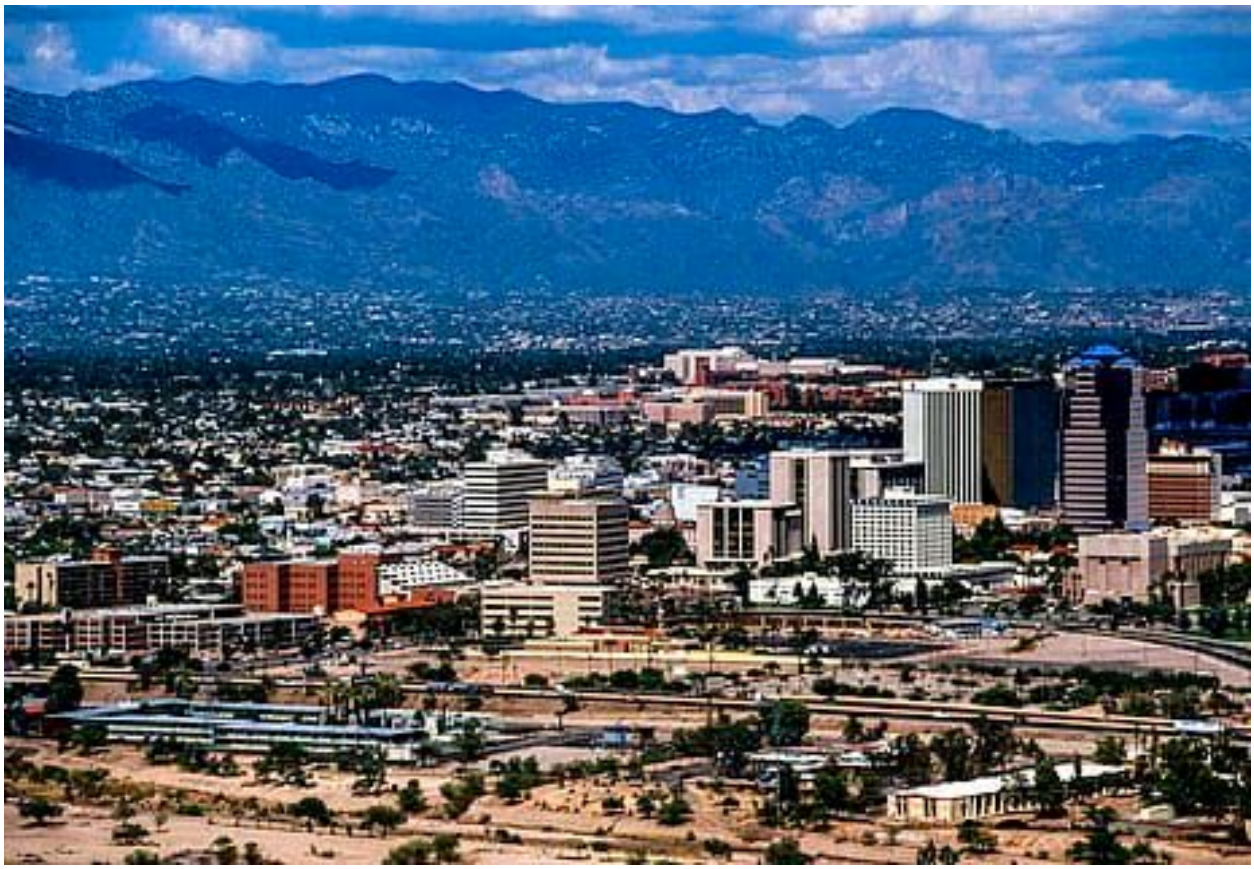
Applied Data Science Capstone Week 5

Relationship of Neighborhoods and The spread of Covid-19 in Santa Clara County



Introduction

Santa Clara County has been one of the first counties hit by Covid-19. While government officials have been vigilant on minimizing the spread of the virus, Covid-19 still spreads at an alarming rate.

This project will try to get the relationship of neighborhoods and Covid-19 cases. The stakeholders of this project are the general public & health workers in Santa Clara County California that needs to understand why certain neighborhoods have more Covid-19 cases than others.

