**IBM Data Science Professional Certificate**

**Capstone Project**

**Peter Cleavland Palmer**

### **Table of Contents**

* Introduction
* Data
* Methodology
* Results
* Discussion
* Conclusion

### **Introduction**

P. James, an electrician with over a decade of experience, working on residential and commercial projects, wants to start an electrical company in Orlando, Florida. He has done electrical work before in Orlando, but he is not sure where in Orlando is the best place to set up an office. He knows that he will need financing and so one criteria that he is looking for is proximity to a lending institution, like a bank or credit union. It would also be great if there are not many registered electrical companies in the area, reducing competition. It is possible, with the relevant licenses, for an electrical company to be located in one county but the electrician does work in another county, or several other counties.

### **Business Problem**

The objective is to help Mr. James make the best decision, about office space, by using the Foursquare location data, data science techniques, webscraping techniques, and machine learning algorithm. With these tools we will try to answer the question: can we determine the best neighborhood in Orlando for Mr. James to set up an electrical company?

### **Data**

In order to solve this problem, we will need the following data:

* a list of neighborhoods in Orlando;
* latitude and longitude of the neighborhoods;
* venue data, from Orlando, identifying banks and electrical companies.

The list of neighborhoods in Orlando can be found on the wikipedia webpage, <https://en.wikipedia.org/wiki/Category:Neighborhoods_in_Orlando,_Florida>. The page has the neighborhoods contained in a bulleted list, showing twenty six neighborhoods. To access this data we will use webscraping techniques which includes the beautifulsoup package. For the coordinates of the neighborhoods, latitude and longitude, we will use the geocoder package, and then join these in a single data frame. We will also need the latitude and longitude of the neighborhoods so maps can be plotted and markers will be placed on these maps so that the different clusters can be visualized. Finally, the Foursquare API will help us with information that can be used to drill down on the neighborhoods and surrounding venues.

### **Methodology**

The methodology can be summarized under eight subsections: import and install libraries, scrape the website for the neighborhoods, get the coordinates of the neighborhoods, create a map of Orlando, Florida, explore the neighborhoods in Orlando, analysis of the neighborhoods, form the clusters and use the clustering algorithm, and examine the clusters.

The necessary libraries must be either imported or installed. Each tool that will be used requires a library, and without this library we will get error messages, in the notebook, and the analysis cannot be completed in this case. The libraries can be grouped as webscraping libraries, data frame and data manipulation libraries, machine learning libraries, and plotting libraries.

The list of neighborhoods in Orlando can be found on the wikipedia webpage, <https://en.wikipedia.org/wiki/Category:Neighborhoods_in_Orlando,_Florida>. The page has the neighborhoods contained in a bulleted list, showing twenty six neighborhoods. To access this data we will use webscraping techniques which includes the beautifulsoup package. We start the webscraping process by sending a GET request to the wikipedia webpage and save the information in variable and use beautifulsoup to parse the data from the webpage. Now we create an empty list to store neighborhood data that will be retrieved from the webpage. The data on the wikipedia page can be found in a div tag with class "mw-category", and then in a list with the li tag. We end the webscraping part by appended the data into an empty list, and create a data frame, which we call df.

We need the coordinate of the neighborhoods in Orlando so that we can create the map of Orlando. Now, to get the coordinates of the neighborhoods, we define a function, then after we call the function the data will be stored in a data frame. Therefore, we now have two separate data frames, one with the neighborhoods and one with the coordinates. The final thing here is to merge these two data frames into one data frame.

We create a map of Orlando. Before the map is created, we find the coordinates of Orlando using geolocater. Now that we have the coordinates, we can use folium to draw the map, and we also superimpose the markers of the neighborhoods, on the map of Orlando.

We explore the neighborhoods in Orlando, Fl. by using the Foursquare API. Foursquare credentials, client id, client secret, and version, are necessary and must be obtained by setting up an account on Foursquare. Use your credentials and version to get the top 100 venues that are within a radius of 1000 meters and store them. Now create a data frame from the list of venues and check the first five rows. Some of the venue categories that we see in the first five rows of the venues data frame are electronic store, Latin American Restaurant, breakfast spot, and bakery. We can determine the number of venues returned for each neighborhood. This gives us an idea of the density of the businesses in the neighborhood. We can determine the number of unique categories and print some of them. Are there any electrical companies nearby? It shows that there is none.

We do an analysis of the neighborhoods in Orlando by using one hot encoding. Include the neighborhoods column in the one hot encoding data frame, make this into a smaller data frame by grouping the rows by the neighborhood, and taking the mean of the frequency of each category. Then we look at the banks only to meet the criteria set by Mr. James.

