Determining Covid Similarities in Orange County Neighborhoods

Ronnie Howard

***Abstract*—**TODO: For now see the Introduction and Data sections below

1 INTRODUCTION

In this project I have decided to look at the problem of COVID-19 outbreaks in the different neighborhoods of Orange County. I will define a **neighborhood** as each area in Orange County, Florida that has a unique zip code. Interestingly, a quick glance on zip-codes.com [1] shows that there are over 50 zip codes in the Orange County area. The plan is to cluster these neighborhoods into like neighborhoods. I will then use a Chloropleth map to show how severe the outbreaks are in each neighborhod. My hope is that similar neighborhoods will show similar amounts of COVID-19 cases so that more research can be done to determine what in particular makes a certain neighborhood have more cases than others. The **null hypothesis** is that there is no relationship between similarly clustered neighborhoods and the number of outbreaks that occur. The **alternative hypothesis** is that similarly clustered neighborhoods will have similar outbreaks of **COVID-19.**

1.1 Audiences

This is important to a variety of different audiences and I will discuss a few now. The first audience that this problem affects is the Department of Health, along with other government officials such as the mayor. This stakeholder is responsible for tracking outbreaks and quarantining different areas based on how bad the pandemic is. This stakeholder will be interested in my analysis because if my hypothesis is correct, it will allow them to zone in on neighborhood clusters that have the most outbreaks. They could use my analysis to begin further research into determining what makes some neighborhood clusters for susceptible than others.

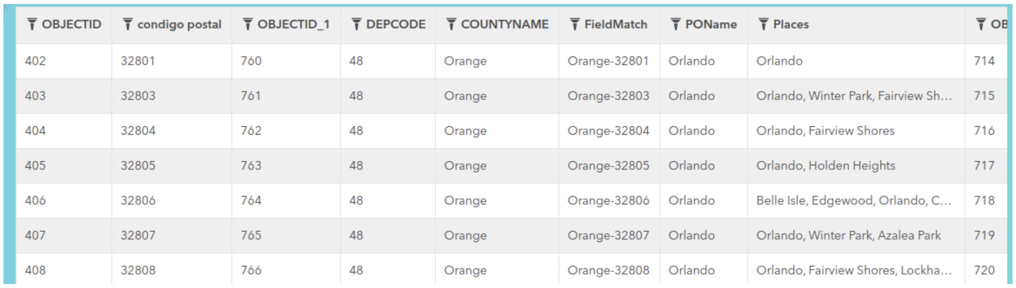
Another audience is the general population. Most people want to avoid areas that have high concentrations of outbreaks. My analysis would tell them specific types of neighborhoods that they should avoid.

Finally, I want to mention that the leading theory on outbreaks is that higher populated areas will have higher outbreaks. I agree that this is most likely true. However, my analysis goes a bit deeper because Orange County is already a largely populated county in Florida. Orange County includes Orlando which implies that Orlando's cases will most likely be higher than the rest of the county but further analysis of the types of neighborhoods may still be important.

2 DATA

There are two main sources where I plan to gather my data but time will tell if I need more as I continue on the project.

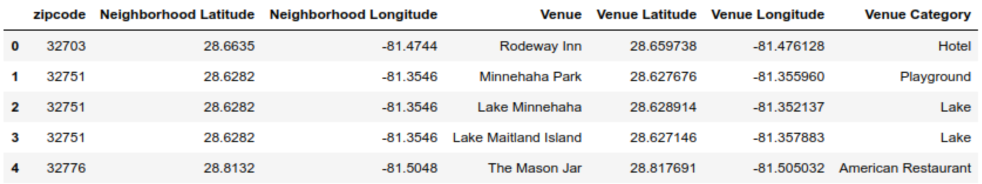
1. The Florida Department of Health maintains an open database of known cases of COVID-19 based on zipcode. This dataset can be accessed through a RESTful API. I will use this database to gather the zipcodes of Orance County and also to gather the number of reported outbreaks based on that zipcode. The COVID-19 data will be used to create a Chloropleth map based on outbreaks. An example of their dataset can be seen below.