We will Study neighborhoods in Santa Clara county using zip codes to represent neighborhoods. After this study we will at least learn a little bit why some neighborhoods have more Covid-19 cases than others.

Data used

Based on our problem, we need to gather data on :

The neighborhoods in Santa Clara County and its coordinates using zip codes

- The number of Covid-19 cases per neighborhood(using zip code)

- The population of each Santa Clara County neighborhood

- The venues near each Santa Clara neighborhood

We will be using :

- Santa Clara County Open Data COVID-19 cases by zip code of residence:

https://data.sccgov.org/COVID-19/COVID-19cases-by-zip-code-of-residence/j2gj-bg6c/data

- Opendatasoft US Zip Code Latitude and Longitude:

https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/table/

- Foursquare venues data:

https://foursquare.com/

Using this data will let us explore, analyze and answer our questions. The Covid-19 cases data will show us the number of infections by zip code, and the Us zipcodes will give us the

coordinates in longitude and latitude of the neighborhoods. The Foursquare api will be used to explore and get the venues in a given neighborhood and cluster it to determine if a certain kind of neighborhood affects Covid-19 spread.

Data Cleaning

We Download the Santa Clara County Open Data COVID-19 cases by zip code of residence dataset , while this data include the number of Covid-19 cases, population and zip code; the longitude and latitude of each entry is missing. Opendatasoft US Zip Code Latitude and Longitude dataset were downloaded to get the latitude and longitude of each zip code. The two datasets need to be merged to get a more complete dataset.

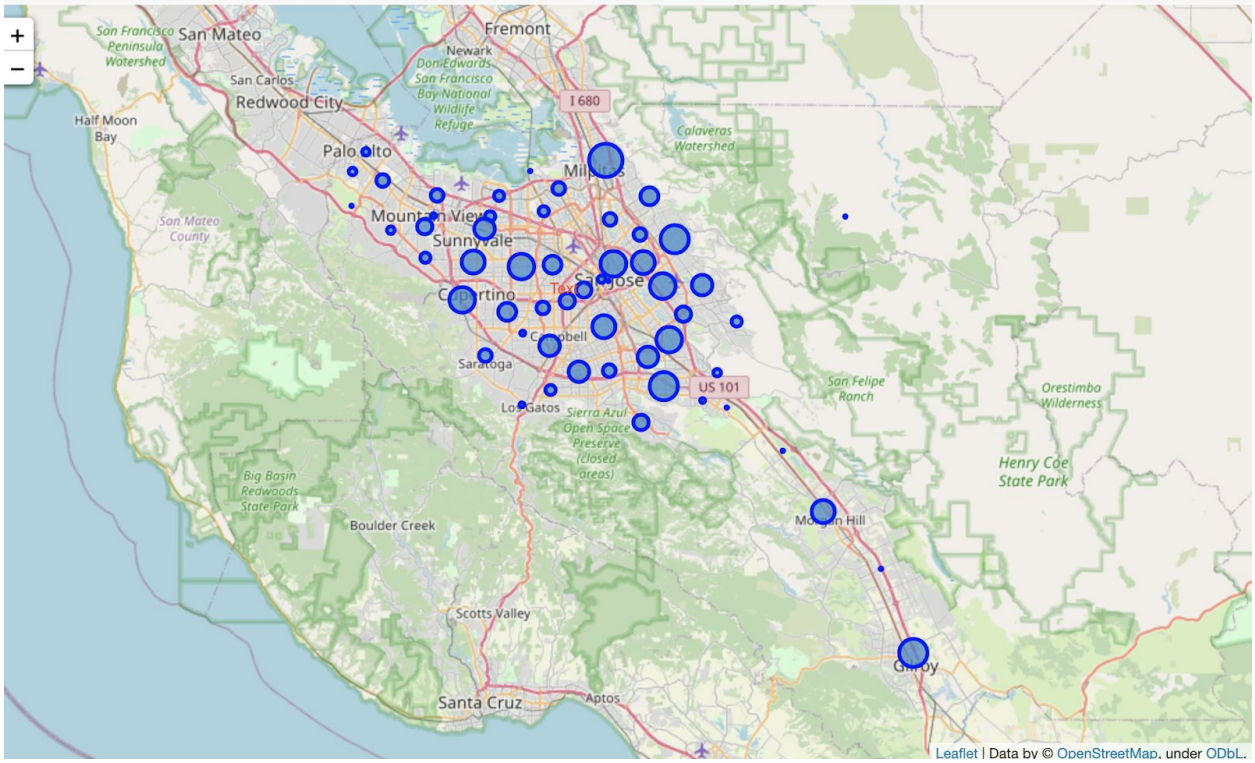
	Unnamed: 0	zipcode	Cases	Population	Rate	City	State	Latitude	Longitude	Timezone	Daylight savings time flag	geopoint
0	NaN	94022	322	19378	1662	Los Altos	CA	37.377140	-122.124120	-8	1	37.37714,-122.12412
1	NaN	94024	347	23961	1448	Los Altos	CA	37.353741	-122.087170	-8	1	37.353741,-122.08717
2	NaN	94040	1052	35845	2935	Mountain View	CA	37.380207	-122.087870	-8	1	37.380207,-122.08787
3	NaN	94041	433	14394	3008	Mountain View	CA	37.389490	-122.078460	-8	1	37.38949,-122.07846
4	NaN	94043	739	31488	2347	Mountain View	CA	37.406790	-122.074610	-8	1	37.40679,-122.07461
5	NaN	94085	876	23612	3710	Sunnyvale	CA	37.388940	-122.017956	-8	1	37.38894,-122.017956
6	NaN	94086	1381	49630	2783	Sunnyvale	CA	37.378341	-122.023980	-8	1	37.378341,-122.02398
7	NaN	94087	966	57219	1688	Sunnyvale	CA	37.350091	-122.036020	-8	1	37.350091,-122.03602
8	NaN	94089	786	22313	3523	Sunnyvale	CA	37.406289	-122.008380	-8	1	37.406289,-122.00838
9	NaN	94301	345	17191	2007	Palo Alto	CA	37.443688	-122.150710	-8	1	37.443688,-122.15071
10	NaN	94304	105	3902	2691	Palo Alto	CA	37.397823	-122.166050	-8	1	37.397823,-122.16605
11	NaN	94305	112	15730	712	Stanford	CA	37.427039	-122.164960	-8	1	37.427039,-122.16496
12	NaN	94306	408	27549	1481	Palo Alto	CA	37.419389	-122.132730	-8	1	37.419389,-122.13273
13	NaN	95002	145	2146	6757	Alviso	CA	37.427439	-121.975210	-8	1	37.427439,-121.97521
14	NaN	95008	1512	46513	3251	Campbell	CA	37.278843	-121.954460	-8	1	37.278843,-121.95446