Form the clusters and use the k-means algorithm. We use three clusters and fit the k-means clustering algorithm. Create a new data frame with the clusters and the top 10 venues for each neighborhood, and include the clustering labels, 0, 1, and 2. We then merge the data frames and add the coordinates of each neighborhood. Next the data frame is sorted by the cluster labels 0, 1, 2. Lastly, in this section, we create a map of Orlando but this time showing the showing the clusters. Finally, we examine the clusters individually.

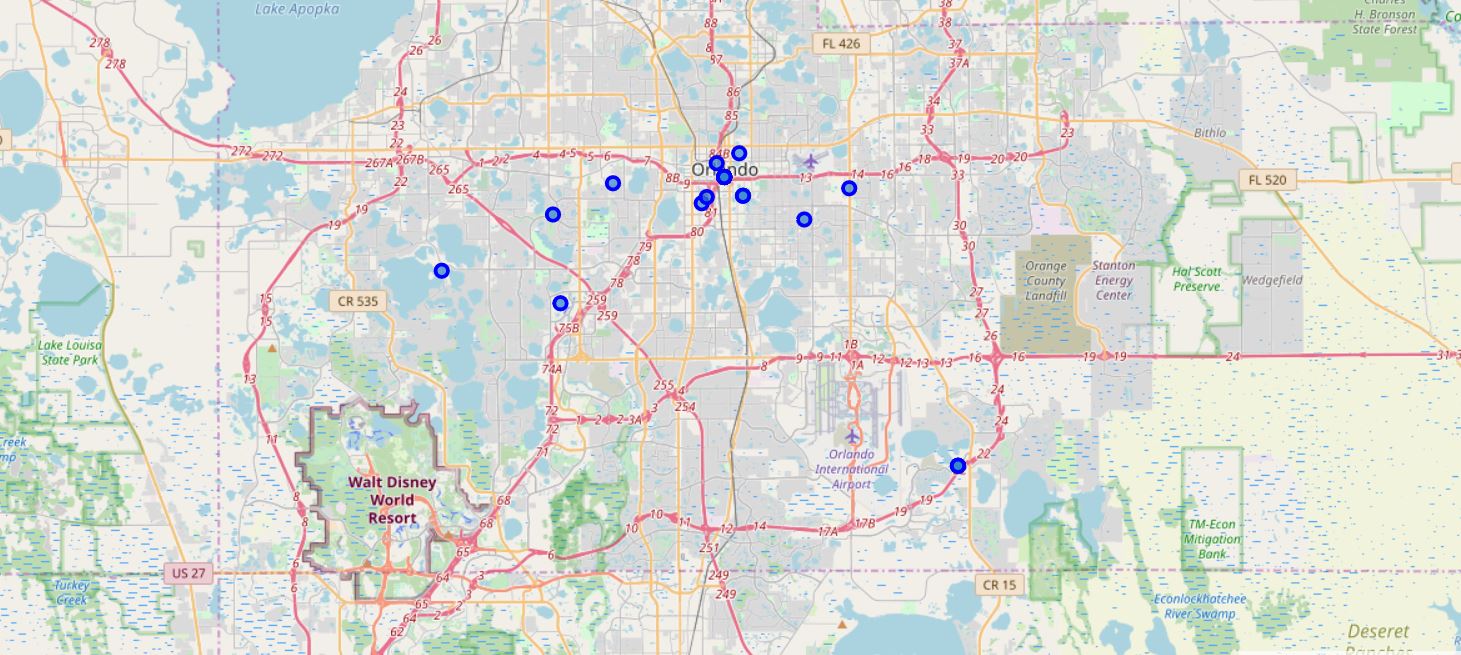
### **Results**

The results obtained by using the k-means clustering algorithm show that we can categorize the neighborhoods, in Orlando, into three categories based on the frequency of occurrence of banks. The following are the categories:

* cluster 0: neighborhoods in Orlando with no bank;
* cluster 1: neighborhoods in Orlando with a low number of banks;
* cluster 2: neighborhoods in Orlando with a moderate number of banks.

Therefore, Mr. James can choose, to set up his company office, in the neighborhoods in cluster 1 or cluster 2.

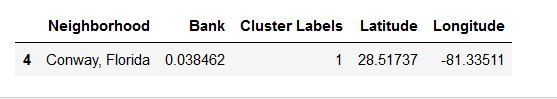
**Map of Orlando with Neighborhood Markers**