OBJECTID	condigo postal	OBJECTID_1	DEPCODE	COUNTYNAME	FieldMatch	POName	Places	OB
402	32801	760	48	Orange	Orange-32801	Orlando	Orlando	714
403	32803	761	48	Orange	Orange-32803	Orlando	Orlando, Winter Park, Fairview Sh...	715
404	32804	762	48	Orange	Orange-32804	Orlando	Orlando, Fairview Shores	716
405	32805	763	48	Orange	Orange-32805	Orlando	Orlando, Holden Heights	717
406	32806	764	48	Orange	Orange-32806	Orlando	Belle Isle, Edgewood, Orlando, C...	718
407	32807	765	48	Orange	Orange-32807	Orlando	Orlando, Winter Park, Azalea Park	719
408	32808	766	48	Orange	Orange-32808	Orlando	Orlando, Fairview Shores, Lockha...	720

Figure 1—FDOH data on COVID-19 cases based on zipcodes in Orange County, FL [2]

2. I will use the **Foursquare API** to gather venue information within each of the aforementioned zipcodes of Orange County, FL. This data will allow me to cluster the neighborhoods in Orange County based on each neighborhood's venues. Below is an example of the type of data Foursquare provides.



	zipcode	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	32703	28.6635	-81.4744	Rodeway Inn	28.659738	-81.476128	Hotel
1	32751	28.6282	-81.3546	Minnehaha Park	28.627676	-81.355960	Playground
2	32751	28.6282	-81.3546	Lake Minnehaha	28.628914	-81.352137	Lake
3	32751	28.6282	-81.3546	Lake Maitland Island	28.627146	-81.357883	Lake
4	32776	28.8132	-81.5048	The Mason Jar	28.817691	-81.505032	American Restaurant

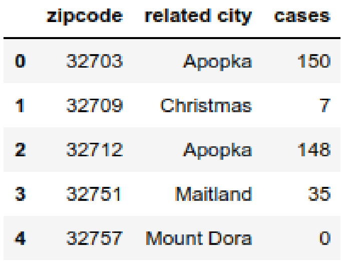
Figure 2—An example of the data Foursquare will return [3]

3 METHODOLOGY

3.1 Initial Data Wrangling

I began my analysis by getting the initial data from the Florida Department of Health's data base on Covid-19 cases in Florida. The data returned included a lot of information as shown in Figure 1 above. It returns a geojson file which includes the boundaries of the zipcodes and the number of covid cases per zipcode.

The covid-19 case data was used first and I trimmed it down so that it only showed relevant information:



	zipcode	related city	cases
0	32703	Apopka	150
1	32709	Christmas	7
2	32712	Apopka	148
3	32751	Maitland	35
4	32757	Mount Dora	0

Figure 3—*Initial dataframe of covid data.*

The plan is to use this data in the Foursquare api but Foursquare requires latitude and longitude coordinates. I decided to use geopy in order to convert the zipcode to these lat/long coordinates. Unfortunately geopy was not smart enough to convert every zipcode. Some of the coordinates it returned were not even in the USA, let alone Orange County, FL. This is shown in the Figure 4.