Feature Selection

While we have a dataset that has all the data we need, there are columns we don't need and have to remove such as Timezone, Daylight savings time flag and geopoint.

	zipcode	Cases	Population	Rate	City	Latitude	Longitude
0	94022	322	19378	1662	Los Altos	37.377140	-122.124120
1	94024	347	23961	1448	Los Altos	37.353741	-122.087170
2	94040	1052	35845	2935	Mountain View	37.380207	-122.087870
3	94041	433	14394	3008	Mountain View	37.389490	-122.078460
4	94043	739	31488	2347	Mountain View	37.406790	-122.074610
5	94085	876	23612	3710	Sunnyvale	37.388940	-122.017956
6	94086	1381	49630	2783	Sunnyvale	37.378341	-122.023980
7	94087	966	57219	1688	Sunnyvale	37.350091	-122.036020
8	94089	786	22313	3523	Sunnyvale	37.406289	-122.008380
9	94301	345	17191	2007	Palo Alto	37.443688	-122.150710
10	94304	105	3902	2691	Palo Alto	37.397823	-122.166050
11	94305	112	15730	712	Stanford	37.427039	-122.164960
12	94306	408	27549	1481	Palo Alto	37.419389	-122.132730
13	95002	145	2146	6757	Alviso	37.427439	-121.975210
14	95008	1512	46513	3251	Campbell	37.278843	-121.954460

A Folium map that shows the location of the neighborhoods with the radius of circles as population