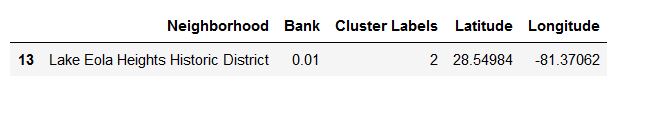
**Neighborhoods in Cluster 0**

****

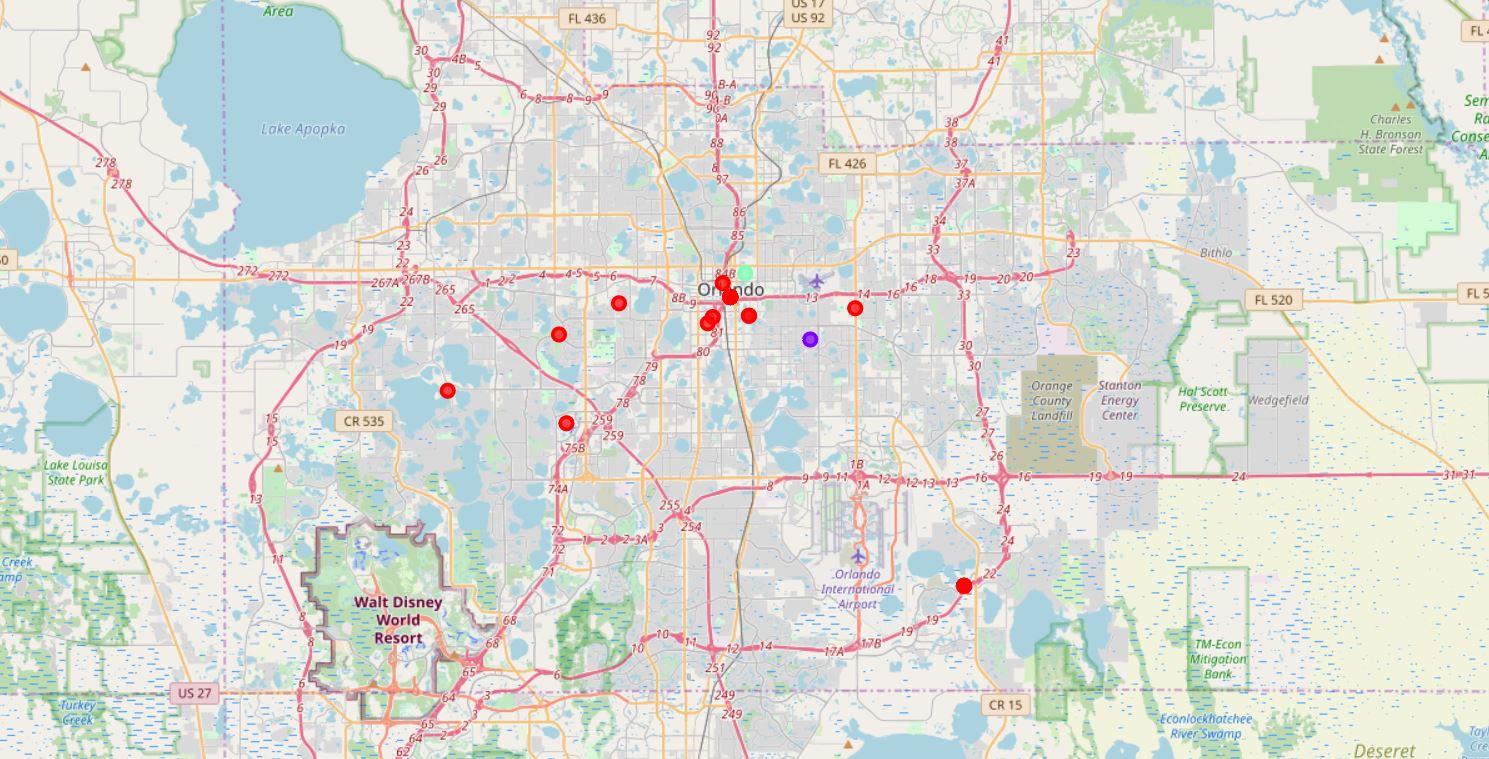
**Neighborhoods in Cluster 1**

****

**Neighborhoods in Cluster 2**



**Map of Orlando Neighborhoods in Clusters**

****

### **Discussion**

We observed that there were no electrical companies in the venues categories. Firstly, the lack of electrical companies in the clusters suggests that there is great potential for Mr. James to establishing his company in these clusters. However, I believe that there may be some error here, that there are no electrical companies in the neighborhoods, and one possibility could be the electrical companies' business names, in the neighborhoods, do not contain the word electrical. With added criteria for selecting the clusters, instead of just financial institutions, we could form better clusters. This research could be extended to dig deeper into the clusters and actually start to look at properties, available for renting. We would need other criteria about the type of space that Mr. James needs, for example price, number of rooms, amenities, and so on.

### **Conclusion**

In this project, trying to find the best neighborhood for Mr. James to set up an office, we were able to suggest two clusters of neighborhoods that would be good candidates. In order to complete the project we scraped a website for a list of the neighborhoods, got the coordinates of the neighborhoods using geolocater, created maps of Orlando, Florida, explored and analyzed the neighborhoods of interest, and used a clustering algorithm to create and examine the clusters. There are opportunities for us to do further work towards finding not just the potential neighborhood for the office but to find an actual rental property, for Mr. James.