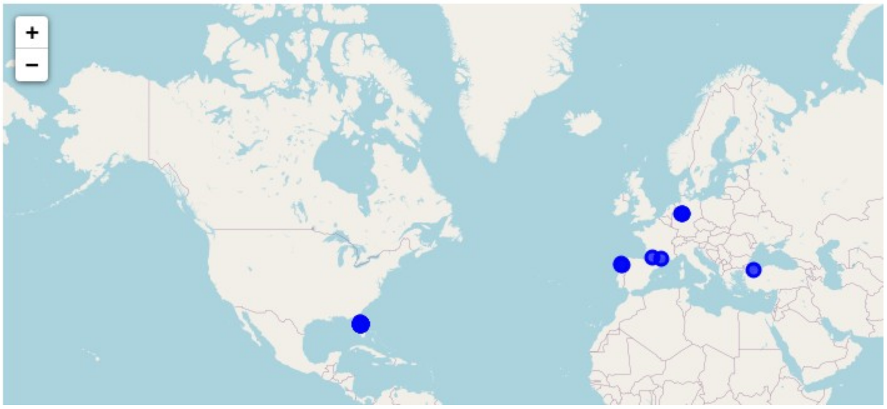


Figure 4—*Some locations are not in Florida*

To combat this, I googled about ten of the zipcodes and manually converted the zipcodes to coordinates. Now all of the zipcodes are correctly converted and displayed on the map as shown in Figure 5.

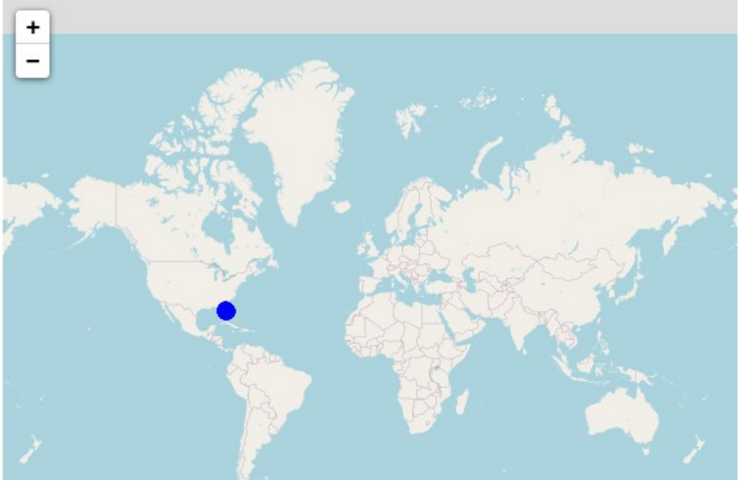
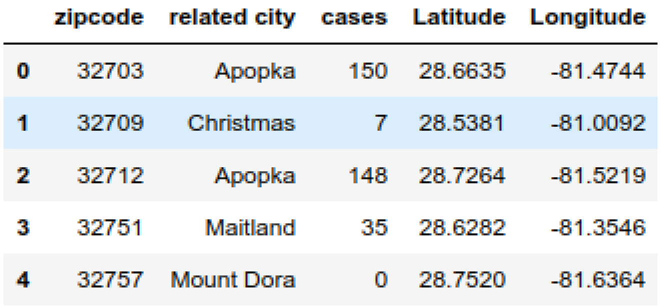


Figure 5*—All Locations are now in Orange County, FL*

We now have fully formed data with no errors.



	zipcode	related city	cases	Latitude	Longitude
0	32703	Apopka	150	28.6635	-81.4744
1	32709	Christmas	7	28.5381	-81.0092
2	32712	Apopka	148	28.7264	-81.5219
3	32751	Maitland	35	28.6282	-81.3546
4	32757	Mount Dora	0	28.7520	-81.6364

Figure 6*—Fully formed dataframe of covid-19 cases*

3.2 Using the Foursquare API

I now used the Foursquare API to get venue information on the above locations. I initially used a radius of 500 meters. However, this returned very few results for each location. About a third of the locations had no venues and another third had less than 10. I initially didn't catch this and when I ran Kmeans clustering later on I found that all locations were grouped into a single cluster. I went back and changed the radius to 3000 meters and got many more results. As can be seen in Figure 7 below, not all zipcodes hit the limit of 100 venues. This is probably due to being low population areas.