Methodology

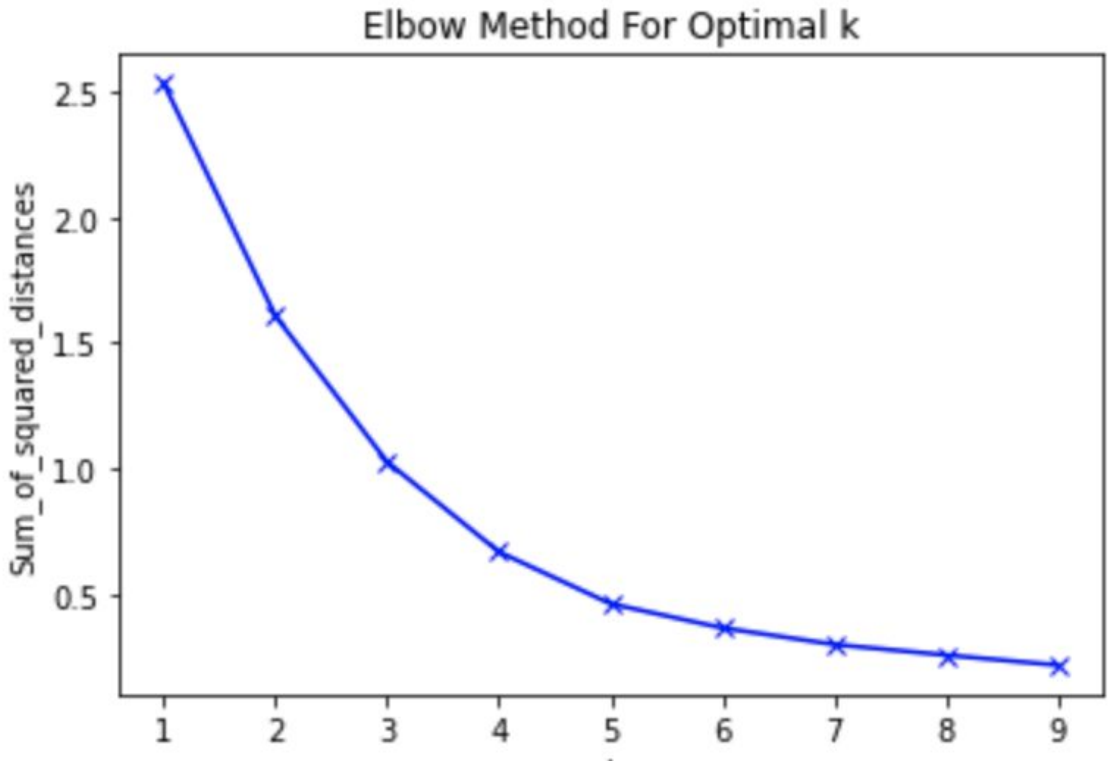
Using Foursquare api we will get all nearby venues for each zipcode using its longitude and latitude. We use One hot encoding to turn venues into columns.

	zipcode	Arts & Entertainment	College & University	Escape Room	Food	Martial Arts School	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
0	94022	3	0	0	45	0	0	24	0	0	25	3
1	94024	3	0	0	46	0	0	25	1	0	24	1
2	94040	2	0	0	50	0	0	18	0	1	28	1
3	94041	3	0	0	51	0	0	20	0	1	25	0
4	94043	5	0	0	42	0	0	28	0	1	24	0
5	94085	3	0	0	56	0	2	22	0	1	15	1
6	94086	2	0	0	57	0	2	23	0	1	15	0
7	94087	2	0	0	62	0	1	18	0	0	16	1
8	94089	7	0	0	50	0	3	23	1	0	14	2
9	94301	6	2	0	53	0	0	17	1	0	20	1

Let's check what venues are most common in each zip code.

	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
0	94022	Food	Shop & Service	Outdoors & Recreation	Travel & Transport	Arts & Entertainment	Residence	Professional & Other Places	Nightlife Spot
1	94024	Food	Outdoors & Recreation	Shop & Service	Arts & Entertainment	Travel & Transport	Professional & Other Places	Residence	Nightlife Spot
2	94040	Food	Shop & Service	Outdoors & Recreation	Arts & Entertainment	Travel & Transport	Residence	Professional & Other Places	Nightlife Spot
3	94041	Food	Shop & Service	Outdoors & Recreation	Arts & Entertainment	Residence	Travel & Transport	Professional & Other Places	Nightlife Spot
4	94043	Food	Outdoors & Recreation	Shop & Service	Arts & Entertainment	Residence	Travel & Transport	Professional & Other Places	Nightlife Spot

We group neighborhoods as clusters, but first we need to know the best amount of clusters to use. Using the elbow method we will get the optimum number of clusters.



