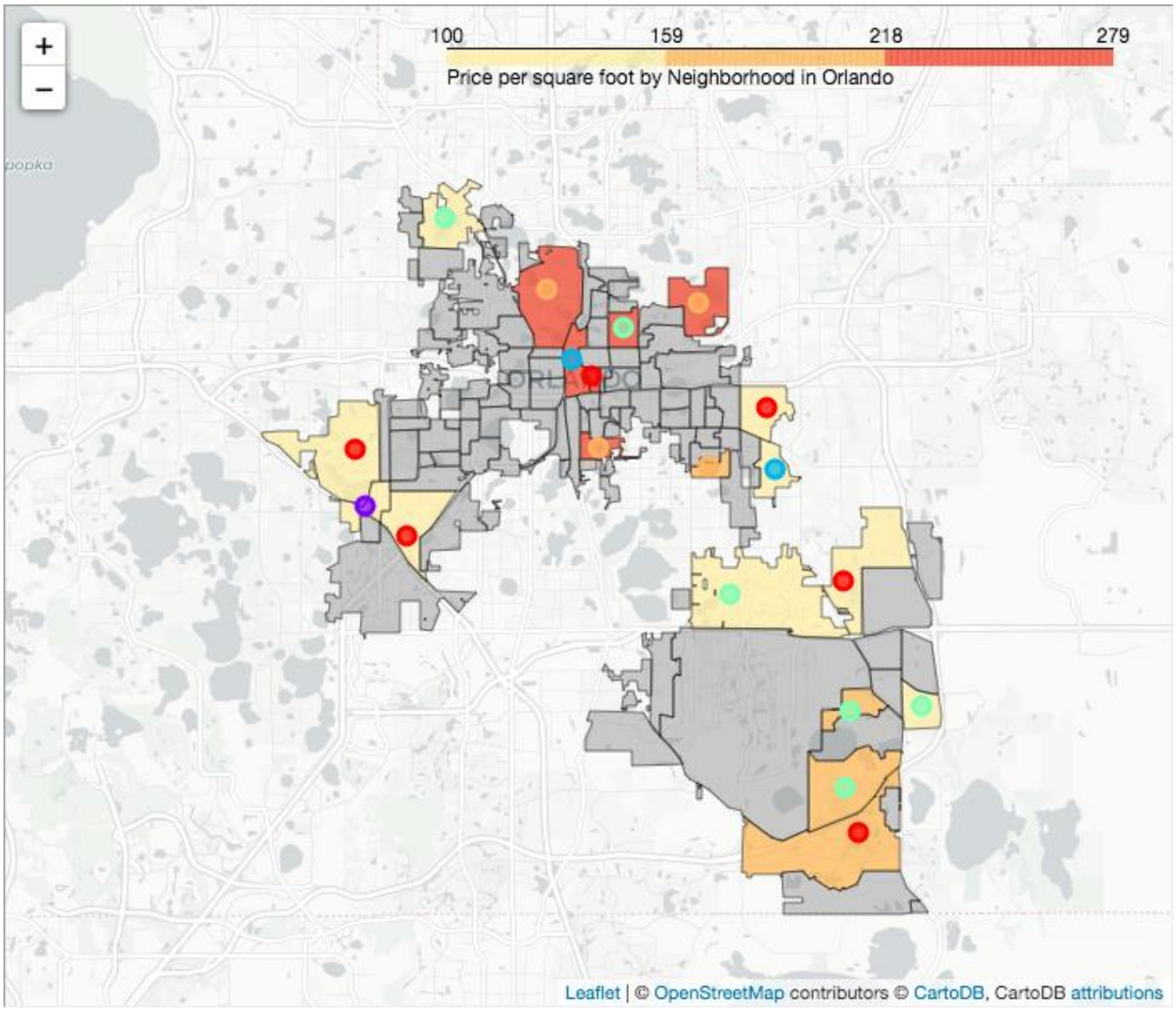
To get an idea of the availability of housing in different neighborhoods in Orlando, I visualized the proportion of houses available for sale or rent in a bar chart.




As seen on the chart, there is a significant amount of homes available for either sale or rent in neighborhoods across Orlando. The mean of houses available for sale is 52 houses while the mean of houses available for rent is 54 houses. Metro West has the highest number of houses available for sale and for rent, indicating that it is a viable option to look into for individuals interested in neighborhoods in the 'Restaurants and Parks' cluster because of its relatively competitive housing market.

### 3.3 Map Visualization

The results of the clustering and the price categorization was then visualized on a map to evaluate geographic trends and the spread of the clustered neighborhoods.




The clusters are visualized in the following manner:

- Cluster 0 (Red): Restaurants and Parks
- Cluster 1 (Purple): Restaurants and Community Services
- Cluster 2 (Blue): Theme Parks
- Cluster 3 (Green): Restaurants and Stores
- Cluster 4 (Orange): Mixed Use

As shown on the map, we can observe some correlation between the geographic location of neighborhoods and their price category. We can observe that the neighborhoods in the high-price category are located in the same general area. Similarly, Neighborhoods in the west and in the south show a similar

pattern. This could indicate that neighborhoods within proximity to each other would to an extent, have a similar price per square foot range. There appears to be a weaker correlation in terms of the proximity of neighborhoods in the same cluster. It is important to note that this visualization does not give a complete picture on the geographic patterns in Orlando due to the limited data available but regardless, the map gives an insight into the aforementioned trends.

# 4. Limitations

## 4.1 Limitations

There are a number of limitations to be considered when attempting to generalize the results of this analysis on neighborhoods in Orlando. Firstly, the data does not include all 121 neighborhoods in Orlando. This significantly limits the results of this analysis as it was performed on only 20 neighborhoods. The clustering obtained from this analysis might not reflect the clustering that would have been obtained if data for all neighborhoods in Orlando was used. Furthermore, the clusters might have been allocated differently if more than one variable was used. The only variable used for clustering was common venue categories. If other variables such as crime data or ratings of nearby schools were included in the clustering parameters, the allocation would be different. Lastly, the data was obtained from a single real estate source: Realtor.com. This means that the metrics are based on only houses listed on their website and may not necessarily reflect the actual condition of the housing market in those neighborhoods.

# 5. Conclusion

## 5.1 Conclusion

Within the scope of the data analyzed, we have uncovered certain trends within the housing market in Orlando. Using K-Means clustering, we were able to cluster neighborhoods into distinct categories to help home seekers narrow down their search by providing them with neighborhoods that are similar to each other in terms of common venues in each neighborhood. As the Orlando housing market continues to remain competitive, home seekers need to be able to quickly narrow down their choices to a select few listings to be able to have a first-mover advantage over other home seekers. Analysis has revealed that Metro West would be a viable option for home seekers interested in the 'Restaurants and Parks' cluster because of its relatively competitive housing market compared to other neighborhoods with a lower listing count. Lastly, visual analysis of the map visualization has revealed that some neighborhoods in the same price range are within proximity of each other, providing evidence that neighborhoods close to one another could be similarly priced regardless of what cluster they're in. The insights from this analysis could be used to help home seekers evaluate their choices better when deciding what neighborhood to settle in based on their preferences and budget.