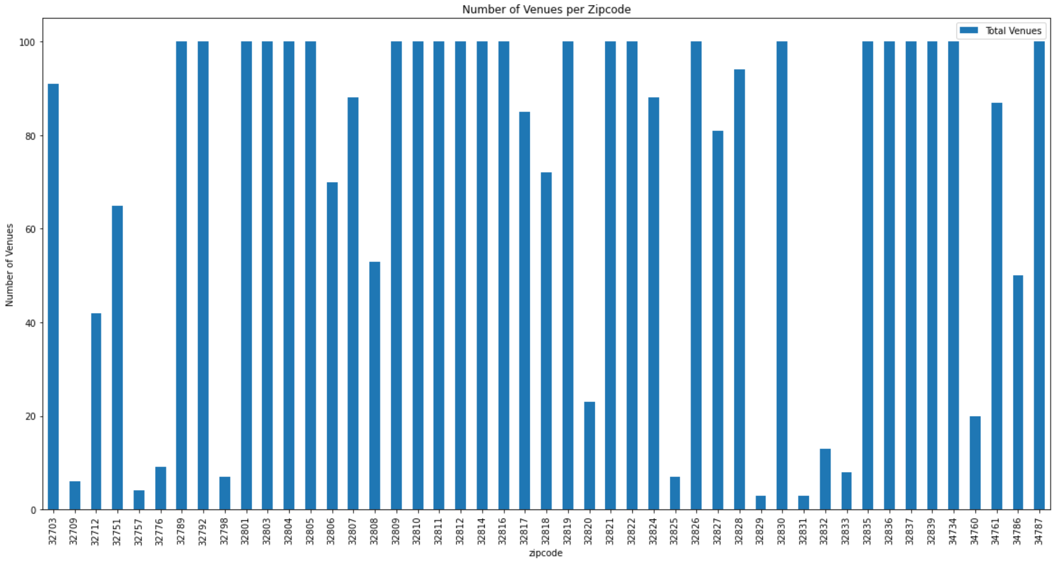
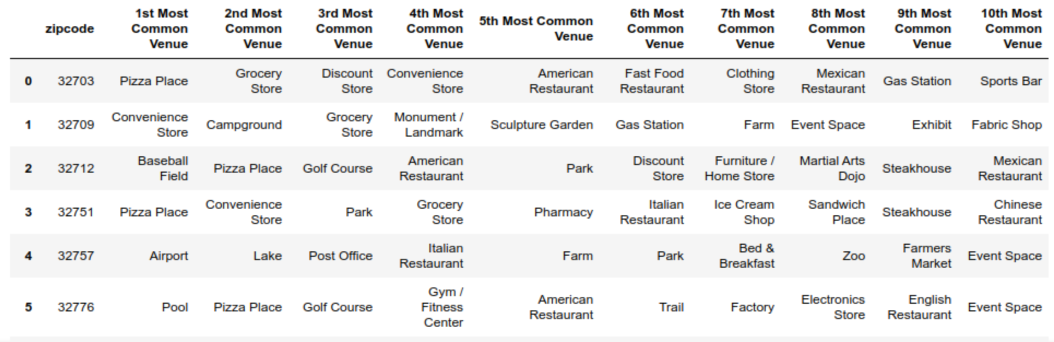


Figure 7—*Number of Returned Venues per zipcode*

For reference, I also decided to look at the types of venues returned and created the ten most common venues for each zipcode. Granted, some zipcodes will have less than ten based on the above chart. Below you can see the a subsection of the top venues in some zipcodes.



	zipcode	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	32703	Pizza Place	Grocery Store	Discount Store	Convenience Store	American Restaurant	Fast Food Restaurant	Clothing Store	Mexican Restaurant	Gas Station	Sports Bar
1	32709	Convenience Store	Campground	Grocery Store	Monument / Landmark	Sculpture Garden	Gas Station	Farm	Event Space	Exhibit	Fabric Shop
2	32712	Baseball Field	Pizza Place	Golf Course	American Restaurant	Park	Discount Store	Furniture / Home Store	Martial Arts Dojo	Steakhouse	Mexican Restaurant
3	32751	Pizza Place	Convenience Store	Park	Grocery Store	Pharmacy	Italian Restaurant	Ice Cream Shop	Sandwich Place	Steakhouse	Chinese Restaurant
4	32757	Airport	Lake	Post Office	Italian Restaurant	Farm	Park	Bed & Breakfast	Zoo	Farmers Market	Event Space
5	32776	Pool	Pizza Place	Golf Course	Gym / Fitness Center	American Restaurant	Trail	Factory	Electronics Store	English Restaurant	Event Space

Figure 8—*Top Venues in each Zipcode*

3.3 Kmeans Clustering

As a reminder, the purpose of this analysis is to see if similar types of neighborhoods fall within the same range of covid-19 cases. In order to determine if this is the case, I need some way of determining what "similar types of neighborhoods" means. I decided to use the unsupervised learning method: Kmeans clustering.