Based on the chart, the optimum cluster count is 4. K=4

Now that we found out the best number of clusters (k=4) it is time to group our neighborhoods into clusters using K means.

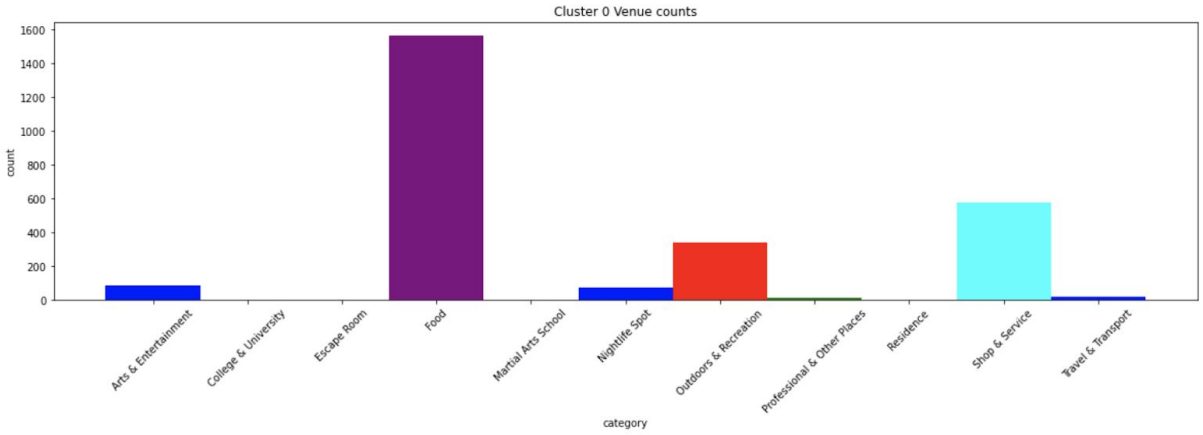
	zipcode	Cases	Population	Rate	City	Latitude	Longitude	Cluster Labels	Arts & Entertainment	College & University	Escape Room	Food	Martial Arts School	Nightlife Spot	Outdoors & Recreation
0	94022	322	19378	1662	Los Altos	37.377140	-122.12412	1	3	0	0	45	0	0	24
1	94024	347	23961	1448	Los Altos	37.353741	-122.08717	1	3	0	0	46	0	0	25
2	94040	1052	35845	2935	Mountain View	37.380207	-122.08787	1	2	0	0	50	0	0	18
3	94041	433	14394	3008	Mountain View	37.389490	-122.07846	1	3	0	0	51	0	0	20
4	94043	739	31488	2347	Mountain View	37.406790	-122.07461	1	5	0	0	42	0	0	28

We show which venues are common

	zipcode	Cases	Population	Rate	City	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
0	94022	322	19378	1662	Los Altos	37.377140	-122.12412	1	Food	Shop & Service	Outdoors & Recreation	Travel & Transport	Arts & Entertainment	Residence
1	94024	347	23961	1448	Los Altos	37.353741	-122.08717	1	Food	Outdoors & Recreation	Shop & Service	Arts & Entertainment	Travel & Transport	Professional & Other Places
2	94040	1052	35845	2935	Mountain View	37.380207	-122.08787	1	Food	Shop & Service	Outdoors & Recreation	Arts & Entertainment	Travel & Transport	Residence
3	94041	433	14394	3008	Mountain View	37.389490	-122.07846	1	Food	Shop & Service	Outdoors & Recreation	Arts & Entertainment	Residence	Travel & Transport
4	94043	739	31488	2347	Mountain View	37.406790	-122.07461	1	Food	Outdoors & Recreation	Shop & Service	Arts & Entertainment	Residence	Travel & Transport