In order to determine the best value of K for the clustering, I used the *Elbow Method*. Using the "yellowbrick" python library, I discovered that the best number for k is 5. This is shown in figure 9 below.

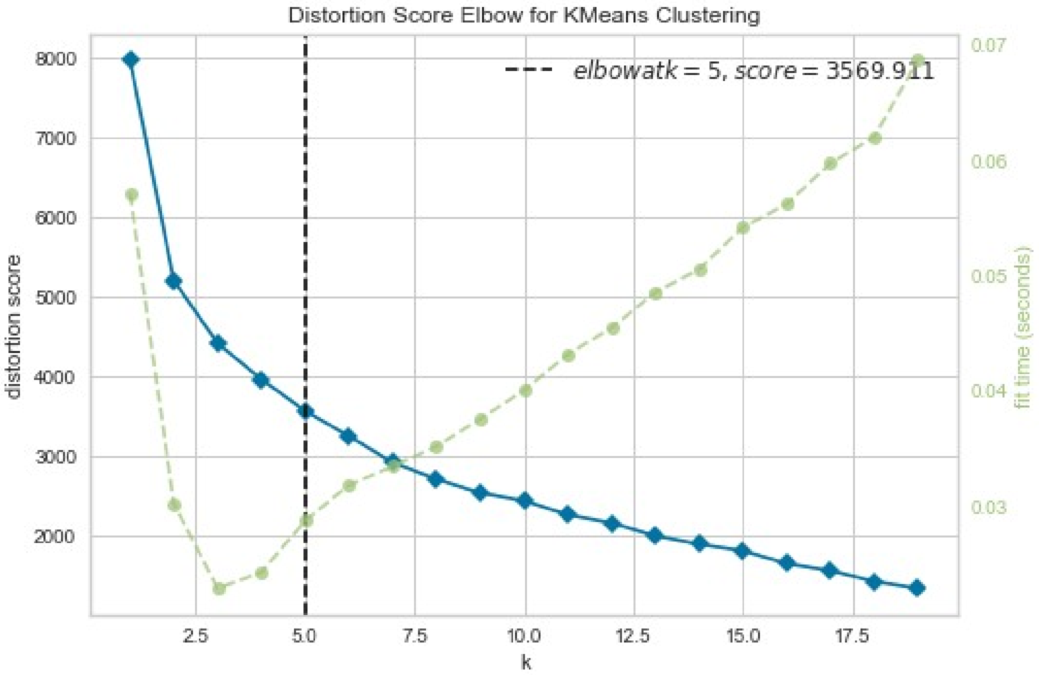


Figure 9—*Performing the elbow method for kmeans clustering*

After performing kmeans, the method gave each zipcode a cluster label. Interestingly one of the clusters had only a single zipcode. The others were more evenly distributed. I decided to display the clusters on a map for reference.

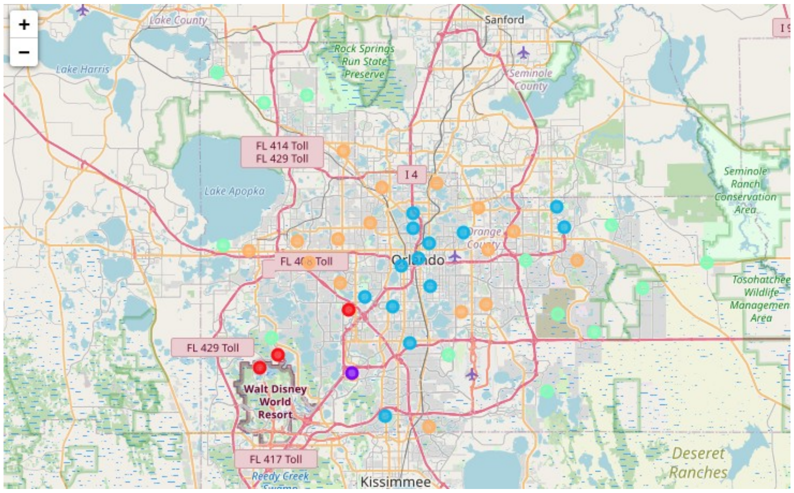
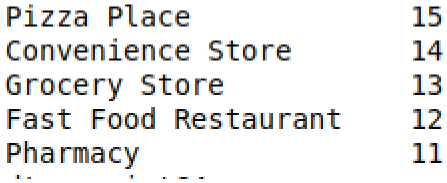


Figure 10—*Visualization of the zipcodes in their clusters*

As noted there are 5 clusters above. In order to give some kinds of name to each cluster, I performed exploratory analysis to see what the 5 most common type of venue was in each location. For brevity, I show the orange cluster shown above in figure 10's 5 most common venues below.



```
Pizza Place              15
Convenience Store        14
Grocery Store            13
Fast Food Restaurant     12
Pharmacy                 11
```

Figure 10—*Top 5 venues in the orange cluster*

After performing the analysis for each cluster, I got the following cluster names:

- **red cluster:** theme parks and gift shops
- **purple cluster:** aquarium
- **blue cluster:** restaurants
- **green cluster:** recreation and farms
- **orange:** pizza and stores

3.4 Investigating Covid-19 cases in each zipcode

At the beginning of the Methodology section, I mentioned that we received a geojson file that included the zipcodes, covid-19 cases and location boundaries of each zipcode. Thus far we had used the zipcodes to gather venue information and cluster the zipcodes. Now we use the covid-19 cases and location boundaries to further explore these zipcodes.

Since my alternative hypothesis states that the type of cluster correlates with the covid-19 cases, I found that the most beneficial thing to do would be to create a choropleth map to visualize how true this statement was. I used the folium library to create a choropleth map that shows covid cases in Orange County by zipcode. I also plotted the the clusters onto this choropleth map. This gave me a quick visual on whether my hypothesis might be true. This can be seen in figure 11 below.

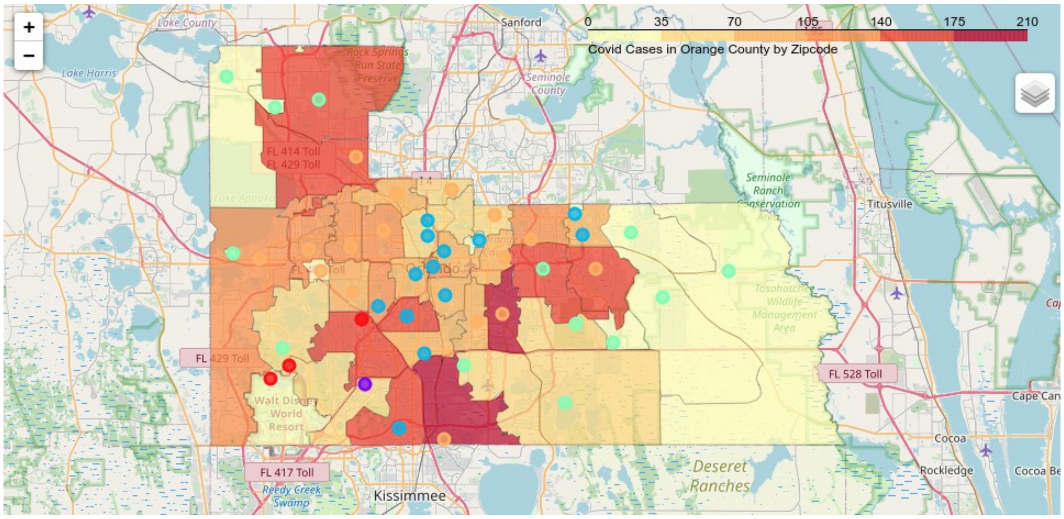


__Figure 11__—Covid Cases plotted with clusters

By interpreting the above map, it seems that the following statements might be true:

- The orange cluster almost entirely composed of zipcodes in high covid-19 areas.
- The blue cluster is almost entirely composed of zicodes between 70-140 cases which shows that it falls within the middle of covid-19 cases.
- The green cluster interestingly falls on both sides of the cases but almost all of the low amounts (less than 35) are in the orange cluster
- The red cluster is mixed
- The purple cluster is between 35-70

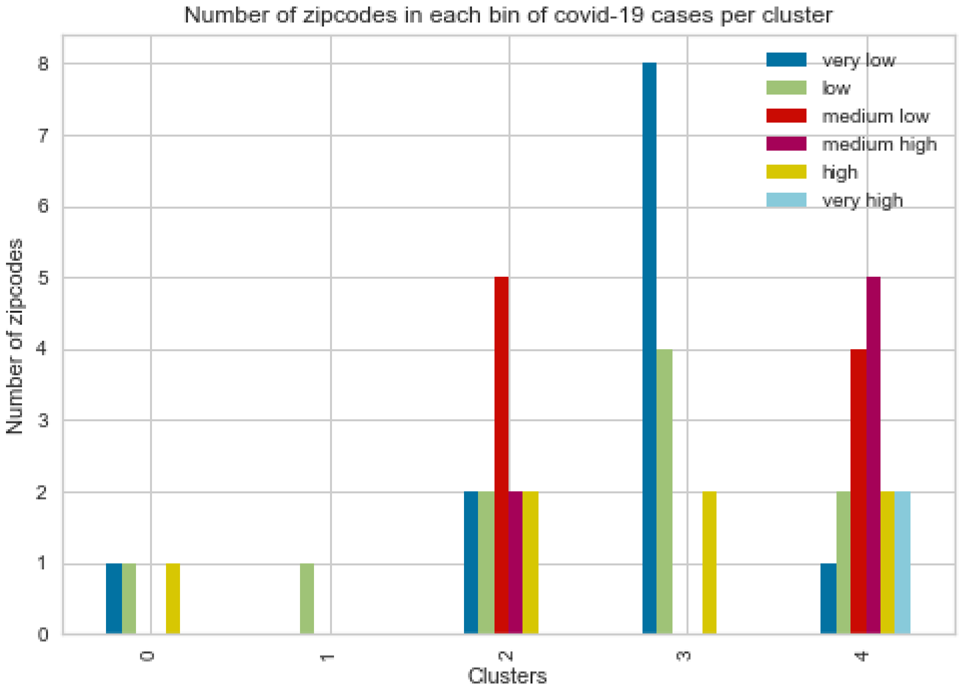
Finally in order to see how true the above assumptions are, I created the same bins for cases as shown above:

- very low: < 35
- low: 35-70
- medium-low: 70-105
- medium-high: 105-140
- high: 140-175
- very high: 175-210

We will discuss these results in the following section.

4 RESULTS

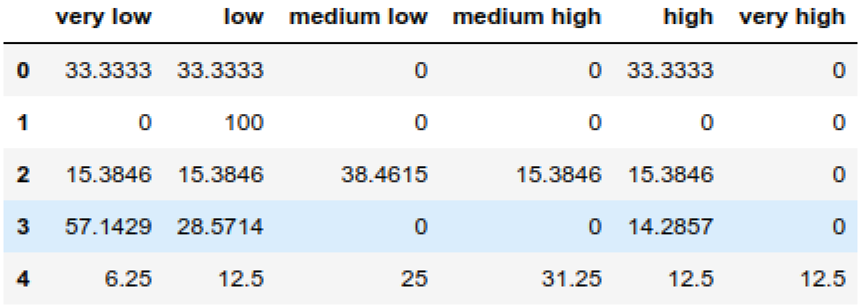
I then calculated, how many zipcodes fell into each bin per cluster as can be seen in figure 12 below.