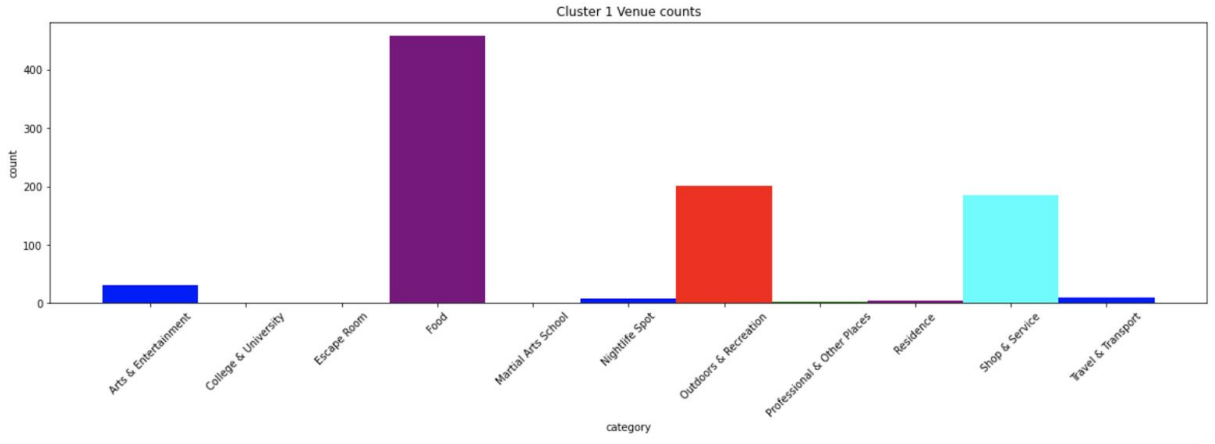
Now we make bar charts for each cluster to show the most common venues in each cluster.

Cluster 0



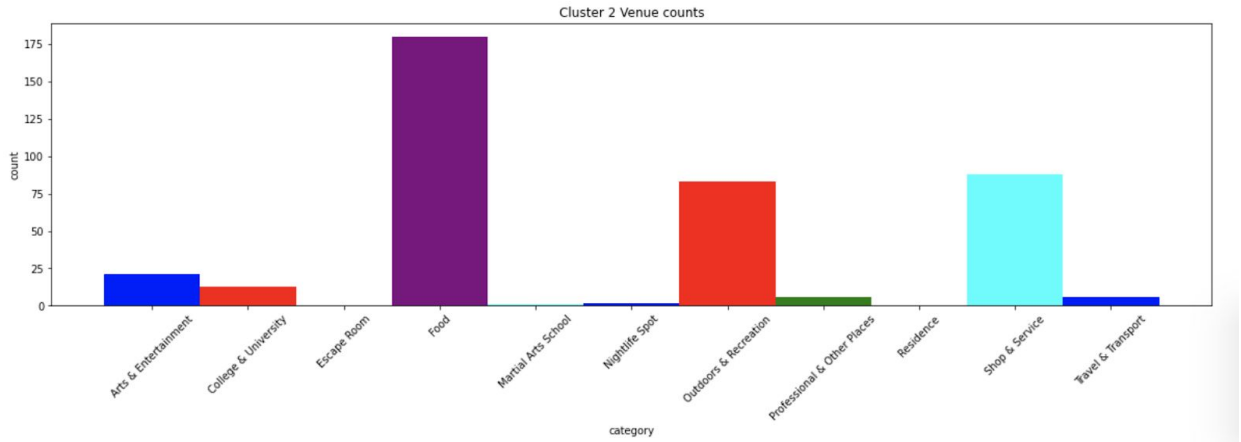
Cluster 0 has lots of restaurants, shops and nightlife spots. Based on the most common venues this cluster is a Downtown Cluster, cluster where more people gather, eat, shop and have fun.

Cluster 1



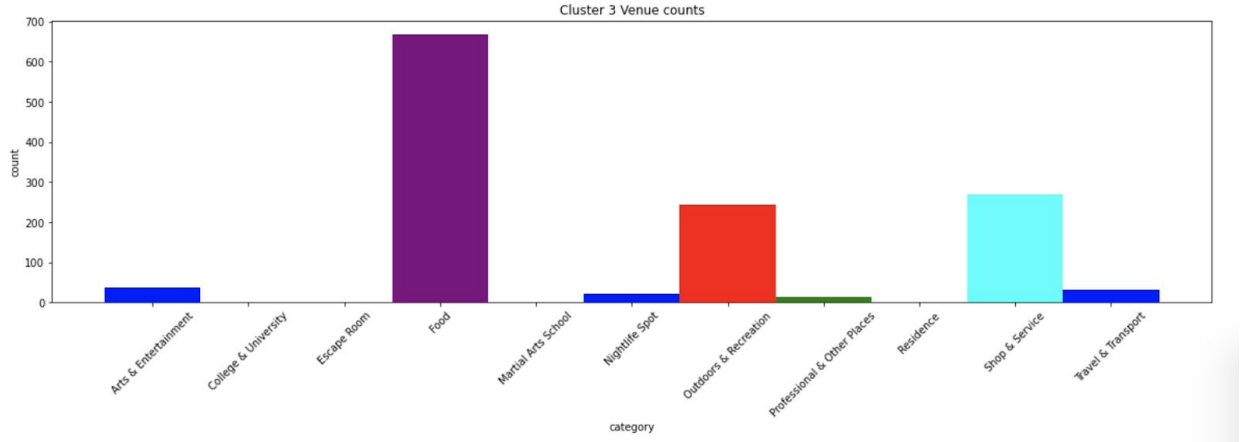
Cluster 1 has more Outdoors and Recreation spots. While it has many restaurants and shops it fails in numbers compared to Cluster 0. We will call it the Outdoors and Recreation cluster.

Cluster 2



Cluster 2 has the most number of College and Universities, it too has many outdoor and recreation spots. We will call this the Academy cluster.

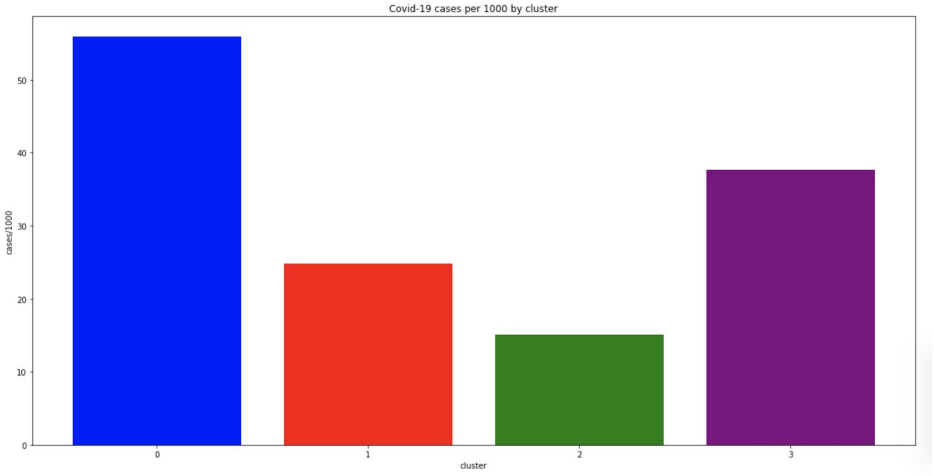
Cluster 3



Cluster 3 has like other clusters many restaurants, shops and outdoor spots, but unlike other clusters it has more Travel and Transport spots. We will call it the Travel cluster.