2

Figure 12—Counts of covid-cases per cluster

Finally, I found the percentage of each bin per cluster:



	very low	low	medium low	medium high	high	very high
0	33.3333	33.3333	0	0	33.3333	0
1	0	100	0	0	0	0
2	15.3846	15.3846	38.4615	15.3846	15.3846	0
3	57.1429	28.5714	0	0	14.2857	0
4	6.25	12.5	25	31.25	12.5	12.5

Figure 13—Stats of covid-cases per cluster

In order to make sense of the results, I remind you that:

- cluster 0: theme parks and gift shops
- cluster 1: aquarium
- cluster 2: restaurants
- cluster 3: recreation and farms
- cluster 4 pizza and stores

The results seem to show that overall there is a slight correlation between clusters and number of covid-19 cases. Cluster 0 and 1, which includes them parks, gift shops and aquariums mostly contain less a *low* amount of covid-19 cases. There is 1 case that is in the high section but I believe this is due to the location of the zipcode moreso than its venues.

Cluster 3 (which is recreation and farms) afalls within the low and very low categories. However 14% of the zipcodes in this cluster due fall in the high category.

Clusters 2 and 4 are the oddballs though, they are both almost evenly distributed in the number of cases that they have.

5 Discussion

5.1 Observations

I reiterate my hypotheses: The **null hypothesis** is that there is no relationship between similarly clustered neighborhoods and the number of outbreaks that occur. The **alternative hypothesis** is that similarly clustered neighborhoods will have similar outbreaks of **COVID-19.**

In the above results, 3 out of 5 clusters were highly correlated with the number of outbreaks that occurred. Cluster2 was evenly distributed around *medium low* cases and did not have any *very high* cases. Cluster 4 skewed slightly to medium-high to very high range. All of the *very high* cases occurred in cluster 4.

This leads me to believe that there is some correlation in similarly grouped clusters and the number of covid-19 cases they have. However, clusters 2 and 4, which included, restaurants, pizza, and stores, are more complex than what can be gathered from this data.

5.2 Recommendations:

Based on my analysis, I would present the following recommendations.

- If you wish to move to Orange County Florida, I'd suggest moving into clusters 0, 1, 3 (theme parks, aquariums and farms and recreation).
- If you wish to be closer to lots of food, then cluster 2 seems safer than cluster 5. I would recommend dining in one of the restaurants in cluster 2 compared to 5.
- I'd suggest staying out of any zipcode in cluster 5 since most of the very high cases appear here.

5.3 Further Analysis:

Based on my analysis, I do believe further analysis should be given to why cluster 2 and 5 have such a large range of covid-19 cases.

Also, theme parks have unusually low cases of reported covid-19 cases. Is this due to being closed throughout most of the crisis? Or is it because people contract the disease here but report it elsewhere?

6 CONCLUSION

Overall, I think it is a fair assumption to make that specific areas in Orange County correlate with the amount of Covid-19 cases. However, this analysis does not show causation, only correlation. As can be seen above two clusters were almost normally distributed int he number of cases that occurred there which implies more factors are at work than just the types of venues within a cluster.

This analysis should be used along with other information to inform a new resident where to visit in Orange County, Florida during this pandemic.

7 REFERENCES

1. Datasheer, LLC. (n.d.). Zip-Codes.com. Retrieved June 15, 2020, from https://www.zip-codes.com/county/fl-orange.asp
2. FDOH. (2020). Florida COVID19 Cases by County. Retrieved June 15, 2020, from https://open-fdoh.hub.arcgis.com/datasets/florida-covid19-cases-by-county/geoservice
3. Foursquare. (2019). Foursquare Developer. Retrieved June 15, 2020, from https://developer.foursquare.com/
4.