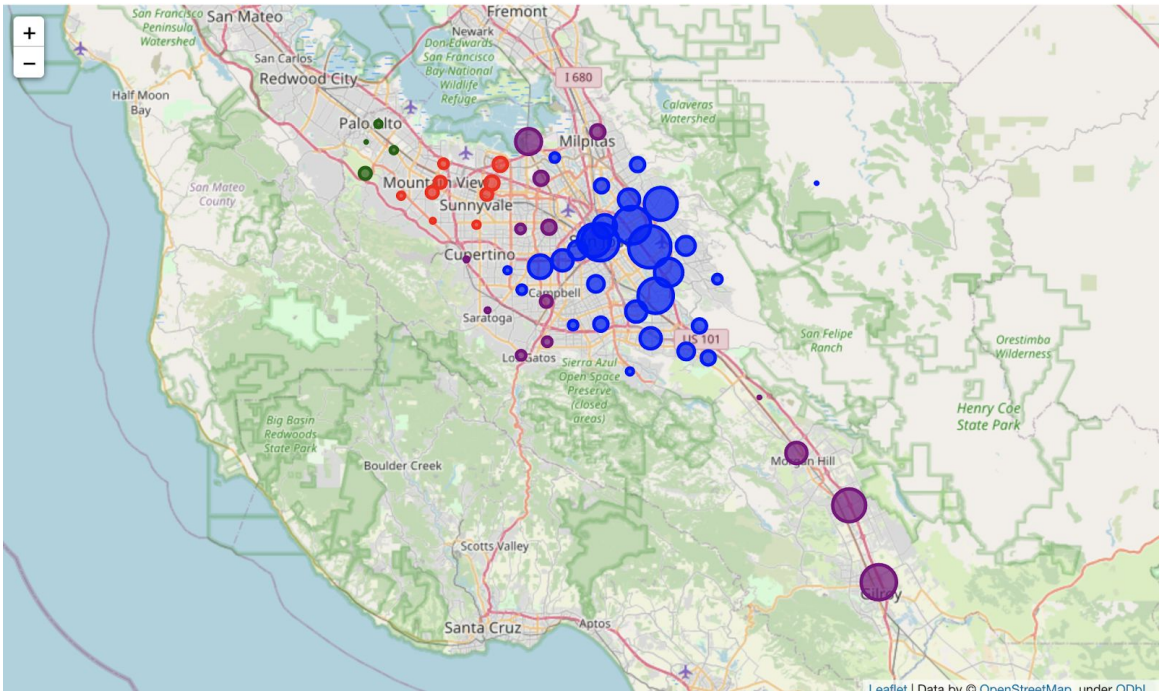
Covid-19 cases per cluster

Let’s check how many Covid-19 cases / 1000 people per cluster.



Based on the above chart cluster 0 has the highest number of cases and cluster 2 has the lowest number of cases.

Let's look at this data spread on a map. The circle’s radius represents Covid-19 cases by population.



Results and Discussion

Based on our study neighborhoods in cities with high populations like San Jose have the most number of Covid-19 cases. Neighborhoods immediately near San Jose like Milpitas, Santa Clara and Morgan hill have high cases, but going further south to like San Martin and Gilroy, cases increase significantly. Going further north west to Palo Alto and Mountain View cases decreased.

The study shows that population and location affect the number of Covid-19 cases, but venues near neighborhoods affect the number of cases?

We grouped the neighborhoods by clusters using K-means clustering. Based on our study, the Downtown cluster or thr cluster with the most number of restaurants, shops and nightlife spots has the most number of Covid-19 cases. The travel cluster which are neighborhoods with the most number of travel and transport spots have the second highest number of cases. The Outdoor and Recreation clusters which are neighborhoods with mostly outdoor and recreation pots have lower numbers of cases. The cluster with the fewest number of Covid-19 cases is the Academy cluster, the cluster that has the most number of Universities and Colleges.

Because of the stay-at-home order, closure of many Universities, College and Outdoor recreation spots lessens the number of Covid-19 cases in neighborhoods with many of those venues. Neighborhoods near where many people gather to eat out, shop and party have a larger chance to have more Covid-19 cases.

Conclusion

We created this project to know why some neighborhoods have more Covid-19 cases than others. By gathering venues near Santa Clara county neighborhoods using Foursquare api and clustering them, we found out that certain venue types affect the number of cases. Population and location affects the number of cases too.

This study is just a guide using the data that are available to us. Many other factors can affect the number of Covid-19 cases that are not used in this study.

This can be a small tool, a small glimpse to help stakeholders in making decisions about the relationship of neighborhoods and Covid-19 cases. This should not be used as a be-all end-all study used to make decisions.

Disclaimer

This study will not be an exhaustive study of Covid-19's spread but just a simple glimpse of the relationship of Covid-19 cases and neighborhoods in Santa Clara county.

Created by James Cuevo for Applied Data Science Capstone Coursera course 2020-2021